Network Working Group
Request for Comments: 3277
Category: Informational

D. McPherson
TCB

April 2002

Intermediate System to Intermediate System (IS-IS) Transient Blackhole Avoidance

Status of this Memo

Copyright Notice

Abstract

This document describes a simple, interoperable mechanism that can be employed in Intermediate System to Intermediate System (IS-IS) networks in order to decrease the data loss associated with deterministic blackholing of packets during transient network conditions. The mechanism proposed here requires no IS-IS protocol changes and is completely interoperable with the existing IS-IS specification.

1. Introduction

When an IS-IS router that was previously a transit router becomes unavailable as a result of some transient condition such as a reboot, other routers within the routing domain must select an alternative path to reach destinations which have previously transited the failed router. Presumably, the newly selected router(s) comprising the path have been available for some time and, as a result, have complete forwarding information bases (FIBs) which contain a full set of reachability information for both internal and external (e.g., BGP) destination networks.

When the previously failed router becomes available again, it is only seconds before the paths that had previously transited the router are again selected as the optimal path by the IGP. As a result, forwarding tables are updated and packets are once again forwarded along the path. Unfortunately, external destination reachability information (e.g., learned via BGP) is not yet available to the router, and as a result, packets bound for destinations not learned via the IGP are unnecessarily discarded.

A simple interoperable mechanism to alleviate the offshoot associated with this deterministic behavior is discussed below.

2. Discussion

This document describes a simple, interoperable mechanism that can be employed in IS-IS [1, 2] networks in order to avoid transition to a newly available path until other associated routing protocols such as BGP have had sufficient time to converge.

The benefits of such a mechanism can be realized when considering the following scenario depicted in Figure 1.

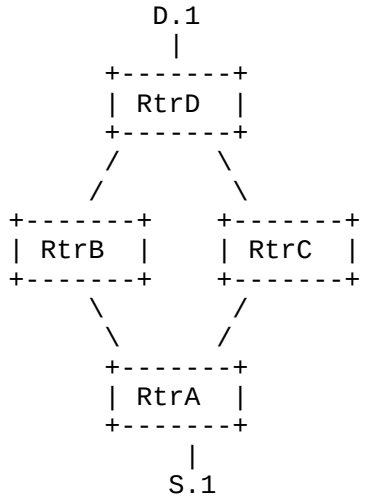

Figure 1: Example Network Topology

Host S.1 is transmitting data to destination D.1 via a primary path of RtrA->RtrB->RtrD. Routers A, B and C learn of reachability to destination D.1 via BGP from RtrD. RtrA's primary path to D.1 is selected because when calculating the path to BGP NEXT_HOP of RtrD, the sum of the IS-IS link metrics on the RtrA-RtrB-RtrD path is less than the sum of the metrics of the RtrA-RtrC-RtrD path.

Assume RtrB becomes unavailable and as a result the RtrC path to RtrD is used. Once RtrA's FIB is updated and it begins forwarding packets to RtrC, everything should behave properly as RtrC has existing forwarding information regarding destination D.1's availability via BGP NEXT_HOP RtrD.

Assume now that RtrB comes back online. In only a few seconds, IS-IS neighbor state has been established with RtrA and RtrD and database synchronization has occurred. RtrA now realizes that the best path to destination D.1 is via RtrB, and therefore updates it FIB appropriately. RtrA begins to forward packets destined to D.1 to RtrB. Though, because RtrB has yet to establish and synchronize its BGP neighbor relationship and routing information with RtrD, RtrB has no knowledge regarding reachability of destination D.1, and therefore discards the packets received from RtrA destined to D.1.

If RtrB were to temporarily set its LSP Overload bit while synchronizing BGP tables with its neighbors, RtrA would continue to use the working RtrA->RtrC->RtrD path, and the LSP should only be used to obtain reachability to locally connected networks (rather than for calculating transit paths through the router, as defined in [1]).

However, it should be noted that when RtrB goes away, its LSP is still present in the IS-IS databases of all other routers in the routing domain. When RtrB comes back it establishes adjacencies. As soon as its neighbors have an adjacency with RtrB, they will advertise their new adjacency in their new LSP. The result is that all the other routers will receive new LSPs from RtrA and RtrD containing the RtrB adjacency, even though RtrB is still completing its synchronization and therefore has not yet sent its new LSP.

At this time SPF is computed and everyone will include RtrB in their tree since they will use the old version of RtrB LSP (the new one has not yet arrived). Once RtrB has finished establishing all its adjacencies, it will then regenerate its LSP and flood it. Then all other routers within the domain will finally compute SPF with the correct information. Only at that time will the Overload bit be taken into account.

As such, it is recommended that each time a router establishes an adjacency, it will update its LSP and flood it immediately, even before beginning database synchronization. This will allow for the Overload bit setting to propagate immediately, and remove the potential for an older version of the reloaded routers LSP to be used.

After synchronization of BGP tables with neighboring routers (or expiry of some other timer or trigger), RtrB would generate a new LSP, clearing the Overload bit, and RtrA could again begin using the optimal path via RtrB.

Typically, in service provider networks IBGP connections are done via peerings with 'loopback' addresses. As such, the newly available router must advertise its own loopback (or similar) IP address, as well as associated adjacencies, in order to make the loopbacks accessible to other routers within the routing domain. It is because of this that simply flooding an empty LSP is not sufficient.

3. Deployment Considerations

Such a mechanism increases overall network availability and allows network operators to alleviate the deterministic blackholing behavior introduced in this scenario. Similar mechanisms [3] have been defined for OSPF, though only after realizing the usefulness obtained from that of the IS-IS Overload bit technique.

This mechanism has been deployed in several large IS-IS networks for a number of years.

Triggers for setting the Overload bit as described are left to the implementer. Some potential triggers could perhaps include "N seconds after booting", or "N number of BGP prefixes in the BGP Loc-RIB".

Unlike similar mechanisms employed in [3], if the Overload bit is set in a router's LSP, NO transit paths are calculated through the router. As such, if no alternative paths are available to the destination network, employing such a mechanism may actually have a negative impact on convergence (i.e., the router maintains the only available path to reach downstream routers, but the Overload bit disallows other nodes in the network from calculating paths via the router, and as such, no feasible path exists to the routers).

Finally, if all systems within an IS-IS routing domain haven't implemented the Overload bit correctly, forwarding loops may occur.

4. Potential Alternatives

Alternatively, it may be considered more appealing to employ something more akin to [3] for this purpose. With this model, during transient conditions a node advertises excessively high link metrics to serve as an indication, to other nodes in the network that paths transiting the router are "less desirable" than existing paths.

The advantage of a metric-based mechanism over the Overload bit mechanism model proposed here is that transit paths may still be calculated through the router. Another advantage is that a metric-based mechanism does not require that all nodes in the IS-IS domain correctly implement the Overload bit.

However, as currently deployed, IS-IS provides for only 6 bits of space for link metric allocation, and 10 bits aggregate path metric. Though extensions proposed in [4] remove this limitation, they have not yet been widely deployed. As such, there's currently little flexibility when using link metrics for this purpose. Of course, both methods proposed in this document are backwards-compatible.

5. Security Considerations

The mechanisms specified in this memo introduces no new security issues to IS-IS.

6. Acknowledgements

The author of this document makes no claim to the originality of the idea. Thanks to Stefano Previdi for valuable feedback on the mechanism discussed in this document.

7. References

[1] ISO, "Intermediate system to Intermediate system routing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)," ISO/IEC 10589:1992.

[2] Callon, R., "OSI IS-IS for IP and Dual Environment," RFC 1195, December 1990.

[3] Retana, A., Nguyen, L., White, R., Zinin, A. and D. McPherson, "OSPF Stub Router Advertisement", RFC 3137, June 2001.

[4] Li, T. and H. Smit, "IS-IS extensions for Traffic Engineering", Work in Progress.

8. Author's Address

Danny McPherson
TCB
Phone: 303.470.9257
EMail: danny@tcb.net

9. Full Copyright Statement

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.