INTERNET-DRAFT

Danny McPherson
Arbor Networks
Naiming Shen
Cisco Systems

Expires: April 2008

October 20, 2007

Intended Status: Proposed Standard

IS-IS Transient Blackhole Avoidance

<draft-ietf-isis-rfc3277bis-00.txt>

Status of this Memo

Copyright Notice

Abstract

This document describes a simple, interoperable mechanism that can be employed in IS-IS networks in order to decrease 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.

The intention of this document is to provide an update to [RFC 3277].

Table of Contents

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 had 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, in only a few seconds 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 outlined below.

1.1. Specification of Requirements

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 [RFC 2119].

2. Discussion

This document describes a simple, interoperable mechanism that can be employed in IS-IS [ISO 8473] [RFC 1195] 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 realized when considering the scenario depicted in Figure 1.

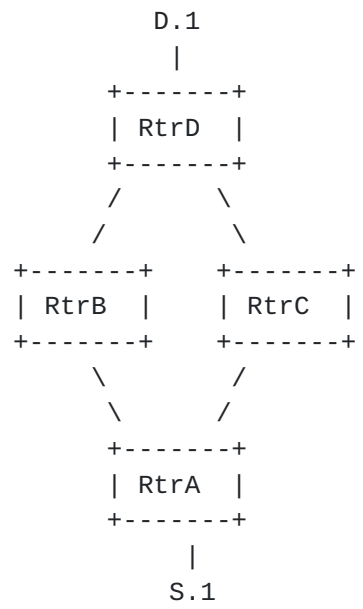

Figure 1: Example Network Topology

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 is used to reach RtrD. 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 subsequently updates it FIB appropriately. RtrA begins to forward packets destined to D.1 to RtrB. However, because RtrB has yet to establish and synchronization it's 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 it's LSP Overload bit while synchronizing BGP tables with it's neighbors, RtrA would continue to

use the operational RtrA->RtrC->RtrD path, and the IS-IS LSP SHOULD only be used to obtain reachability to locally connected networks (rather than for calculating transit paths through the router, as defined in [ISO 8473]).

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 transmitted it's 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's LSP (the new one has not yet arrived). Once RtrB has finished establishing 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 (and other routers in the routing domain) could again begin using the optimal path via RtrB.

Typically, in service provider networks IBGP connections are done via peering sessions associated with 'loopback' addresses. As such, the newly available router must advertise it's 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's because of this requirement for local destinaion reachability 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 [RFC 3137] have been defined for OSPF, 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, and a variety of techniques to configure and trigger overload bit setting and clearing are available in many implementations. Such 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 [RFC 3137], 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).

It should also be noted that if all systems within an IS-IS routing domain haven't implemented this Overload bit behavior correctly, forwarding loops may occur.

Alternatively, it may be considered more appealing to employ something more akin to [RFC 3137] 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 alternative paths.

The advantage of a metric-based mechanism over the Overload bit mechanism 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 handling procedures.

As traditionally specified, IS-IS provided for only 6 bits of space for link metric allocation, and 10 bits aggregate path metrics. Though extensions provided in [RFC 3784] remove this limitation, they may not yet be fully deployed in many networks. As such, there's possibly less flexibility when using link metrics for this purpose. Of course, both methods proposed in this document are backwards-compatible.

Two other more more recent techniques can help to alleviate these

transient network conditions further. Graceful restart [rfc 4724] [RFC 4781] with a control plane only restart, and "BGP free cores". Furthur discussion of these techniques is beyond the scope of this document.

4. Manageability Considerations

These extensions which have been designed, developed and deployed for many years do not have any new impact on management and operation of the IS-IS protocol via this standardization process.

5. Security Considerations

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

6. Acknowledgments

The original efforts and corresponding acknowledgements provided in [RFC 3277] have enabled this work.

Others to be provided....

7. IANA Considerations

This specification introduces no new IANA considerations and therefore requires no actions on the part of IANA.

8. References

8.1. Normative References

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

8.2. Informative References

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

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

[RFC 3137] Retana et al., "OSPF Stub Router Advertisement", RFC 3137, June 2001.

[RFC 3277] McPherson, D., "Intermediate System to Intermediate System (IS-IS) Transient Blackhole Avoidance", RFC 3277, April 2002.

[RFC 3784] Li, T., Smit, H., "IS-IS extensions for Traffic Engineering", RFC 3784, June 2004.

[RFC 4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., Rekhter, Y., "Graceful Restart Mechanism for BGP", RFC 4724, January 2007.

[RFC 4781] Rekhter, Y., Aggarwal, R., "Graceful Restart Mechanism for BGP with MPLS", RFC 4781, January 2007.

9. Authors' Addresses

Danny McPherson
Arbor Networks, Inc.
EMail: danny@arbor.net

Naiming Shen
Cisco Systems, Inc.
EMail: naiming@cisco.com

Intellectual Property Statement

Copyright Statement

IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Acknowledgment

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