INTAREA
Internet-Draft
Updates: 6296 (if approved)
Intended status: Experimental
Expires: April 6, 2012

R. Bonica, Ed.
Juniper Networks
F. Baker
Cisco Systems
M. Wasserman
Painless Security
October 4, 2011

Multihoming with IPv6-to-IPv6 Network Prefix Translation (NPTv6)
draft-bonica-v6-multihome-00

Abstract

This memo describes an architecture for sites that are homed to multiple upstream providers. The architecture described herein uses IPv6-to-IPv6 Network Prefix Translation (NPTv6) to achieve redundancy, transport-layer survivability, load sharing and address independence.

This memo updates Section 2.4 of RFC 6296.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 6, 2012.

Copyright Notice

Table of Contents

1. Introduction

[RFC3582] establishes the following goals for IPv6 site multihoming:

> Redundancy - A site's ability to remain connected to the Internet, even when connectivity through one or more of its upstream providers fails.
>
> Transport-Layer Survivability - A site's ability to maintain transport-layer sessions across failover and restoration events. During a failover/restoration event, the transport-layer session may detect packet loss or reordering, but neither of these cause the transport-layer session to fail.
>
> Load Sharing - The ability of a site to distribute both inbound and outbound traffic across its upstream providers.

[RFC3582] notes that a multihoming solution may require interactions with the routing subsystem. However, multihoming solutions must be simple and scalable. They must not require excessive operational effort and must not cause excessive routing table expansion.

[RFC6296] explains how a site can achieve address independence using IPv6-to-IPv6 Network Prefix Translation (NPTv6). In order to achieve address independence, the site assigns an inside address to each of its resources (e.g., hosts). Nodes outside of the site identify those same resources using a corresponding Provider Allocated (PA) address.

The site resolves this addressing dichotomy by deploying an NPTv6 translator between itself and its upstream provider. The NPTv6 device maintains a static, one-to-one mapping between each inside address and its corresponding PA address. That mapping persists across flows and over time.

If the site disconnects from one upstream provider and connects to

another, it may lose its PA assignment. However, the site will not need to renumber its resources. It will only need to reconfigure the mapping rules on its local NPTv6 device.

Section 2.4 of [RFC6296] describes an NPTv6 architecture for sites that are homed to multiple upstream providers. While that architecture fulfils many of the goals identified by [RFC3582], it does not achieve transport-layer survivability. This memo describes an alternative architecture for multihomed sites that require transport-layer survivability. It updates Section 2.4 of [RFC6296].

2. NPTv6 Deployment

This section demonstrates how NPTv6 can be deployed in order to achieve the goals of [RFC3582].

2.1. Topology

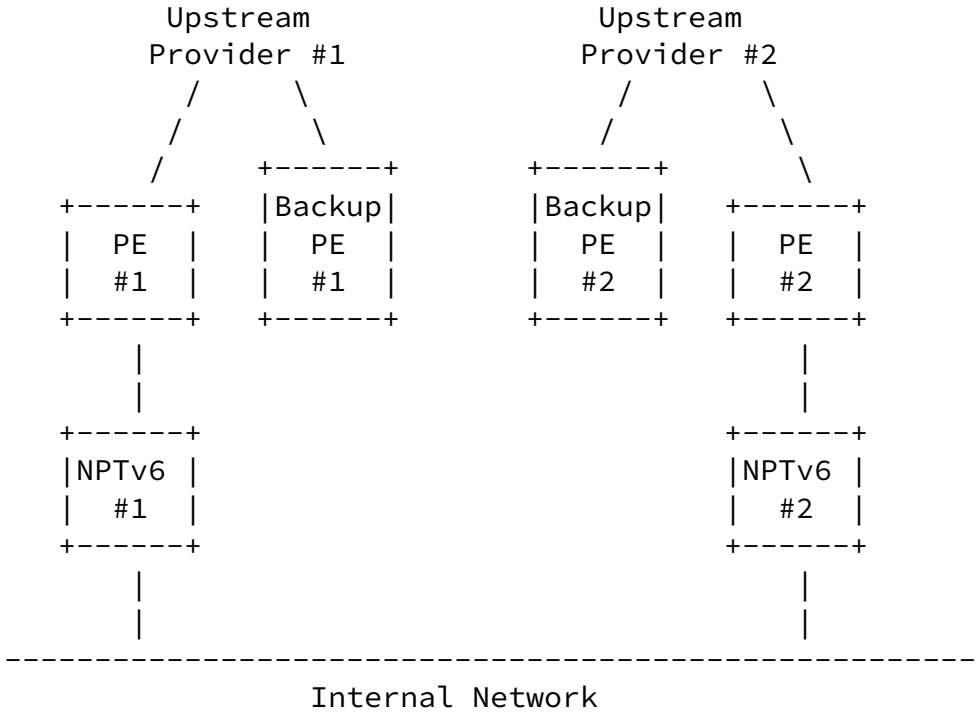

Figure 1: NPTv6 Multihomed Topology

In Figure 1, a site attaches all of its assets, including two NPTv6 translators, to an Internal Network. NPTv6 #1 is connected to Provider Edge (PE) Router #1, which is maintained by Upstream Provider #1. Likewise, NPTv6 #2 is connected to PE Router #2, which is maintained by Upstream Provider #2.

Each upstream provider also maintains a Backup PE Router. A forwarding tunnel connects the loopback interface of Backup PE Router #1 to the outside interface of NPTv6 #2. Likewise, another forwarding tunnel connects Backup PE Router #2 to NPTv6 #1. Network operators can select from many encapsulation techniques (e.g., GRE) to realize the forwarding tunnel. Tunnels are not depicted in Figure 1.

In the figure, NPTv6 #1 and NPTv6 #2 are depicted as separate boxes. While vendors may produce a separate box to support the NPTv6 function, they may also integrate the NPTv6 function into a router.

During periods of normal operation, the Backup PE routers is very lightly loaded. Therefore, a single Backup PE router may back up multiple PE routers. Furthermore, the Backup PE router may be used for other purposes (e.g., primary PE router for another customer).

2.2. Addressing

2.2.1. Upstream Provider Addressing

A Regional Internet Registry (RIR) allocates Provider Address Block (PAB) #1 to Upstream Provider #1. From PAB #1, Upstream Provider #1 allocates two sub-blocks, using them as follows.

Upstream Provider #1 uses the first sub-block for its internal address assignments. It also uses that sub-block for numbering both ends of the interfaces between itself and its customers.

Upstream Provider #1 uses the second sub-block for address allocation to its customers. We refer to a particular allocation from this sub-block as a Customer Network Block (CNB). A CNB allocated for a particular customer must be large enough to provide addressing for the customer's entire Internal Network. In our example, Upstream

Provider #1 allocates a /60, called CNB #1, to its customer.

The customer configures translation rules that reference CNB #1 on NPTv6 #1 and NPTv6 #2. This makes selected hosts that are connected to the Internal Network accessible using CNB #1 addresses. See Section 2.3 for details.

In a similar fashion, a Regional Internet Registry (RIR) allocates PAB #2 to Upstream Provider #2. Upstream Provider #2, in turn, allocates CNB #2 to the multihomed customer.

2.2.2. Site Addressing

The site obtains a Site Address Block (SAB), either from Unique Local Address (ULA) [RFC4193] space, or by some other means. The SAB is as large as all of the site's CNBs, combined. In this example, because CNB #1 and CNB #2 are both /60's, the SAB is a /59.

The site divides its SAB into smaller blocks, with each block being exactly as large as one CNB. It also associates each of the resulting sub-blocks with one of its CNBs. In this example, the site divides the SAB into a lower half and an upper half. It associates the lower half of the SAB with CNB #1 and the upper half of the SAB with CNB #2.

Finally, the site assigns one SAB address to each interface that is

connected to the Internal Network, including the inside interfaces interfaces of the two NPTv6 translators. The site also assigns a SAB address to the loopback interface of each NPTv6 translator. During periods of normal operation, interfaces that are assigned addresses from the lower half of the SAB receive traffic through Upstream Provider #1. Likewise, interfaces that are assigned addresses from the upper half of the SAB receive traffic through Upstream Provider #2.

Selected interfaces, because they receive a great deal of traffic, must receive traffic through both upstream providers simultaneously. Furthermore, those interfaces must control the portion of traffic arriving through each upstream provider. The site assigns multiple addresses to those interfaces, some from the lower half and others from the upper half of the SAB. For any interface, the ratio of

upper half to lower half assignments roughly controls the portion of traffic arriving through each upstream provider. See Section 2.3 and Section 2.5 for details.

2.3. Address Translation

Both NPTv6 translators are configured with the following rules:

> For outbound packets, if the first 60 bits of the source address identify the lower half of the SAB, overwrite those 60 bits with the 60 bits that identify CNB #1
>
> For outbound packets, if the first 60 bits of the source address identify the upper half of the SAB, overwrite those 60 bits with the 60 bits that identify CNB #2
>
> For outbound packets, if none of the conditions above are met, either drop or pass the packet without translation, according to local security policy
>
> For inbound packets, if the first 60 bits of the destination address identify CNB #1, overwrite those 60 bits with the 60 bits that identify the lower half of the SAB
>
> For inbound packets, if the first 60 bits of the destination address identify CNB #2, overwrite those 60 bits with the 60 bits that identify the upper half of the SAB
>
> For inbound packets, if none of the conditions above are met, either drop or pass the packet without translation, according to local security policy

Due to the nature of the rules described above, NPTv6 translation is

stateless. Therefore, traffic flows do not need to be symmetric across NPTv6 translators. Furthermore, a traffic flow can shift from one NPTv6 translator to another without causing transport-layer session failure.

2.4. Domain Name System (DNS)

In order to make all site resources reachable by domain name

[RFC1034], the site publishes AAAA records [RFC3596] associating each resource with all of its CNB addresses. While this DNS architecture is sufficient, it is suboptimal. Traffic that both originates and terminates within the site traverses NPTv6 translators needlessly. Several optimizations are available. These optimizations are well understood and have been applied to [RFC1918] networks for many years.

2.5. Routing

Upstream Provider #1 uses an Interior Gateway Protocol to flood topology information throughout its domain. It also uses BGP [RFC4271] to distribute customer and peer reachability information.

PE #1 acquires a route to CNB #1 with NEXT-HOP equal to the outside interface of NPTv6 #1. PE #1 can either learn this route from a single-hop eBGP session with NPTv6 #1, or acquire it through static configuration. In either case, PE #1 overwrites the NEXT-HOP of this route with its own loopback address and distributes the route throughout Upstream Provider #1 using iBGP. The LOCAL PREF for this route is set high, so that the route will be preferred to alternative routes to CNB #1. Upstream Provider #1 does not distribute this route to CNB #1 outside of its own borders.

NPTv6 #1 acquires a default route with NEXT-HOP equal to the directly connected interface on PE #1. NPTv6 #1 can either learn this route from a single-hop eBGP session with PE #1, or acquire it through static configuration.

Similarly, Backup PE #1 acquires a route to CNB #1 with NEXT-HOP equal to the outside interface of NPTv6 #2. Backup PE #1 can either learn this route from a multi-hop eBGP session with NPTv6 #2, or acquire it through static configuration. In either case, Backup PE #1 overwrites the NEXT-HOP of this route with its own loopback address and distributes the route throughout Upstream Provider #1 using iBGP. Distribution procedures are defined in [I-D.ietf-idr-best-external]. The LOCAL PREF for this route is set low, so that the route will not be preferred to alternative routes to CNB #1. Upstream Provider #1 does not distribute this route to CNB #1 outside of its own borders.

Even if Backup PE #1 maintains an eBGP session NPTv6 #2, it does not

advertise the default route through that eBGP session. During failures, Backup PE #1 does not attract outbound traffic to itself.

Finally, the Autonomous System Border Routers (ASBR) contained by Upstream Provider #1 maintain eBGP sessions with their peers. The ASBRs advertise only PAB #1 through those eBGP sessions. Upstream Provider #1 does not advertise any of the following to its eBGP peers:

any prefix that is contained by PAB #1 (i.e., more specific)

PAB #2 or any part thereof

the SAB or any part thereof

Upstream Provider #2 is configured in a manner analogous to that described above.

2.6. Failure Detection and Recovery

When PE #1 loses its route to CNB #1, it withdraws its iBGP advertisement for that prefix from Upstream Provider #1. The route advertised by Backup PE #1 remains and Backup PE #1 attracts traffic bound for CNB #1 to itself. Backup PE #1 forwards that traffic through the tunnel to NPTv6 #2. NPTv6 #2 performs translations and delivers the traffic to the Internal Network.

Likewise, when NPTv6 #1 loses its default route, it makes itself unavailable as a gateway for other hosts on the Internal Network. NPTv6 #2 attracts all outbound traffic to itself and forwards that traffic through Upstream Provider #2. The mechanism by which NPTv6 #1 makes itself unavailable as a gateway is beyond the scope of this document.

If PE #1 maintains a single-hop eBGP session with NPTv6 #1, the failure of that eBGP session will cause both routes mentioned above to be lost. Otherwise, another failure detection mechanism such as BFD [RFC5881] is required.

Regardless of the failure detection mechanism, inbound traffic traverses the tunnel only during failure periods and outbound traffic never traverses the tunnel. Furthermore, restoration is localized. As soon as the advertisement for CNB #1 is withdrawn throughout Upstream Provider #1, restoration is complete.

Transport-layer connections survive Failure/Recovery events because both NPTv6 translators implement identical translation rules. When a

traffic flow shifts from one translator to another, neither the source address nor the destination address changes.

2.7. Load Balancing

In the architecture described above, site addressing determines load balancing. If a host is numbered from the lower half of the SAB, its address is mapped to CNB #1, which is announced only by Upstream Provider #1 (as part of PAB #1). Therefore, during periods of normal operation, all traffic bound for that host traverses Upstream Provider #1 and NPTv6 #1. Likewise, if a host is numbered from the upper half of the SAB, its address is mapped to CNB #2, which is announced only by Upstream Provider #2 (as part of PAB #2). Therefore, during periods of normal operation, all traffic bound for that host traverses Upstream Provider #2 and NPTv6 #2.

Hosts that receive a great quantity of traffic can be assigned multiple addresses, with some from the lower half and others from the upper half of the SAB. The address chosen for any particular flow determines the path of inbound traffic for that flow. For flows initiated outside of the Internal Network, the site influences the probability that a particular address will be used by manipulating the type and number of PAB addresses advertised in DNS.

3. Discussion

This section discusses the merits of the proposed architecture, as compared with other multihoming approaches [I-D.ietf-lisp] [I-D.rja-ilnp-intro]. The following are benefits of the proposed architecture:

Address mapping information is required only at the NPTv6 translator. There is no need to distribute mapping information beyond the boundaries of the multihomed site.

Because only a small number of mapping rules are required at each multihomed site, there is no need to cache these rules.

During periods of normal operation, packets do not need to be encapsulated. Inbound traffic traverses a tunnel only during failure periods and outbound traffic never traverses a tunnel.

The proposal can be realized using a wide variety of existing encapsulation methods. It does not require a new encapsulation method.

The failover/restoration mechanism is localized to a single autonomous system. Once updated routing information has been distributed throughout the autonomous system, the failover/restoration event is complete.

Benefit can be derived from incremental, partial and even minimal deployment.

The cost of the solution is born by its beneficiaries (i.e., primarily the multihomed site and secondarily multihomed site's upstream provider).

The following are disadvantages of the proposed architecture:

By modifying IPv6 addresses, this architecture violates the end-to-end principle.

The load balancing capabilities described in this memo may not suffice for all sites. Those sites might be required to fall back upon other load balancing solutions (e.g., advertising multiple prefixes)

The time required to redistribute traffic from one path to another is determined by DNS TTL

4. IANA Considerations

This document requires no IANA actions.

5. Security Considerations

As with any architecture that modifies source and destination addresses, the operation of access control lists, firewalls and intrusion detection systems may be impacted. Also many users may confuse NPTv6 translation with a NAT. Two limitations of NAT are that a) it does not support incoming connections without special configuration and b) it requires symmetric routing across the NAT

device. Many users understood these limitations to be security features. Because NPTv6 has neither of these limitations, it also offers neither of these features.

6. Acknowledgements

Thanks to John Scudder and Yakov Rekhter for their helpful comments, encouragement and support. Special thanks to Johann Jonsson, James

Piper, Ravinder Wali, Ashte Collins, Inga Rollins and an anonymous donor, without whom this memo would not have been written.

7. References

7.1. Normative References

[RFC1034] Mockapetris, P., "Domain names - concepts and facilities", STD 13, RFC 1034, November 1987.

[RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.

[RFC3582] Abley, J., Black, B., and V. Gill, "Goals for IPv6 Site-Multihoming Architectures", RFC 3582, August 2003.

[RFC3596] Thomson, S., Huitema, C., Ksinant, V., and M. Souissi, "DNS Extensions to Support IP Version 6", RFC 3596, October 2003.

[RFC4193] Hinden, R. and B. Haberman, "Unique Local IPv6 Unicast Addresses", RFC 4193, October 2005.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, June 2010.

[RFC6296] Wasserman, M. and F. Baker, "IPv6-to-IPv6 Network Prefix Translation", RFC 6296, June 2011.

7.2. Informative References

[I-D.ietf-idr-best-external]
Marques, P., Fernando, R., Chen, E., Mohapatra, P., and H. Gredler, "Advertisement of the best external route in BGP", draft-ietf-idr-best-external-04 (work in progress), April 2011.

[I-D.ietf-lisp]
Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "Locator/ID Separation Protocol (LISP)", draft-ietf-lisp-15 (work in progress), July 2011.

[I-D.rja-ilnp-intro]
Atkinson, R., "ILNP Concept of Operations", draft-rja-ilnp-intro-11 (work in progress), July 2011.

Authors' Addresses

Ron Bonica (editor)
Juniper Networks
Sterling, Virginia 20164
USA

Email: rbonica@juniper.net

Fred Baker
Cisco Systems
Santa Barbara, California 93117
USA

Email: fred@cisco.com

Margaret Wasserman
Painless Security

356 Abbott Street
North Andover, MA 01845
USA

Phone: +1 781 405 7464
Email: mrw@painless-security.com
URI: http://www.painless-security.com