INTAREA
Internet-Draft
Updates: 6296 (if approved)
Intended status: Experimental
Expires: October 15, 2012

R. Bonica
Juniper Networks
F. Baker
Cisco Systems
M. Wasserman
Painless Security
G. Miller
Verizon
W. Kumari
Google, Inc.
April 13, 2012

Multihoming with IPv6-to-IPv6 Network Prefix Translation (NPTv6)
draft-bonica-v6-multihome-03

Abstract

RFC 6296 introduces IPv6-to-IPv6 Network Prefix Translation (NPTv6). By deploying NPTv6, a site can achieve addressing independence without contributing to excessive routing table growth. Section 2.4 of RFC 6296 proposes an NPTv6 architecture for sites that are homed to multiple upstream providers. By deploying the proposed architecture, a multihomed site can achieve access redundancy and load balancing, in addition to addressing independence.

This memo proposes an alternative NPTv6 architecture for hosts that are homed to multiple upstream providers. The architecture described herein provides transport-layer survivability, in addition to the benefits mentioned above. In order to provide transport-layer survivability, the architecture described herein requires a small amount of additional configuration.

This memo updates RFC 6296.

Status of this Memo

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 translator 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 translator.

Section 2.4 of [RFC6296] describes an NPTv6 architecture for sites that are homed to multiple upstream providers. While that architecture fulfills many of the goals identified by [RFC3582], it does not achieve transport-layer survivability. Transport-layer survivability is not achieved because in this architecture, a PA address is usable only when the multi-homed site is directly connected to the allocating provider.

This memo describes an alternative architecture for multihomed sites that require transport-layer survivability. It updates Section 2.4 of [RFC6296]. In this architecture, PA addresses remain usable, even when the multihomed site loses its direct connection to the allocating provider.

The architecture described in this document can be deployed in sites that are served by two or more upstream providers. For the purpose of example, this document demonstrates how the architecture can be deployed in a site that is served by two upstream providers.

1.1. Terminology

The following terms are used in this document:

inbound packet - A packet that is destined for the multi-homed site

outbound packet - A packet that originates at the multi-homed site and is destined for a point outside of the multi-homed site

NPTv6 inside interface - An interface that connects an NPTv6 translator to the site

NPTv6 outside interface- An interface that connects an NPTv6 translator to an upstream provider

2. NPTv6 Deployment

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

2.1. Topology

```
                       Upstream                Upstream
                      Provider #1             Provider #2
                        /     \                 /       \
                       /       \               /         \
                      /     +------+       +------+       \
                 +------+   |Backup|       |Backup|   +------+
                 |  PE  |   |  PE  |.     .|  PE  |   |  PE  |
                 |  #1  |   |  #1  |   .   |  #2  |   |  #2  |
                 +------+   +------+ .   . +------+   +------+
                     |              .      .              |
                     |. . . . . . .          . . . . . .  |
                 +------+                             +------+
                 |NPTv6 |                             |NPTv6 |
                 |  #1  |                             |  #2  |
                 +------+                             +------+
                     |                                    |
                     |                                    |
              ------------------------------------------------------
                               Internal Network
```

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. 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 tunnels. Tunnels are depicted as dotted lines 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 to number its internal interfaces. It also uses that sub-block to number the interfaces that connect it to its customers.

Upstream Provider #1 uses the second sub-block for customer address assignments. We refer to a particular assignment from this sub-block as a Customer Network Block (CNB). In this case, Upstream Provider #1 assigns CNB #1 to the multihomed site. For the purpose of example, assume that CNB #1 is a /59.

In a similar fashion, a Regional Internet Registry (RIR) allocates PAB #2 to Upstream Provider #2. Upstream Provider #2, in turn, assigns CNB #2 to the multihomed site. For the purpose of example, assume that CNB #2 is a /60.

The multihomed site does not number any of its interfaces from CNB #1 or CNB #2. Section 2.3 describes how the multihomed site uses CNB #1 and CNB #2.

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. For the purpose of example, assume that the site obtains a /48 from ULA space.

The site then draws a /59 prefix and a /60 prefix from the SAB. In this document, we call those prefixes SAB #1 and SAB #2. Note that SAB #1 and CNB #1 are both /59 prefixes. Likewise, SAB #2 and CNB #2 are both /60 prefixes. In Section 2.3, the site will map SAB #1 to CNB #1 and SAB #2 to CNB #2. Mapped prefixes must be of identical size.

The site then numbers its resources from SAB #1 and SAB #2. SAB #1 and SAB #2 are the only usable portions of the SAB, because they are the only prefixes that will be mapped to CNB addresses.

During periods of normal operation, hosts that are numbered from SAB #1 receive inbound traffic from Upstream Provider #1. Hosts that are numbered from SAB #2 receive inbound traffic from Upstream Provider

#2. Selected hosts receive inbound traffic from both upstream providers, balancing the load between them. These hosts have multiple addresses, with at least one address being drawn from SAB #1 and at least one address being drawn from SAB #2.

Section 2.7 explains how load balancing is achieved.

2.3. Address Translation

Both NPTv6 translators are configured with the following rules:

> For outbound packets, if the first 59 bits of the source address identify SAB #1, overwrite those 59 bits with the 59 bits that identify CNB #1
>
> For outbound packets, if the first 60 bits of the source address identify SAB #2, overwrite those 60 bits with the 60 bits that identify CNB #2
>
> For outbound packets, if none of the conditions above are met, silently discard the packet
>
> For inbound packets, if the first 59 bits of the destination address identify CNB #1, overwrite those 59 bits with the 59 bits that identify SAB #1
>
> 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 SAB #2
>
> For inbound packets, if none of the conditions above are met, silently discard the packet

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 because it is part of the larger aggregate PAB #1, which is itself advertised.

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.

PE #2 acquires a route to CNB #1 with NEXT-HOP equal to the outside interface of NPTv6 #2. PE #2 can either learn this route from a single-hop eBGP session with NPTv6 #2, or acquire it through static configuration. PE #2 uses this route to enforce source address filtering [RFC2827] on the interface through which it is connected to NPTv6 #2. PE #2 does not advertise this route to CNB #1 to any or its routing peers.

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.

Because both NPTv6 gateways are configured with identical translation rules, and because both PE routers maintain routes to CNB #1 and CNB #1, outbound packets can traverse either NPTv6 gateway. Outbound routing is controlled by the site and therefore, is beyond the scope of this document.

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. Because PE #2 maintains routes to both CNB #1 and CNB #2, it does not discard any traffic from CNB #1 or CNB #2 as a result of source address filtering. 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

Outbound load balancing is controlled by the site and is beyond the scope of this document.

For inbound traffic, addressing determines load balancing. If a host is numbered from SAB #1, its address is mapped into 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 SAB #2, its address is mapped into 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.

Selected hosts receive inbound traffic from both upstream providers, balancing load between them. These hosts have multiple addresses, with at least one address being drawn from SAB #1 and at least one address being drawn from SAB #2. The number of addresses drawn from each range determines how connections originating outside of the multihomed site distribute inbound load.

Recall that all CNB addresses associated with a host are published in DNS. When a remote host initiates a TCP connection, it selects from among these addresses in a relatively random manner. If it selects an address from CNB #1, inbound packets belonging to that connection will traverse Upstream Provider #1 and NPTv6 #1. If it selects an address from CNB #2, inbound packets belonging to that connection will traverse Upstream Provider #2 and NPTv6 #2.

When the multiply addressed host initiates a connection, it associates one of its own addresses with the connection. If the address that it chooses is from SAB #1, that address will be mapped to a CNB #1 address and return traffic will traverse Upstream Provider #1 and NPTv6 #1. Alternatively, if the host selects an address from SAB #2, that address will be mapped to a CNB #2 address and return traffic will traverse Upstream Provider #1 and NPTv6 #2.

3. Discussion

When compared to the multihoming architecture described in Section 2.4 of [RFC6296], the proposed architecture achieves transport-layer survivability at the cost of backup PE hardware and additional configuration. The cost of backup PE hardware is minimal, because backup PE routers are very lightly loaded during periods of normal operation. However, in the example provided above, Upstream Provider #1 must configure the following additional items:

- an interface to the multihomed site on Backup PE #1

- a forwarding tunnel connecting that interface to NPTv6 #2

- either a multi-hop eBGP session between Backup PE #1 and NPTv6 #2, or a static route to CNB #1 on Backup PE #1

Furthermore, if PE #1 does not maintain an eBGP session with NPTv6 #1, Upstream Provider #1 must configure a static route to CNB #2 (as well as CNB #1) on PE #1. However, if PE #1 does maintain an eBGP session with NPTv6 #1, Upstream Provider #1 must configure policy on that session causing it to accept, but not readvertise a path to CNB #2.

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. Acknowledgments

Thanks to Holger Zuleger, 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.

[RFC2827] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.

[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-05 (work in progress), January 2012.

Authors' Addresses

Ron Bonica
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, Massachusetts 01845
USA

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

Gregory J. Miller
Verizon
Ashburn, Virginia 20147
USA

Email: gregory.j.miller@verizon.com

Warren Kumari
Google, Inc.
Mountain View, California 94043

Email: warren@kumari.net