Secure Network Tunneling Using 6rd and GETVPN

Detailed Network Report

September 2019

Prepared by:

Nicholas Russo

CCIE #42518 (RS/SP)

CCDE #20160041

Change History

Version and Date	Change	Responsible Person
20190902 Version 1.0	Initial Draft	Nicholas Russo
20190907 Version 1.1	Technical corrections, spelling, and grammar	Nicholas Russo

Contents

Figures

Tables

No table of figures entries found.

1. Overview

1.1. Problem Statement

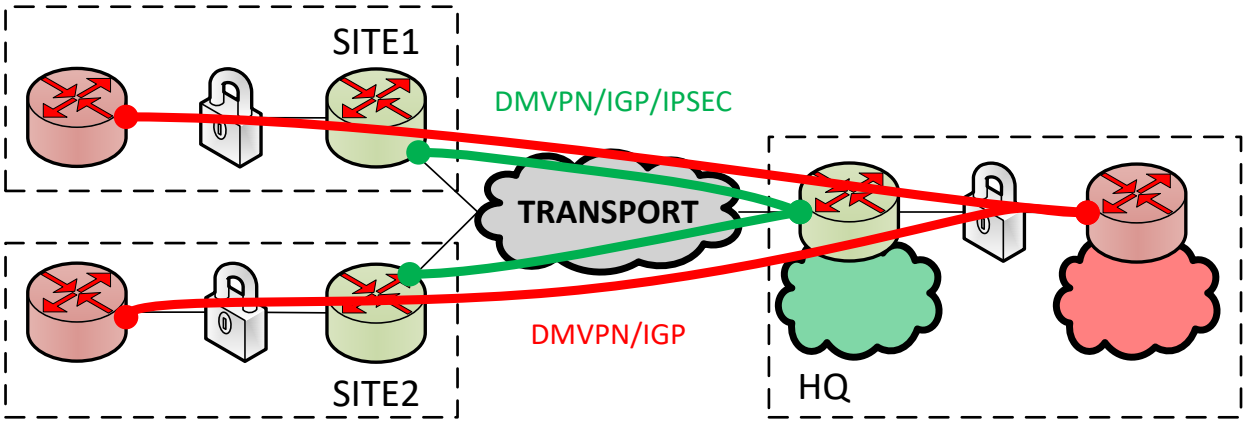
Providing secure and scalable communications over wide area networks (WANs) is a problem with many potential solutions. All of these solutions have a variety of trade-offs, and this document outlines an alternative solution using a mix of technologies. This design was originally developed for tactical military networks which have the following general requirements:

1. **Scalability**: Some networks can grow to thousands of nodes over low-bandwidth, high-latency transports. The overhead of routing and security protocols severely limits scaling. This is especially true for organizations that have many small sites that join and leave the network regularly, as opposed to a smaller number of large, highly capable sites. Scalability is a nuanced topic with many considerations which are discussed later in this document.
2. **Convergence speed**: Tactical networks tend to have relatively low levels of availability given the reliance on satellite communication and complex hardware complement required for transport. After the inevitable failures, resuming regular network forwarding is critical. Layering more protocols into the network creates more dependent events with longer convergence times.
3. **Mobility**: Military units are often reorganized and reassigned based on mission need. Moving between WAN overlay meshes or disparate transport systems is a challenge that requires interworking through a central site. Over low-bandwidth, high-latency networks, this is highly ineffective in facilitating timely communications.
4. **Security**: Given the sensitivity of most military operations, security cannot be an afterthought. Both applications and networks must provide confidentiality, integrity, and availability. Ideally, these security requirements should be met without degrading the scalability, convergence speed, or mobility requirements.

These issues are not limited to tactical military networks, as many of these considerations hold true in the private sector. These multi-layered network tend to have multiple, disparate networks representing various security levels based on data sensitivity. A typical, nondescript tactical network may look similar to the diagram below. Highly sensitive networks may be secured behind dedicated IP encryption devices, while less sensitive networks may use commercial encryption technology. In this document, red routers represent higher sensitivity networks while green routers represent lower sensitivity networks. This document is primarily concerned on explaining low sensitivity case, although the design is valid for both. Often, there is a hub node that provides access to the rest of the network, including centralized data centers or the Internet.

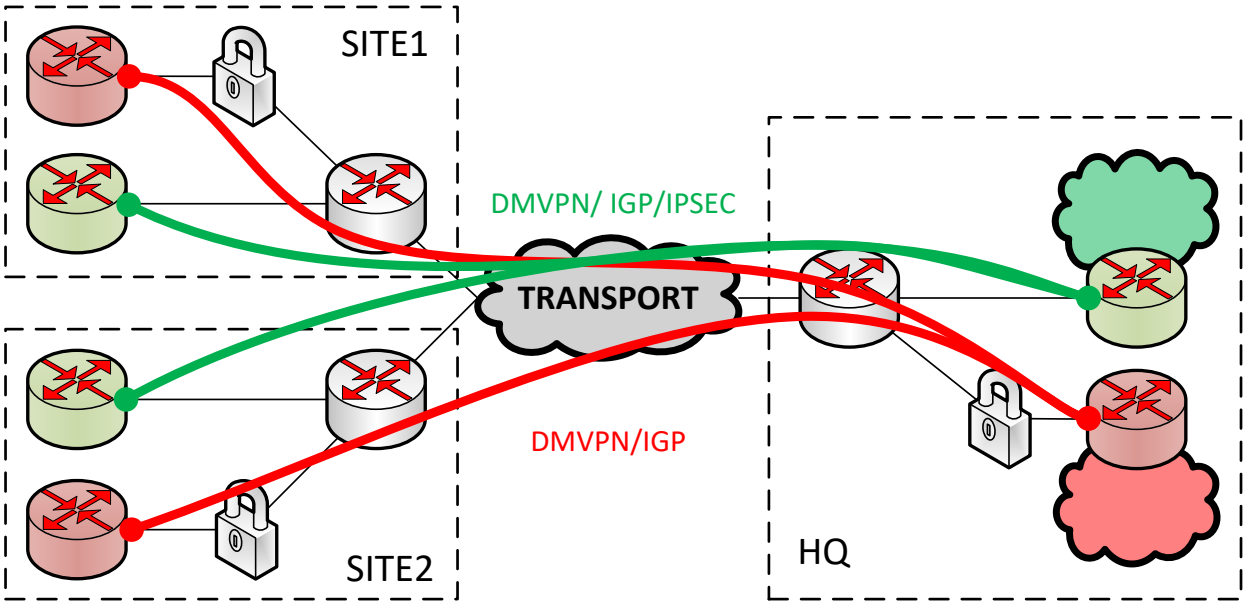
The diagram below depicts a typical tactical network using generic technologies. Often times, Cisco Dynamic Multipoint Virtual Private Network (DMVPN) is combined with Internet Key Exchange (IKE) based IPsec VPNs to build these networks, much like corporate WANs.

Figure 1 - Generic Tactical Network with Hub/Spoke Overlay



Some tactical networks introduce a "black core" for transport with dedicated data-bearing networks for each sensitivity level. The concepts discussed in this document are applicable whether a dedicated transport router exists or not.

Figure 2 - Generic Tactical Network with Hub/Spoke Overlay with Black Core



While hub/spoke WANs tend to meet the aforementioned requirements nicely, they have several drawbacks which are outlined throughout this document.

1.2. Solution Summary

As an alternative to the high-scale, hub/spoke WAN overlay design, this document details how IPv6 Rapid Deployment (6rd) and Group Encrypted Transport VPN (GETVPN) are combined to offer a viable solution. Supplementary technologies such as Network Address Translation 6-to-4 (NAT64) and anycast IP routing are used to further enhance the design.

The solution statelessly provides IPv6 unicast connectivity between all sites in the network. This is accomplished using standard 6rd design and implementation techniques. No routing protocols are required over 6rd tunnel, which provides unlimited routing scalability. There is no concept of a hub which aggregates routing information with 6rd, although the concept of a border relay (BR) is used to reach outside of the 6rd network.

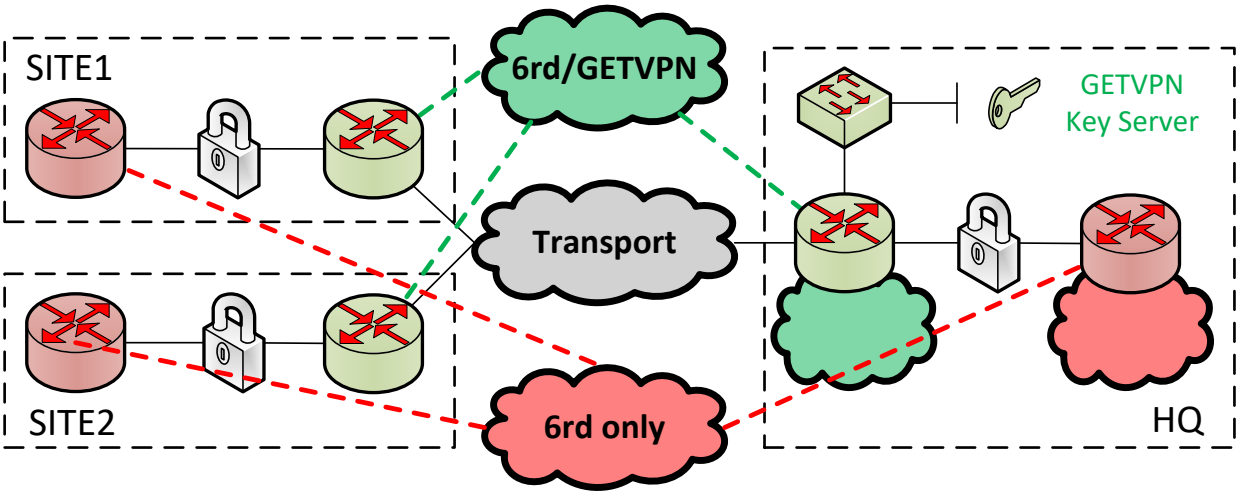
6rd is inherently insecure as it encapsulates IPv6 traffic inside of IPv4 packets, so GETVPN is applied to the 6rd tunnel endpoints. This provides any-to-any connectivity using a single IPsec security association (SA) that is proactively available, rather than reactively constructed on-demand. This improves both convergence speed after a failure and inter-site communications. The GETVPN control-plane traffic is the only persistent traffic over the network, which flows between the group members (GM) and key servers (KS). The KS nodes are likely deployed large, fixed locations.

Anycast IP routing provides rapid failover between 6rd sites and is often deployed to add availability of BRs. GETVPN works seamlessly with this design, ultimately combining network resilience with fast convergence speed.

Acknowledging that some users may need to access resources on the IPv4 Internet (or other internal IPv4 networks beyond the 6rd network), stateful NAT64 provides inside-to-outside access. This will likely need to be combined with Domain Name system 6-to-4 (DNS64). The NAT64/DNS64 components are unnecessary in IPv6-only networks and therefore not strictly required to implement the design.

The diagram below illustrates the major design components. For simplicity, all future design discussions will assume that absence of a black core. Because the placement and upstream design of NAT64 and Internet access is highly variable, it is omitted from this high-level diagram.

Figure 3 - Generic Tactical Network with 6rd/GETVPN Overlay



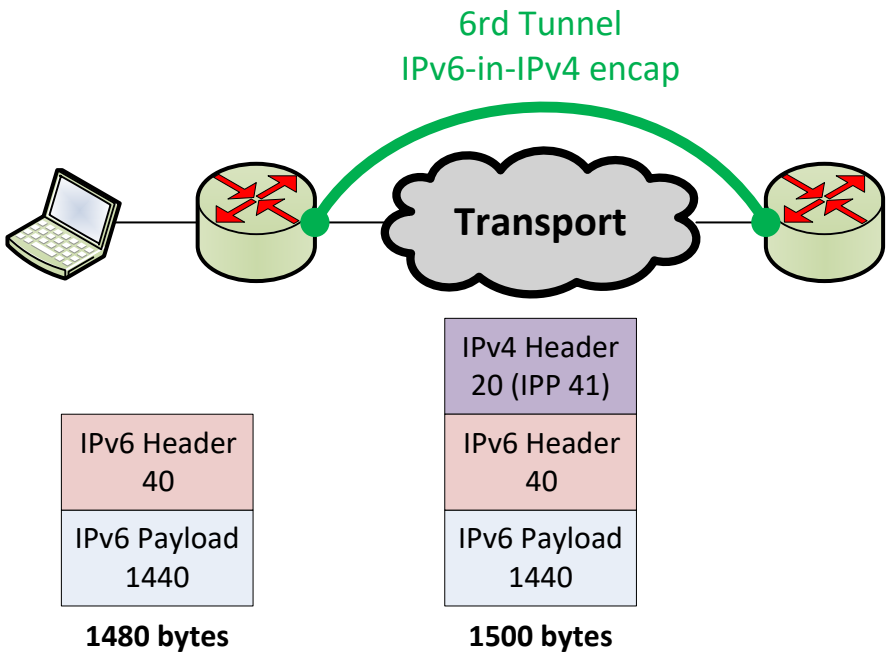
2. Architecture

This section describes the solution in greater technical depth. It examines each individual component in depth, adding new components as it progresses. This document is not a training tutorial on the technologies, but does explain how they work within the context of the design.

2.1. Stateless Tunneling with 6rd

6rd is defined in RFC 5969 and was originally designed to help residential service providers deploy IPv6 to their customers without needing to build IPv6 access and aggregation networks. RFC 4213 describes the basic IPv6-in-IPv4 encapsulation technique, which uses IP protocol 41 and adds 20 bytes of encapsulation given the IPv4 transport header. Like most worthwhile tunneling protocols, 6rd did not reinvent a new encapsulation and simply recycled the standard IPv6-in-IPv4 encapsulation. Not including any additional encapsulation that may be added after 6rd, this reduces the tunnel maximum transmission unit (MTU) to 1480 bytes. The diagram below illustrates this encapsulation.

Figure 4 - 6rd Tunnel using IPv6-in-IPv4 Encapsulation



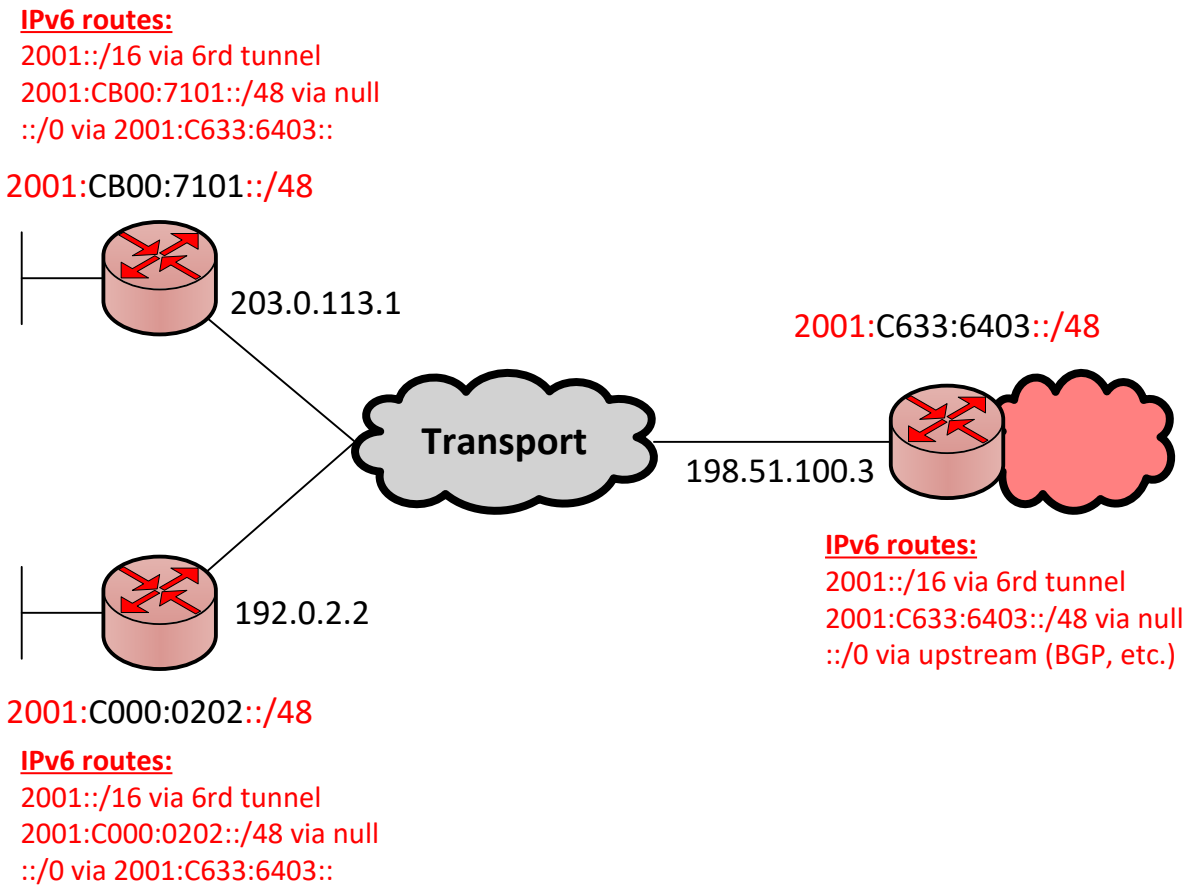
6rd generally replaces 6to4 automatic tunneling, described in RFC 3056. Unlike 6to4, 6rd is not required to use the global 2002::/16 prefix. It also does not have to rely on the global anycast prefix of 192.88.99.0/24 for IPv6 Internet access. Using complex bit-boundary logic, the 6rd prefix can be variable length with only parts of the IPv4 tunnel source being included in the site prefix. For simplicity, this document uses a prefix-length of 0 and a suffix-length of 0, ensuring

all 32-bits of the IPv4 tunnel source used in the site prefix. Organizations can adjust this based on 6rd standard implementation rules.

The real advantage of 6rd in this design is stateless any-to-any connectivity. Consider the simpler case where GETVPN is not required to protect the 6rd tunnel. In tactical environments, this is true when there is a dedicated IP encryption hardware appliance in front of the 6rd tunnel endpoint. In service provider environments, encryption is unnecessary in the first place, especially for consumer-grade Internet service. Every 6rd node has a static route covering the 6rd prefix, which is 2001::/16 in this document. This route does not specify a next-hop IPv6 address, but identifies the 6rd tunnel as the egress interface. The destination IPv4 address is specified in the next 4 bytes of the IPv6 destination, which guarantees delivery to the correct 6rd endpoint. The 6rd customer edge (CE) devices, which represent the remote sites, have a ::/0 default route pointing towards the 6rd BR tunnel address. Last, adding a static null route at each site for the local 6rd prefix is a best practice to avoid sending unnecessary traffic back out over the tunnel. This locally blackholes any traffic for which there is not a more specific prefix.

The design from the 6rd BR up towards the IPv6 Internet is not the focus of this document, but relies on basic IPv6 routing techniques using whichever routing protocols the organization deems appropriate. This is made more complex if NAT64 is introduced, which is discussed later in the document. The diagram illustrates the basic 6rd site design. Transport components such as encryptors, black core routers, and other unnecessary components are removed for clarity.

Figure 5 - General 6rd Site and Prefix Design



The next subsections detail a non-exhaustive set of remote site designs. Attributes from each option can be mixed and matched based on organizational needs.

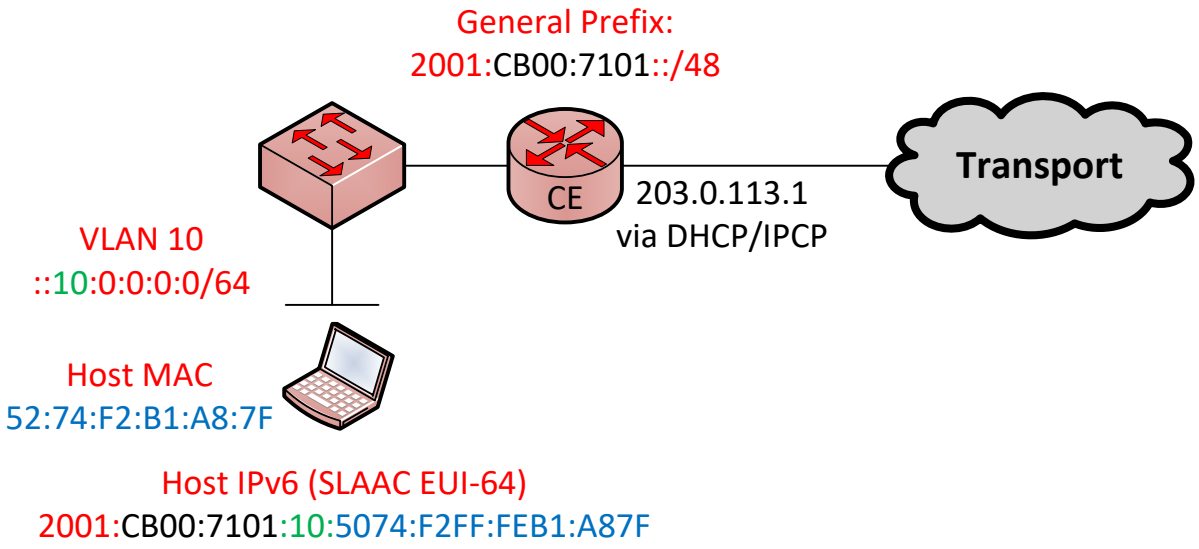
2.1.1. Example Case 1: IPv4 Dynamic Address + Single Router

In some cases, nodes may have their 6rd tunnel source IPv4 addresses allocated dynamically. Common sources include Dynamic Host Configuration Protocol (DHCP) and Point to Point Protocol (PPP) using IP Configuration Protocol (IPCP). This latter option is commonly used in broadband aggregation networks when PPP over Ethernet (PPPoE) is used for connectivity. In either case, the IPv6 prefix used at the site will change based on the IPv4 address, and therefore cannot be statically configured.

Different vendors use different names, but Cisco calls this a “general prefix”. The general prefix is bound to the 6rd tunnel and is automatically updated whenever the 6rd tunnel source changes. As such, all interfaces on the CE should inherit their IPv6 addressing from the general prefix. It follows that clients behind the CE will need to use some kind of dynamic addressing technique. Most commonly, stateless address auto-configuration (SLAAC) is used for addressing. The extended unique identifier 64-bit (EUI-64) specification defined in RFC 4921 uses the Ethernet

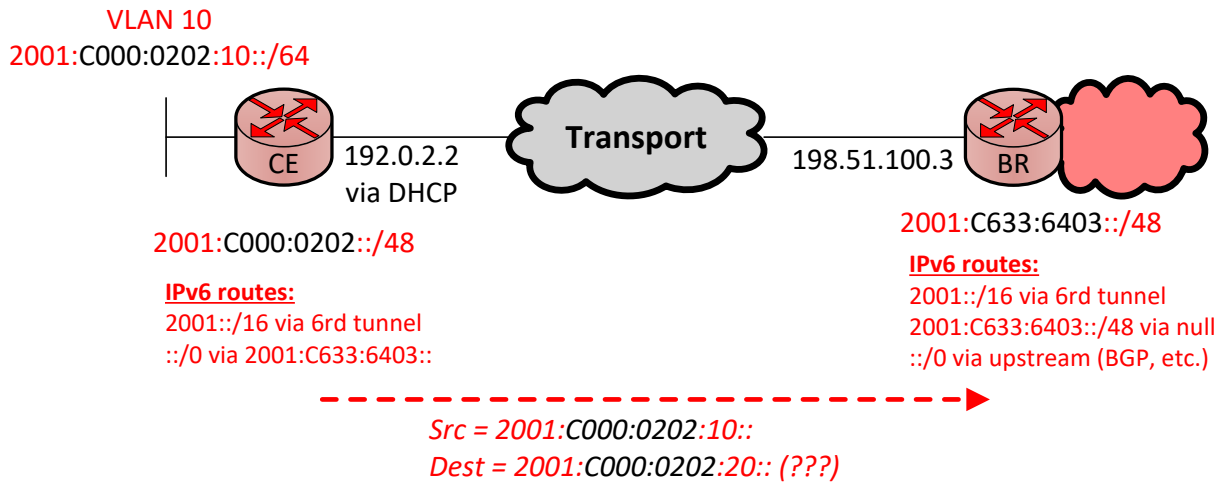
Media Access Control (MAC) address to construct a globally unique IPv6 address. To provide auxiliary configuration parameters, such as DNS servers and domain name, stateless DHCPv6 can be hosted on every CE. The diagram below shows how these dynamic components work in concert. Different colors are used to illustrate the different components that make up the host IPv6 addressing. The 6rd prefix makes up the first part of the address, namely 2001. The next 32 bits represent the embedded IPv4 address, which can vary based on the 6rd configuration.

Figure 6 - IPv6 Generic Prefix, SLAAC, and EUI-64



On Cisco IOS platforms today, you cannot configure a null route to the IPv6 general prefix automatically. This is a minor limitation as the router is smart enough not to send traffic for its local prefix out of the 6rd tunnel towards the BR. Some implementation may not have this built-in safety mechanism. Those platforms that do not may exhibit the behavior illustrated in the diagram below. In those cases, some organizations may be willing to accept this risk by gaining speed during new deployments or reorganizations.

Figure 7 - Potentially Incorrect Forwarding when Null Route is Absent

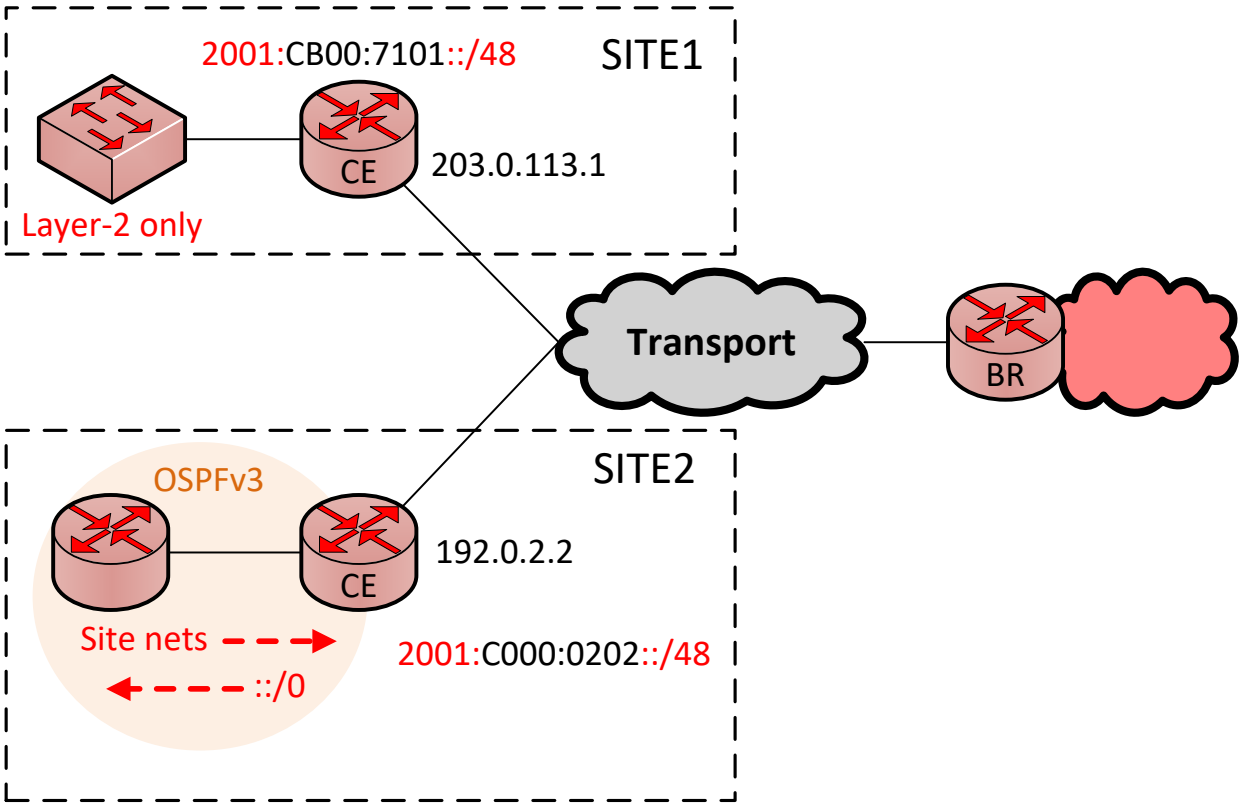


2.1.2. Example Case 2: IPv4 Static Address + LAN Routing

In environments where the IPv4 underlay addresses are configured statically, the general prefix is not required (though it could still be used). The router's interfaces can be statically configured with IPv6 addressing based on the 6rd prefix, rather than referencing the dynamically-computed general prefix. Adding a null route is also easy because the 6rd prefix is known ahead of time. The main drawback of this approach is that if the underlay address ever changes, the entire site must be renumbered to match the new 6rd prefix. This is easily done with automation, and assuming clients are using SLAAC, their IPv6 addressing will be updated as well. In a completely static configuration, such a renumbering is difficult.

In the diagram below, Site 1 is using DHCP and the IPv6 generic prefix method, while Site 2 is using a static IPv4 address assignment. The Site 1 switch is a layer-2 device that inherits an IPv6 address and default gateway via SLAAC. The same is true for all clients at this site. The Site 2 LAN-side router and the 6rd CE run Open Shortest Path First version 3 (OSPFv3) between one another. Any routing protocol can work, but OSPFv3 (RFC 5340) is used for the examples in this document. Local interior gateway protocol (IGP) islands between routers within a site is supported, provided all nodes in the site use the correct 6rd prefix. The 6rd CE can originate a default route into OSPFv3 because it already has a static default route towards the 6rd BR in the routing table. No route redistribution is required.

Figure 8 - Using IGP within a Statically-addressed Site



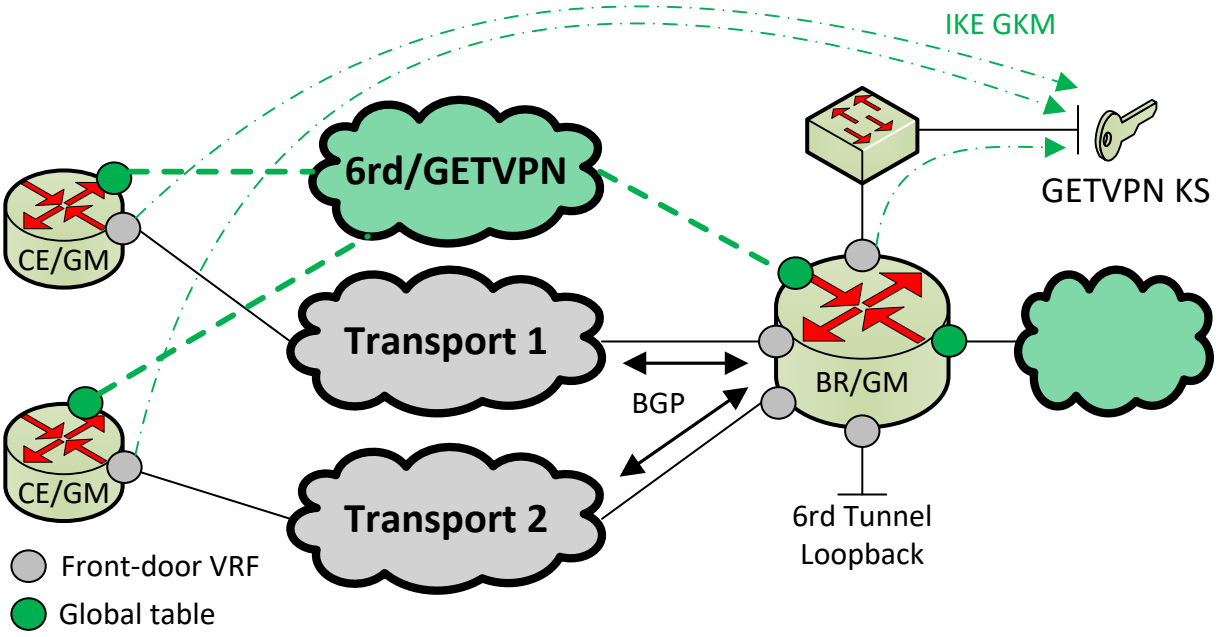
2.2. Security with GETVPN

So far, the examples have illustrated networks that either do not require encryption or provide it using an additional hardware component. In cases where encryption is desired and standalone appliances are unavailable or undesirable, GETVPN is the best choice to secure the 6rd network. The GETVPN KS must be reachable within the IPv4 underlay and therefore must have IPv4 addressing. Each 6rd endpoint, both CEs and BRs, are GETVPN GMs that establish IKE SAs with the KS. This session also provides IPsec transform details such as cipher and hash algorithms to be used. The single shared IPsec SA covers all 6rd endpoints, improving scalability and eliminating the IKE process between individual nodes. This uses a modified IKE process based on group key management (GKM). While GETVPN is Cisco-specific, GKM is standardized in RFC 4046. Although not relevant to the design, designers should always prefer IKEv2 over IKEv1 given its superior security qualities.

To keep things simple, designers should use a simple encryption policy to identify what traffic should be encrypted. Only IP protocol 41 representing IPv6-in-IPv4 should be encrypted. This implicitly denies UDP port 848 which must remain unencrypted as it carries the control-plane traffic between GMs and the KS. GETVPN will only ever see traffic after 6rd encapsulation, so individual applications never need to be classified.

VPN Routing and Forwarding (VRF) assignments can be applied to tunnel sources to further separate underlay from overlay. This places the IPv4 routing in a separate table than the IPv6 routing, and while not strictly required, is a best practice for routing segmentation. Often times, the KS will be placed at a central site along with a 6rd BR. The 6rd BR would therefore have multiple interface in the underlay VRF. At least one of them would connect into the transport network(s), while another connects to the KS LAN. The BR will likely run Border Gateway Protocol (BGP) to exchange routing with the service provider, and the KS LAN will need to be included in the list of advertised networks. When multiple transports are used, the 6rd BR can use a loopback as its tunnel source. The diagram below illustrates the complete design thus far.

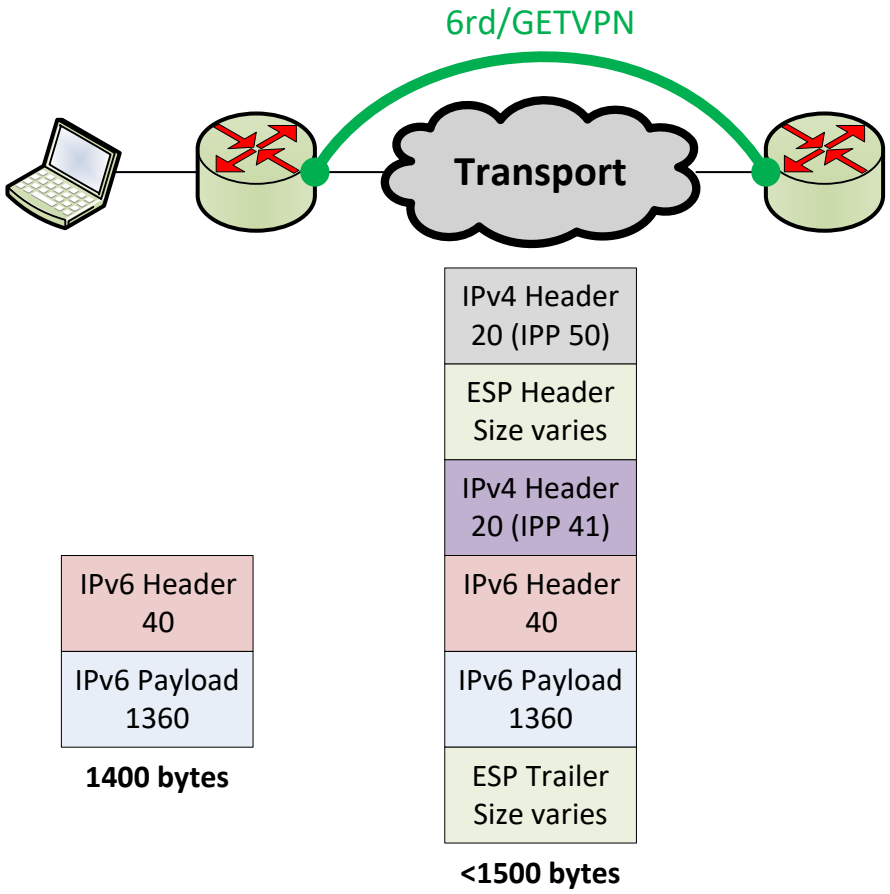
Figure 9 - Combining GETVPN (IKEv2/GKM) and 6rd



There are some minor points worth discussing as well. It is sometimes said that "GETVPN cannot work over the public Internet". Interpreted literally, this is false. GETVPN cannot work across NAT devices. While many Internet transport services use NAT in their transport paths, some do not. These NAT-free transports are suitable for GETVPN. Additionally, GETVPN is sometimes called a "tunnel-less" VPN technology. This is because the inner IP addresses are copied to the outermost IPsec encapsulation. The GETVPN IPsec SA still uses tunnel mode, and therefore adds 20 bytes of encapsulation via an additional IPv4 header. The Encapsulating Security Payload (ESP) header is at least 8 bytes, consisting of a 4 byte security parameter index (SPI) and 4 byte sequence number. It may also embed an initialization vector (IV) to the beginning of the payload, which is cipher-dependent. The ESP trailer may contain padding, next header, and authentication digest fields, making its length variable. A good rule of thumb is to use an MTU of 1400 which accounts for 6rd, IPsec, and additional underlay encapsulations such as PPPoE that may exist. Additional accuracy is possible for specific cipher/hash combinations, but MTU optimization computations are not the focus of this document. See below for a visual depiction of these MTU considerations. The two IPv4 headers contain the same source and destination IP addresses despite being a tunnel mode IPsec SA, which is expected when using

GETVPN. For highly sensitive networks that use hardware-based encryptors, a 1400 byte MTU is likely to still work. These devices tend to use similar (but not identical) cipher/hash algorithms and also utilize tunnel mode IPsec SAs.

Figure 10 - Additional Encapsulation added by GETVPN



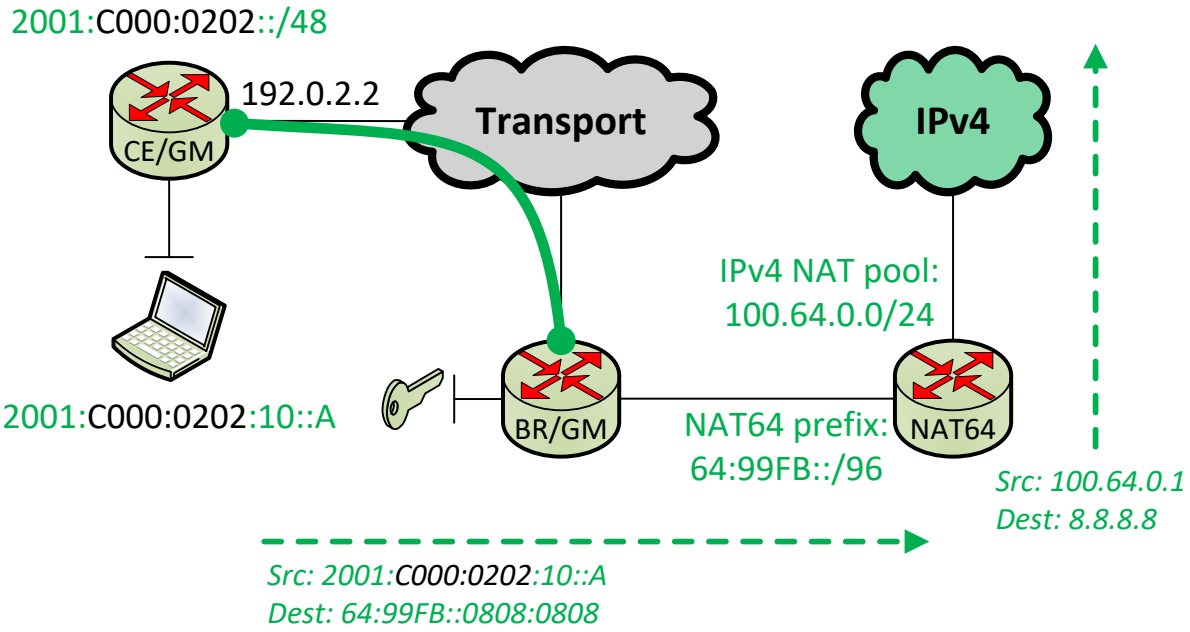
2.3. IPv4 Internet Access with Stateful NAT64

While this 6rd design only supports IPv6 unicast transport to the remote sites, it is likely that there are resources only available on the IPv4 Internet that CEs must access. For the sake of completeness, this document explains some NAT64 design options. NAT64 is defined in RFC 6146 and a common addition to it, DNS64, is defined in RFC 6147.

Before continuing, reference RFC 6052 which covers the NAT64 Well-known Prefix (WKP) of 64:FF9B::/96. This is the destination prefix that IPv6-only clients use to reach IPv4 destinations. Any prefix can be used here, even prefixes shorter than /96, but these examples use the WKP for simplicity. The IPv4 destination is encoded in the last 32 bits of the destination in the WKP. This document uses this WKP as the NAT64 prefix in all illustrations. When traffic destined for the NAT64 prefix arrives at the NAT64 device, it is routed into a local virtual interface that changes the packet to IPv4. The source address will be selected from a configured IPv4 NAT pool, and

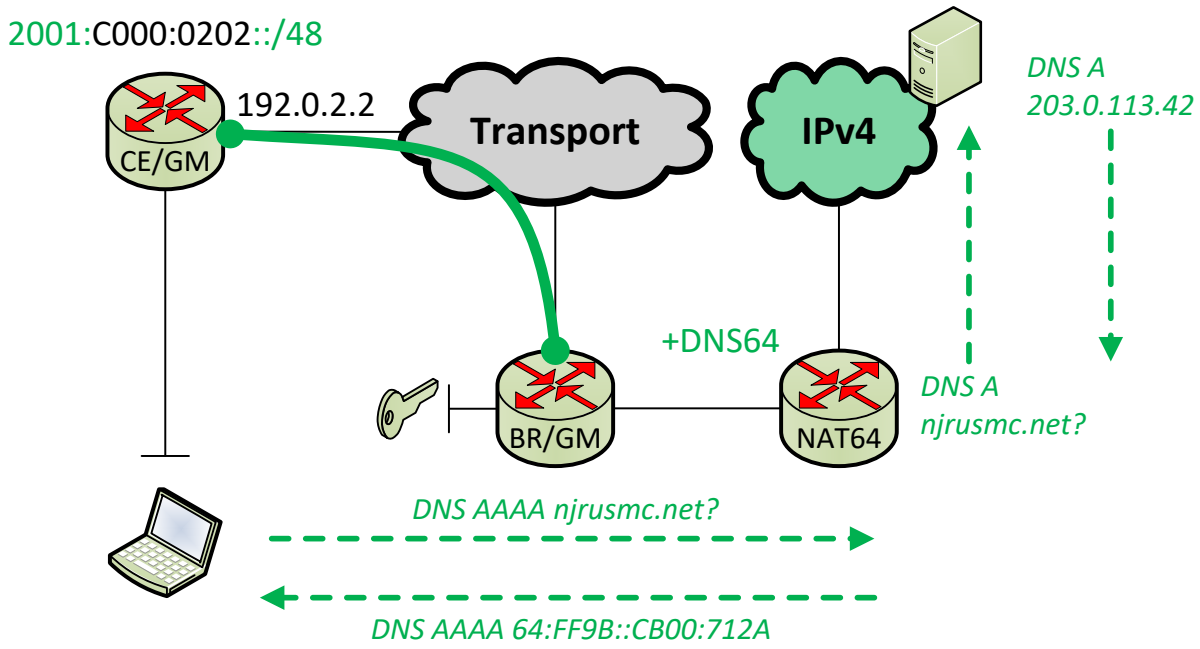
the destination is revealed by stripping the NAT64 prefix from the destination IPv6 address. In the WKP case, it strips the first 96 bits, revealing only the IPv4 destination. The diagram below illustrates the high-level stateful NAT64 operation.

Figure 11 - Using NAT64 to Provide IPv4 Internet Access to IPv6-only Clients



Consider adding DNS64 to the design. This technology allows IPv6-only clients to resolve IPv6 address for hosts on the IPv4 Internet. DNS64, which can be run on a standalone server or integrated into some NAT64 appliances, will convert the DNS AAAA query for a hostname into a DNS A query for consumption on the IPv4 Internet. Upon receiving the DNS response, the DNS64 server translates the A record into an AAAA record using the configured NAT64 prefix. In this way, DNS64 must be aware of the NAT64 prefix in use in order to generate the correct AAAA record responses for clients in the IPv6 network. The diagram below illustrates the high-level DNS64 operation. Note that DNS64 does not have to operate on the same physical device as NAT64.

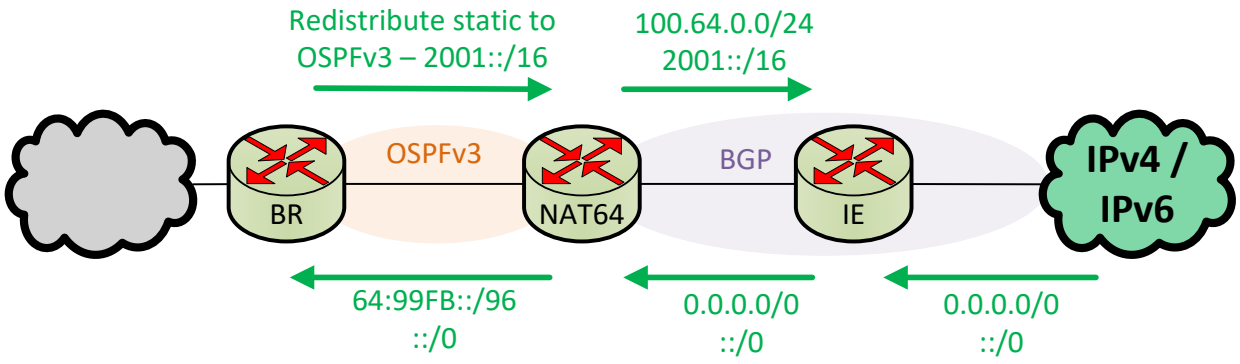
Figure 12 - Adding DNS64 to NAT64 Deployments



2.3.1. Combined NAT64 and Native IPv6 Routing Path

This is the simplest option that effectively dual-stacks the path from 6rd BR to the Internet. The 6rd BR connects directly to the IPv6 side of the NAT64 device. The 6rd BR must learn the WKP and the NAT64 device must learn the 6rd prefix, often using IGP or BGP. Additionally, the NAT64 device runs both IPv4 and IPv6 on its upstream Interface. For IPv6, the NAT64 device advertises the 6rd prefix but not the WKP. This latter point is explained more in the next subsection. For IPv4, the NAT64 device must advertise its IPv4 NAT pool towards the IPv4 Internet. The 6rd CE hosts will be represented by these addresses. To reach the IPv4 Internet, the NAT64 device also needs to receive an IPv4 default route. This routing design is illustrated below. The presence of a dedicated "Internet Edge" router is optional in this design, though would generally be desirable.

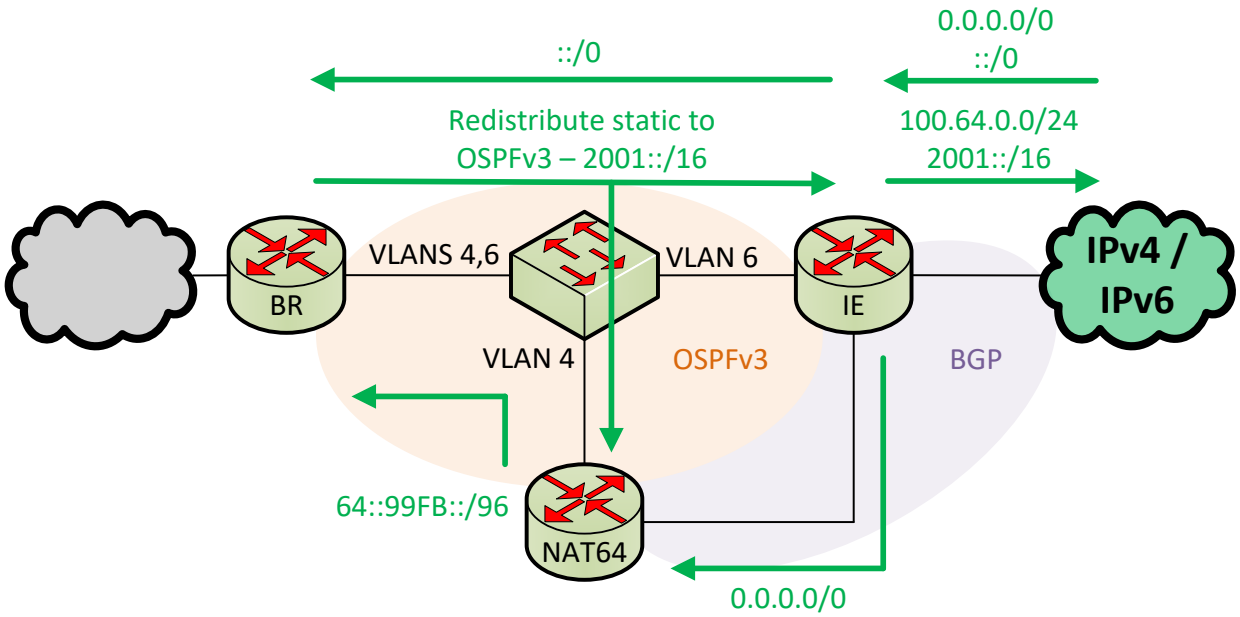
Figure 13 - Traversing a NAT64 Router for IPv4 and IPv6 Internet Access



2.3.2. Separate NAT64 and Native IPv6 Routing Path

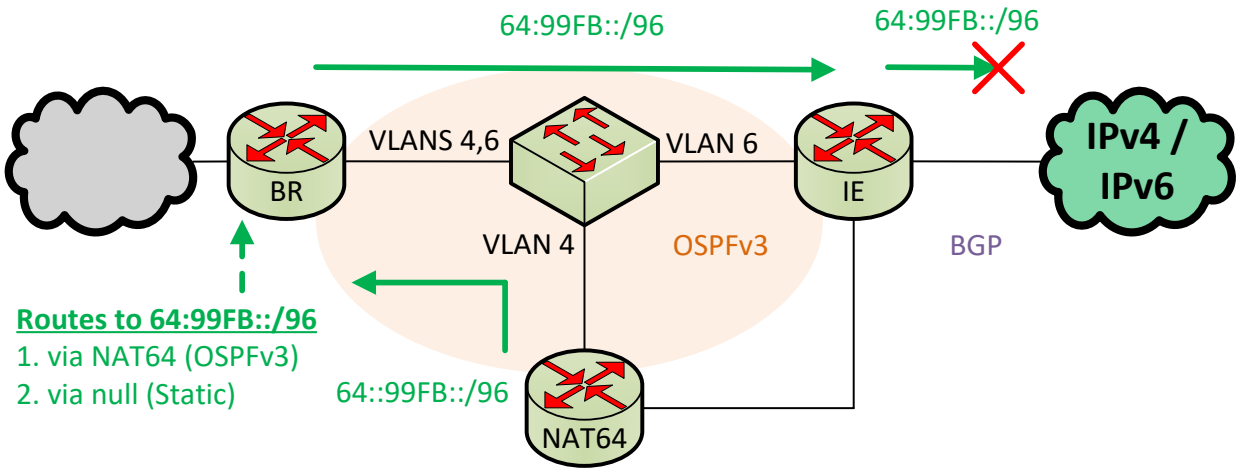
Because stateful NAT64 is a computationally-intensive operation, it is common to use a dedicated NAT64 appliance while the ordinary IPv6 traffic takes a different path. The precise implementation with respect to layers 1 and 2 may vary, but the general design is as follows. The 6rd BR connects to a switch carrying at least two virtual LANs (VLANs). This example uses VLAN 4 for the IPv4 connection to the NAT64 appliance and VLAN 6 for the native IPv6 connection to the Internet edge router. The NAT64 device becomes a "service" in the routing chain and the 6rd routes all WKP traffic towards the NAT64 device. Any IPv6 traffic not destined for the WKP will be matched by the IPv6 default route coming from the Internet edge router. For IPv4 Internet access, the Internet edge router receive an IPv4 default route and passes it down to the NAT64 router. The NAT64 router advertises its IPv4 NAT pool in the reverse direction. The diagram below illustrates this routing exchange.

Figure 14 - Separating NAT64/IPv4 and IPv6 Internet Access



There are some precise caveats with this design. First, the NAT64 prefix (in this case, the WKP) should not be advertised to the Internet. Competent service providers should filter it on ingress, but it is better to handle it internally. Second, the 6rd BR should blackhole the NAT64 prefix if the NAT64 appliance fails. If the 6rd BR does not blackhole the NAT64 prefix, 6rd CE traffic destined towards the WKP will flow towards the default route. Rather than introduce complex firewall filters, use a low-preference static route on the 6rd to prevent any unnecessary forwarding. Using Cisco terminology, a static route to null0 with a high administrative distance (AD) should be configured. See the diagram below to understand these caveats. Note that there are no problems associated with the NAT64 device learning the IPv6 default route via the 6rd BR. This may even be desirable if the NAT64 router needs IPv6 Internet access for software updates or licensing. For cleanliness, this default route learning is not depicted as it is not highly relevant.

Figure 15 - Additional Considerations when using a Separate NAT64 Routing Path



2.4. High Availability

Providing redundancy for the various technologies in the network is accomplish using the generally-accepted design principles for each component. This section details how such enhancements can be made.

2.4.1. Multiple 6rd Routers using IP Anycast

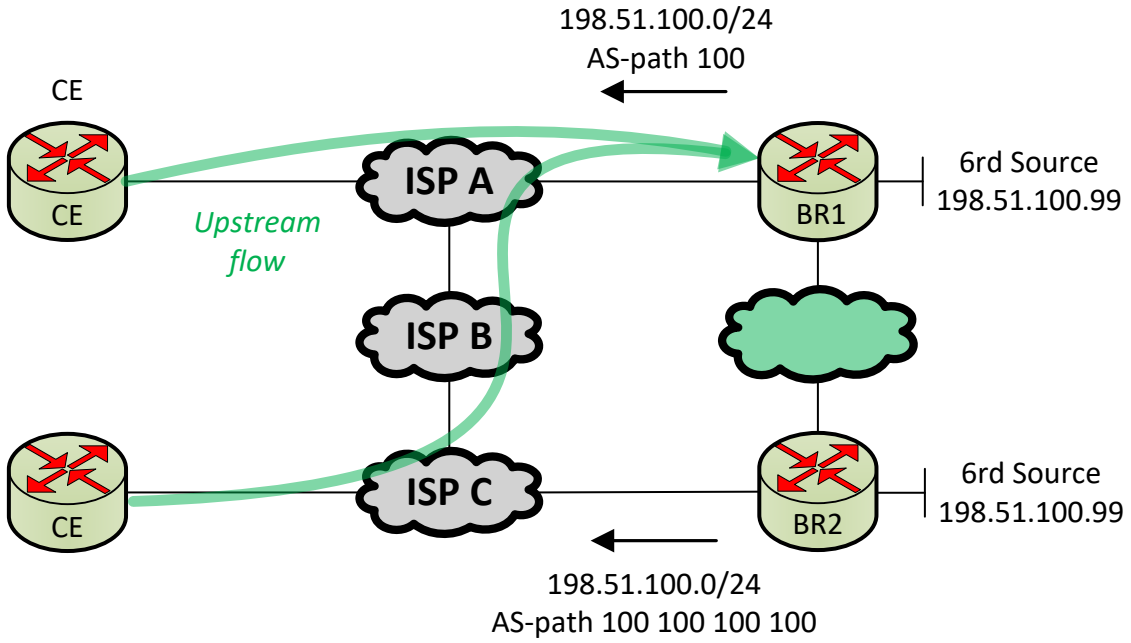
IP anycast is a routing technique that enables the advertisement of overlapping routes into a common routing domain. When traffic is destined for a given prefix, whichever target is closest in terms of routing cost is chosen. In addition to providing optimal traffic flow, this also provides high availability in a stateless way.

Because each 6rd site must be numbered using the IPv4-derived 6rd prefix, one cannot simply add additional CEs or BRs to a site with unique addressing. Doing so would require a different 6rd prefix, implying that the two 6rd routers would not be mutually supporting in the event of node or underlay link failures. Protecting against such failures is the primary goal of a 6rd high availability design, so this document does not explore this ineffective solution.

Instead, sites hosting BRs or large remote sites can use IP anycast. This example illustrates 6rd BRs as using IP anycast for 6rd BR failover is common. Each 6rd BR connects to the underlay network and advertises the 6rd tunnel source, which is commonly a loopback IPv4 address. The prefix-length of this loopback may vary. Over private WAN transports, a /32 could be used, but over the public IPv4 Internet, /24 or larger must be used. The underlay transport then determines what the best path is to reach the 6rd tunnel IP using anycast. Some remote sites may choose

BR1 while others choose BR2 in the diagram below. If BGP is used between the 6rd BRs and transport provider edge routers, standard BGP path selection attributes can be applied. If the 6rd network wants to prefer BR1 over BR2, it can use autonomous system (AS) path prepending, longest-match routing, or multi-exit discriminator (MED) to influence egress traffic. Longest-match routing is the only guaranteed ingress traffic engineering option as the others are only hints. Note that MED only makes sense when all 6rd BRs connect into the same transport AS. The service provider network can use local-preference to influence egress traffic towards the 6rd BRs as well. The diagram below illustrates a stable network following this design. Also note that the 6rd network can disable certain 6rd BRs for maintenance by making the path undesirable or filtering the tunnel source prefix entirely.

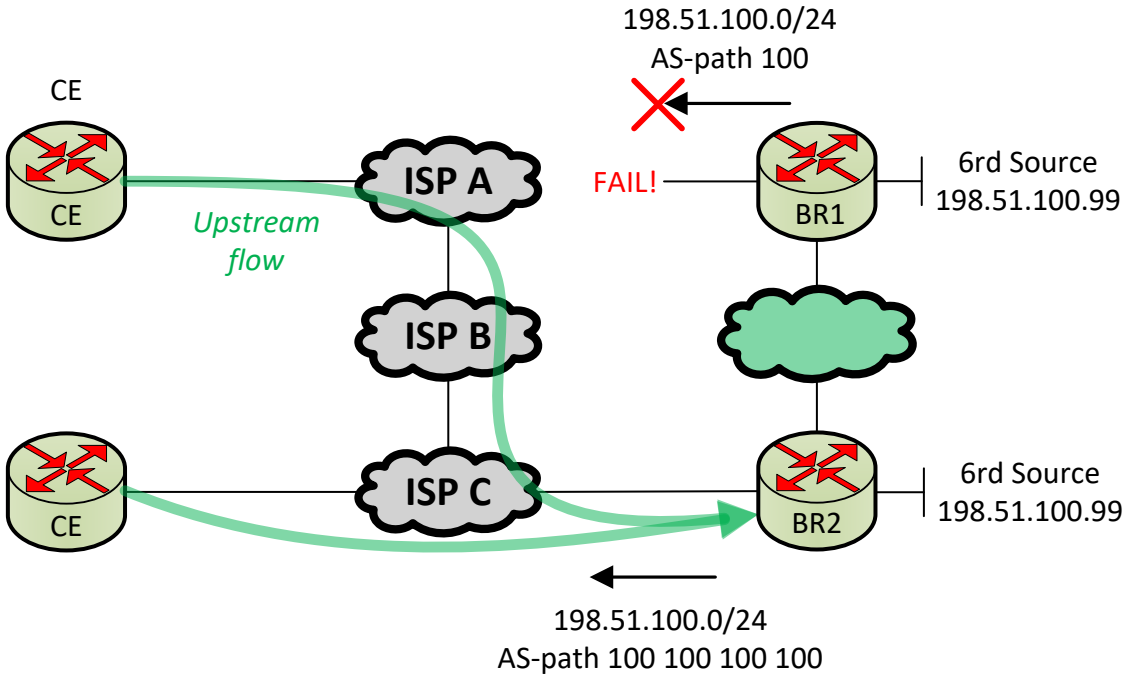
Figure 16 - Using IP Anycast for 6rd BR High Availability



Consider three failure cases:

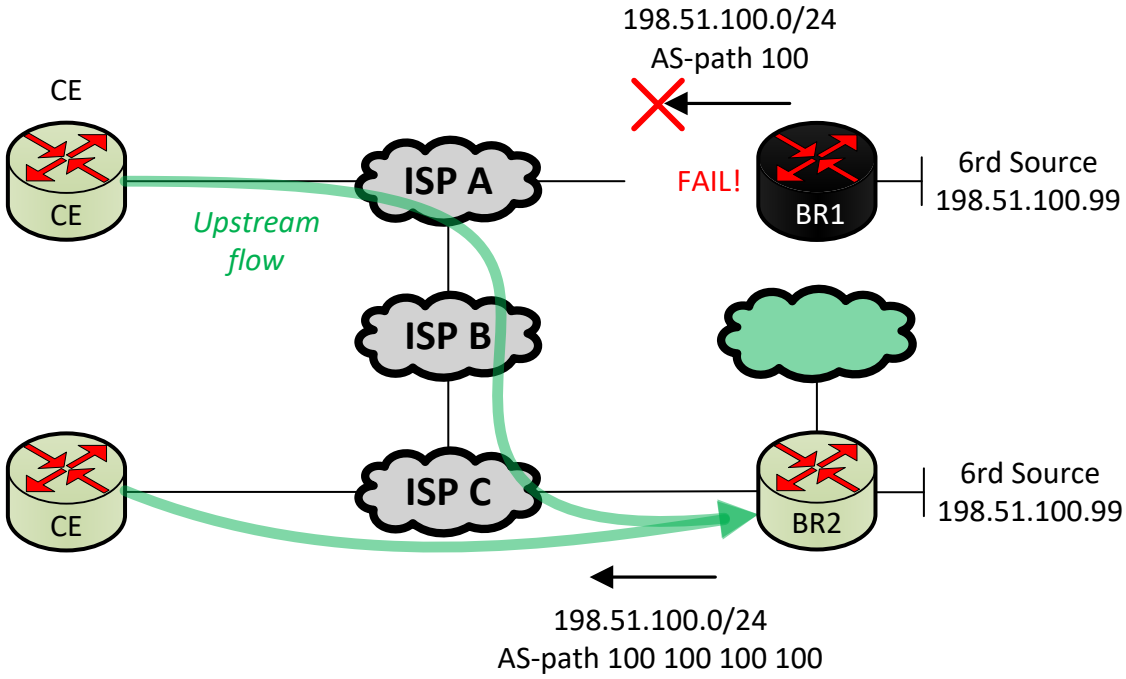
1. **PE-BR link failure**: If the link between 6rd BR and transport PE fails, the transport network will reconverge and select one of the alternative paths to the 6rd BR anycast IP. This is wholly dependent on the service provider's convergence time, which is often dependent upon the failure detection time. No additional routing convergence is required in the overlay.

Figure 17 - BR Anycast Resilience: PE to BR Link Failure



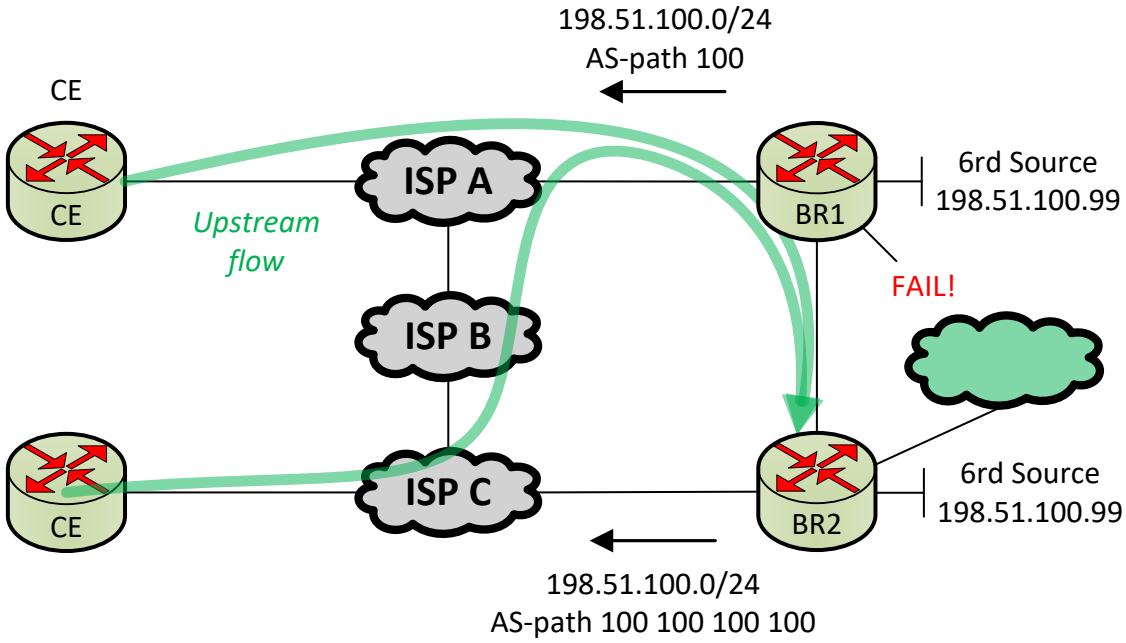
2. **BR node failure**: If the 6rd BR itself fails, and the physical interface to the PE remains up, then convergence will be slower since the PE-BR routing session needs to fail before convergence can begin. If bidirectional forwarding detection (BFD) is used, this additional time penalty can be reduced to a few seconds. BFD would also be useful in the PE-BR link failure scenario described above.

Figure 18 - BR Anycast Resilience: BR Node Failure



3. **BR internal upstream link or node failure**: The 6rd BR will have uplinks to NAT64 appliances, Internet edge routers, firewalls, and other components unrelated to 6rd. Consider a design with two 6rd BRs in a strict active/standby design. If the active 6rd BR is cut off from these resources, it is possible that traffic could be blackholed. All upstream traffic destined for the NAT64 prefix or IPv6 Internet would flow through a site with no uplink. To prevent against this, connect all the 6rd BRs together so traffic can be laterally shuffled between sites. While this is suboptimal, it is better than blackholing. If the uplink failure is expected to be long-term, filter the anycast IP prefix from being advertised from the 6rd BR suffering the failure, removing it from service. The same symptom exists if an upstream node, such as a NAT64 appliance or IPv6 Internet edge router, has failed.

Figure 19 - BR Anycast Resilience: Upstream Link/Node Failure

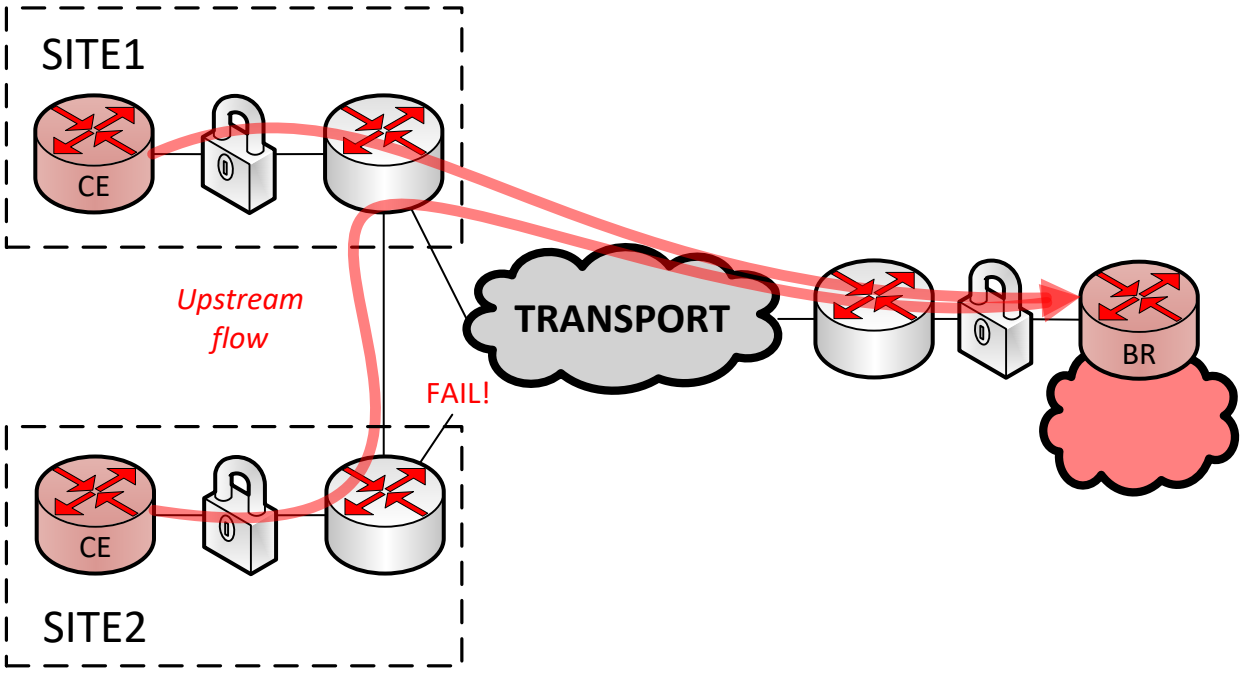


2.4.2. Multiple 6rd Routers using Network Layering

In tactical networks specifically, there could be many layers of networks separated by hardware encryption devices. This separation can be advantageous to 6rd resilience, although it is focused on addressing transport resilience rather than 6rd overlay resilience.

In the diagram below, the untrusted transport network is separated from the trusted data-bearing network using encryption appliances. 6rd is configured only on the data-bearing network. In order for Site 1 and Site 2 to mutually support one another, they must be connected on the transport side. This connection is completely transparent to the data-bearing network, and assuming a transport IGP is present, the transport network can convergence quickly when failures occur. If Site 1 loses its upstream connectivity, it can use Site 2's link instead, and the 6rd network is unaffected. The diagram below illustrates this resilience technique.

Figure 20 - 6rd Resilience with Network Layering



Note that this option does not address the failures described in the previous subsection. Failures occurring within the data-bearing network related to 6rd are unrecoverable. This option is suitable when the PE-CE or PE-BR links are links within the same system as opposed to circuits over a carrier's access network. Also, note that these solutions are not mutually exclusive. At the BR, the anycast IP solution is most appropriate given that CEs may not have multiple routers, multiple links, or the capability to peer BGP with a carrier. At the CE, the network layering solution provides some resilience with little added effort and complexity.

2.4.3. Multiple GETVPN Key Servers

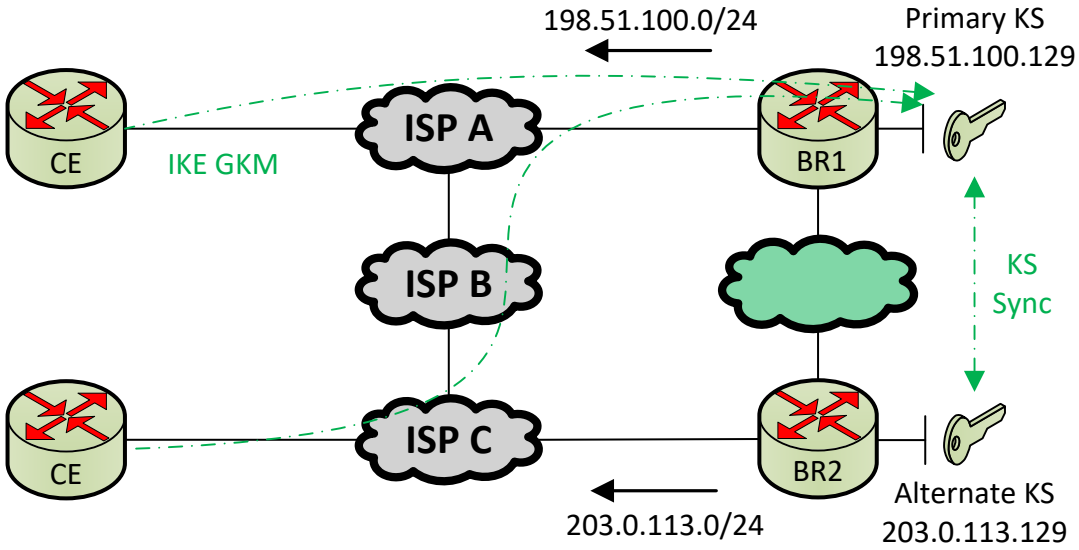
With respect to GETVPN, there are no additional resilience considerations beyond the standard design best practices. Using a continuity of operations (COOP) key server is well-documented and won't be discussed in detail. The GETVPN GMs simply need to identify both key servers in their configuration to form the GETVPN control-plane sessions.

However, consider a design in which a large regional site has two 6rd BRs for high availability along with two key servers. There are two general designs, and note that both require that the KS LAN be placed in the front-door VRF per the standard GETVPN design in this document.

1. **Tie one KS to one BR**: Behind each BR, create a small VRF-aware LAN segment to host the KS and other transport-reachable services, such as management stations. Each KS should use a different IP address as GETVPN key servers will communicate laterally to maintain state. This underlay LAN should be advertised to the PE using BGP, along

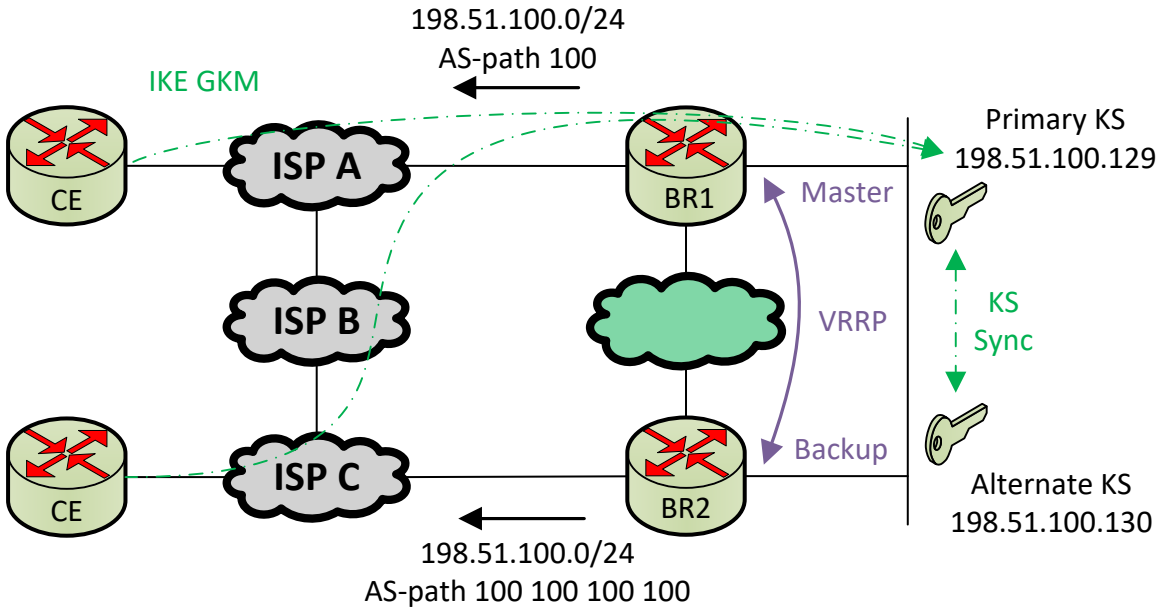
with the 6rd anycast IP. The advantage of this design is simplicity and the ability to separate BRs. This design would work if there were two remote BR sites, provided that BRs were still connected virtually (tunneling, etc.). The disadvantage is that the failure of a BR would mean unreachability for a KS. The other KS failing, even if the other BR is intact, would mean a complete failure of the GETVPN control-plane over time.

Figure 21 - GETVPN Key Server Resilience; One per BR



2. **Both KS and both BR on shared LAN**: Introducing a first hop redundancy protocol (FHRP) such as Virtual Router Redundancy Protocol (VRRP) can improve the availability of GETVPN. In this design, both 6rd BRs will share a LAN segment with both GETVPN KS nodes. One of the 6rd BRs will be the master hosting the virtual IPv4 address serving as the default gateway off the LAN segment. The other serves as the backup. Both 6rd BRs advertise the LAN route to the PE, and to maintain symmetric routing, the VRRP master should advertise the more preferred route. This isn't strictly required unless there is an underlay stateful firewall hosted on the 6rd BR or on the PE-BR link. VRRP should track the PE-BR interface and quickly fail over if that uplink fails. The PE-BR routing protocol should also be tracked. Notwithstanding multi-site layer-2 extensions, this design requires that both 6rd BRs be collocated with access to the same switched network. It offers higher resilience but with increased complexity by introducing VRRP to the design. It is not recommended to extend layer-2 between sites to achieve this; this option is best used when both BRs are collocated and the Ethernet LAN is actually local.

Figure 22 - GETVPN Key Server Resilience; Two KS on Shared BR LAN



2.4.4. Multiple NAT64 Translators

Upstream from the 6rd network, adding multiple NAT64 devices could add even more resilience. This is not directly related to 6rd or GETVPN, but for completeness, this section provides an example of deploying stateful NAT64 in pairs.

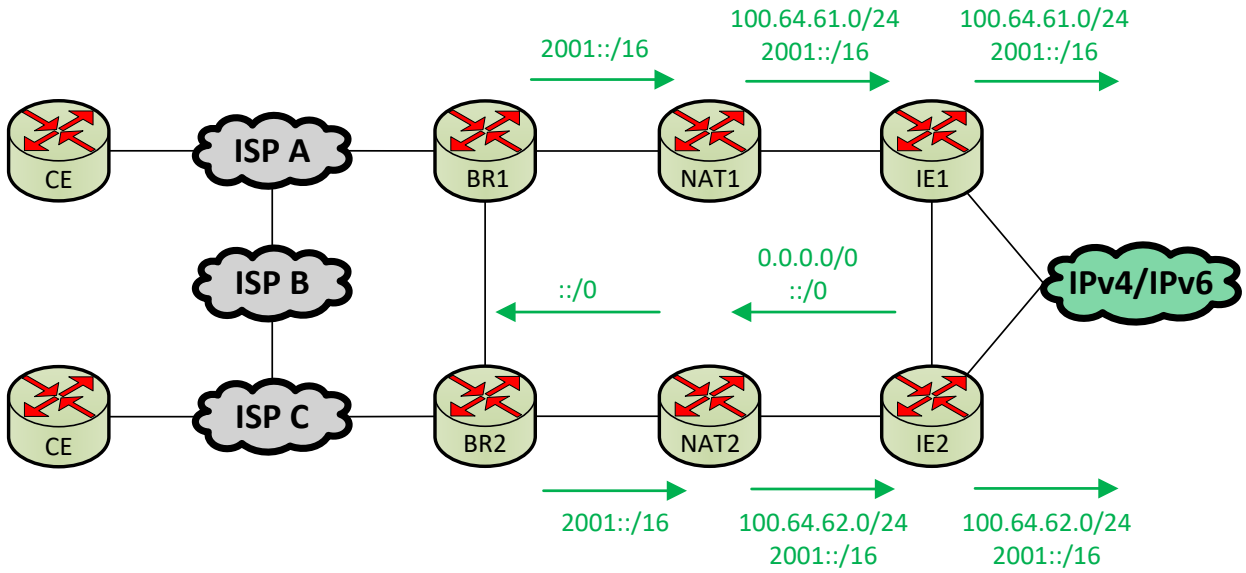
As discussed earlier, NAT64 routers can perform dual-purpose forwarding functions, providing access to both the IPv4 Internet via NAT64 and to the IPv6 Internet via native routing. Alternatively, the NAT64 device can be treated as an appliance and only handle IPv4 Internet traffic. Both of these options can be expanded to operate in high availability designs, but for simplicity, this example shows dual-purpose functionality. Also note that it is possible to deploy multiple NAT64 routers behind a single 6rd BR, or a single NAT64 router behind multiple 6rd BRs.

A general best-practice for any NAT design, be it NAT44 or NAT64, is to use different NAT pools at every Internet access point. This guarantees symmetric routing for return IPv4 traffic coming in from the Internet. This document uses the carrier-grade NAT (CGN) prefix of 100.64.0.0/10 to demonstrate these NAT pools, although in real life, this prefix would not be used in this context. It is important that all NAT64 appliances be configured with the same NAT64 prefix. Likewise, all DNS64 servers, if they exist, should reference the same NAT64 prefix. Both NAT64 routers advertise their NAT64 prefixes and the IPv6 default route into IGP, providing IPv4 and IPv6 reachability to the 6rd network. The diagram below shows a stable network with multiple NAT64 routers.

If symmetric routing for IPv6 ingress traffic is desired, consider using standard BGP AS-prepending, longest-match routing, or Network Prefix Translation v6 (NPTv6) as defined in RFC

6296. This stateless and symmetric NAT technique operates on a whole-prefix basis and would be necessary if creating a lateral link between Internet edge routers is not possible.

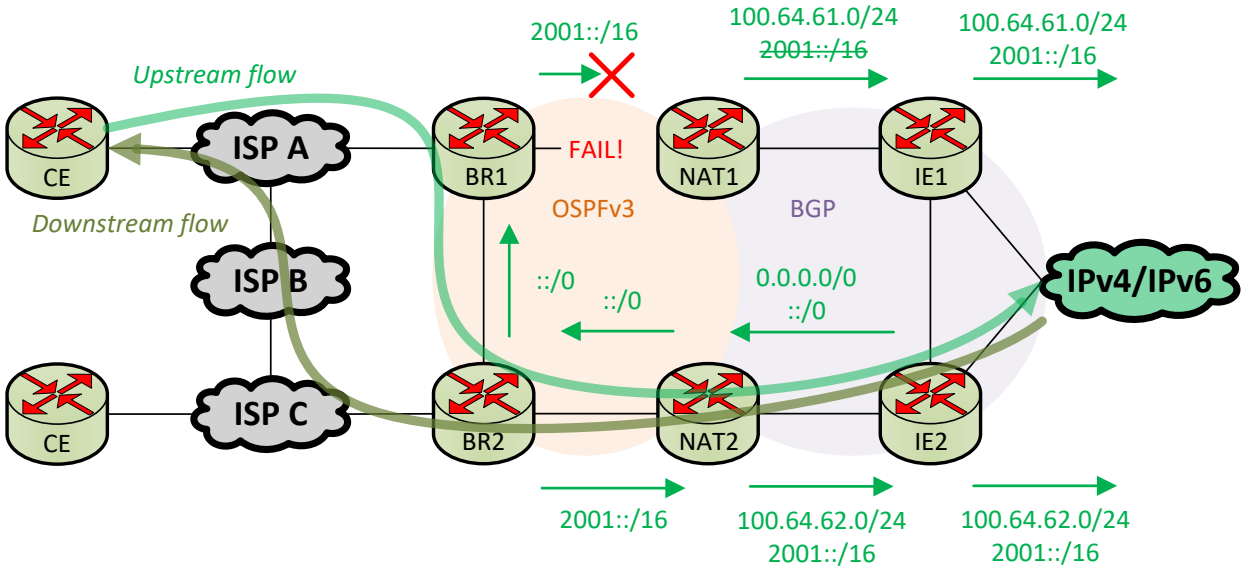
Figure 23 - NAT64 Resilience; General Design



In addition to the general design, consider these failure cases. Note that some of these routing functions (BR + IE, NAT64 + IE, etc.) could be combined into a single device. This paper uses dedicated devices for each one to simplify the explanations.

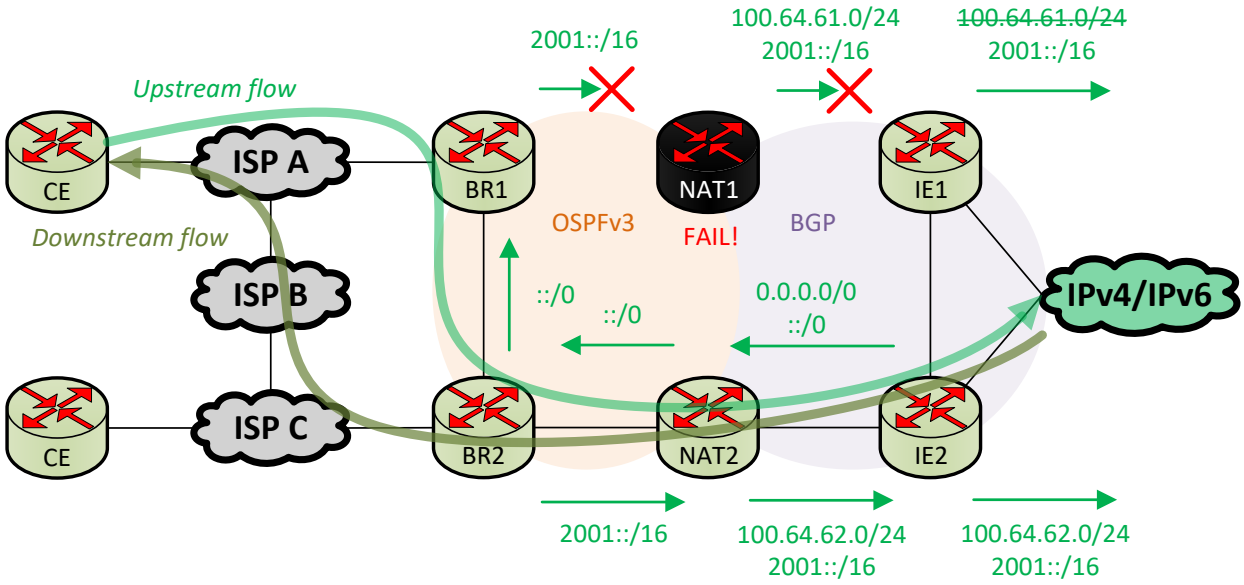
1. **BR to NAT64 link**: If the link between a 6rd BR and NAT64 router fails, the IGP neighbor session between them will fail soon thereafter. Any 6rd CEs still using this 6rd BR for upstream forwarding can continue to use it, assuming the 6rd BRs are still laterally connected. The NAT64 prefix and default route are exchanged over this lateral link using IGP. If the link is expected to be down for an extended period of time, administrators should consider filtering 6rd BR anycast IP from the underlay at this failed node, which will force traffic through the other site. Also note that convergence on the Internet side is unaffected, because all traffic now flows through a single NAT64 appliance and thus has an IPv4 source address for a route the IPv4 Internet already has installed. Last, downstream traffic may be asymmetric across the WAN. The first 6rd BR to receive the downstream flow will forward it into the 6rd network, which may be different than the original ingress 6rd BR.

Figure 24 - NAT64 Resilience; BR to NAT64 Link Failure



2. **NAT64 node failure**: This case has the same behavior as the previous case. In addition to those details, the IPv4 NAT pool advertised by the failed NAT64 router will be withdrawn from the Internet routing tables. This isn't significant at the time of failure, but is a consideration for when the router comes back online. Even if the NAT64 process works again, it may take the Internet some time to learn and install the NAT64 pool after re-activation.

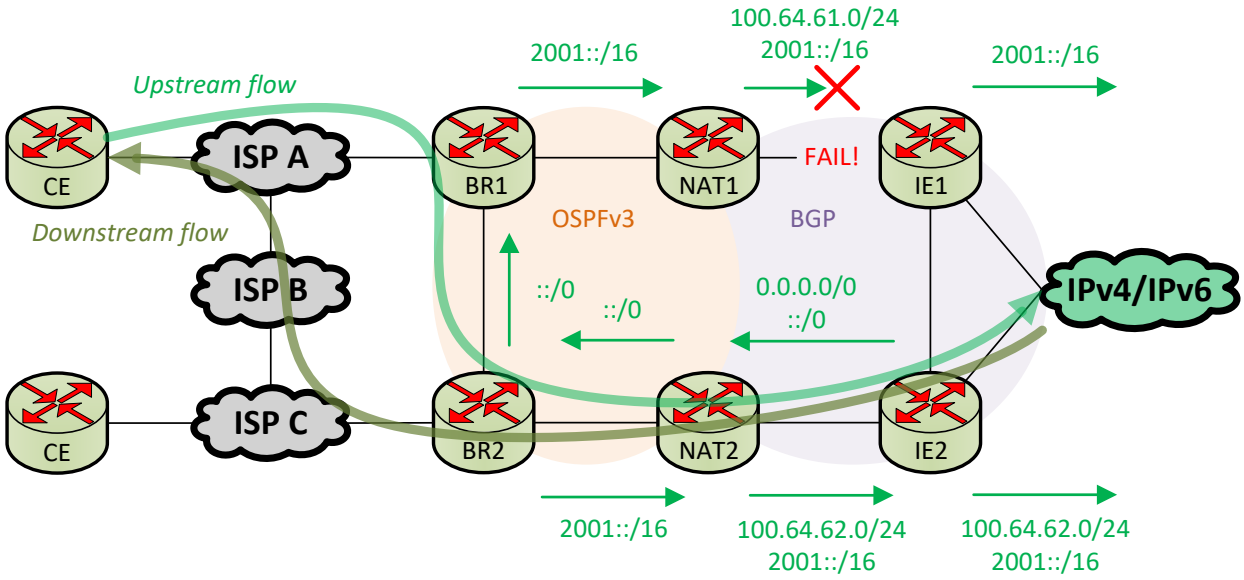
Figure 25 - NAT64 Resilience; NAT64 Node Failure



3. **NAT64 to Internet edge failure**: If the dual-stack IPv4/IPv6 uplink between the NAT64 router and the Internet edge router fails, the NAT64 router will stop advertising the IPv6

default route. It is important not to advertise the NAT64 prefix into IGP. This NAT64 prefix would still be sent into IGP and potentially blackhole upstream traffic destined for the IPv4 Internet. By only relying on the default route, a failure upstream from the NAT64 device will ensure that no traffic is forwarding towards it. This is still imperfect because the ideal solution is to advertise the NAT64 prefix into IGP only if the IPv4 default route is presentation. Most commercial routers do not support conditional route advertisement between address families, although this functionality could be implemented using custom on-box scripts or other vendor-specific route tracking features. Given that the IPv4 and IPv6 default routes are usually received at the same time from the ISP, the imperfection of this simplified solution of relying on the IPv6 default route is minimized.

Figure 26 - NAT64 Resilience; NAT64 to Internet Edge Link Failure



2.5. Limitations

This design has a number of limitations that make it ill-suited in some environments. In summary, networks that host IPv4-only applications, rely extensively on IP multicast, or demand "anyone, anytime, anywhere" plug-and-play connectivity are poor candidates to deploy this design. The design is best deployed in homogenous, greenfield environments over a privately owned transport network where scalability and convergence speed are the primary drivers. To summarize, using Cisco DMVPN or a comparable technology can overcome many of these limitations and may be suitable for a small number of special-purpose sites.

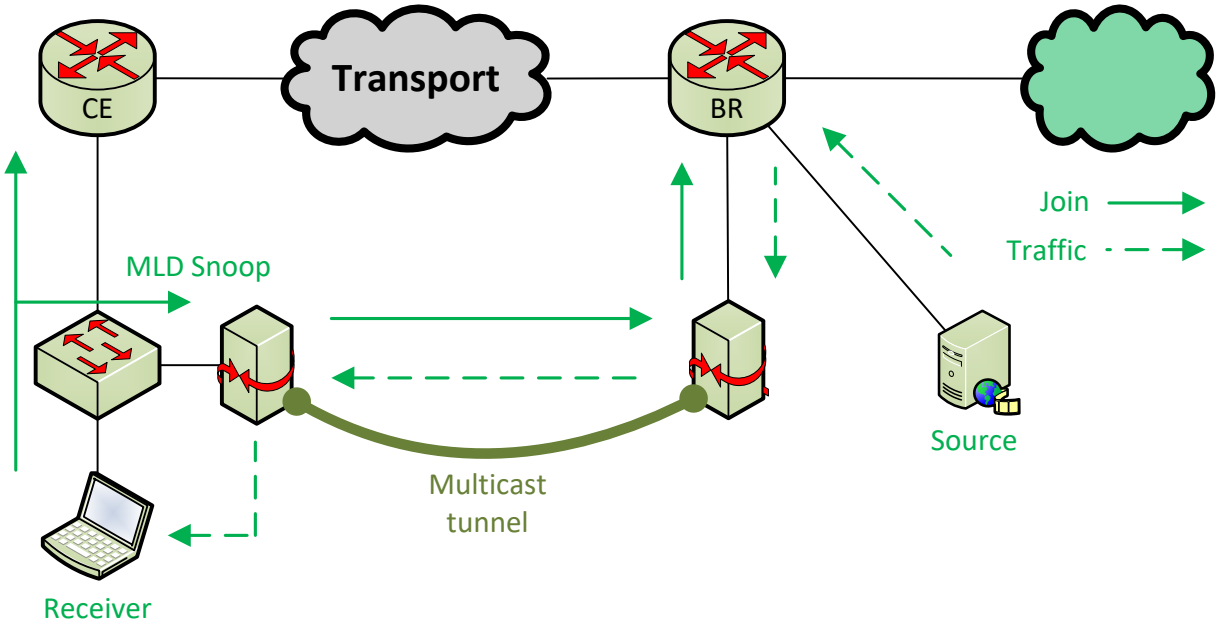
2.5.1. No IPv6 Multicast Support between Sites

Some organizations make extensive use of IP multicast. In tactical networks, this is less commonly used for streaming video and often used for massively distributed applications with an unknown number of participating endpoints. At the time of this writing, 6rd does not support

multicast, although the RFC does permit this extension in the future. An expired IETF draft in the "References" section details one possible solution, but it has not seen widespread adoption.

Deploying a multicast-to-unicast stream converter is one solution to this problem. This could be a custom application or a network device feature. Near the IPv6 receivers, the converter could snoop Multicast Listener Discovery (MLD) membership reports or Protocol Independent Multicast (PIM) join messages, then communicate this interest to another converter near the source. The converter would then issue an MLD membership report into the network. The traffic flow could be encapsulated in a unicast tunnel between converters for transport over 6rd. It could also be translated from multicast to unicast, then back, to avoid adding encapsulation. The diagram below conceptually illustrates the tunneling approach. Effectively, it creates a multicast-only tunnel between the sites, which is automatically wrapped inside 6rd and GETVPN.

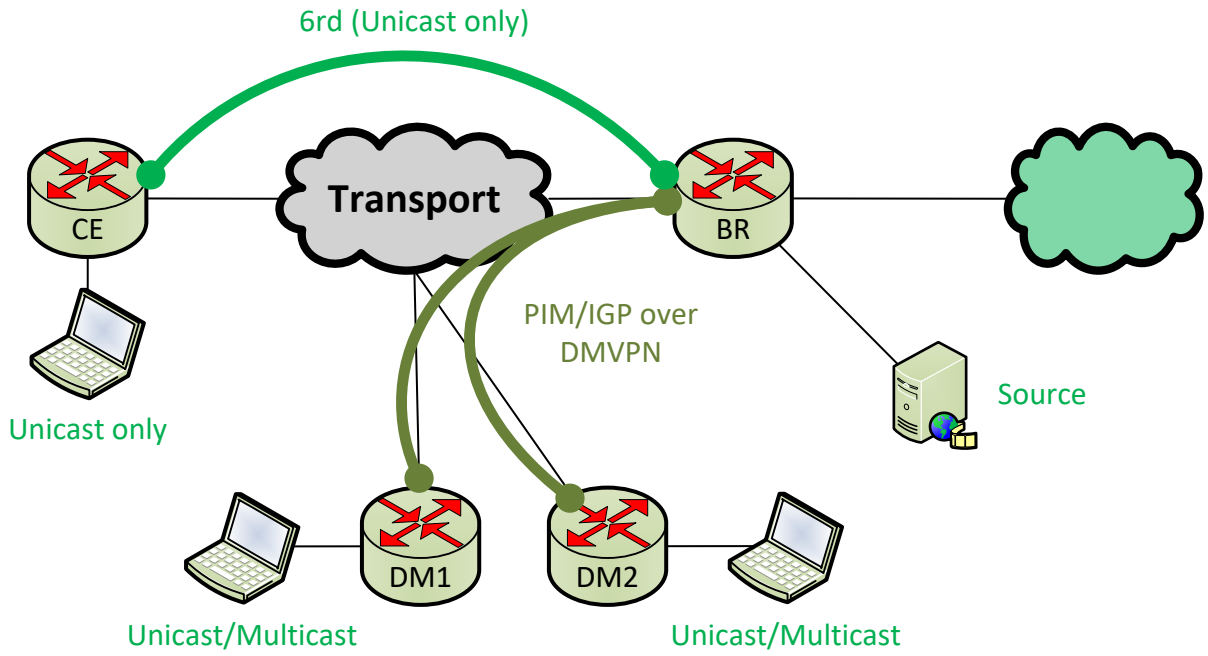
Figure 27 - Multicast Support; Dedicated Conversion Application



Also consider networks that only require multicast transport locally within a site. This is fully supported in the design given that, for any source multicast (ASM), there is a rendezvous point (RP) within each site.

If neither of these solutions are feasible and traditional multicast transport is required to a small subset of locations, consider using a traditional WAN overlay technology to provide it. Cisco's DMVPN and other comparable technologies are good choices here. The existing 6rd BRs can be used as DMVPN hubs, or additional nodes can be deployed for greater separation of duties. The routing protocol used over this new overlay isn't relevant, but ideally should match what is used between the 6rd BR and NAT64 devices. That will reduce redistribution and potential routing loops. This document does not detail DMVPN design considerations.

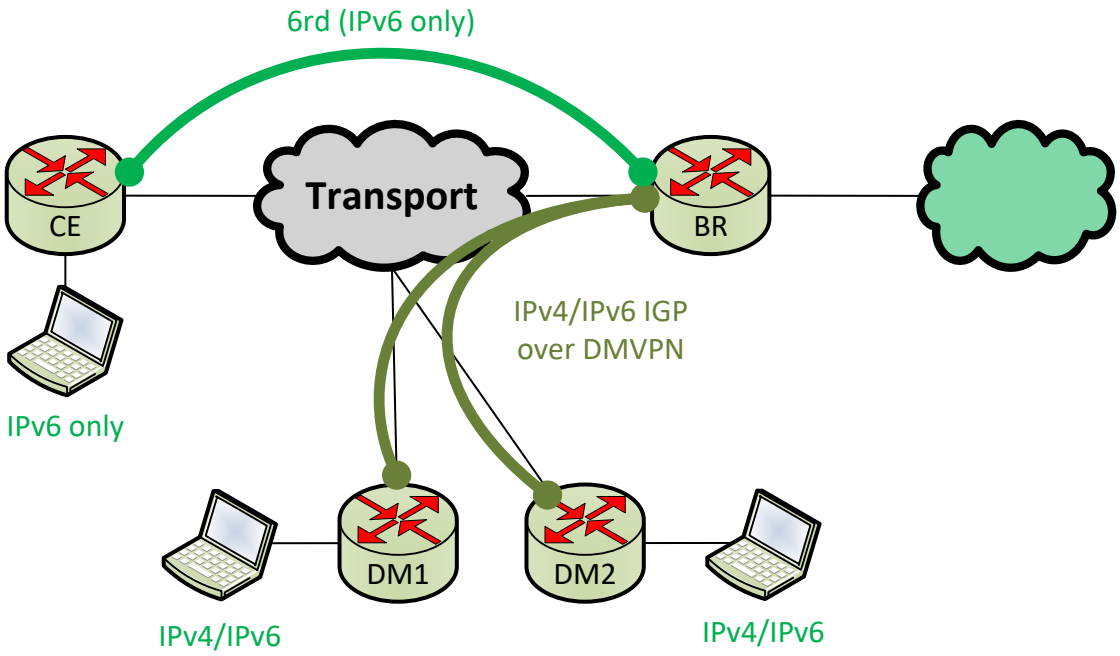
Figure 28 - Multicast Support; Alternative Overlay (e.g. DMVPN)



2.5.2. No IPv4 Support at Remote Sites

Some sites may also require IPv4 support for legacy applications. 6rd can never support this. Much like the multicast requirement described above, the simplest solution is to use DMVPN for this sites instead of 6rd, providing dual-stack support via IGP. DMVPN will support both IPv4 unicast and IPv4 multicast flows, making it a good temporary solution for legacy application support. Once these applications are removed from the network, sites using DMVPN solely for IPv4 support should be migrated to 6rd if possible. More sites on DMVPN means less optimal traffic flows, more network state retained, and lower scale, so it is recommended to use this workaround sparingly.

Figure 29 - IPv4 Support; Alternative Overlay (e.g. DMVPN)



An alternative to using DMVPN is to configure IPv4 support within each site, much like the multicast recommendation in the previous subsection. This is useful for small, distributed legacy applications hosted in close proximity. It could also be used for integrations with extranets that only need access to the resources at a given site and not the greater network. Naturally, this alternative supports IPv4 multicast within a site.

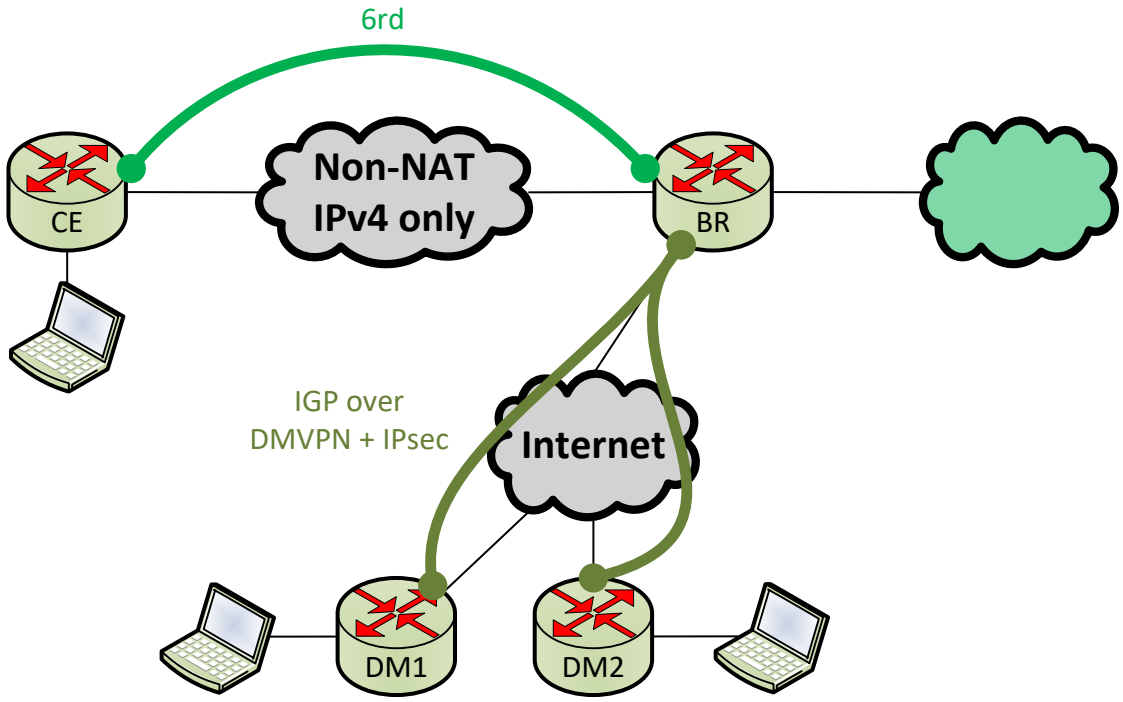
2.5.3. Cannot Traverse NAT-enabled Transports

Remote sites cannot be placed behind any kind of NAT as this breaks the shared SA functionality of GETVPN. If remote sites are placed on the public IPv4 Internet, for example, they must be allocated publicly routable IPv4 addresses. This is sometimes true for directly connected wireline services, but almost never true for wireless or broadband consumer grade Internet access. Note that using the public IPv6 Internet is never supported with 6rd as the underlay must be IPv4 for 6rd to derive its local prefix. For highly sensitive networks that use dedicated IP encryption appliances, this isn't an issue since the 6rd underlay is a LAN segment connected to the encryption device. For less sensitive networks using GETVPN for encryption, it is possible that the 6rd tunnel source ends up being a private IP address.

Like the previous limitations, the best solution is to use DMVPN for sites that exist behind NATs. However, DMVPN must be paired with IPsec using ESP, even if ESP-null is used. IPsec Authentication Header (AH) cannot traverse any kind of NAT, and Generic Routing Encapsulation (GRE), the base encapsulation used by DMVPN, only works over 1:1 NAT. Building standard point-to-point IPsec SAs from spokes to hubs, rather than consuming the shared GETVPN IPsec SA, will still provide security for these sites. This negatively affects scale as the hubs must account for additional IGP neighbors and IPsec SAs.

In addition to NAT problems, ISP-issued private IP addressing creates 6rd problems. In carrier-grade NAT (CGN) environments, multiple CEs could be issued the same IP address in different parts of the world, and the 6rd prefixes would therefore be identical. This would cause connectivity problems to and from these sites with duplicated IPv6 addressing.

Figure 30 - NAT-enabled Transport Support; Alternative Overlay (e.g. DMVPN)



2.5.4. Fixed Remote Site Addressing

All IPv6 addressing behind a 6rd CE is based on the underlay IPv4 address. One cannot arbitrarily add new IPv6 prefixes to a site as they must be derived from the 6rd prefix. If the IPv4 tunnel source changes, the entire site's IPv6 addressing must change. As discussed earlier in the document, using the IPv6 "general prefix" or other comparable features can make provisioning new 6rd nodes easier and more dynamic.

On the topic of extranets providing lateral connectivity to foreign, non-6rd sites, there are two general designs based on the reason for the extranet:

1. **Local access only**: When the foreign node only needs access to local resources, using any routing protocol between local and foreign sites is appropriate. The foreign site doesn't have access to the greater 6rd network, IPv4 Internet, or IPv6 Internet through the connected 6rd site. This might be used for somewhat larger 6rd CEs that host a data center or other useful services.

2. **Full network access**: If the foreign node expects to use the 6rd CE as transport for remote network services, such as Internet access, the foreign node must use addressing from the 6rd prefix. Renumbering could be difficult if done manually, but if good network automation tooling exists and all clients are using SLAAC, this renumbering is not terribly difficult. If renumbering is forbidden, consider using NPTv6 to translate the existing foreign prefixes into site-specific 6rd prefixes. Many NPTv6 platforms do not support application level gateways (ALG), so some applications may not function correctly across it. This is a general consideration for all NATs, not just NPTv6.

3. Complexity Assessment

This section objectively addresses the complexity of each solution using the State/Optimization/Surface (SOS) model. This model was formalized by White and Tantsura *("Navigating Network Complexity: Next-generation routing with SDN, service virtualization, and service chaining", R. White / J. Tantsura Addison-Wesley 2016)* and is used as a quantifiable measurement of network complexity. This section is relevant when comparing this solution to more traditional hub/spoke WAN designs, such as those using DMVPN

3.1. State

State quantifies the amount of control-plane data present and the rate at which state changes in the network. While generally considered something to be minimized, some network state is always required. The manner in which a solution scales, typically with respect to time and/or space complexity, is a good measurement of network state.

This solution has extremely low state in a few key areas:

1. **Control plane**: Traditional hub/spoke WAN technologies tend to scale linearly from the perspective of the hubs. Each spoke will have a fixed number of peers, typically two, while the hubs will have N peers, one per spoke. The word "peer" applies to many different protocols, such as IKE, IPsec, IGP/BGP, and tunnel management (in the case of Cisco DMVPN, this is next-hop resolution protocol or NHRP). The 6rd solution has a fixed number of IPv6 static routes at each site, typically three, and the 6rd design maintains no state with any peers. If GETVPN is used, each GM has one or two IKE SAs to the key servers, and the key servers scale linearly as this is a hub/spoke control plane. There are no routing protocols over 6rd and the only constraint is how many IKE sessions the GETVPN key servers can handle. Networks not using GETVPN are constrained only by the IPv4 address space (effectively, no constraints whatsoever).
2. **Network growth:** Just like in the commercial space, networks are growing quickly as technology is made more available and affordable. IPv4 has reached the point of exhaustion, and while 6rd does rely on IPv4 underlay networks, it consumes a single IPv4 address per site. Expansion of the 6rd network, especially in private environments where IPv4 allocation is not a concern, is very easy. Given the control-plane scalability evaluation, growing the network is not difficult.

3.2. Optimization

Unlike state and surface, optimization has a positive connotation and is often the target of any design. Optimization is a general term that represents the process of meeting a set of design goals to the maximum extent possible; certain designs will be optimized against certain criteria. Common optimization designs will revolve around minimizing cost, minimizing convergence

time, minimizing network overhead, maximizing utilization, maximizing manageability, maximizing user experience, etc.

The design is highly optimized for a relatively narrow set of use cases, described at the very beginning of this document. Organizations that require high performance but are willing to part with IPv4 and multicast are good candidates for this design.

In terms of optimization of routing and traffic flow, this design is nearly perfect. Traffic between 6rd endpoints always flows directly between the two points as governed by the underlay routing. There is no concept of a "hub" as the 6rd BR is just an egress point out of the 6rd network, not a point of traffic aggregation or route distribution. Often times in networking, using prefix summarization leads to suboptimal routing given the reduction in state. With 6rd, this is false, as a single static route covers all other sites in the 6rd network without any trade-offs with respect to traffic forwarding. Those trade-offs are instead focused on the wholesale loss of capability (IPv4, multicast, etc.)

3.3. Surface

Surface defines how tightly intertwined components of a network interact. Surface is a two-dimensional attribute that measures both breadth and depth of interactions between said components. The breadth of interaction is typically measured by the number of places in the network some interaction occurs, whereas the depth of interaction helps describe how closely coupled two components operate.

The main drawback of this design is the tight integration between the 6rd underlay and overlay. The 6rd tunnel source heavily influences the 6rd prefix, which is a very deep surface interaction. This is a leaky abstraction; the underlay and overlay networks are not so separate after all. This surface interaction is endlessly broad as it is true for every 6rd node in the network. As such, this design is classified as having very high surface interaction, generally a negative attribute.

Another aspect to surface interaction is dependency. The underlay must be IPv4-capable and generally be NAT-free. Usage of any IPv6 transport, be it the public Internet or a private WAN, or the introduction of NATs along the path, will break the design. The tight coupling with IPv4 makes any possible integration with alternative transports impossible. This is why 6rd is often deployed as a "quick fix" to add IPv6 support to remote sites, versus being a long-term solution as this design proposes.

Appendix A – Acronyms

Acronym	Definition
6rd	IPv6 Rapid Deployment
ABR	Area Border Router
AD	Administrative Distance
AH	Authentication Header
ALG	Application Level Gateway (NAT)
AS	Autonomous System
ASM	Any Source Multicast
ASN	AS Number
BFD	Bidirectional Forwarding Detection
BGP	Border Gateway Protocol
BR	Border Relay (6rd)
CE	Customer Edge router
CGN	Carrier Grade NAT
COOP	Continuity of Operations (GETVPN KS)
DMVPN	Dynamic Multiple VPN
DNS	Domain Name System
ESP	Encapsulating Security Payload
EUI-64	Extended Unique Identifier, 64-bit (IPv6)
FHRP	First Hop Redundancy Protocol
GETVPN	Group Encrypted Transport VPN
GKM	Group Key Management (GETVPN)

Acronym	Definition
GM	Group Member (GETVPN)
GRE	Generic Routing Encapsulation
IGP	Interior Gateway Protocol
IKE	Internet Key Exchange
IP	Internet Protocol
IPCP	IP Configuration Protocol (IPCP)
IV	Initialization Vector (IPsec ESP)
KS	Key Server (GETVPN)
LAN	Local Area Network
MAC	Media Access Control (Ethernet)
MED	Multi-exit Discriminator (BGP)
MLD	Multicast Listener Discovery (IPv6 ICMP)
MTU	Maximum Transmission Unit
NAT	Network Address Translation
NPTv6	Network Prefix Translation for IPv6
OSPF	Open Shortest Path First
PE	Provider Edge router
PIM	Protocol Independent Multicast
PPP	Point to Point Protocol
PPPoE	PPP over Ethernet
RP	Rendezvous Point
SA	Security Association (IPsec)
SLAAC	Stateless Address Auto-configuration (IPv6)

Acronym	Definition
SOS	State Optimization Surface
SPI	Security Parameter Index (IPsec ESP)
VLAN	Virtual LAN
VPN	Virtual Private Network
VRF	VPN Routing and Forwarding
VRF	Virtual Routing and Forwarding
VRRP	Virtual Router Redundancy Protocol
WAN	Wide Area Network
WAN	Wide Area Network
WKP	Well-known Prefix (NAT64)

Appendix B – References

IPv6 Rapid Deployment (IETF RFC 5969)

6rd Multicast Proposal (Expired Draft)

6to4 Automatic Tunneling (IETF RFC 3056)

Border Gateway Protocol (IETF RFC 4271)

Domain Name System 64 (IETF RFC 6147)

Group Key Management (IETF RFC 4046)

Internet Key Exchange version 2 – IKEv2 (IETF RFC 5996)

IP Security – IPsec (IETF RFC 4301)

IPv6-in-IPv4 Encapsulation (IETF RFC 4213)

Navigating Network Complexity (White and Tantsura)

Network Prefix Translation for IPv6 (IETF RFC 6296)

Open Shortest Path First version 3 (IETF RFC 5340)

Operation of Anycast Services (IETF RFC 4786)

Stateful Network Address Translation 64 (IETF RFC 6146)