Reliably Scaling OSPF over Hub/Spoke with Cisco DMVPN

Technical Whitepaper

Version 1.7

Authored by:

Nicholas Russo
CCDE #20160041
CCIE #42518 (EI/SP)

THE INFORMATION HEREIN IS PROVIDED ON AN "AS IS" BASIS, WITHOUT ANY WARRANTIES OR REPRESENTATIONS, EXPRESS, IMPLIED OR STATUTORY, INCLUDING WITHOUT LIMITATION, WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Change History

Version and Date	Change	Responsible Person
20180706 Version 1.0	Initial Draft	Nicholas Russo
20180711 Version 1.1	Clarification around hub capacity	Nicholas Russo
20180817 Version 1.2	Corrections from community	Nicholas Russo
20181127 Version 1.3	Spelling and grammar cleanup	Nicholas Russo
20190221 Version 1.4	Corrections from community	Nicholas Russo
20190902 Version 1.5	Spelling and grammar cleanup	Nicholas Russo
20200819 Version 1.6	Spelling and grammar cleanup	Nicholas Russo
20201205 Version 1.7	Legal disclaimers and cleanup	Nicholas Russo

Contents

Figures

Tables

No table of figures entries found.

1. OSPF over Hub/Spoke Overview

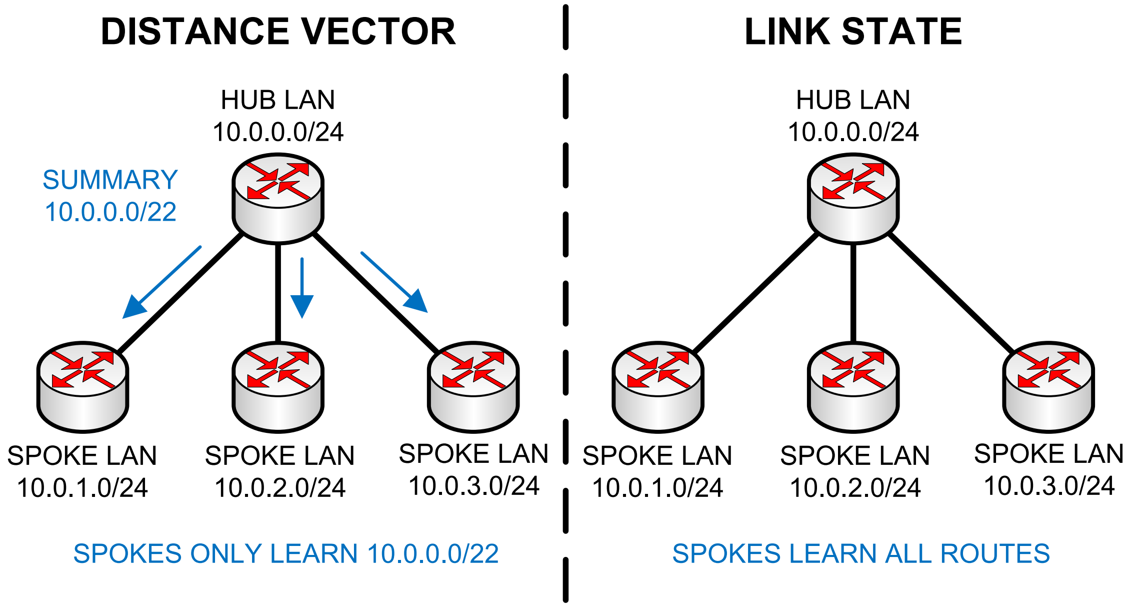
Open Shortest Path First (OSPF) is a link-state interior gateway protocol (IGP) commonly used for routing within an autonomous system (AS). Using a variety of link state advertisements (LSA), the protocol communicates the entire topology to all nodes within a flooding domain. To scale, OSPF areas segment the network into different flooding domains, with the area border routers (ABR) originating LSAs to support routing between areas. This document assumes that the reader has a strong understanding of OSPF.

Link-state protocols are not well suited to most hub/spoke network designs. In these designs, spokes only form routing adjacencies with the hubs and thus receive all their routing updates from said hubs. There is little value in one spoke learning all of the routing information from all other spokes. Short of deploying a variety of network segmentation techniques (different hub/spoke networks in different OSPF areas, etc.), scaling OSPF over these designs has historically been challenging.

Cisco's dynamic multipoint virtual private network (DMVPN) solution is a popular hub/spoke WAN overlay technology used today. It uses a control-plane protocol known as Next Hop Resolution Protocol (NHRP) which is primarily used to resolve underlay addresses for a given overlay address. Using NHRP, spokes register to a statically configured set of hubs, and the hubs dynamically accept spokes. Once this registration is complete, routing adjacencies form across the overlay. Despite not directly exchanging control-plane information, spokes can form dynamic, on-demand tunnels between each other. This is particularly useful for latency sensitive applications such as voice over IP (VoIP) as it removes the hub as a transit node. This whitepaper specifically discusses DMVPN phase 3, and readers are expected to have a basic understanding of this technology.

The diagram below compares OSPF to a generic, unnamed distance vector (DV) routing protocol. The latter minimizes state by its very nature, only relying on information supplied by each neighbor. Hubs can therefore issue one IP summary route which covers all spokes, effectively hiding the entire topology from each spoke. In link-state protocols, this is not inherently possible.

Figure 1 – Generic Comparison between Distance Vector and Link-State over Hub/Spoke



This design was motivated by organizations constrained to using OSPF over their hub/spoke networks, specifically DMVPN phase 3. While a technically inferior choice in almost every way, using OSPF over hub/spoke WANs is influenced by the following business drivers. This is not an exhaustive list but provides some real-world examples:

1. Deployment of a common IGP everywhere to reduce cost through minimal staff training, common troubleshooting, simpler automation (less to change), and less change-related risk overall.
2. Minimization of non-standard protocols and technologies in the network to increase business agility. As it relates to OSPF over DMVPN, this driver makes less sense. OSPF over a generic hub/spoke network, perhaps one of point-to-point overlays in a hub/spoke fashion, would be better supported by this argument.
3. Political pressure to remain conformant within legacy designs and their associated frameworks. This is especially true in massive corporations or bureaucracies where stability, not growth or profitability, is the main organizational driver.

2. DMVPN Anycast Architecture

This section describes the proposed solution to OSPF scalability and availability over DMVPN phase 3. The term "anycast" refers to a design where a common IP subnet exists in more than one place. This technique is a simple, low cost, and stateless way to provide failover capability for networks and applications. It forms the basis of the entire solution discussed herein.

Given the highly technical nature of this solution, a demonstration lab is used to reinforce key points using Cisco IOS devices. The underlay network for this demonstration uses traditional multi-protocol label switching (MPLS) layer-3 VPN (L3VPN) service, although other transports could be used, such as the public Internet.

2.1. Solution Basics

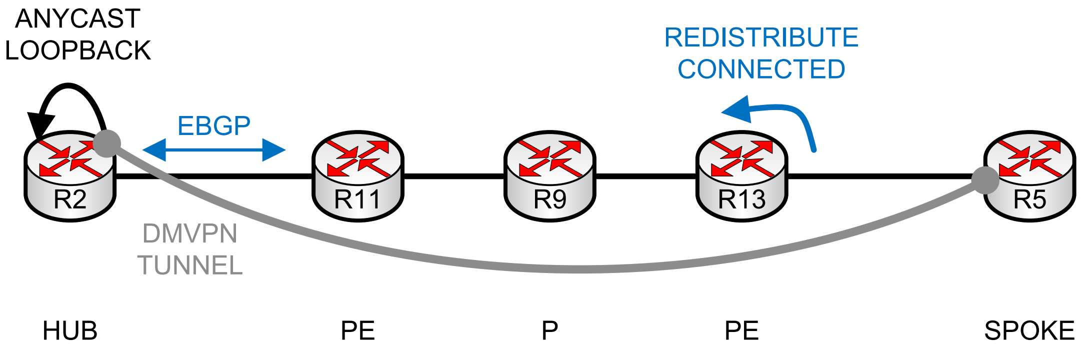
First, consider a single DMVPN hub in a design with no requirement for high availability. The DMVPN hub connects to the service provider at their provider edge (PE) router. The DMVPN hub is therefore a customer edge (CE) device. The DMVPN hub does not have to be a CE as it could be deeper in the customer network, but this is less common. The DMVPN hub has upstream connections into the rest of the network, including the data center.

The DMVPN hub runs external Border Gateway Protocol (eBGP) on its underlay interfaces towards the PE. It advertises its tunnel source network via BGP into the carrier's network so that spokes will be able to send traffic upstream to it. This network is typically a loopback interface on the DMVPN hub. In private WAN environments, the subnet mask of this tunnel source prefix may vary, though /32 is commonly used for IPv4. If the Internet is used for transport, a prefix length /24 or shorter must be used for IPv4 unless the provider handles public IP aggregation upstream in their network. The uplink CE IP address should not be the anycast address because:

1. When adding new hubs for availability, those hubs must use the exact same CE IP, implying the same PE-CE subnet. Few carriers will allow this.
2. Even if the carrier did allow it, changing service providers in the future may require a renumbering, which could be challenging in environments where spokes do not have Domain Name System (DNS) capability.
3. When there are multiple hubs, availability could be broken. If there is a local layer-1 or layer-2 failure at the CE, it is possible that the PE interface remains up, resulting in a black hole as upstream traffic is absorbed by the PE then discarded due to the local CE-side failure.

Spokes that connect into the network, presumably via the same MPLS provider, can use their connected interfaces. It's unlikely that spokes, which are typically small outstations/branch sites, would run BGP to a service provider, especially with Internet as a transport. When MPLS L3VPN is used, carriers often redistribute the connected PE-CE subnet into BGP for transport across the VPN. When this isn't the case, BGP could be used to advertise the PE-CE link to from the CE the PE. However, the aforementioned comments about using a loopback as the tunnel source apply primarily to hubs. The diagram below depicts the network described thus far.

Figure 2 – Basic DMVPN Deployment over an MPLS Underlay



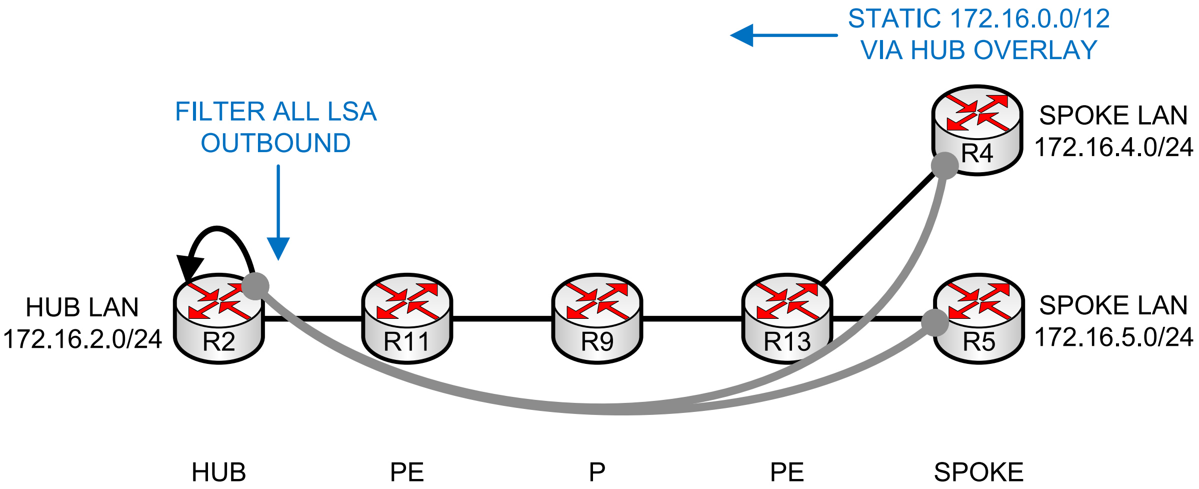
Suppose OSPF is enabled over this tunnel mesh between one hub and several spokes. Assuming the network discussed so far is built according to specification, the result is unspectacular. The hub forms neighbors with all spokes, and each spoke has exactly one neighbor (the hub). All routers have the full topology view for all other routers, along with their corresponding IP routes. While the design is functional, it scales poorly, as there is little value in providing a full topology view to spokes.

To overcome this, two critical but simple modifications are applied:

1. **At the hub:** Enable outbound LSA filtering for all LSAs. This has the effect of halting all OSPF flooding downstream from the hub to the spokes.
2. **At the spokes:** Configure a set of static upstream summary routes to the hub's overlay IP address. These routes should cover the main data center subnets (behind the hub) and all of the other spoke networks to which a given spoke must communicate. This provides full reachability within the overlay network.

Ignoring the technologies in play and the configuration just applied, the resulting logic is identical to that enabled by using DV protocols over hub spoke. In DV networks, the hub dynamically sends down a set of summary routes rather than all network state. In the OSPF example, the spokes use a small and fixed set of static routes to achieve the same purpose. Concurrently, the hub sends no network state down to the spokes. Without considering the specific implementation of internal IGP data structures, this allows OSPF to scale similarly to DV protocols. The diagram below explains this point. Note that these configuration modifications are not specific to DMVPN, but are generic to any hub/spoke network running OSPF.

Figure 3 – Optimizing OSPF for Scale over Hub/Spoke

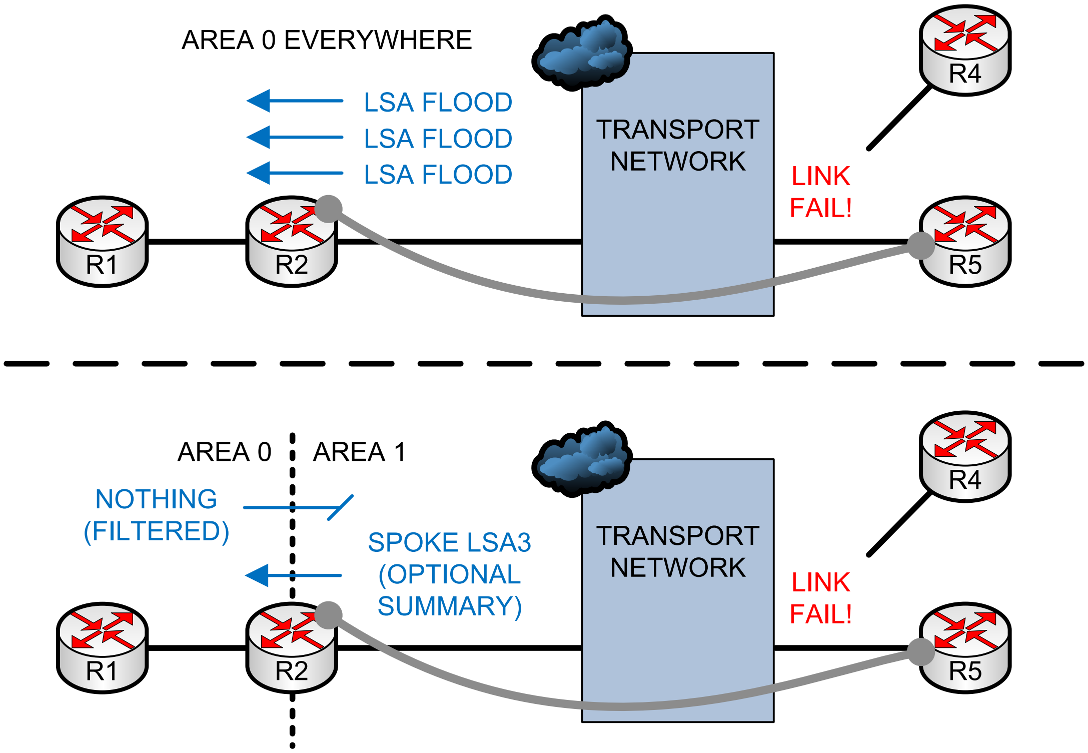


As it relates to area design, it is generally wise to assign the DMVPN tunnel to a non-zero area while the hubs upstream links to the rest of the network are placed in area 0. Additionally, the area is best left as a standard area, not a form of stub area. These decisions are explained below.

1. The hub, as an ABR, can't send any information down to the spokes due to LSA filtering. Making the area a form of stub has no positive impact. Instead, it could add complexity by restricting redistribution and interoperability with third-party devices that do not support these special area types.
2. In large-scale hub/spoke WANs, at least some spokes regularly join and leave the network. The rest of the OSPF domain upstream from the DMVPN hubs should be minimally affected. That is to say, the upstream backbone routers should not need to re-flood, re-compute, and reinstall routing information. This is true in any OSPF network, WAN or otherwise, but the WAN is unique in that the transport cost is high and available bandwidth is low by comparison.
3. In single hub networks, OSPF prefix summarization at the ABR from the non-zero area into area 0 is low risk and could further reduce state retention in the backbone. This technique does not require trading off optimal routing or increasing the risk of blackholes/loops, as summarization often does.

The diagram below compares the flat area 0 design versus the proposed enhancement. Note that technically LSAs from area 0 are "flooded" into area 1, but since the only area 1 interface has LSA filtering enabling, there is no consequence other than some increased memory requirements at the DMVPN hub. Thus, it makes sense to filter all LSA3s from being sent from area 0 into area 1. This could be achieved by using a totally stubby area, but such a design would unnecessarily restrict future redistribution at the remote sites.

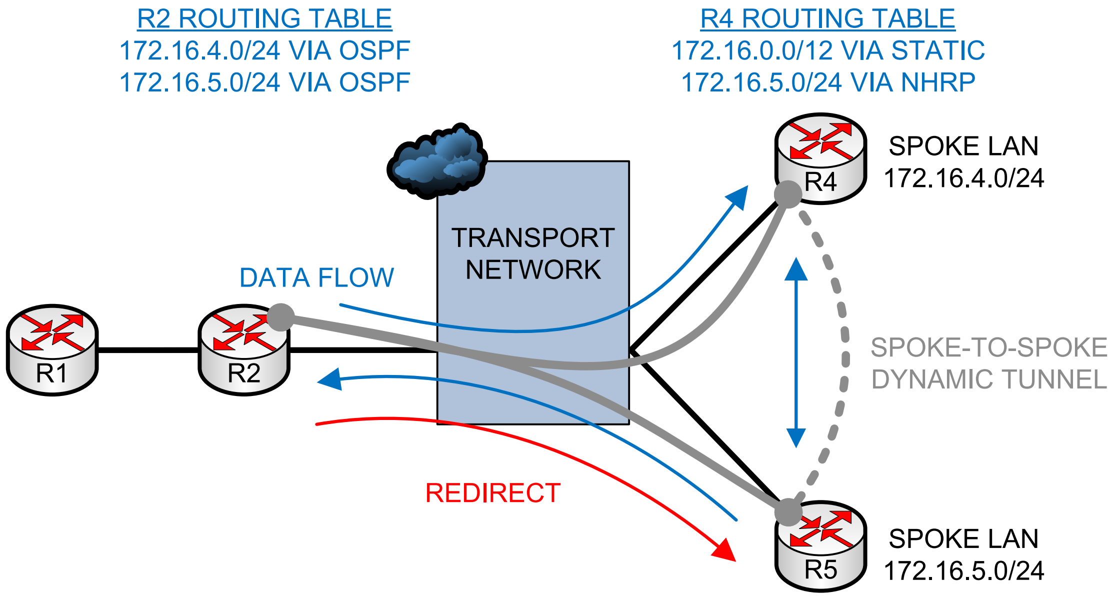
Figure 4 - Using OSPF Areas to Further Reduce State



When DMVPN phase 3 is used in these designs, spoke-to-spoke dynamic tunnel formation is still supported. Such traffic will initially traverse the hub, but when the hub realizes it is hairpinning traffic, it issues a redirect message towards the originating spoke. This triggers the spoke to issue a resolution request to the hub, effectively asking for the underlay address of the target spoke. The hub forwards the request to the target spoke, who responds directly to the originating spoke. The process occurs in the reverse direction as well, assuming the traffic flows bidirectionally. Provided that the two spokes actually have reachability to one another directly, the spoke-to-spoke dynamic tunnel will form and traffic can be exchanged directly. The spokes still never have any overlay-learned OSPF routes, but instead install NHRP routes to describe their active shortcut paths. The diagram below illustrates spoke-to-spoke tunnel formation in this design at a high level.

While this document does not detail the specifics of DMVPN over network address translation (NAT), note that this prevents spoke-to-spoke tunnel formation. The examples in this document assume NAT is not in play to keep focus on the DMVPN anycast and OSPF technologies.

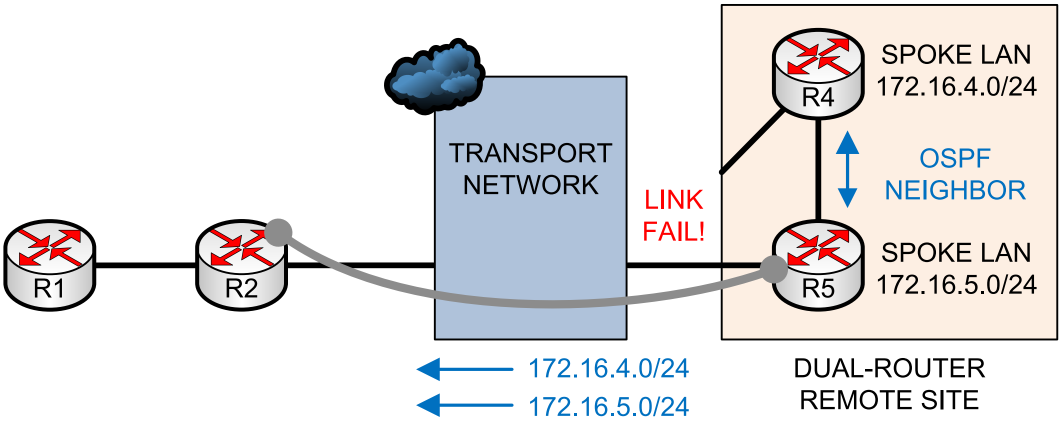
Figure 5 - DMVPN Phase 3 Dynamic Tunnels (High Level)



The solution also provides support for two common design use-cases.

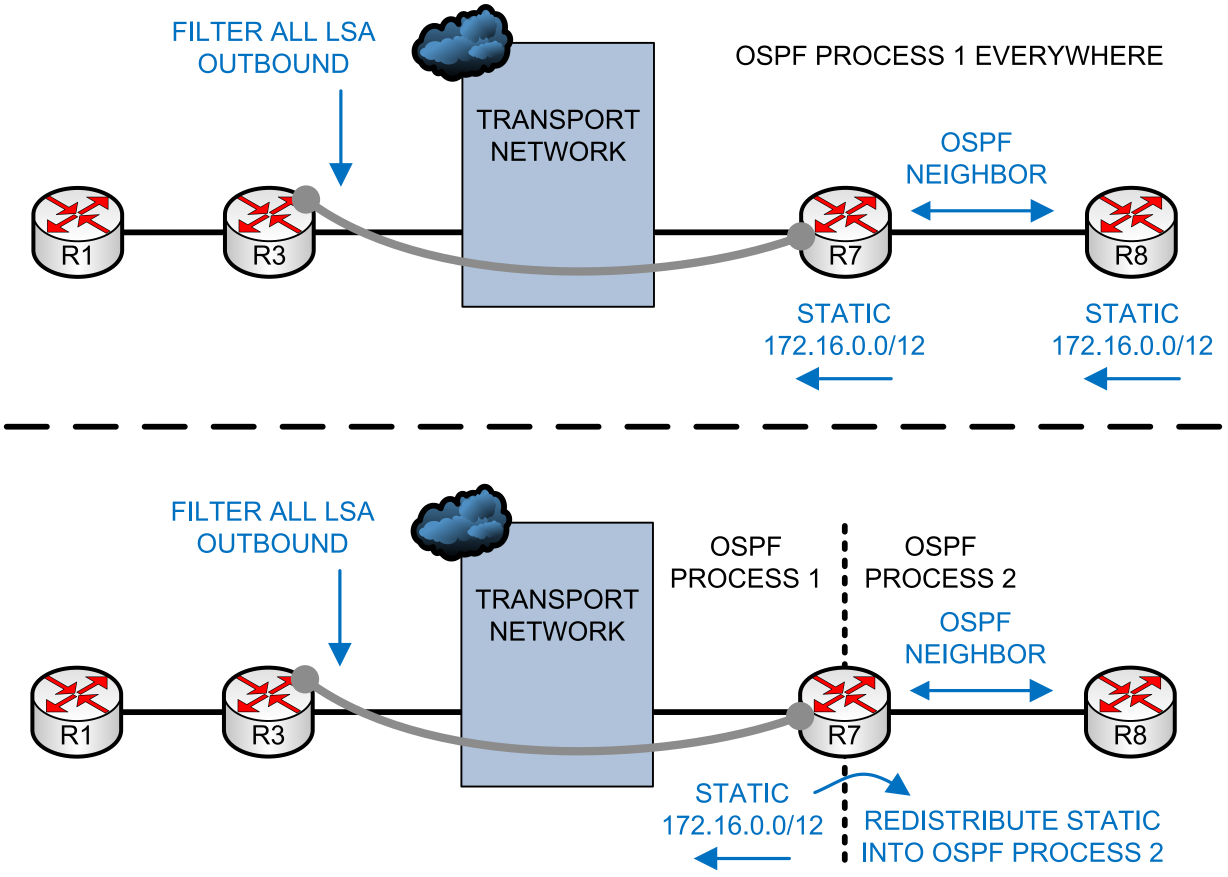
First, a dual-node remote site can have a transit link between the two DMVPN spoke routers running OSPF as usual. This transit link is a dedicated routing link between the two remote site CEs that does not service clients, unlike the user LANs deeper in the site. OSPF routing is still advertised normally from spoke to hub, so the typical resilience requirements are satisfied. Both spokes will have static routes up to the hub via the overlay IP.

Figure 6 - Failover with a Dual-Router Remote Site



Second, it provides support for downstream routers inside the branch site. For example, a layer-3 access switch or firewall may be running OSPF inside of a branch site. This device is not part of the DMVPN overlay, so it will need upstream reachability to destinations reachable over the WAN. This is accomplished by copying the set of static routes (or using a default route) from the DMVPN spokes to the downstream router, or redistributing the static routes from the DMVPN spokes into a separate OSPF process local to the site. The diagram below depicts some of these options. This document does not explore every option as it is beyond the scope of the general solution.

Figure 7 - Options for Handling Intra-Branch Routing Devices



2.2. High Availability

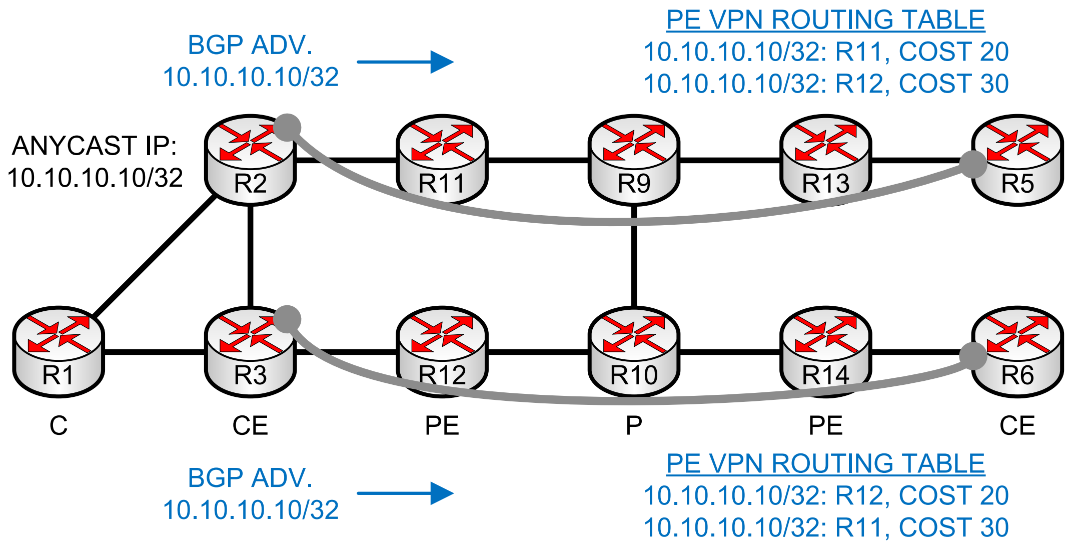
Most organizations cannot operate with only one DMVPN hub. Additionally, all the anycast-related work put into the aforementioned design is meaningless in a single hub design. This section explores how to add additional hubs and discusses the importance of keeping the hubs connected over the OSPF backbone upstream from the anycast tunnels.

Perhaps the simplest and most powerful part of the entire design is how it scales. When additional hubs are needed, follow this simple process:

1. Copy the exact overlay/underlay configuration from the existing hub.
2. Deploy that configuration on the new hub, along with any differing PE-CE link specifics.
3. Connect the new hub to the network in the desired region of operation.

Consider the simplest possible example which has two hubs, each serving a different geographic region. All nodes are connected to the same MPLS L3VPN provider which is assumed to use a single BGP AS and have internal IGP link costs representative of the distance between sites. That is to say, intra-region path costs between PEs are less than inter-region path costs between PEs. R2 serves one region containing R4 and R5 spokes while R3 serves another region containing R6 and R7 as spokes. All intra-SP links have the same IGP cost of 10. Further assume that all PEs have access to all VPN routes (i.e., any route reflection in the SP network conceals nothing). The spoke-facing PEs in each region, R13 and R14, each learn two copies of the anycast tunnel sources, one from each hub-facing PE, R11 and R12. The spoke-facing PE chooses the closer one based on the carrier's shorter IGP cost.

Figure 8 - Basic Anycast Tunnel Hub Selection

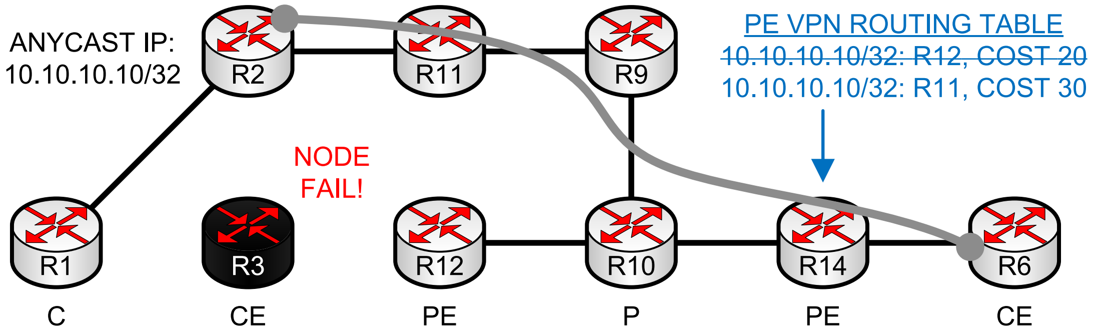


Failover is automatic and stateless. When the anycast IP address from one hub is no longer available, the spoke-facing PE simply chooses the next best path. Because this failover is stateless, the NHRP registration and optional IPsec session are torn down and established again with the new hub. With some moderate timer modification, this entire process can occur in about 45 seconds. These details are discussed in the upcoming "Tuning the WAN behavior" section.

There are many failure points that would trigger a spoke to switch from one hub to another:

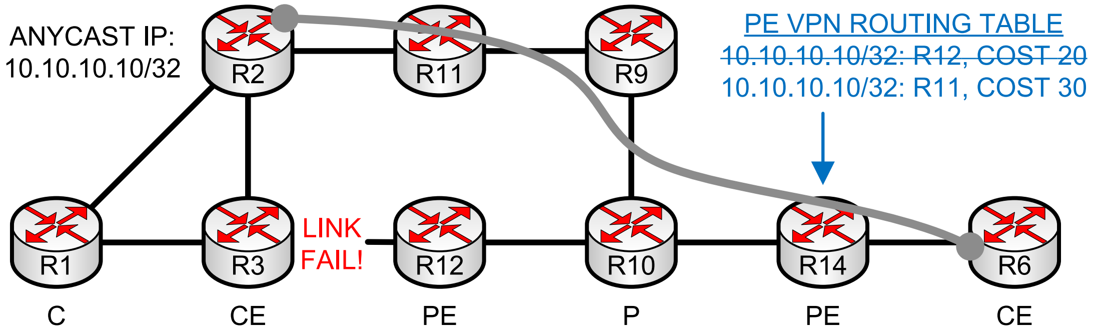
1. **Hub router failure:** The PE-CE link will immediately go down if there are no intermediate media conversion or layer-2 devices on the PE-CE circuit. If there are, the BGP session to the PE will eventually time out. In both cases, the PE will no longer learn the anycast tunnel endpoint, and will issue BGP withdraw messages into the carrier network. The remote, spoke-facing PEs will remove this route from their routing tables and prefer the next best path.

Figure 9 - Anycast Tunnel Failover; Hub Node Failure



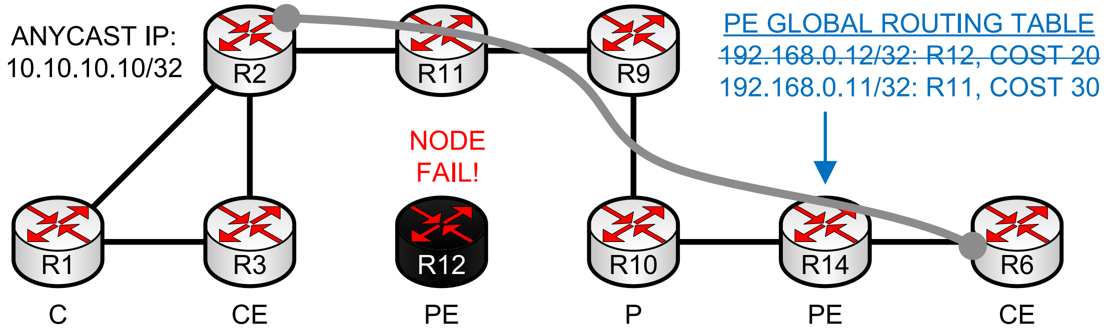
2. **PE-CE link failure:** The series of events in case of a PE-CE link failure is similar to the previous case. Either the PE-CE link failure triggers an immediate closure of the BGP session on the PE, or the session times out if the PE-CE link appears to stay up from the PE's perspective. The PE follows the same BGP withdrawal process described above, triggering an alternative path to be considered on the spoke-facing PEs.

Figure 10 - Anycast Tunnel Failover; Hub PE-CE Link Failure



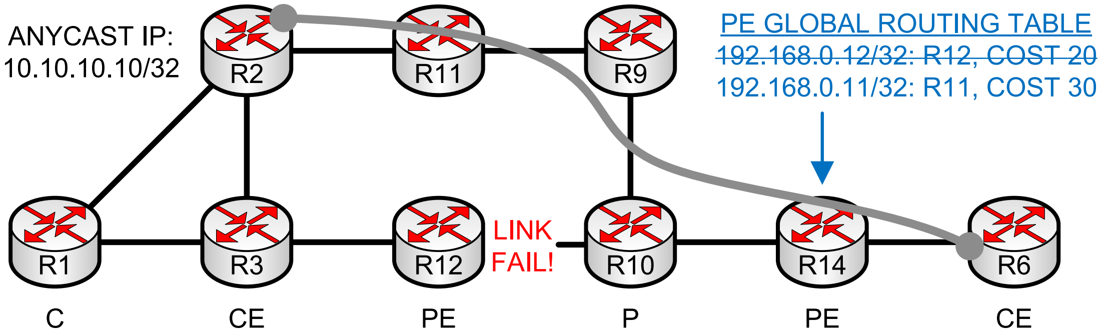
3. **Hub-facing PE failure:** If the PE itself fails, the spoke-facing PE will lose IGP reachability to that PE's loopback, which is likely the BGP next-hop for all VPN routes originated from it. Assuming BGP next-hop tracking (NHT) is used, the remote PE will immediately purge all BGP routes with that next-hop and consider alternative paths.

Figure 11 - Anycast Tunnel Failover; Hub-facing PE Node Failure



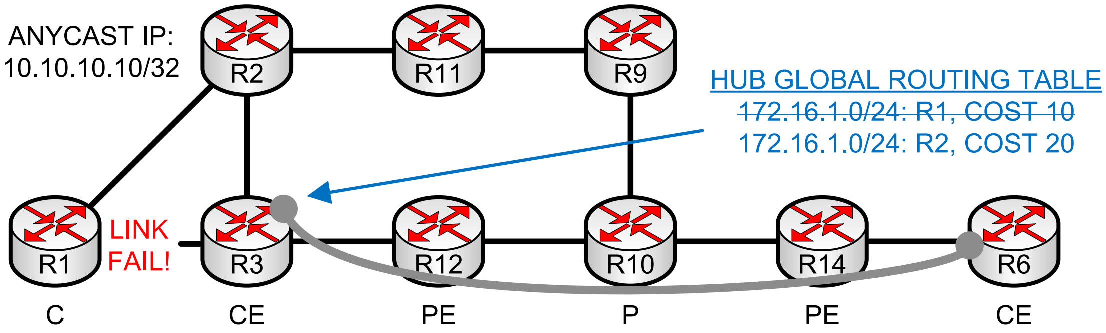
4. **PE uplinks to SP core failure:** The series of events in case of a PE-CE link failure is similar to the previous case. If the PE is isolated from the core network, the BGP session to the DMVPN hub remains up, but the remote PEs lose reachability to the isolated PE's loopback. This causes those devices to purge all routes from the isolated PE and consider alternative paths. Shortly thereafter, the internal BGP (iBGP) sessions to that PE within the carrier's network will collapse, prompting a withdrawal of any spoke routes advertised up to the headend.

Figure 12 - Anycast Tunnel Failover; PE to Core Link Failure



Failures upstream from the hub routers are also protected, but not as a function of the anycast design. Consider a failure between R1 and R3. The spokes registered to R3 would maintain their anycast tunnel to R3, but any spoke-originated traffic destined for R1 would traverse laterally from R3 to R2. This explains the need for the lateral link between the two hubs. This does not have to be a physical circuit, which could be expensive over long distances. It can be a tunnel interface that traverses the carrier architecture or the internal links, perhaps data center interconnection (DCI) points. These failures include all manner of link and node failures, but since this is beyond the scope of the anycast design, it is only discussed briefly. The diagram below illustrates one example of such an upstream failure, which is supported by the inter-hub backbone connection.

Figure 13 - Upstream Failure Handling; Post-Tunnel Reroute



2.3. Tuning the hub selection

The anycast solution typically works well when deployed without much tuning. However, larger organizations may wish to horizontally scale the solution to a much larger number of hubs. Consider an organization with 3 regions, each with 4 hubs, for a total of 12 hubs. Assume that, upon plugging in all of the hubs and spokes, each of the 12 hubs has at least some spokes attached to it. By default, this design is technically valid and needs no adjustment.

Some organizations may find it worrisome to have this lack of predictability built into their daily network operations. If an SP changes some costs on their links, adds new links, suffers an outage, or otherwise changes their internal architecture, it could cause a short WAN outage as remote PEs may choose newer optimal egress points from the carrier's AS. Other organizations may simply want to reduce complexity by defining a strict failover sequence: Prefer hub "A" first, then hub "B", etc. Besides predictability, other reasons for defining a hub preference order may include transient events. Examples include taking a hub offline for maintenance or temporarily avoiding poorly performing hot spots.

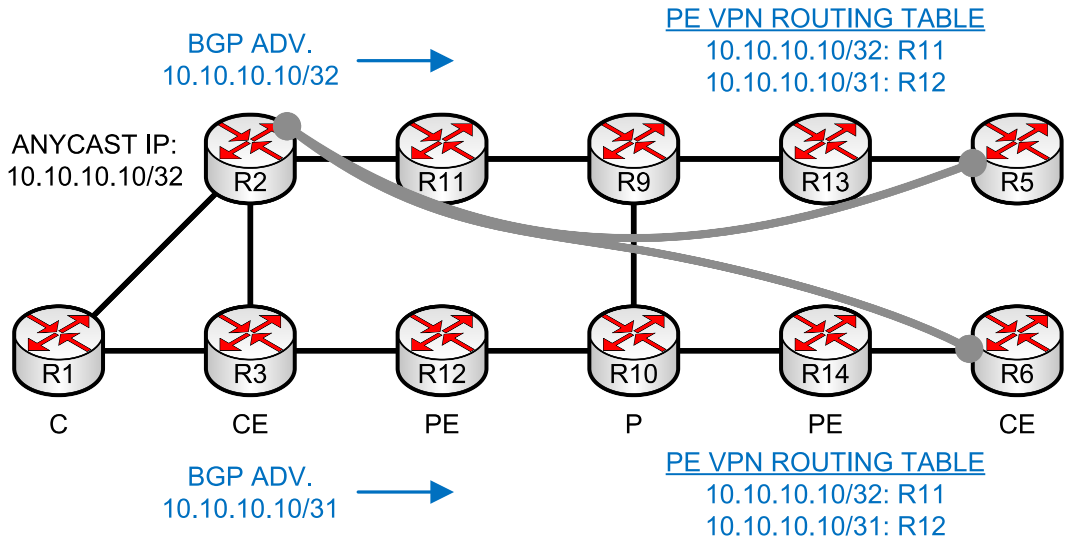
There are many ways to influence hub selection. Three straightforward approaches are discussed next. Before discussing the options, the term "hub class" must be defined. Within this document, a hub class is a set of hubs that, from the carrier's perspective, have an equal level of preference and thus are determined by the carrier's lowest IGP cost. When no tuning is applied, all hubs are in the same hub class. When there are many hubs, it is possible to segment the hubs into hub classes, such as a set of primary hubs and a class of secondary hubs. Hubs in the same class have the same preference level, hence the name.

In any design where high availability is a requirement, the standby system should be sized appropriately to carry the entire load of the primary system. If the load is distributed across multiple systems, the system must be able to tolerate an appropriate number of failures in accordance with the business needs. Consider two examples:

1. A simple active/standby design whereby all clients tie into the primary system. The standby system should be able to accommodate all clients currently tied into the primary system.
2. A complex load sharing design with 10 hubs in a region, each carrying approximately equal shares of the client base. It is unlikely that 1 of these 10 hubs would have to carry the entire client base at once, however the system should be able to tolerate some failures. For example, up to 3 hubs can fail concurrently, leaving the remaining 7 to each absorb some of the work from the failed hubs. The precise numbers will be determined by the business drivers and current capabilities of the network.

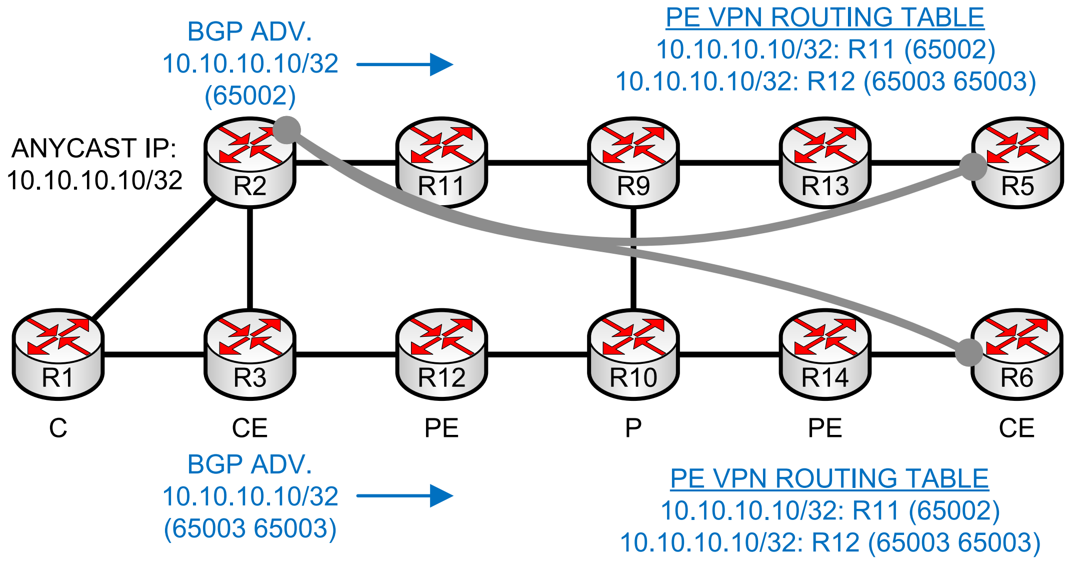
The first option uses longest-match routing to the anycast address as a mechanism of steering traffic towards certain hubs. The preferred hubs will advertise more specific prefixes to the carrier, such as IPv4 /32 host routes over a private WAN service or a public IPv4 /24 over the Internet. The backup hubs would advertise less specific prefixes, such as IPv4 /31 over private WAN or a public IPv4 /23 over the Internet. In inter-AS scenarios, this solution is guaranteed to work in implementing an organization's hub class policy. The main drawback of this approach is the massive waste of IPv4 addressing when Internet is used as a transport. As the number of hub classes increases, progressively larger amounts of space must be wasted when constructing shorter match prefixes for advertisement.

Figure 14 - Hub Selection; Longest Match Anycast Prefix



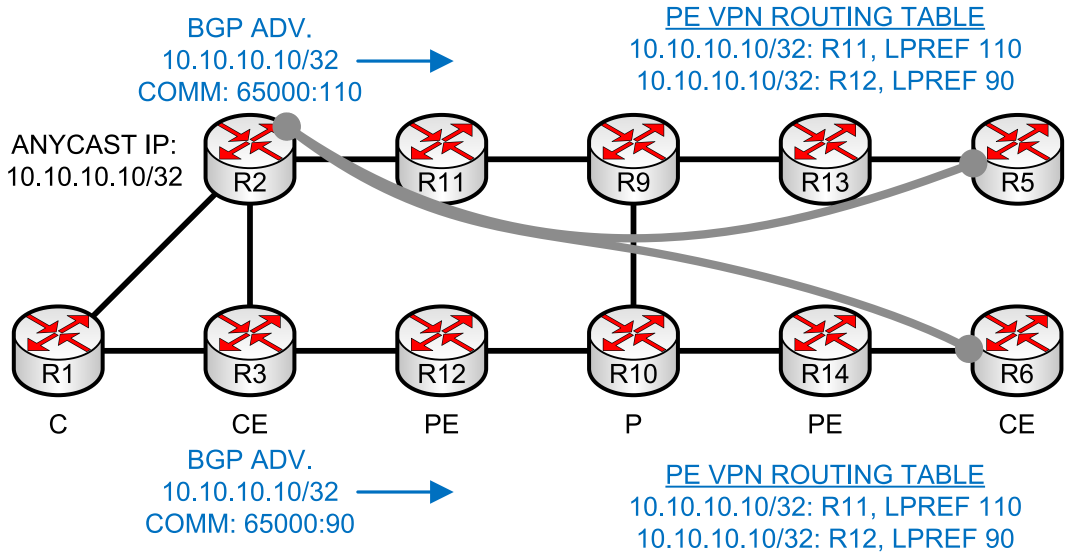
The second option uses outbound BGP AS-path prepending from the hub (CE) to the carrier (PE). The non-preferred hubs will append the local BGP AS more times than the preferred ones. Consider the example in the diagram below. R2 is preferred over R3, so it will prepend fewer times, or perhaps not at all. This technique may work in an inter-AS scenario with multiple providers. Its success is not guaranteed, unlike the longest-match option.

Figure 15 - Hub Selection; Anycast Prefix AS-Path Prepending



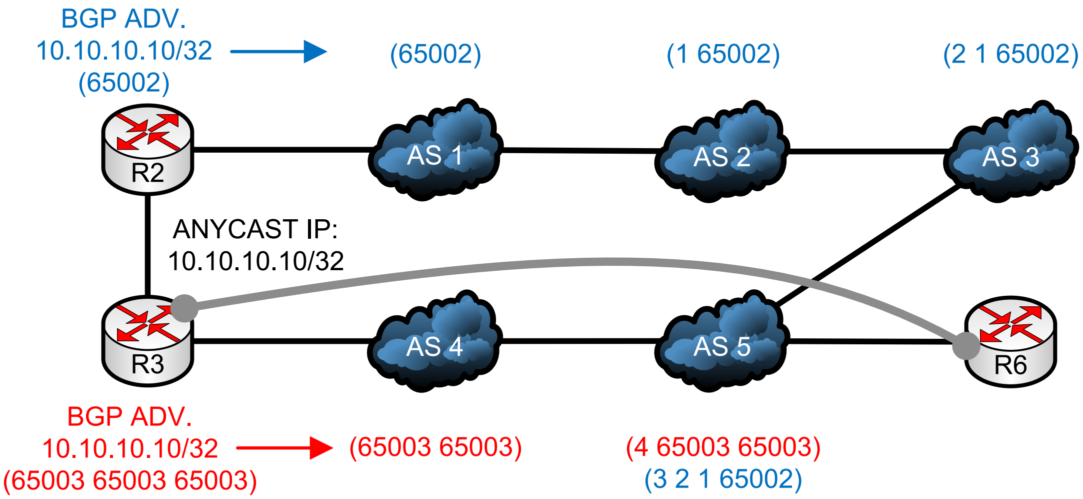
The third option employs RFC 1998 BGP community-based preference. Rather than directly apply BGP policy or change the prefix length of the anycast address, the hubs will affix designated BGP communities to their anycast routes. Many large carriers publish their community policies which allow customers to control, on a per-prefix basis, how traffic should egress from the carrier's network. This is generally simpler and more effective than AS-path prepending, but requires explicit support from the SP. In the diagram below, the SP publishes that communities of 65000:X will set the local preference to X upon receipt at the PE. As such, the preferred hub R2 sends its anycast loopback with community 65000:110 while R3 uses community 65000:90. Like AS-path prepending, it is most effective within a single AS, but could have positive effects in inter-AS scenarios depending on how other carriers are routing.

Figure 16 - Hub Selection; RFC 1998 Community Applied to Anycast Prefix



The diagram below better illustrates why the AS-path prepending and RFC 1998 community options may not work in multi-carrier environments. This particular example uses AS-path prepending as a hub selection mechanism. Depending on the density of connections between Internet service providers, the backup hub (in this case, R3) could be chosen as the primary hub. AS 5 has two paths to the anycast loopback, and the preferred blue path has a longer AS path length than the backup red path. Continuing to prepend more AS numbers to R3's outbound AS path could solve this particular issue, but such a brute-force solution is still not guaranteed to work in all cases.

Figure 17 - Why AS-path Prepending May Not Work for Hub Selection



2.4. Tuning the WAN behavior

This section discusses further enhancements to the anycast DMVPN design. Specifically, it describes IPsec design considerations and tuning timers for fast convergence. Note that all combinations of these options are fully supported with the anycast DMVPN solutions.

There are four general strategies in determining what flavor of IPsec to deploy in conjunction with most WAN designs.

1. **Nothing:** DMVPN with only GRE encapsulation is a valid design and is commonly used in environments that have a dedicated encryption device (either for scale or specialized security). It is also used in environments where the underlay transport is owned by the same organization managing the overlay, and security is guaranteed from some other means. GRE can only traverse 1:1 network address translation (NAT) devices, making it a poor choice for Internet transport. Its lack of built-in packet authentication and encryption capability adds to this unsuitability.
2. **Authentication Header (AH):** IPsec AH is not commonly used but offers authentication of traffic in transit. With a fixed encapsulation size and inability to traverse NAT devices, AH is a good choice when the public Internet is not used. It may be appropriate for WANs specifically designed to carry data that is not secret but is sensitive to tampering.
3. **Encapsulating Security Payload (ESP) with null cipher:** Also uncommonly used, ESP null is coupled with an authentication mechanism. Unlike AH, ESP can traverse NAT, making it a good choice for Internet transport. This option is excellent when the traffic profile is similar to the AH use-case (not secret, but sensitive to tampering) but it is likely that NAT devices exist in the path.
4. **ESP with a secure cipher**: Choosing ESP with a standards-based encryption algorithm is the most common option. Computing maximum transmission unit (MTU) is often challenging as the exact packet size depends upon the algorithm used. The solution can traverse NAT devices and is commonly used for data that is considered secret.

If an organization decides to use IPsec, special case must be taken when choosing between transport mode and tunnel mode. In the majority of cases, transport mode can be used, since the DMVPN endpoint and IPsec endpoint are the same IP address. This even works across NAT devices in some cases. If carrier-grade NAT (CGN) architecture is in play, then tunnel mode must be used.

The reason why IPsec tunnel mode must be used to traverse CGN is the behavior of NHRP. By using tunnel mode, only the outermost address is translated across the NAT devices, which is handled and processed by the IPsec endpoints. The middle header is the GRE header used by DMVPN to uniquely identify spoke underlay addresses; the DMVPN hub will see the original inside local addresses configured on the spokes. Failing to use tunnel mode will cause an overlap of NBMA addresses within DMVPN since both spokes will have the same public underlay address offered by CGN.

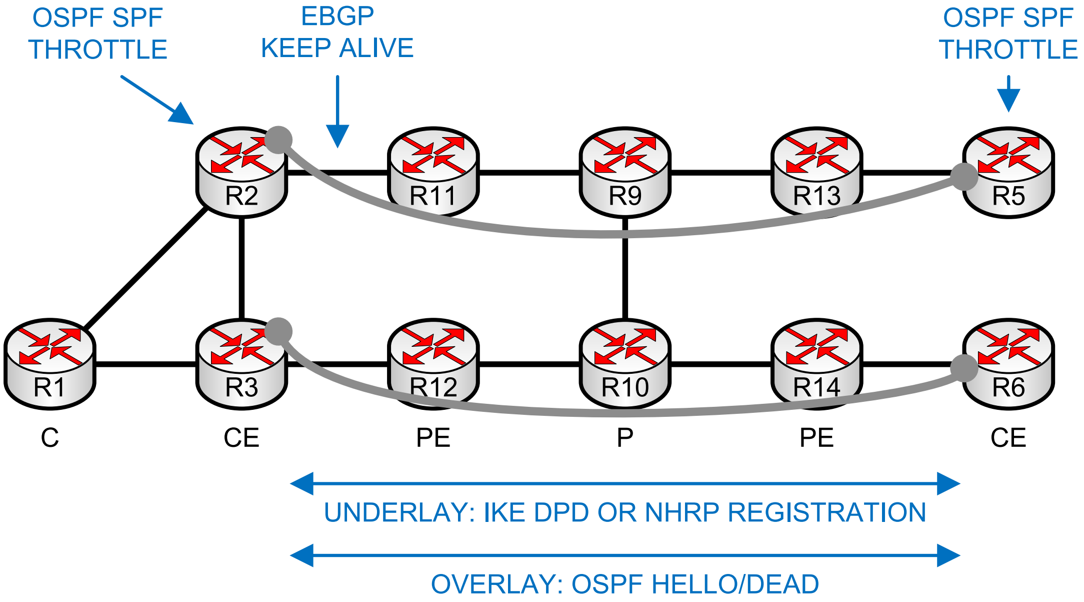
While IPsec is not required for the general design, it is commonly used and is therefore the basis for the following analysis. Upon failover, the tunnels will have to reform, both from an IPsec and DMVPN/IGP routing perspective. Suppose a particular customer has an availability requirement requiring the network to converge in less than 60 seconds. This requires careful tuning of five main timers, which are discussed below.

1. **PE-CE eBGP session keepalives:** These timers are configured with 10 second keep alive time and a 30 second hold time. This ensures that, in the event of a PE-CE link or CE node failure at the hub, the PE is aware of the failure within 30 seconds (worst case). As discussed earlier, the failure of the layer-1 circuit would shrink this time.
2. **IKEv2 Dead Peer Detection (DPD):** IKE converges slightly slower than the previous timer. This is deliberate; ensuring that the underlay reconverges before the failed IPsec SA is torn down guarantees that the new IPsec SA can immediately be negotiated. This speeds up the convergence process considerably. DPD is configured for periodic probes every 20 seconds. When a DPD probe is lost, 3 retransmissions occur, each of which take 5 seconds. The total time spent on DPD retries is 15 seconds. As such, DPD can detect a failed IPsec peer within 35 seconds (worst case). This time does not exist in non-IPsec environments and can be skipped.
3. **NHRP registration timer:** When IPsec is used, the DPD timer helps detect whether the underlay path is healthy. Without IPsec, the NHRP registration timer accomplishes the same thing. These probes are sent every 30 seconds with retransmissions sent after 1, 2, and 4 seconds (7 seconds total). This leads to a 37 second (worst case) failure detection time, approximately the same as using DPD. This only needs to be adjusted at the spokes and only needs to be used when IPsec is not deployed. However, having it configured with IPsec DPD at the same time did not cause any observable negative effects.
4. **OSPF hello/dead:** These timers are set to the classic “fast” values of 10 seconds hello and 40 seconds dead. The neighbor formation process, given the absence of a pseudonode over the point-to-multipoint network, takes only seconds. This is slower than the DPD timer which allows the underlay to fail before the overlay. There is no harm in having a stale OSPF neighbor with the old hub for a few seconds while the OSPF neighbor with the new hub forms.

5. **OSPF LSA throttle:** By tuning the LSA throttle timer only on the spokes, upstream updates are sent more rapidly rather than waiting several seconds after a neighbor forms. In the worst case, this process adds 2 seconds. This timer is not configured on hubs because hubs are filtering all their updates outbound. If the LSA throttling is configured on the hubs to support fast convergence in area 0, that is acceptable, but is unrelated to this particular design consideration. After the first change, the router reacts after 50 ms, then 200 ms for the next change, and continues to double the time (400 ms, 800 ms, etc.) until 5000 ms is reached.
6. **OSPF SPF throttle:** This timer is tuned only on the hubs so that LSAs received from downstream spokes are processed immediately instead of waiting several seconds for the next scheduled SPF run. In the worst case, this process adds 2 seconds. This timer is not configured on the spokes because they do not rely on OSPF for upstream routing. Notwithstanding a large intra-site OSPF domain at a remote site, there is no compelling reason to configure this timer at the spokes. The same values of 50 ms minimum, 200 ms next, 5000 ms maximum described in the LSA throttle section above are suggested for SPF throttling as well. Note that this setting should be consistent within all routers in the headend, such as those in area 0. This dampens the impact of microloops.

The diagram below summarizes the timer tuning needed to reduce failover times.

Figure 18 - Visualization of Necessary Time Adjustments



The following diagrams are timelines from actual testing to validate that these timers are sufficient. The caption of each diagram describes the test case. In each case, the eBGP session between R2 and R11 was blocked via access-list to create the longest possible failure time. Note that the IPsec convergence times are typically a few seconds longer (closer to 50 seconds) despite all the timers being closer to 35 or 40 seconds. The extra packet exchanges needed to form IKE and IPsec sessions takes time, which is why timers closer to 50 seconds were not chosen. Failback times are also dependent on how quickly the carrier's BGP bestpath calculation runs, making it difficult to measure in real life.

Figure 19 - Non-IPsec Failover from R2 to R3

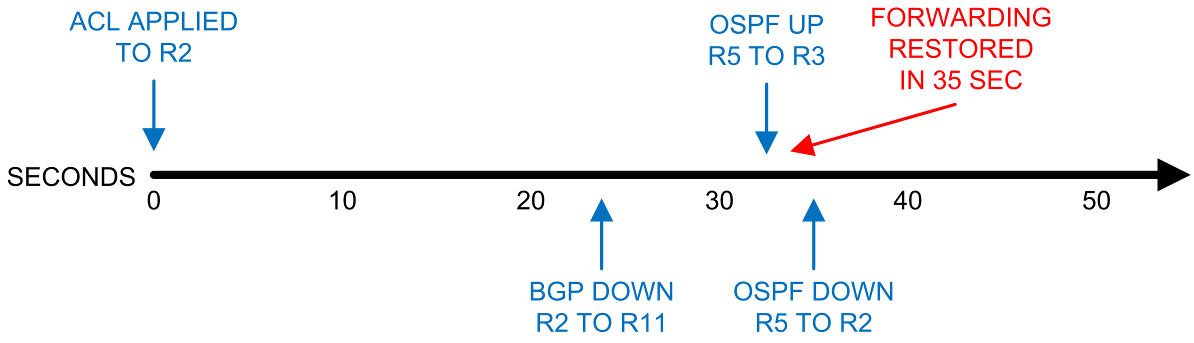


Figure 20 - Non-IPsec Failback from R3 Back to R2

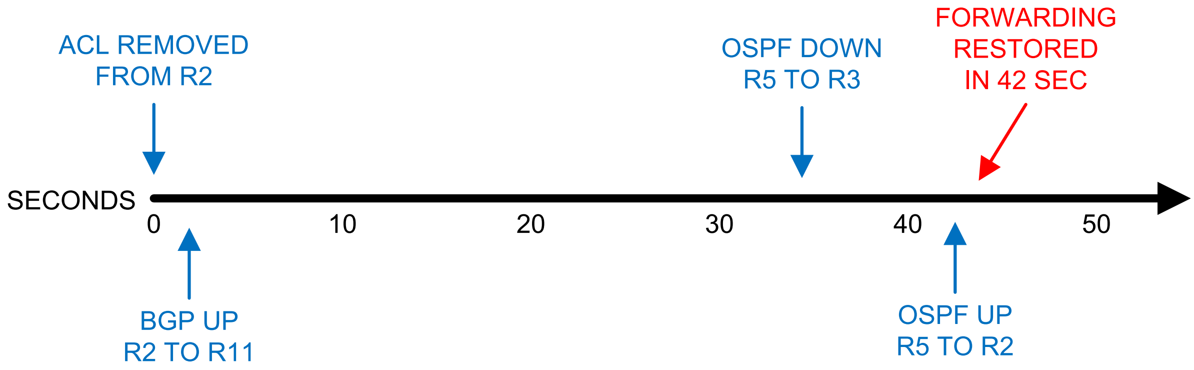


Figure 21 - IPsec Failover from R2 to R3

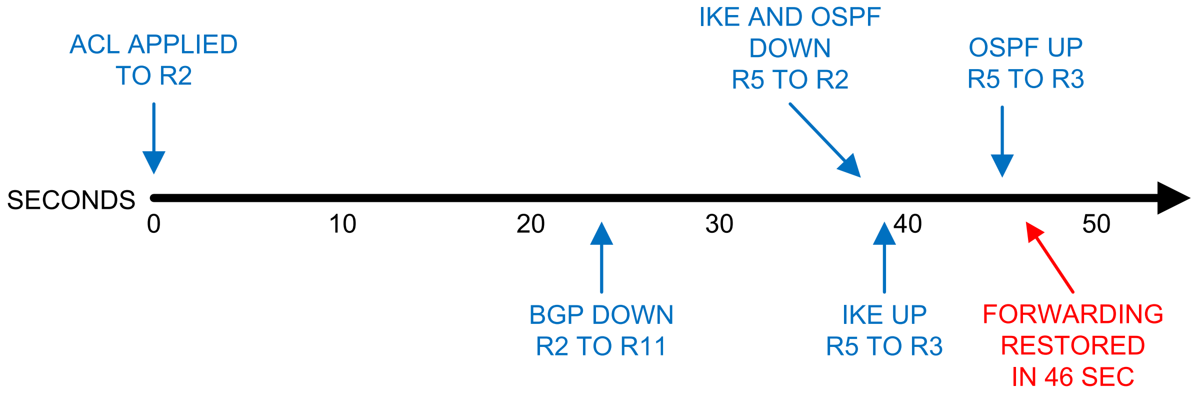
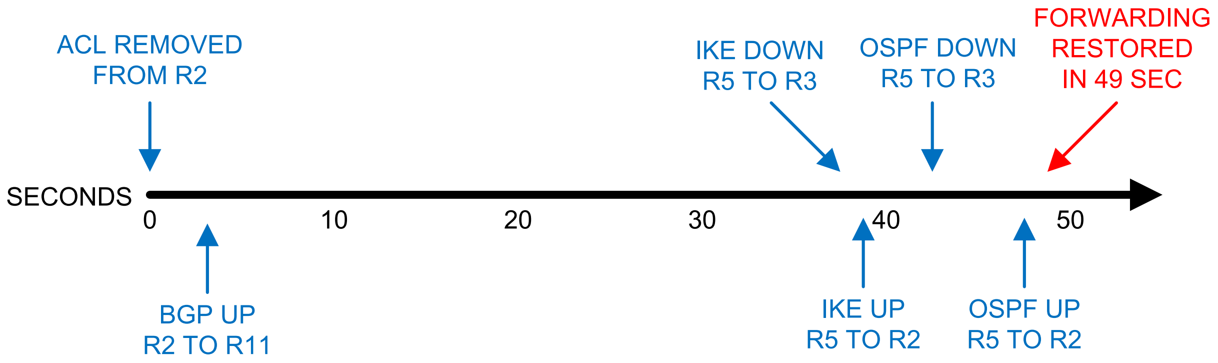


Figure 22 - IPsec Failback from R3 Back to R2



These time measurements will be slightly different each time, as the statistical variation and dependency (sequencing) of the timers will influence the outage times. The transport carrier's timer tuning and fast reroute capabilities will also impact it, although these will differ across transport types, geographical regions, and service provider business drivers. Readers are encouraged to test the architecture in their own environment for benchmarking purposes.

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 OSPF hub/spoke WAN designs, such as those not performing any LSA filtering.

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.

The state retained at the anycast hubs scales linearly as spokes are added to the WAN, the overlay network scales linearly as spokes are added. The hub serving a specific region in which a spoke is added will carry slightly more state than the other hubs serving other regions. The former will form a new OSPF neighbor, complete with all manner of intra-area flooding and state retention. The latter will only receive inter-area routing information from the hub in the form of some summary prefix LSAs. Notwithstanding any LSA filtering or aggregation between anycast meshes, each hub will have all routes from all spokes. To summarize:

a. Adding a spoke to region A causes the hub for region A to form one new neighbor.
b. Adding a spoke to region A causes all other hubs to learn only the routing information from this new spoke via summary LSAs, not the topology information via router LSAs.
c. Adding a spoke to region A has absolutely no impact on any other spoke, be it inside region A or elsewhere.

We conclude that, in general, the routing state on hubs scales in linear space and time as spokes are added to the network. The routing state on spokes scales in constant space and time as spokes are added.

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 primary driver of the design described in this document is to minimize the undesirable effects of using link-state protocols in hub/spoke networks, such as excessive protocol traffic, increased path calculation computational cost, slower convergence, and increased fragility. As state decreases in the network to achieve scale, optimization is directly proportional and likewise decreases. The small set of IP summary routes on the spokes, while generally harmless, qualify as “un-optimized” in this context. In a mixed network with other unfiltered links back to the hub, this lack of optimization would guarantee that the “other” links were always preferred for upstream routing compared to the DMVPN overlay. The spokes still follow the longest match routing principle which would make the anycast overlay less preferable to other paths.

The biggest optimization trade-off with this design is the amount of downtime between hub switchovers. While this can be tuned down to a few seconds, it is still disruptive to tear down and rebuild the overlay, especially with IPsec enabled. The design is optimized for organizations requiring scalable and stable OSPF hub/spoke deployments that have only modest availability requirements in terms of downtime.

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.

Surface interactions with respect to the control-plane are minimally deep. The usage of front-door VRFs separate the control planes of the customer overlay and carrier underlay. Other complex surface interactions (such as redistribution) are likewise absent unless some overriding business concern introduces them. For example, a spoke site is running another IGP and needs to redistribute into the OSPF process running in the anycast overlay. The interaction between static routing and OSPF is a surface interaction on the spokes as upstream routing follows a static route while return traffic follows an OSPF learned route. This interaction is shallow but has wide breadth as it spans the entire set of spokes.

Surface interactions with respect to the various protocols in the network are relatively deep and very wide. The timer tuning section of this document illustrates this point. The multitude of different protocols and implementations, to include the dependencies between them, requires careful planning, coordination, and testing. Because these interactions are spread everywhere, they are wide.

It is worth noting that while almost all of the surface interactions described above are wide, they are also consistent. The design does not encourage one-off modifications, such as one hub not using eBGP to the service provider or some spokes being on a separate overlay with LSA filtering disabled.

4. Other Considerations

4.1. Extension over multiple transports

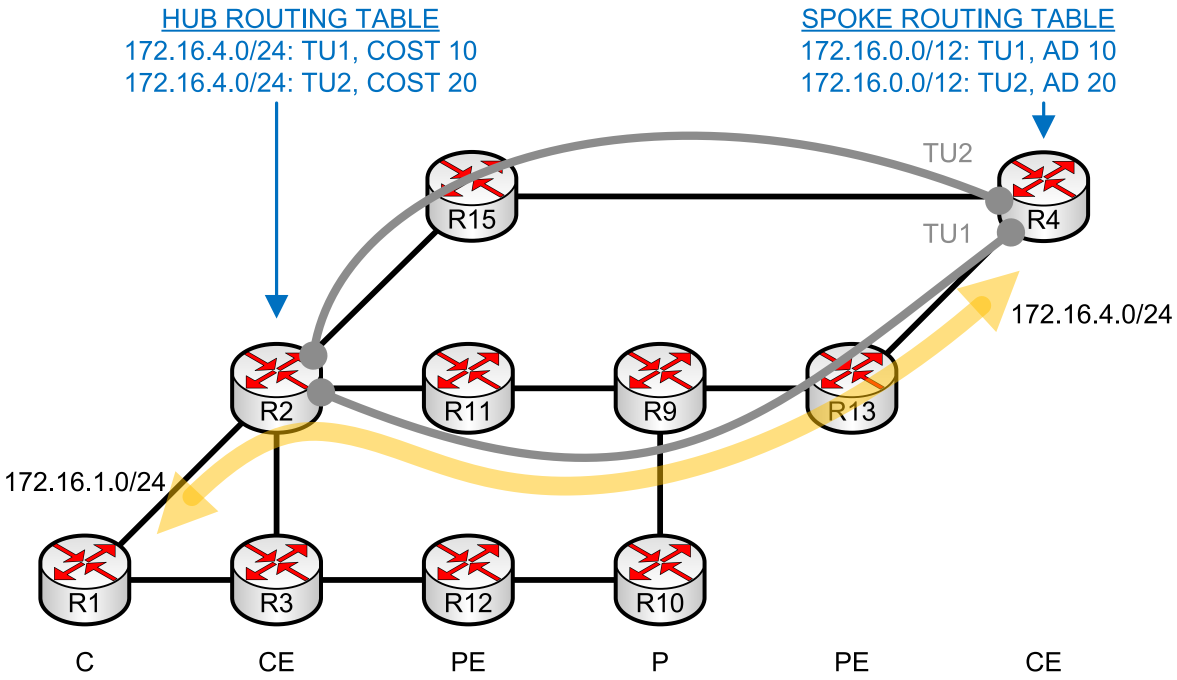
The anycast solution can be replicated across many additional transports simply by deploying the same solution multiple times. For example, a remote site may have an Internet connection as a backup to its MPLS connection. Each of these links may be placed in separate front-door VRF with underlay routes pointing to different anycast addresses. Just like the original design, the hub uses outbound LSA filters to prevent any OSPF routing information from being sent to the spokes. The spoke uses a set of upstream static summary routes to reach remote LANs. There are two key routing considerations introduced with multiple overlays:

1. The hub should use explicit OSPF cost adjustments on each overlay to indicate preference for the downstream routing path. The spoke does not need to make any cost adjustments.
2. The spoke adds static routes upstream out of the new overlay. These routes can be longer matches to always prefer the new overlay, shorter matches to never prefer it, or equal matches with variable administrative distance (AD) for a combination approach. Assuming the overlays are configured to go down when NHRP registration fails, static routes with unreachable next-hops will be withdrawn, so failover works.

This solution scales to many transports with separate, concurrent overlays with the same scaling characteristics described earlier in this document. However, it is unlikely that this design would be deployed with multiple tunnels across multiple transports. Any organization that can afford multiple links at their remote sites likely has a requirement for higher availability than this solution provides. Such organizations would be more likely to deploy a protocol like BGP as it is better suited for that environment. While this section has demonstrated that it is possible, it is not common and not recommended.

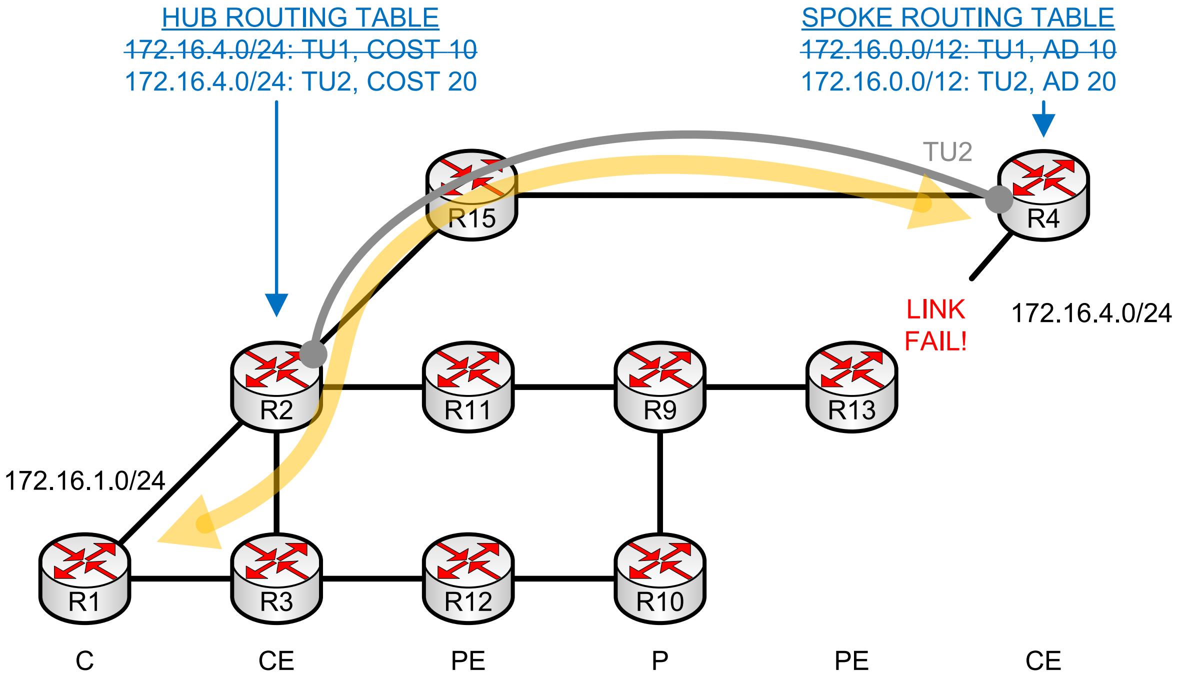
Making both of the aforementioned routing adjustments concurrently results in a symmetrically-routed, multi-transport resilience WAN design. The diagram below illustrates the steady-state operation of this design with no failures. R1 sees a path to R4's LAN via tunnel 1 given a lower OSPF cost. R4 uses the upstream static route over tunnel 1 given a lower administrative distance.

Figure 23 - Multi-Overlay Symmetric Routing; Steady State


HUB ROUTING TABLE
172.16.4.0/24: TU1, COST 10
172.16.4.0/24: TU2, COST 20
SPOKE ROUTING TABLE
172.16.0.0/12: TU1, AD 10
172.16.0.0/12: TU2, AD 20
TU2
TU1
R15
R4
172.16.4.0/24
R2
R11
R9
R13
172.16.1.0/24
R1
R3
R12
R10
C
CE
PE
P
PE
CE

Consider a failure state. At R2, the OSPF routes to R4's networks via tunnel 1 are withdrawn upon failure, revealing the higher-cost backup path. Likewise on R4, the failure of tunnel 1 removes the lower administrative distance static route, making the tunnel 2 path preferred.

Figure 24 - Multi-Overlay Symmetric Routing; Failure State


HUB ROUTING TABLE
~~172.16.4.0/24: TU1, COST 10~~
172.16.4.0/24: TU2, COST 20
SPOKE ROUTING TABLE
~~172.16.0.0/12: TU1, AD 10~~
172.16.0.0/12: TU2, AD 20
TU2
R15
R4
LINK
FAIL!
172.16.4.0/24
R2
R11
R9
R13
172.16.1.0/24
R1
R3
R12
R10
C
CE
PE
P
PE
CE

4.2. IP Multicast

As it relates strictly to the deployment of IP multicast, there are no specific considerations. Simply enabling Protocol Independent Multicast (PIM) on all non-underlay interfaces will result in a functional multicast transport network. Traditional hub/spoke multicast design considerations around rendezvous point (RP) placement remain unchanged. There are no known issues with reverse path forwarding (RPF) path resolution in any multicast data flow.

There is a scalability trade-off for two reasons, both of which relate to increased state. These drawbacks adversely affect the DMVPN hubs by consuming additional resources not required by the unicast-only design.

1. Enabling new protocols always requires more compute resources. This implies one new protocol neighbor at each spoke but many at the hub. The latter point can be problematic.
2. The hubs must dynamically create and maintain multicast mappings for each spoke. When PIM is not enabled, OSPF can work with an OSPF point-to-multicast nonbroadcast network type at the hub and an OSPF point-to-multicast network type at the spokes. This allows the hub to respond to the spoke-initiated hello packets with a unicast reply, obviating the need for dynamically generated multicast mapping entries at the hub.

Additionally, note that NHRP routes can be used for RPF, but because there will never be a PIM neighbor between two spokes, traffic will not flow spoke-to-spoke. Because there is only one hub overlay IP, a static multicast route (mroute) from spoke to hub overlay IP is sufficient to resolve this problem.

To optimize for state, consider disabling shortest path tree (SPT) switchover on the spokes as there is only one path towards the RP. In some cases, this is undesirable, such as when the RP is not the DMVPN hub but is much farther away.

4.3. IPv6

IPv6 has no special considerations other than feature availability. Like the previous section on multicast, the designer must consider the impact of effectively doubling the number of protocol neighbors at all nodes. Unlike the multicast example, the DMVPN hub does not need to create dynamic multicast mappings. The network-types can be configured intelligently as described above. IPv4 is demonstrated here for simplicity, but using OSPFv3 for both IPv4 and IPv6 address-families is fully functional.

4.4. Security

This solution has some security advantages over a more traditional DMVPN-based WAN. This section does not detail generic DMVPN security techniques such as front-door VRFs, underlay access-lists, NHRP authentication, etc. These are beyond the scope of the document.

First, the reliance on DNS is reduced. While it is still possible to use DNS to resolve the anycast IP address, it is generally unnecessary as the IP address will be universal throughout each region, or perhaps the whole world. This reduces the reliance on DNS security (DNSSEC) being deployed and operational, ultimately preventing your hub-destined traffic from the spokes from being redirected towards an attacker via DNS spoofing.

Second, the spokes can never be transit nodes in multi-overlay designs. Other protocols like BGP have explicit controls to prevent nodes from becoming transit nodes, but OSPF only has a "hinting" mechanism via IGP cost adjustment. Using separate OSPF areas or processes would also accomplish this, but the specific design allows a single area, single process to remain non-transit simply by virtue of its design. The spoke receives no OSPF routes from the hub, and thus cannot advertise any to any other node.

Third, the lack of topology flooding information over the WAN reduces an attacker's ability to map the internal topology. Suppose IPsec with an encrypting ESP cipher is not used given a lack of secrecy requirement or preference for high performance at low cost. This implies that the OSPF traffic inside the overlay is visible to the underlying transport via packet sniffers. Since the hubs send no LSAs downstream, intermediate listeners cannot visualize the core topology. The most they could see is the intra-remote site topology, which is typically very few devices.

Appendix A – Acronyms

Acronym	Definition
ABR	Area Border Router
AD	Administrative Distance
AH	Authentication Header
AS	Autonomous System
ASN	AS Number
BGP	Border Gateway Protocol
C	Customer (core) router
CE	Customer Edge router
CGN	Carrier Grade NAT
DCI	Data Center Interconnection
DMVPN	Dynamic Multiple VPN
DNS	Domain Name System
DPD	Dead Peer Detection
DV	Distance Vector
ESP	Encapsulating Security Payload
GRE	Generic Routing Encapsulation
IGP	Interior Gateway Protocol
IKE	Internet Key Exchange
IP	Internet Protocol
L3VPN	Layer-3 VPN
LAN	Local Area Network

Acronym	Definition
LSA	Link State Advertisement
MPLS	Multi-Protocol Label Switching
NAT	Network Address Translation
NBMA	Non Broadcast Multi Access
NHRP	Next Hop Resolution Protocol
OSPF	Open Shortest Path First
P	Provider (core) router
PE	Provider Edge router
PIM	Protocol Independent Multicast
RP	Rendezvous Point
RPF	Reverse Path Forwarding
SOS	State Optimization Surface
SPF	Shortest Path First
VPN	Virtual Private Network
VRF	Virtual Routing and Forwarding
WAN	Wide Area Network

Appendix B – References

BGP-4 (IETF RFC 4271)

BGP Communities for Multi-homed Routing (IETF RFC 1998)

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

IP Security – IPsec (IETF RFC 4301)

Operation of Anycast Services (IETF RFC 4786)

MPLS VPNs (IETF RFC 4364)

Navigating Network Complexity (White and Tantsura)

NHRP (IETF RFC 2332)

OSPF Version 2 (IETF RFC 2328)