Scaling CSfC Mobile Access using DHCPv6 Prefix Delegation

Technical Whitepaper

Version 1.2

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
20201117 Version 0.1	Initial Draft	Nicholas Russo
20201203 Version 0.2	Technical corrections	Nicholas Russo
20201205 Version 1.0	Legal disclaimers and cleanup	Nicholas Russo
20210123 Version 1.1	Spelling and grammar corrections	Nicholas Russo
20211012 Version 1.2	Updated ipv6-tools to net-tools link	Nicholas Russo

Contents

Figures

Tables

1. Solution Overview

This section explains the Commercial Solutions for Classified (CSfC) program, why it exists, how consumers can implement it, and the specific architecture in scope for this document. Before continuing, users should have a basic understanding of IP routing, wireless technologies (WiFi, cellular, etc.), network security techniques, and IPv6 protocol operations.

1.1. CSfC Program and its Purpose

The CSfC program was created by the US National Security Agency (NSA) in response to requests from US Government organizations to provide a more flexible security framework. Traditionally, classified information transmitted over untrusted IP networks must be encrypted using a High Assurance IP Encryptor (HAIPE) device, which is a purpose-built hardware device. These devices are expensive, cryptographically controlled, and relatively simple devices with minimal feature sets. Fundamentally, HAIPEs are IPsec endpoints except that the HAIPE cipher suites, although based on the Advanced Encryption Standard (AES), are proprietary to the US Government.

The CSfC program empowers the NSA to certify various commercial products, such as routers, firewalls, and computers to act as substitutes for HAIPEs with respect to data encryption. This enables the US Government community to leverage their existing network equipment to meet NSA security requirements without new capital investments in HAIPEs. In tactical networks especially, the absence of HAIPEs results in reduced size, weight, power consumption, and cost. Such advantages often translate to improvements in troop mobility and combat effectiveness.

1.2. CSfC Capability Packages

In order to implement CSfC, the NSA has provided a variety of reference architectures known as capability packages. Each capability package details an arrangement of CSfC-approved products in a prescriptive manner that satisfies the NSA's security requirements (linked in Appendix B).

Although the exact specifications vary between capability packages, the underlying theme for each is that of a "dual encryption" design. To compensate for the relative weakness of commercial encryption and the increased likelihood of product bugs (at least compared to the proprietary US Government algorithms), all classified traffic must be encrypted twice for transport across untrusted networks. Such networks include the public Internet, a private WAN service, military satellite communications networks, line-of-sight radio meshes, and more. These untrusted networks are called "black" while the classified networks are called "red". These are not new terms, and traditional HAIPEs had black and red interfaces to securely interconnect these networks.

Given the "dual VPN" designs described within most of the CSfC capability packages, the intermediate "gray" carries traffic that has only been encrypted once. The size and scope of the gray network varies based on the capability package selected and the organization's connectivity requirements. At the time of this writing, the NSA offers the following capability packages:

a. **Mobile Access Capability Package:** Describes a "dual VPN" design across a generic black transport network. The client, using some mix of hardware and software, must establish two VPNs to various IPsec gateways at the black/gray (outer) and gray/red (inner) boundaries to securely access the red network. This capability package is discussed in depth within this document.
b. **Campus Wireless LAN (WLAN) Capability Package:** Like the previous capability package, this option applies only to WiFi transports. It allows WiFi Protected Access 2 (WPA2) to qualify as the first layer of encryption, forming the outer black/gray VPN. The end host, typically using a software-only VPN client, can then establish the inner gray/red VPN rather easily. Other wireless technologies, such as cellular, cannot rely on their native encryption techniques to qualify as a CSfC encryption layer.
c. **Multi-site Connectivity Capability Package:** Formerly known as the "VPN Capability Package", this option describes a site-to-site connectivity model. Imagine setting up a basic site-to-site VPN inside of another basic site-to-site VPN. This is among the easiest to implement because it doesn't require end host modifications, but it often demands the most hardware while providing the least mobility.
d. **Data At Rest (DAR) Capability Package:** This option isn't related to network transport security, but rather data-on-disk security It substitutes two layers of encryption for the two layers of VPNs described in the previous packages, relying on both platform-level (operating system) and file-level technologies working in concert to dual-encrypt a given file. The solution, like all the others, requires various forms of authentication for additional security.

1.3. Architecture Overview

The design outlined in this document uses the Mobile Access Capability Package (abbreviated as MACP), which combines network-based site-to-site VPNs (black/gray) and host-based remote-access VPNs (gray/red). Both VPNs are IPsec-based. The terms "black/gray" and "gray/red" are used interchangeably with the terms "outer" and "inner" throughout this document, respectively. This document also focuses primarily on the gray network architecture. The NSA has mandated many black and red requirements, such as the presence of firewalls, various protocol limitations, and more. These topics are non-controversial and are of little interest in this document, so they are discussed only briefly.

This design is unique among MACP implementations in several ways. First, it uses a dedicated device for each VPN, which is permitted by the MACP specification. Rather than utilize a fully wireless endpoint, clients connect to a small router with embedded switchports using short Ethernet cables. The router serves as the black/gray gateway and forms a site-to-site IPsec connection to the outer VPN headend. This simplifies the clients as there is no custom software needed for "dual VPN" connectivity. Instead, a basic remote access VPN client, such as Cisco

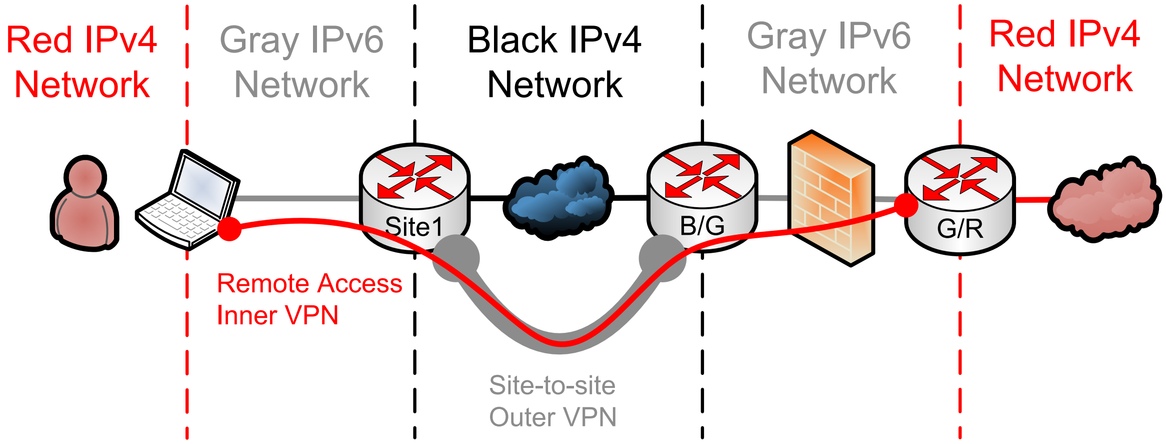
AnyConnect, forms the gray/red IPsec connected to the inner VPN headend. Given that clients are physically tethered to the remote routes, this architecture is best suited for pop-up sites where users do most of their classified work while seated, such as disaster relief facilities and command posts. Additionally, it is suitable for vehicular kits whereby each vehicle has a single router with pre-installed cabling to the various vehicular communication systems. The solution can be installed in land, air, sea, and space-based vehicles.

In addition to the hardware differences, the solution mixes IPv4 and IPv6 in a way that concurrently improves both security and scalability. In our environment, the red and black networks were IPv4-only, and although either one could use IPv6, such designs were irrelevant due to lack of demand. The gray network was a new concept and thus was not burdened by decisions and constraints from decades past. As a result, the gray network design leveraged IPv6 exclusively.

Deploying IPv6 improved security because of the gray network's position in the architecture. In this way, IPv6 serves as a buffer zone between untrusted (black) and trusted (red) networks, implying that any black/red leakage would have to traverse the gray network. When the gray network runs IPv6, the likelihood of red IPv4 traffic leaking into the black IPv4 network, or vice versa, is near zero. This is yet another layer contributing to the "defense in depth" architecture described by all the CSfC capability packages. Put another way, the usage of IPv6 in this design complements the MACP requirements by enabling a new dimension of network segmentation.

The diagram below depicts the high-level architecture using roughly the same graphical layout as the MACP documentation. Not all required components are explicitly depicted (such as the various firewalls and management services) for cleanliness. While this layout omits many technical details, it is symmetric and easy to understand, providing a good solution overview. The gray/red remote access VPN tunnel is transported inside of the site-to-site black/gray VPN tunnel, guaranteeing that any listeners on the black network must break two layers of strong IPsec encryption to compromise the secured data.

Figure 1 - High-level MACP Design



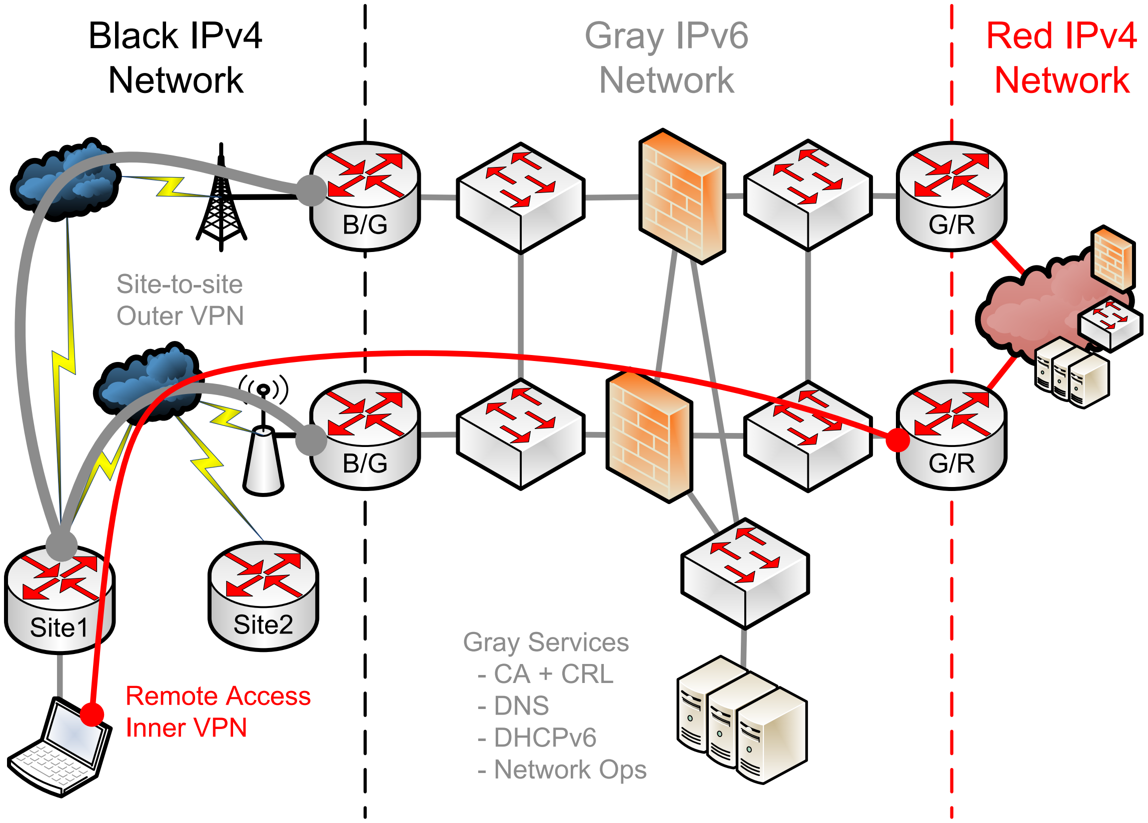
Deploying IPv6 by itself did not improve scalability, but instead enabled the use of Dynamic Host Configuration Protocol for IPv6 (DHCPv6) prefix delegation (PD). Rather than simply hand out individual IPv6 addresses, DHCPv6 PD hands out entire prefixes. In PD designs, the

DHCPv6 client is often a router, not an end host. The router can apply the prefix to its LAN interfaces in a dynamic fashion. The prefix can then be advertised to clients in an ICMPv6 router advertisement (RA) message, allowing clients to perform stateless address autoconfiguration (SLAAC). Each client on the LAN then has a unique and routable IPv6 address.

Additionally, DHCPv6 supports a stateless operation whereby SLAAC clients can receive non-address information from DHCPv6, such as the domain name and DNS servers. Each remote router can "import" this information from the DHCPv6 PD response and serve it to local SLAAC clients, saving WAN bandwidth and participating in a hierarchical DHCPv6 architecture.

For this document to be technically accurate, some additional components will be depicted on most of the diagrams from this point forward. The diagram below drills deeper into the architecture, depicting a high-availability setup with "two of everything", including both WiFi and cellular transports, gray firewalls, and various gray management services. It is assumed that comparable red services already exist and are not discussed in detail. Although not symmetrically depicted, the dual-tunnel IPsec design is still in effect.

Figure 2 - Medium-level MACP Design



1.4. Scope and Disclaimers

This document details a design that was built for a specific US Government customer but was never fully implemented nor deployed into production. As such, nothing in this document is a direct or indirect risk to any US Government network. The CSfC capability packages are marked as “resources for everyone” by the NSA and all information in this document is fit for public consumption. No customer data, explicit or implicit, is included.

This document does not detail a complete NSA-approved CSfC design instance. As mentioned earlier, this document gives extra focus to the gray networking and scaling aspects at the expense of black and red security aspects. Readers are encouraged to use the architectural concepts of this document in their own designs but must also comply with NSA guidance with respect to hardening, accreditation, and daily operations. This document is not an authoritative source on NSA security policies or CSfC requirements.

2. Black Network Design

This section details the black transport network, which consists of untrusted transport links between the remote/mobile sites and the upstream headend.

2.1. Transport Options

Any kind of IP-based transport is supported by this design and by the MACP at large, but this implementation accounted for three types of transport. They are listed in sequence from fastest and least mobile to slowest and most mobile. These transport types are depicted in isolation for clarity but can (and should) be combined to increase transport diversity and availability. Such integration designs are discussed later in the context of gray overlay networking.

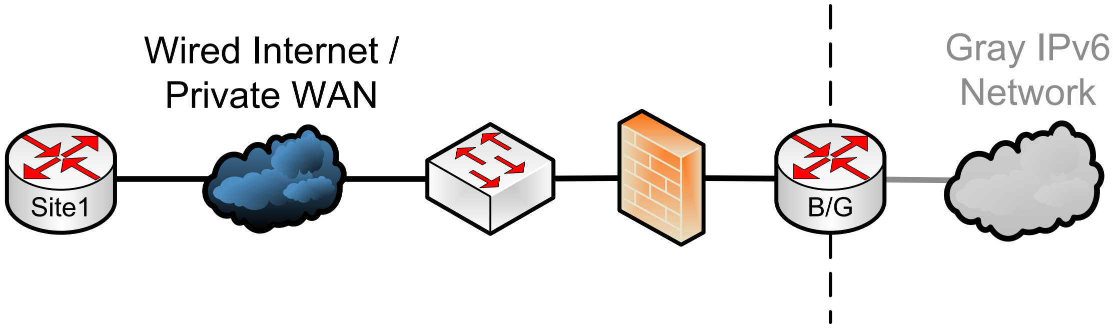
2.1.1. Wireline

This transport category includes all manner of wired connectivity to the headend, be it through the public Internet, private WAN service, dark fiber, or anything else. This option is easy to set up as it usually involves plugging in a few cables, receiving an IPv4 address through DHCP, and establishing a connection from the remote site router to the black/gray VPN headend.

In such a design, the only additional black network component required is a firewall. This firewall is deployed at the headend and should be used to screen incoming VPN sessions from clients, allowing only Internet Key Exchange (IKE) signaling, Encapsulating Security Payload (ESP) representing IPsec bearer traffic, and any required management traffic. Note that if Network Address Translation (NAT) occurs in the transport network, IPsec NAT Traversal (NAT-T) will add additional User Datagram Protocol (UDP) encapsulation atop the ESP traffic. This exposes a layer-4 port which NAT can use for translation.

The diagram below illustrates a generic wireline connection. Other auxiliary equipment, such as modems or media converters, may exist in the physical transit path. These items are not depicted as they have no effect on the logical topology.

Figure 3 - Black Wireline Connectivity



2.1.2. WiFi

Using WiFi as a transport mechanism is ideal for mobile sites, such as those mounted in vehicles or sites where rapid setup/teardown is critical. On the client side, a wireless access point (AP) is configured as a workgroup bridge (WGB) to perform the duties of a wireless client. It bridges the wired clients connected behind it, and since there is only one host device (the router), the vendor agnostic universal WGB (uWGB) can be used instead of the Cisco proprietary multi-host WGB feature. The Cisco Industrial Router 829 (IR829) is one example of a device that comes with a router and AP embedded in a single product.

At the headend, regular APs must be deployed to accept connections from the WGBs. In terms of security, several new components are required for WiFi connectivity. A Wireless LAN Controller (WLC) is often deployed at the headend to manage all the APs deployed there. The WLC centralizes the management of the APs, allowing all devices to share a common configuration. Some designs may also opt for centralized forwarding through the WLC using Control and Provisioning of Wireless Access Points (CAPWAP) tunnels, especially if the black headend network is large. In our case, two constraints prevented us from implementing a centralized forwarding design:

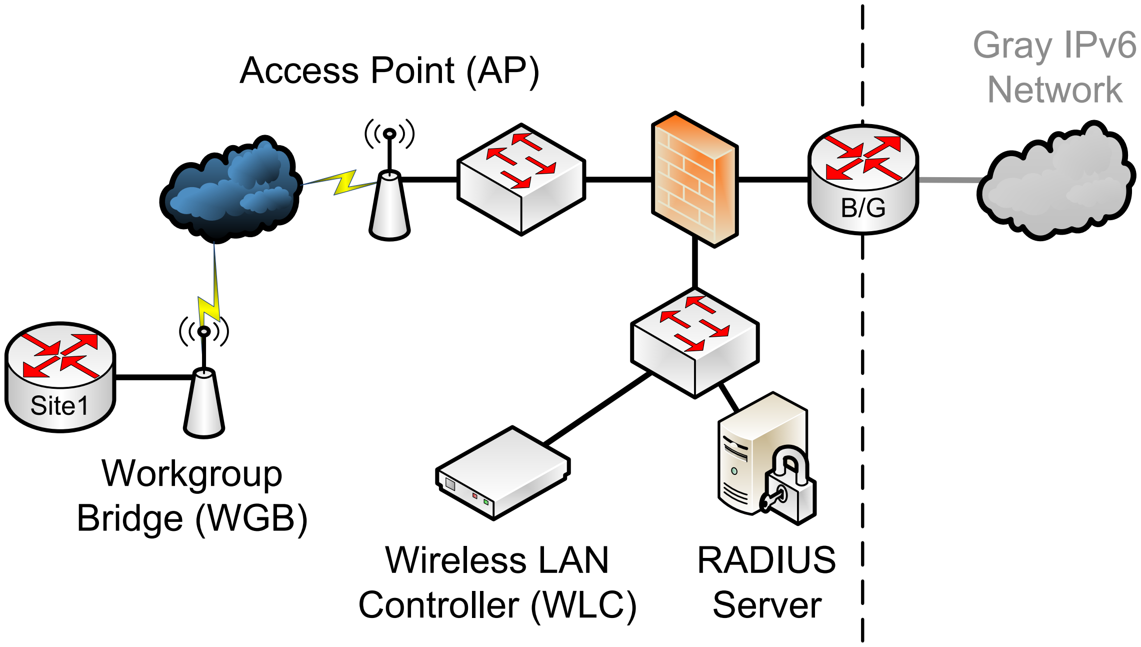
a. Our black network has only a single layer-2 switch, so the advantage of tunneling to a centralized WLC was minimal
b. The WLC was a virtual machine that does not support CAPWAP tunneling for data traffic. Even if it did, performance would likely be poor compared to direct Ethernet forwarding between the APs and the upstream black network devices

In addition to the WLC and headend APs, a Remote Authentication Dial-In User Service (RADIUS) server is required to implement 802.1X for WPA2 Enterprise authentication. This technique is stronger than a WiFi Pre-Shared Key (PSK) and enables a variety of authentication methods within the Extensible Authentication Protocol (EAP) family of protocols. Our design used an outer method of Protected EAP (PEAP) which establishes a TLS connection to the RADIUS server. Inside that secure communications channel, the WGB supplied a static username/password combination using EAP-MSCHAPv2. The WGB was configured to trust the RADIUS server's Transport Layer Security (TLS) certificate ahead of time which allowed the TLS connection to be established in the first place. Only one Service Set ID (SSID), which represents a specified WiFi network, needs to be created. All WGBs connect into this single network for simplicity, although more SSIDs could be created as needed.

Note that using a more secure approach, such as EAP-TLS (along with various other CSfC requirements), could allow WiFi to qualify as the outer VPN layer. However, this WiFi-specific solution cannot be applied to non-WiFi transports and so was rejected from our consideration.

The diagram below illustrates how all these pieces fit together, which still includes a black firewall for traffic filtering between the upstream AP and the black/gray VPN headend.

Figure 4 - Black WiFi Connectivity



2.1.3. Cellular

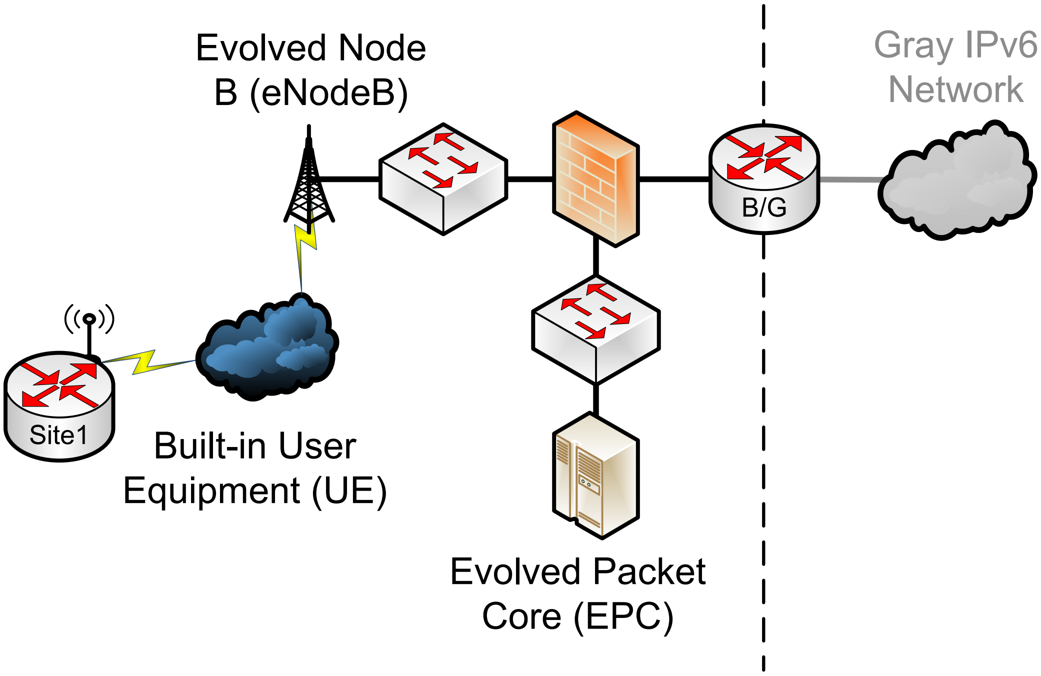
Cellular connectivity generally falls into one of two categories. The first and simplest approach leverages commercial cellular networks based on preexisting infrastructure. In these cases, the headend typically connects to the public Internet using a wireline connection while remote devices, such as Cisco IR829 cellular-capable routers, connect over the cellular networks. In commercial cellular networks, data encryption is often applied to the transport data, but the MACP specifications do not expect nor count it as a VPN layer.

The second and more complex approach is to deploy a private cellular network, which is detailed below. In general, cellular architecture of any generation is complex and is beyond the scope of this document. The list below broadly summarizes the components required in the context of fourth generation cellular, better known as "4G" or Long Term Evolution (LTE):

a. **User Equipment (UE):** The LTE client device, such as a mobile phone or the aforementioned Cisco IR829. This is comparable to the WGB from the WiFi section.
b. **Evolved Node B (eNodeB):** The cellular device that accepts connections from UEs, much like the APs from the WiFi section. Note that unlike WiFi, LTE operates in Government-regulated frequency bands which must be coordinated and deconflicted before use. These frequency bands vary between countries.
c. **Evolved Packet Core (EPC):** The EPC is effectively the LTE control-plane headend that connects to the existing wired network and provides connectivity for UEs. eNodeB devices register to the EPC, which is comparable to the WLC from the WiFi section.
d. **Access Point Name (APN):** The APNs are individual virtual networks that provide cellular access. They are configured and managed on the EPC and are announced over the air via eNodeBs, much like SSIDs from the WiFi section.

Although the exact products, configurations, and operations are different than WiFi, the general architecture is the same. UEs connect to eNodeBs over cellular transport by joining a specific APN (it makes sense to only create one for this design) which connects to the larger black network via the EPC. The diagram below illustrates the high-level architecture.

Figure 5 - Black Cellular (LTE) Connectivity



2.2. Underlay Routing Security Techniques

Regardless of the transport used, there are some best practices with respect to security that should be implemented. The MACP relies heavily on various, dedicated firewalls for traffic filtering, especially at the headend. Given that our MACP design used separate physical devices for the black/gray VPN (a router) and the gray/red VPN (a client), there are opportunities for additional hardening throughout the client's equipment stack.

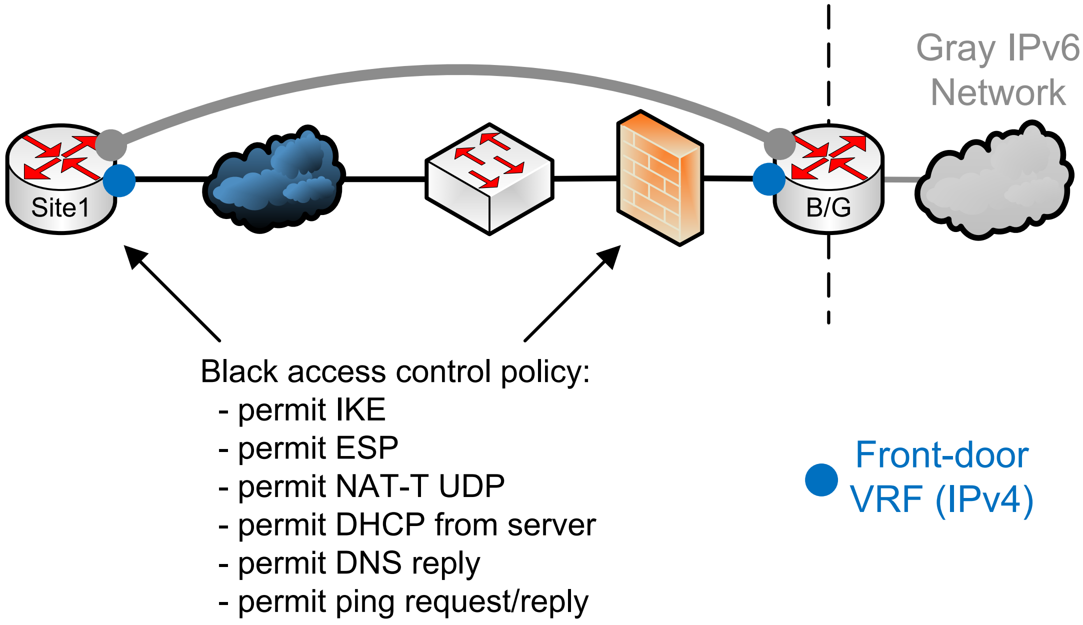
First, consider using front-door VPN routing and forwarding (VRF) instances on all transport links. Assigning router interfaces to different VRFs places them into different virtual routing tables, preventing any leakage across tables. Since the outer VPN routers are both gray devices and black devices, separating gray from black using virtualization is an extra layer of security that comes at zero cost. In our design, the black network is IPv4 and the gray network is IPv6, which makes leakage nearly impossible even without the VRF. The addition of the VRF makes the likelihood of such leakage infinitesimally small.

Second, consider using access-control lists (ACLs) on the router's transport interfaces. Since these interfaces only serve as VPN tunnel endpoints, the list of traffic types allowed to reach into and originate from these interfaces is very short. That list must include IKE and ESP at a minimum. Depending on the connection type, it may include DHCP, DNS, and NAT-T UDP.

Operators may optionally permit ICMP echo-request and echo-reply messaging for ping testing, but other ICMP messages such as unreachables and redirects should be blocked.

The same concepts apply at the headend, except the ACLs should be offloaded from the black/gray VPN headend to a dedicated black firewall. The black/gray VPN headend should continue to use front-door VRFs for routing separation. If the black firewall supports advanced Intrusion Prevent System (IPS) features, consider enabling those on permitted, unencrypted traffic types such as DHCP, DNS, and ICMP. This will further protect against packets with malicious payloads. The diagram below illustrates how these technologies work in concert to create a strong barrier between black and gray networks on the outer VPN endpoints.

Figure 6 - Black Security Techniques



Black access control policy:
- permit IKE
- permit ESP
- permit NAT-T UDP
- permit DHCP from server
- permit DNS reply
- permit ping request/reply

3. Gray Network Design

This section details the gray network architecture, which is the primary purpose of this document. Like the black network, the gray network is generally "transport only" with the exception of some additional components required for black/gray VPN establishment and gray/red VPN traffic filtering.

3.1. Required Headend Services

Combining the requirements of MACP with the required components on this specific design leads to a sizable list of gray network services. First, a certificate authority (CA) is required to issue certificates for all black/gray VPN endpoints. Within this design, the outer VPN is always an IPsec tunnel operating between two routers (site-to-site). Each black/gray gateway will receive its own certificate specifying whatever cipher and hashing algorithms are required by the NSA at the time.

Additionally, black/gray VPN headends will reach back to the CA to check the certificate revocation list (CRL) to ensure clients that connect are valid. If a client certificate appears to be valid but is marked as revoked in the CRL, the client is not authenticated and is denied access to the gray network. This document does not detail the low-level details of CRL operations as they are identical to standard CRL operations in traditional VPN designs.

This specific design requires two key IPv6 servers: DHCP and DNS. DHCPv6 plays a major role in the gray network as it relates to onboarding new devices. DNS is used by the gray/red VPN endpoints to resolve the IPv6 addresses of the gray/red VPN headends using generic hostnames. Both services are relatively basic but important; both are discussed more later.

Like the black and red networks, the gray network requires a traffic filtering firewall. The firewall in this design has 3 zones: inside, outside, and servers. This is like a traditional Internet Edge design with a demilitarized zone (DMZ) of servers that are universally accessible with some restrictions. The difference here is that the inside zone never reaches into the server zone. The firewall design is discussed in greater detail later.

Other MACP-mandated services and devices, such as management laptops, should also be included in the final package. These are not discussed in this document as they are not relevant to the key topics at hand.

3.2. Remote Site Connectivity (Black/Gray VPN)

Connecting to remote sites via a black/gray IPsec VPN is the most innovate aspect of this design. This document details four validated connectivity options.

3.2.1. Single Hub and Single Transport

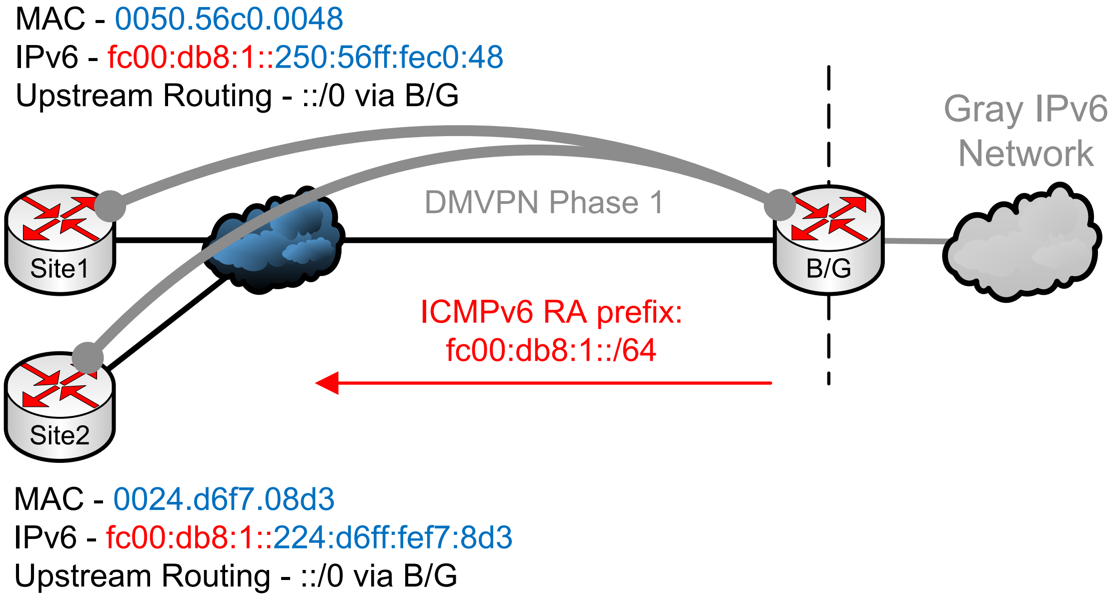
In the simplest case, there is a single black/gray VPN headend with a single transport technology connecting the remote sites to the headend. Only a single black/gray tunnel interface is needed, and Cisco's Dynamic Multipoint VPN (DMVPN) was best suited for this overlay as it allows spokes to dynamically establish connections to hubs. Because the gray network only exists to connect gray/red VPN clients to the gray/red VPN headend, there is no need for spoke-to-spoke connectivity. In fact, spoke-to-spoke connectivity is a security risk that should be prevented entirely within this design. Using DMVPN Phase 1, a strict hub/spoke design, guarantees that all traffic must traverse through the hub. Applying an inbound ACL on the hub tunnel is a simple technique to prevent spoke-to-spoke traffic.

With respect to routing, there are two general solutions:

a. Use a standards-based routing protocol, such as Open Shortest Path First (OSPF) or Border Gateway Protocol (BGP)
b. Enable ICMPv6 Router Advertisements (RA) to be sent from the hub so spokes dynamically learn an IPv6 default route via the hub

Relying on ICMPv6 RAs is the simpler option for the single hub/single transport design. The remote sites act like SLAAC clients and will use their tunnel source interface MAC addresses (which are Ethernet interfaces) to generate EUI-64 addresses for the tunnels. Typically, these will create both link-local addresses (LLA) and unique local addresses (ULA) on the tunnel. The diagram below illustrates the tunnel IPv6 addressing and upstream routing.

Figure 7 - Remote Site Tunnel SLAAC and Upstream Routing

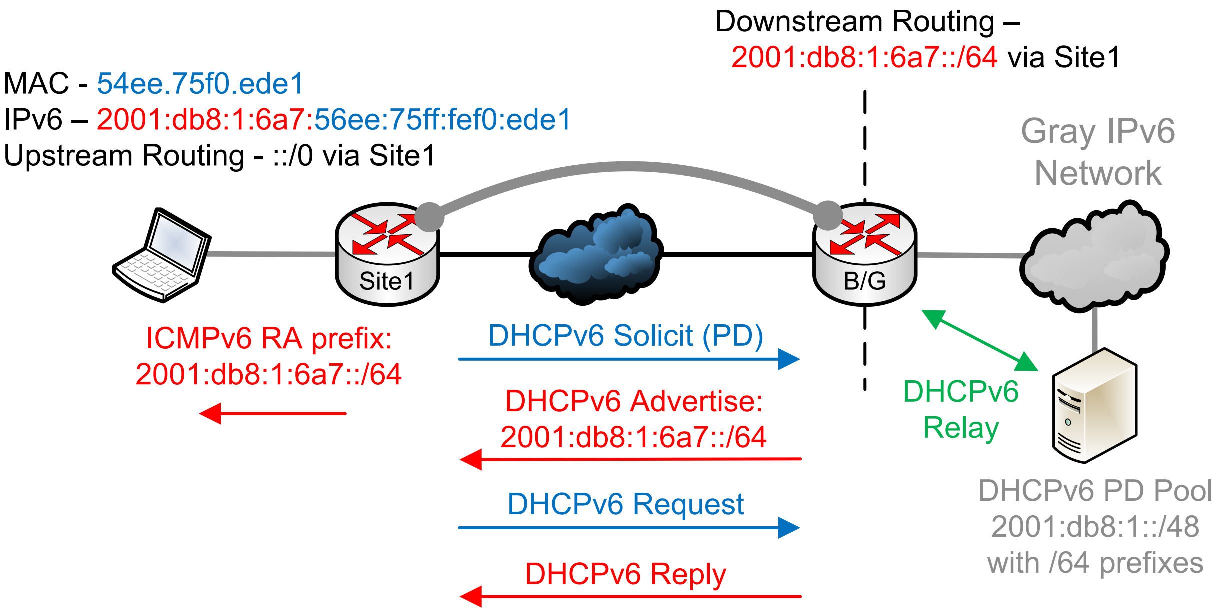

MAC - 0050.56c0.0048
IPv6 - fc00:db8:1::250:56ff:fec0:48
Upstream Routing - ::/0 via B/G
Gray IPv6 Network
DMVPN Phase 1
Site1
B/G
ICMPv6 RA prefix:
fc00:db8:1::/64
Site2
MAC - 0024.d6f7.08d3
IPv6 - fc00:db8:1::224:d6ff:fef7:8d3
Upstream Routing - ::/0 via B/G

The DHCPv6 design and subsequent downstream routing is more complex. Assuming that the DHCPv6 server is not collocated on the hub router, the hub router must act as a DHCPv6 relay. Upon receiving the DHCPv6 Solicit message from the client, which requests a delegated prefix, the hub will relay it to the DHCPv6 server. This server, along with all other gray servers, are protected behind the gray firewall, which must permit DHCPv6 relay traffic. The DHCPv6 server then allocates an IPv6 prefix from the available pool and communicates this to the relay (the hub). The hub then sends it to the client in a DHCPv6 Advertise message.

Like DHCP for IPv4, there are two additional messages used to finalize the DHCPv6 exchange. Upon receiving the Advertise message, the client sends a Request message and the server responds with a Reply message, terminating the DHCPv6 PD process. The only difference is that the client Request message contains a server identifier option identifying the server to which it is making the PD request.

The hub also dynamically installs a static route for the delegated prefix with a next-hop of the DHCPv6 client. This allows the upstream network to reach the delegated prefix and often the hub will redistribute these static routes into whatever interior gateway protocol (IGP) is running within the headend gray network. If there is only one hub and if the headend gray network is very small, the entire design can leverage static routes. These are more secure than any IGP or BGP within the gray network because route injection/spoofing is impossible unless devices themselves are compromised. The diagram below illustrates the DHCPv6 PD process and additional routing details related to it.

Figure 8 - Gray DHCPv6 and Routing Updates



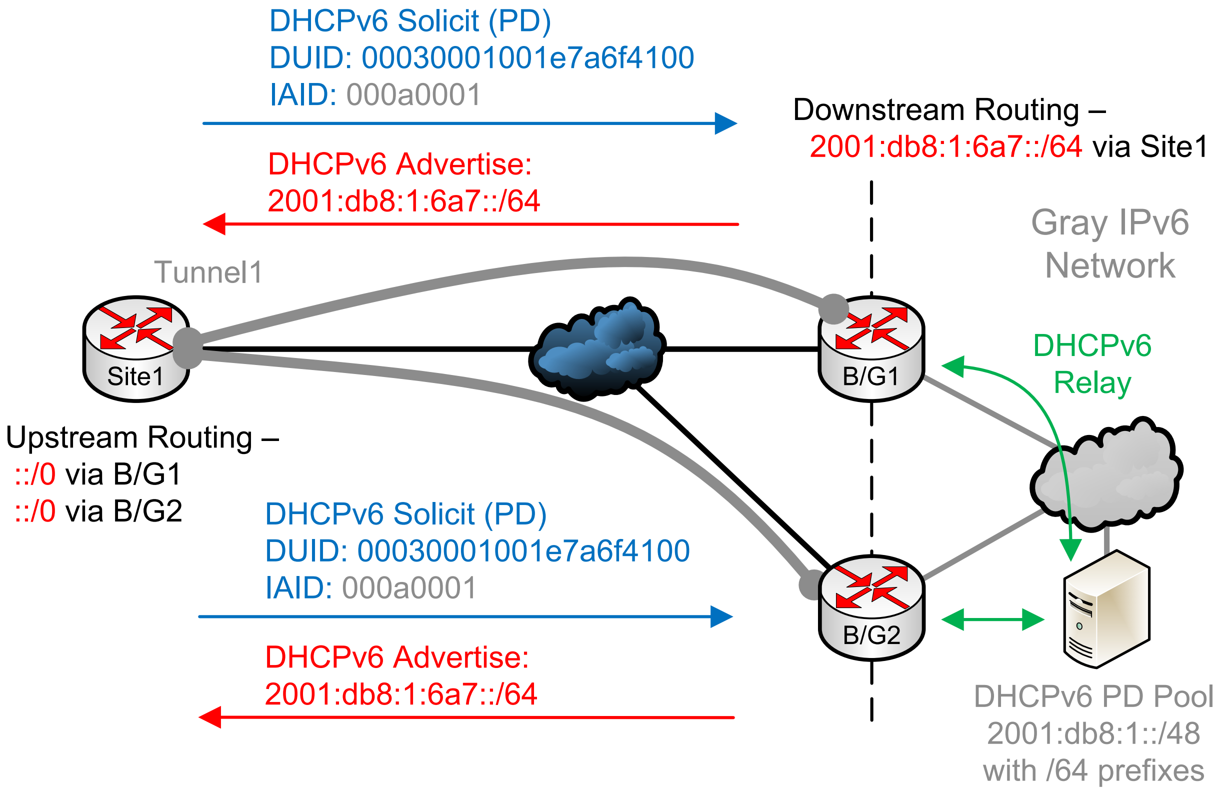
3.2.2. Multiple Hubs and Single Transport

To improve availability, consider adding multiple hubs (two is usually sufficient). If the gray network is larger than a single tactical kit, these hubs should be geographically separated to protect against regional network transport or power outages. This design still assumes there is only one transport type being used. Therefore, each client still only has one tunnel, but with multiple upstream hubs defined for DMVPN registration.

The DHCPv6 process becomes a bit more complicated. There are two important hexadecimal strings used to identify DHCPv6 clients. First, the DHCPv6 Unique Identifier (DUID) is a device-level string that uniquely identifies a device. Vendors compute it differently, but it typically includes the lowest on-device MAC address and may include some additional data as well. Second, every interface has an Identity Association Identifier (IAID) which differentiates between interfaces on a device. Thus, the 2-tuple of DUID + IAID uniquely identifies a device interface, each of which can receive a unique delegated prefix.

Because there is only a single tunnel interface with multiple upstream hubs, the DHCPv6 server will respond with the same delegated prefix to each hub. While this does add some extra DHCPv6 traffic over the network, it is generally harmless and operationally insignificant. The DHCPv6 client (the router) stores the delegated prefix, assigns it to its LAN interface, and the ICMPv6 RA and SLAAC processes occur normally. The diagram below illustrates this process; some of the DHCPv6 messaging is omitted for cleanliness.

Figure 9 - Multiple Hubs Servicing a Single Transport



Designers still must choose whether to use a routing protocol over the DMVPN mesh or not. If a routing protocol is not used, all hubs should enable the transmission of RAs over the tunnel (without any prefixes) so that clients can learn upstream default routes. The RA interval and lifetime timers should be tuned in such a way that the availability and failover requirements of the organization are met. ICMPv6 signaling was not meant as a substitute for routing protocols, and while relying on ICMPv6 RAs is simple, it will converge more slowly.

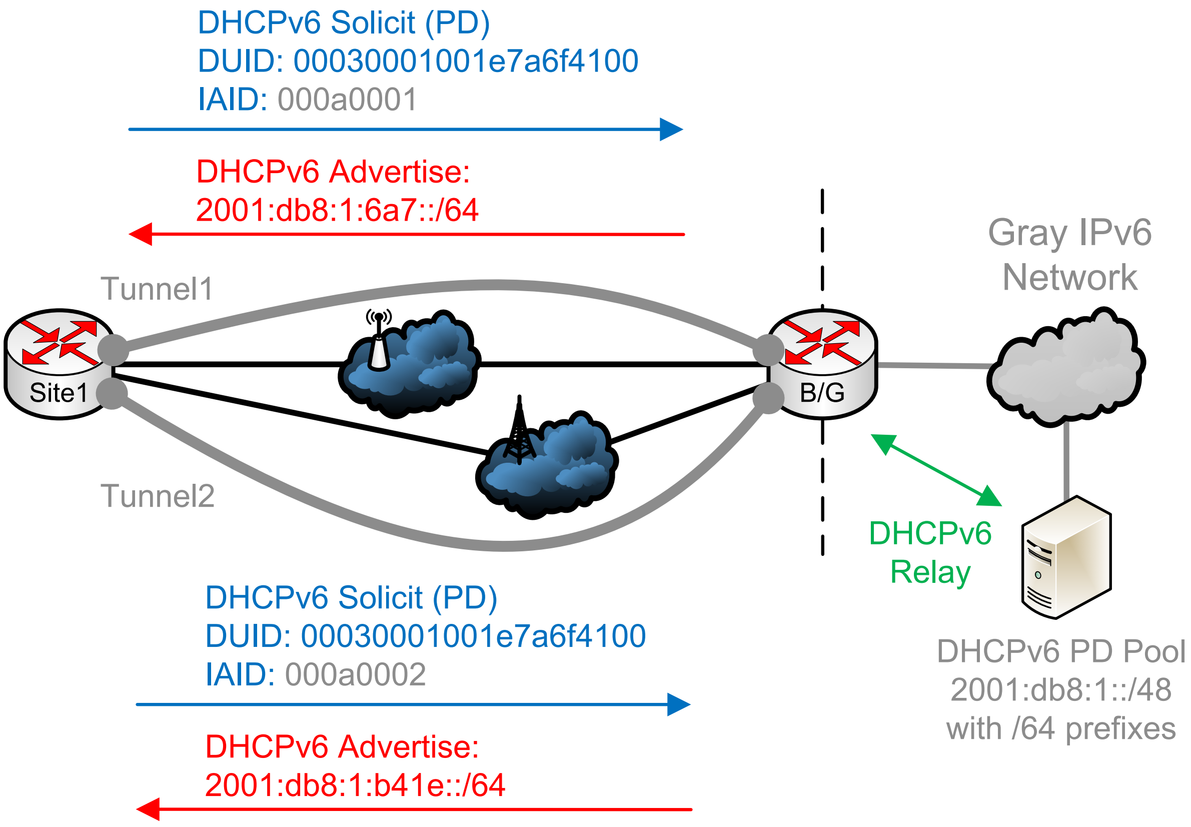
Alternatively, enabling a routing protocol over the tunnel, such as OSPFv3 or BGP, could provide faster convergence. The hubs would send default routes (or, for improved security, more specific routes for only the gray/red VPN headend) down to the spokes. The spokes would not need to send anything to the hubs since their delegated prefixes are already installed on the hubs via static routes, thanks to the intelligent DHCPv6 relay process. These static routes should be redistributed into the gray headend routing protocol as required, much like in the previous example.

3.2.3. Single Hub and Multiple Transports

In networks where the headend footprint is small (single hub), high availability can be achieved by using multiple transports. While this does not protect against a headend node failure, it protects against the far more likely case of transport link failures.

The design is significantly more complex than the "single transport" options because it requires multiple DMVPN meshes, typically one per transport type. Because there are multiple tunnel interfaces, there will be multiple DUID + IAID tuples with respect to DHCPv6 PD. This implies that each remote site will receive N delegated prefixes where N is the number of available transports. In this design, the gray/black VPN headend connects to each transport, ideally using a separate front-door VRF, and supports multiple DMVPN hub tunnels. Each tunnel will relay DHCPv6 client messaging to the DHCPv6 server and receive individual responses with different delegated prefixes. The diagram below illustrates this process.

Figure 10 - Remote Site Receiving Multiple Delegated Prefixes



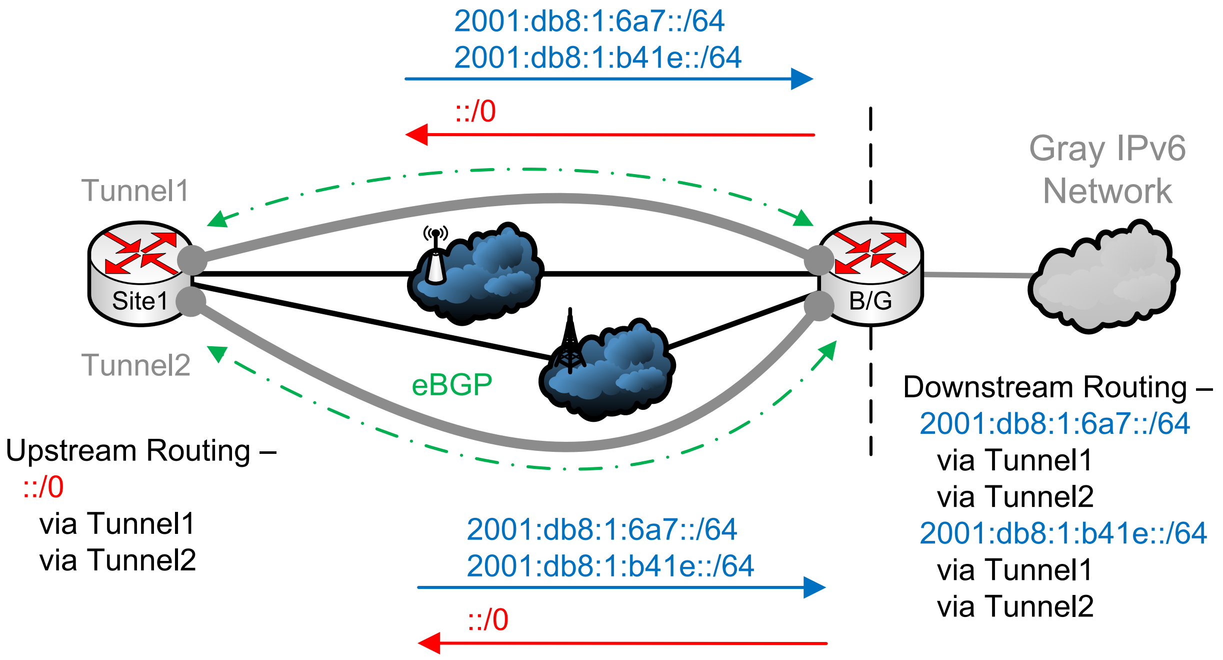
Designs with multiple transports require the use of a dynamic routing protocol. BGP is recommended because it scales over large hub/spoke networks and provides additional filtering/summarization control. The need for dynamic routing is best explained with a failure example. Suppose a router comes online over WiFi and cellular, then receives two separate delegated prefixes. Because IPv6 allows a router interface to announce multiple IPv6 prefixes in an RA message, the client performing SLAAC will derive one address per delegated prefix. Only one of these addresses will be selected as the source address for the gray/red remote access VPN session. If the tunnel from which that specific prefix was received fails, the remaining tunnel must be able to route traffic from the failed tunnel's prefix.

Some network vendors, such as Cisco, support BGP dynamic neighbors. In this design, the DMVPN hubs passively listen for BGP sessions initiated by clients and respond appropriately. This means the hubs do not need to enumerate every remote BGP session, reducing the configuration burden on network operators. The hub addressing is static, so the clients all connect to the same address per each tunnel type.

To simplify the overall routing architecture, the DHCPv6-installed static routes on the hubs should be disabled. Instead, the clients should advertise their delegated prefixes via BGP back to the hubs dynamically. This seems counterintuitive at a glance; the hubs relayed the delegated prefixes in the first place, so why do they need the clients to communicate those prefixes back upstream? Doing this unifies all the routing within BGP and reduces the need for complex redistribution, filtering, and route selection at the hubs.

On the topic of BGP policy application, external BGP (eBGP) should be used for the BGP sessions between hubs and spokes. All spokes can be placed in the same AS since they never need to exchange routes with one another. This also simplifies configuration management and the overall network design. Consider again the failure case from earlier. If one transport tunnel fails and the client is using an IPv6 address from that tunnel's delegated prefix for its gray/red VPN, that prefix can be safely routed over the second tunnel using BGP. The diagram below illustrates how BGP can be used to provide dynamic routing for delegated prefixes.

Figure 11 - Using eBGP for Dynamic Failover across Multiple Transports



In addition to the failover requirements just described, designers should carefully consider the lifetime timer on delegated prefixes. Per RFC 8415 describing DHCPv6 operations, these timers are also known as T1 and T2, but this document uses their English names for simplicity.

There are two timers:

1. **Preferred lifetime (T2):** When this expires, the DHCPv6 clients tries to renew its delegated prefix with the DHCPv6 server.
2. **Valid lifetime (T1):** The upper-bound on how long a DHCPv6 client waits before deciding that its PD renewal has failed.

Using some combination of non-infinite values means that after a period of silence (i.e., when a transport link fails), prefixes can be returned to the pool and re-issued to new clients. This would be desirable in two cases:

1. When the available pool of prefixes is small and recycling is important.

2. When users tolerate having their gray/red VPNs disconnected on occasion and are willing to re-establish these sessions after failures occur.

Alternatively, using infinite values for both timers means that a delegated prefix will be retained forever on both the DHCPv6 server and the client (the router). This reduces the likelihood that clients will need to reconnect their gray/red VPNs after prolonged failures but also consumes more prefixes.

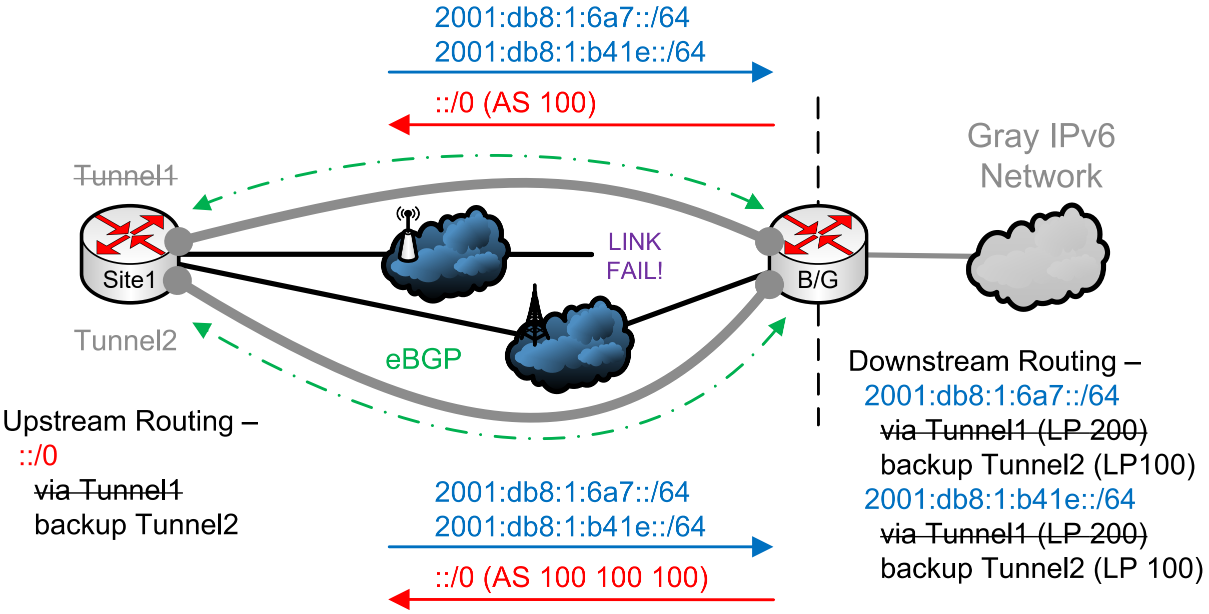
As a final consideration, consider whether the design should be active/active or active/standby with respect to tunnel selection. For example, if one tunnel is WiFi and the other is cellular, it's highly likely that the WiFi tunnel is better performing in terms of bandwidth, latency, jitter, and packet loss. If the cellular network is commercially operated, the WiFi transport is probably less financially expensive as well. The main drawback of WiFi is the limited range and mobility compared to cellular (in most cases, at least). Therefore, it makes sense to exclusively leverage WiFi when it is available, then switch to cellular only when necessary.

To accomplish this in the simplest and most scalable way, use BGP policy attributes to influence routing. It is recommended to only apply BGP policies on the hub nodes, not on the spokes, to simplify configuration. To select one tunnel over the others, configure these BGP policy attributes on the hubs:

1. Use inbound local-preference on all delegated prefixes learned from spokes. On the preferred hub, use a higher value than the backup tunnel (e.g., 200 versus 100). This attribute is preserved within an AS, and if the entire gray upstream network uses the same AS, all BGP speakers will prefer the primary hub as the egress gateway towards the remote sites.
2. Use outbound AS-path prepending on all upstream gray routes to which the spokes expect to send traffic (e.g., the gray/red VPN endpoints). On the backup hub, append the local AS several times, and simply do nothing on the preferred hub. Note that this approach is more robust than using multi-exit discriminator (MED) because the AS-path length is compared even when the peer AS numbers are different. This design decision eliminates one point of misconfiguration if hubs use different AS numbers, perhaps in a large-scale regionalized design.

The diagram below illustrates the application of this BGP policy in action during a failover. Tunnel1 uses a higher local preference and shorter AS path length, meaning it would be preferred for both upstream and downstream traffic. When Tunnel1 fails, all devices switch to Tunnel2, and it doesn't matter which IPv6 source address the inner VPN clients selected. Both prefixes are routable over both tunnels which permits seamless failover and minimal outages.

Figure 12 - Using BGP Policies to Build Active/Standby Overlays



As a final note, both the hubs and remote sites can use device-specific BGP fast-reroute techniques to install both the primary and alternate routes to a given destination. If a transport link fails, ultimately leading to an eBGP session failure, the routers can immediately install backup paths. This bypasses the BGP best-path selection process, improving convergence time.

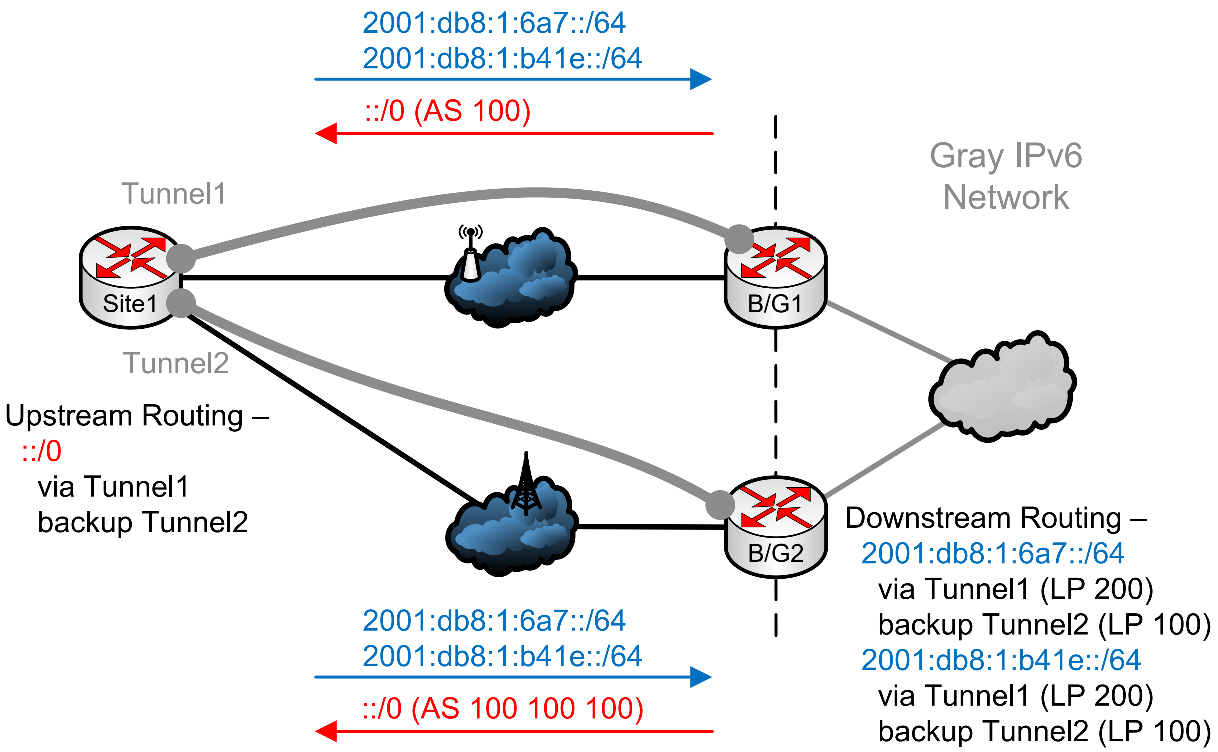
3.2.4. Multiple Hubs and Multiple Transports

This design is the most complex but also offers the highest availability. It requires dynamic routing between hubs and spokes given that there are multiple transports. Dynamic routing is also required in the upstream network as there are multiple hubs and thus multiple ingress/egress points with respect to the remote sites.

The main design consideration with this option is the placement and purpose of each hub. The most common design would be to assign each hub to a single transport type. For example, one hub services WiFi tunnels while another hub services cellular tunnels. This simple "divide and conquer" approach provides the best availability with the minimum number of cross-connecting links. The design can tolerate a single transport failure or a single hub node failure, but not both at the same time. To increase availability even further, all hubs could connect to all transport networks. In practice, this tends to be overly complicated and is not recommended (as evidenced by decades of enterprise WAN and Internet Edge implementations).

The diagram below illustrates this design. The BGP routing design and failover considerations described in the previous section still apply here.

Figure 13 - High Availability using Multiple Hubs and Multiple Transports



3.3. IPv6 SLAAC for VPN Clients

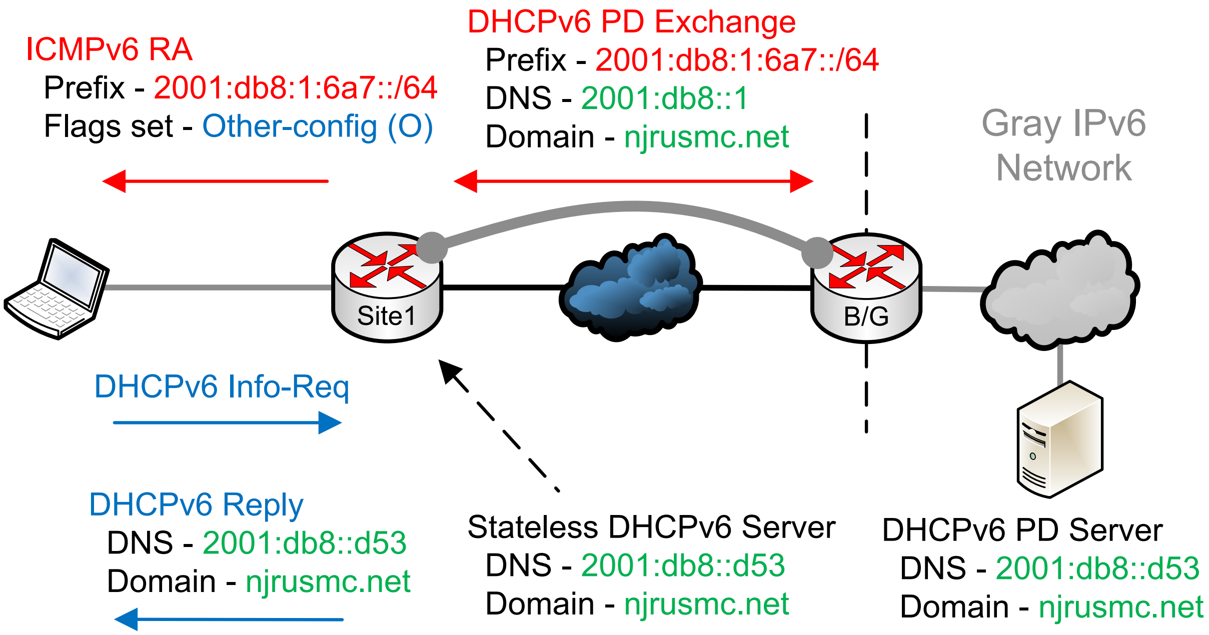
Regardless of which gray transport option is selected, the DHCPv6 Advertise and Reply messages containing the delegated prefix may also carry “other configuration” information such as the domain name and DNS servers. This information is retained within the DHCPv6 PD client (the router) itself, which adds no value to the black/gray VPN formation. Instead, each router should serve as a stateless DHCPv6 server for gray/red VPN clients. Rather than hand out delegated prefixes or managed addresses, the router can communicate the domain name and DNS server information to those clients. In this way, the design leverages a hierarchical DHCPv6 design whereby the top-level DHCPv6 server identifies a stateful PD pool and various “other configuration” options. The second level DHCPv6 servers (remote routers) only distribute stateless information to gray/red VPN clients to which they connect.

To signal that “other configuration” information is available, the remote router sets the O-flag on its ICMPv6 RA messages sent onto the client LAN. RAs are sent both in response to client-originated Router Solicitations (RS) and periodically based on a fixed interval. Regardless, all RAs carry the delegated prefix and O-flag (along with many other minor values such as the hop count, and router preference). Upon receiving the RA, clients will perform SLAAC to derive a unique IPv6 address from which they can source their VPN connections to the gray/red gateway.

Leveraging SLAAC only requires the reception of a prefix within the RA, but clients also need to learn the domain name and DNS servers. This information cannot be carried natively in an RA message, but when the O-flag is set, this instructs the client to send a DHCPv6 Information-Request message to the router. Because the remote routers are stateless DHCPv6 servers themselves, they respond with the requested information using a Reply message. This almost always includes the domain name and DNS servers, but may also include Simple Network Time Protocol (SNTP) servers and other vendor-specific information.

The diagram illustrates the hierarchical DHCPv6 architecture. Note that the stateless DHCPv6 between gray/red client and black/gray router only requires two messages, not four.

Figure 14 - Sending DNS and Domain Information to Clients using Stateless DHCPv6



While utilizing stateless DHCPv6 in this way is useful for most environments, it does increase the complexity within the overall solution. When the gray network is very small and the gray/red VPN headend uses a fixed gray IPv6 address for VPN termination, stateless DHCPv6 becomes unnecessary. Clients do not need any DNS connectivity on the gray network if their remote access VPN profiles have been hardcoded with the correct destination IPv6 address representing the gray/red VPN headend. However, such hardcoding limits future growth and flexibility, leading to poor scale and manageability in the future.

3.4. IPv6 Network Allocation Recommendations

The previous explanations about gray networking were conceptual and design oriented. Given that many engineers lack operational experience with IPv6 in general, this section aims to provide some logical, concrete IPv6 allocation examples.

One approach is to focus on the three major “areas” to the gray network within each kit:

a. **The core network:** This encompasses any gray components upstream from the black/gray gateway. This could be as small as a firewall, a switch, and a server, or as expansive as an enterprise network spanning multiple continents. In general, this includes gray/red gateways, gray management services, gray firewalls, and any other non-remote gray devices or services that remote clients may need to access.

b. **The transit tunnels:** Each black transport type serves as an underlay for a black/gray outer VPN tunnel. Normally, this addressing isn't terribly relevant, but given the DHCPv6 relay functionality implemented on each hub, these networks must be reachable within the core network (at least from the DHCPv6 server's perspective).
c. **The DHCPv6 PD pool:** When black/gray VPNs are formed, each remote tunnel is assigned a prefix (often a /64) from a pre-identified pool. While the routing of these prefixes is design dependent, let's assume the most complex option was chosen whereby they are routed to upstream hubs using eBGP. These prefixes must have reachability to the DNS servers and gray/red VPN headends at a minimum.

Another dimension to consider is the usage of unique local addressing (ULA) versus global unicast addressing (GUA). The former is comparable to RFC1918 private addressing for IPv4 while the latter is comparable to Internet-routable, public addressing. Since there is no such thing as a "gray Internet", even in the context of Government networks, one could use ULA for the entire gray design, even across the enterprise. The remainder of this section will use ULA for that reason.

Working through that example, we begin with fc00::/7, encompassing the entire ULA range. First, carve up the enormous ULA range into a few blocks representing each major area. Keep in mind that the gray network may be used for other things completely unrelated to this design, so don't be too greedy. The table below illustrates a simple example.

Table 1 - Allocating Prefixes to Gray Network Areas/Functions

IPv6 Network Range	**Purpose**
fc00::/16	Unrelated to this design
fc01::/16	Core network services
fc02::/16	DMVPN transit links
fc03::/16	DHCPv6 PD pools

Now that large chunks of address space have been allocated across gray network areas, consider additional granularity using a site identifier. In highly distributed networks, there are many disparate kits, each servicing their own set of remote sites/clients. As such, don't be overly conservative with addressing; allocate a large portion, perhaps 16 bits, of the IPv6 address to contain this information. The following table illustrates some examples on how to further refine the addressing plan.

Table 2 - Area-based, Per-site Allocation Example

IPv6 Network Range	Location	Purpose
fc01:d::/32	Site 13	Core network services
fc02:d:2::/64	Site 13	WiFi DMVPN tunnel
fc02:d:3::/64	Site 13	Cellular DMVPN tunnel
fc03:d::/32	Site 13	DHCPv6 PD pool
fc01:2e::/32	Site 46	DHCPv6 PD pool
fc02:2e:1::/64	Site 46	Wireline DMVPN tunnel
fc03:2e::/32	Site 46	DHCPv6 PD pool

The main advantage of this area-based approached is simplified firewall rulesets and routing policy application. For example, only traffic from fc03::/16 should be permitted towards gray/red VPN gateways. Only traffic from fc02::/16 should be permitted towards the DHCPv6 server. Some exceptions may apply, but these rules are generally true and may simplify long-term network operations and configuration management.

The main drawback is that it aggregates poorly from a routing perspective. It guarantees that each site will originate at least three prefixes (and probably more if contiguity isn't possible). An alternative approach would be to carve up the available IPv6 address space by region. Suppose that this design is deployed across three regions whereby a large, global-scale gray network interconnects everything, as indicated below. As always, it's a good idea to account for any legacy or unforeseen sub-networks.

Table 3 - Allocating Prefixes to Gray Geographic Regions

Region	IPv6 Network Range
Legacy/unforeseen/experimental	fc00::/16
Americas	fc01::/16
Europe, Middle East, and Africa	fc02::/16
Asia Pacific	fc03::/16

Within each region, each site could receive a /32 prefix from the regional /16 prefix. Within each site, each network area could receive a /48 prefix from the site-specific /32 prefix. This hierarchical approach, especially when performed at clean 16-bit boundaries, is relatively easy to understand. The following table illustrates one such example.

Table 4 - Regional, Hierarchical Allocation Example

IPv6 Network Range	Location	Purpose
fc01:d::/32	Site 13	Site-specific prefix (Americas)
fc01:d:1::/48	Site 13	Core network services
fc01:d:2::/48	Site 13	DMVPN transit links
fc01:d:2:2::/64	Site 13	WiFi DMVPN tunnel
fc01:d:2:3::/64	Site 13	Cellular DMVPN tunnel
fc01:d:3::/48	Site 13	DHCPv6 PD pool
fc03:2e::/32	Site 46	Site-specific prefix (Asia Pacific)
fc03:2e:1::/48	Site 46	Core network services
fc03:2e:2::/48	Site 46	DMVPN transit links
fc03:2e:2:1::/64	Site 46	Wireline DMVPN tunnel
fc03:2e:3::/48	Site 46	DHCPv6 PD pool

These examples are not meant to be exhaustive. Individuals may find that GUA addressing (or addressing supplied by a centralized Government authority) is the better fit for their environment. The same prefix allocation concepts are applicable regardless of the prefix range.

3.5. Gray Firewall Design

This section describes the two key aspects of gray firewall design; device placement/availability and policy rules.

3.5.1. High Availability

All CSfC capability packages permit and encourage high availability for any device or service in each design, provided the overall architectural constraints are met. Because most firewalls are stateful, deploying them in parallel requires some additional planning. There are three approaches for implementing highly available firewall designs:

1. **Use NAT to ensure symmetric routing:** This technique is most used at the Internet edge where firewall operate independently (no state sharing). Flows that egress through a given firewall must also ingress (return) through the same firewall. NAT is a crude tool used to guarantee routing symmetry in these designs.

2. **Pair the firewalls using a stateful connection:** Most firewalls, including the Cisco ASA, support "HA pairing" whereby two firewalls are directly connected using a dedicated link for state sharing. This is the most commonly deployed design when two firewalls both service a single site.
3. **Use vendor-proprietary clustering features:** For additional scale, some vendors offer a clustering capability where many firewalls can be parallelized. The logic is like HA pairing except scales to more than two nodes.

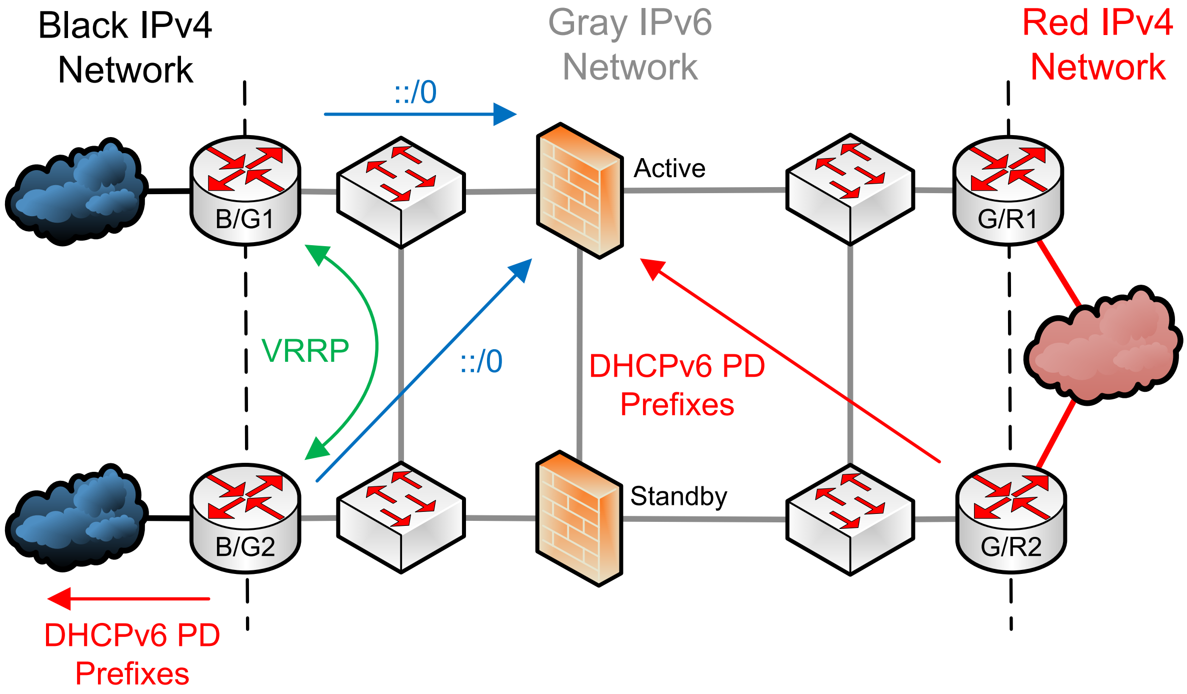
In this MACP instance, the HA pairing option was depicted earlier in the document. This is the recommended option for the design because the firewalls both service the same "site", meaning that they are not geographically separated. Additionally, HA pairing is well-known, relatively easy to troubleshoot, and provides adequate scale and resilience for the design. For extremely high scale designs, clustering may be suitable.

In terms of switching architecture, many firewalls need to communicate across their data-forwarding interfaces as well. This allows them to detect failures in the layer-2 network to trigger a failover event, necessitating the addition of several switches. For high availability, two external and two internal switches should be deployed to "sandwich" the firewalls.

The recommendation to use IGP/BGP to connect black/gray gateways to gray firewalls varies based on the distance between the components. If the gray network is very small, as depicted in both the conceptual MACP architecture and the diagrams specific to this design, then using IGP/BGP with the firewalls may be more trouble than it is worth. Using static routing on the firewall towards a Virtual Router Redundancy Protocol (VRRP) virtual address shared by the routers is likely simpler. The routers would also use static routes towards the firewall to reach the gray management networks and gray/red VPN headends. When the firewall failover event occurs, the active IPv6 address migrates to the standby unit. This simplifies the static routing on the black/gray gateways as they only must specify a next-hop of the active IPv6 address.

Note that there is only one LAN segment on the inside (trusted) zone of the firewall. This is a "connected" or "direct" network from the perspective of both firewalls and both gray/red gateways. The only routing required across it is the downstream connectivity for the gray/red gateways, which must cover all DHCPv6 PD prefixes. Since there is no reason for the gray/red gateways to communicate with the gray management services or site local DHCPv6 relays, a default IPv6 route should not be used. The diagram below depicts this solution.

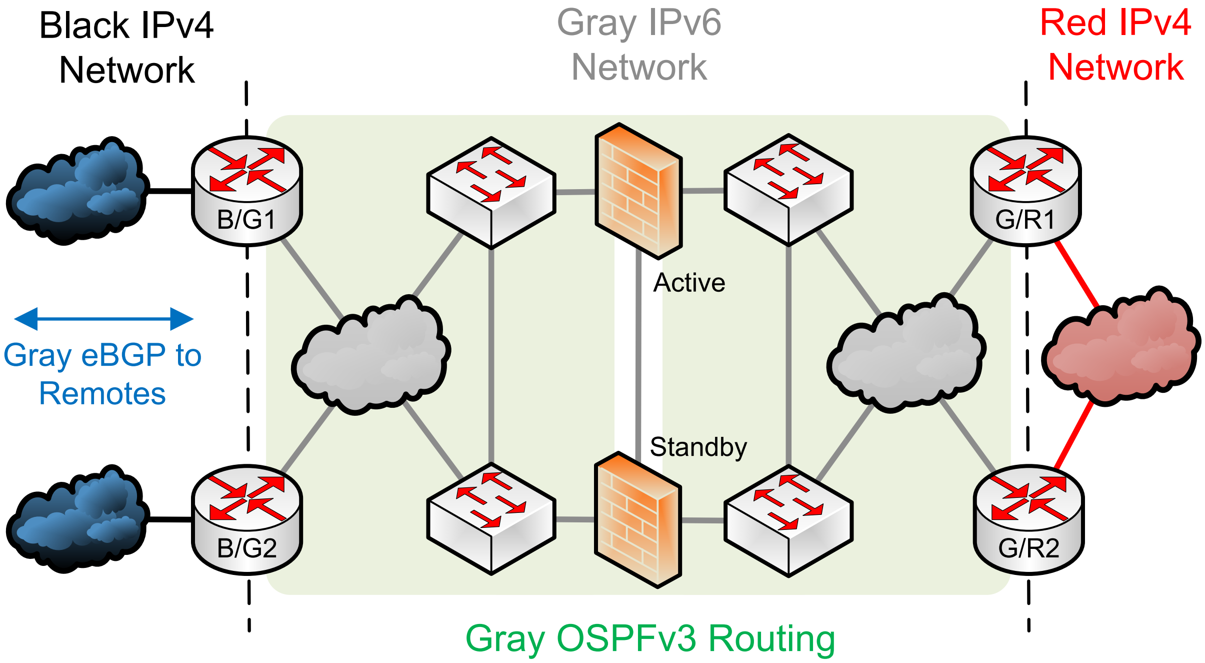
Figure 15 - Gray Firewall HA Routing using Static Routes and VRRP



If there is a sizable gray network between the black/gray gateways and the gray firewalls (and/or between the firewalls and the gray/red gateways), using a dynamic routing protocol is likely the better choice. IGP/BGP failover operates differently on various firewalls. On Cisco ASAs, the standby firewall maintains all the IGP routing state but does not allow data plane traffic to flow. Upon a failover, the standby device assumes the IGP/BGP identity (router ID, IP address, MAC address, etc.) used by the active device so that the adjacent routers do not trigger a reconvergence event. The black/gray gateways no longer need VRRP for gateway failover, and in larger network, these gateway routers may not even share a LAN segment.

The diagram below depicts this solution using a single, contiguous OSPF domain. The active/standby HA pairing link does not run IGP/BGP as this is for intra-firewall communications only. This document does not detail all the internal routing design considerations and optimizations that may arise in different environments. eBGP is still recommended for connectivity between black/gray gateways and remote sites regardless how the upper-level gray routing is designed.

Figure 16 - Gray Firewall HA Routing using OSPFv3



3.5.2. Policy Rules

The gray firewall policy should be designed in such a way that it works seamlessly with all the remote site connectivity options described earlier. Since they all leverage a common set of capabilities, this document details a relatively generic policy. The firewall is stateful so it is safe to assume that returning traffic is permitted once it has been inspected.

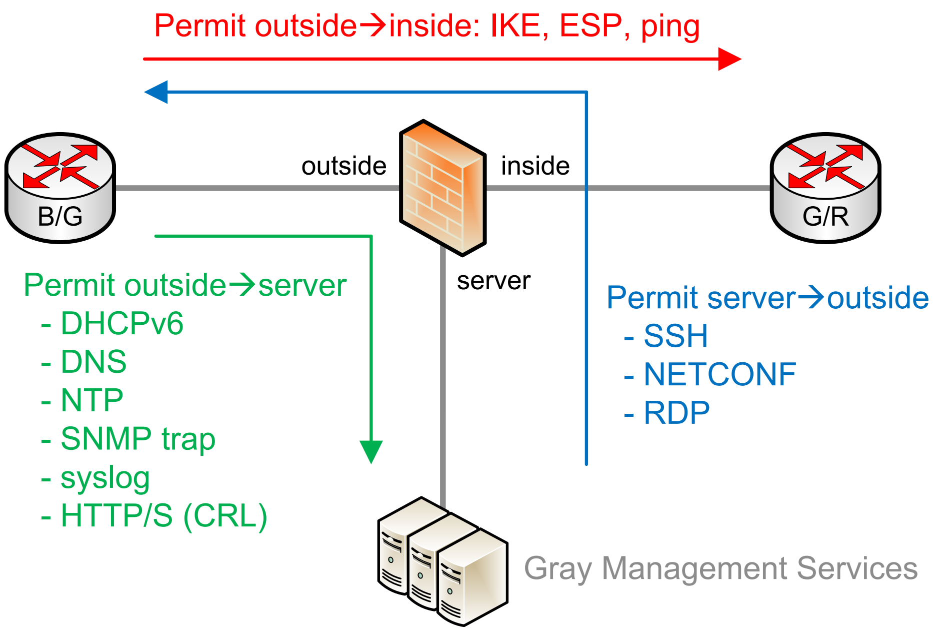
From the outside zone to the inside zone, the only traffic that must be permitted should be IKE and ESP for gray/red VPN establishment. While the MACP does permit the use of TLS instead of IPsec for the inner VPN, using IPsec for both VPNs is operationally simpler. It helps unify firewall policies and design decisions across the black and gray networks by reducing the number of technologies in use. Designers may optionally want to permit ICMPv6 echo-request and echo-reply messages between gray/red VPN clients and their headends for connectivity testing.

From the outside zone to the server zone, both DHCPv6 and DNS traffic must be allowed. The DMVPN hubs originate the DHCPv6 messaging using relay messages on behalf of DHCPv6 clients asking for delegated prefixes. The gray/red VPN endpoints, once they've learned the domain name and DNS servers via stateless DHCPv6, will use DNS AAAA requests to discover the IPv6 addresses of the gray/red VPN headends. Other upstream management services such as SNMP, NTP, and syslog may also flow from the outside zone to the server zone. Also, be sure to include certificate revocation checking (CRL downloads) which is typically carried over HTTP or HTTPS. Adjust the firewall policy as necessary to account for these additional services.

From the server zone to the outside zone, there may be additional downstream management services in use. For example, the gray management laptop may need SSH and/or NETCONF access to the remote routers for standard device management. It may also use RDP or HTTPS for managing various gray/red client operation systems. Be sure to include these management protocols as required.

There is no reason for the inside and server zones to exchange any traffic. Doing so could even pose a security risk; this connectivity should be completely blocked. The diagram below summarizes there gray firewall recommendations as they relate to this design.

Figure 17 - Multi-zone Gray Firewall Policy



3.6. Device Management and Automation

This section details the various automation techniques that can improve the manageability of this design. The code referenced in this section is open-source and is available at https://github.com/nickrusso42518/net-tools on GitHub in the “ipv6_tools” directory.

Note that this repository is likely to change in the future as new tools are introduced and existing tools are improved based on reader feedback. Readers are encouraged to use the GitHub “Issues” feature to submit bug reports or feature requests.

3.6.1. On-box Scripting

One of the major advantages of the DHCPv6 PD design, generally speaking, is that all the remote sites of a given type use an identical configuration. For example, if you have a mix of Cisco 819 and 829 routers, but have 1,000 sites, you only need to manage two configurations, one “golden” configuration per device type. This leads to a few small problems.

The first problem is regarding device hostnames. Even though the entire IPv6 routing and VPN discovery processes are dynamic, the hostnames on the black/gray routers are the same given a common "golden" configuration. To overcome this, a simple on-box script can generate a unique hostname on each device using the system serial number. In Cisco IOS, this is known as Embedded Event Manager (EEM). The high-level logic of the script is as follows:

1. On boot, look for a syslog message indicating that the system has powered on.
2. Check to see if the hostname is set to the default string identified in the golden configuration (e.g., "Router"):
    a. If no, the hostname has been updated already, so exit.
    b. If yes, continue running the script.
3. Collect the device's serial number using regular expressions.
4. Set the device's hostname to a static string prepended to the serial number (e.g., REMOTE-SN12345).
5. Save the configuration.

This script only runs one time (or whenever the hostname is reset to the default string at boot time) to update the device hostname. By embedding this code into the golden configuration, it guarantees all remote sites will have unique hostnames which can simplify SSH, SNMP, and syslog-based management.

The second problem is more challenging and involves configuring a static IPv6 address for downstream management. Note that the DHCPv6 delegated prefixes are dynamic and subject to change; they are not good candidates for populating sources of truth or automation inventory files. Assuming dynamic routing is used, remote sites could advertise non-DHCPv6 prefixes back to the hubs but only if they are guaranteed to be unique.

To accomplish this, each remote site can define a loopback interface which is set to the same /64 prefix. The interface will use SLAAC (and specifically, the EUI-64 process) to derive a unique IPv6 address for the loopback. This results in a unique IPv6 address on each device, but the prefix-length is 64 everywhere, which is unsuitable for advertisement into BGP. Another EEM script can perform the following steps to convert it to a static /128 address for management:

1. On boot, look for a syslog message indicating that the system has powered on.
2. Check to see if the hostname is set to the default string identified in the golden configuration (e.g., "Router"):
    a. If no, one can reasonably assume the IPv6 address has been updated.
    b. If yes, continue running the script.
3. Collect the loopback's IPv6 address using regular expressions
4. Remove the EUI-64 configuration and hardcode the same IPv6 address on the loopback interface as a /128.
5. Save the configuration.

Like the previous script, it only runs once (generally). Unfortunately, Cisco IOS does not allow EUI-64 to work on /128 addresses natively; if it did, this entire script would be unnecessary. At the headend, be sure to identify this uniformly configured /64 range as the "management network" for permissions through the firewall. Each remote site will consume a single IPv6 address from this prefix.

3.6.2. Centralized Automation Tools

Building on the previous section, suppose an administrator wants to use a centralized automation tool like Ansible to manage the remote sites. Each site now has a static IPv6 address assigned to a loopback and advertised into BGP. That loopback was derived from the device's lowest MAC address, which is easy to determine either by looking at the shipping box (on some products) or collecting it from the device's command line. Therefore, supplying all the remote client MAC addresses into an offline script can output a complete list of IPv6 management address for each node by manually applying the EUI-64 arithmetic to each address. Once these IPv6 addresses are computed, they can be written to an Ansible inventory file.

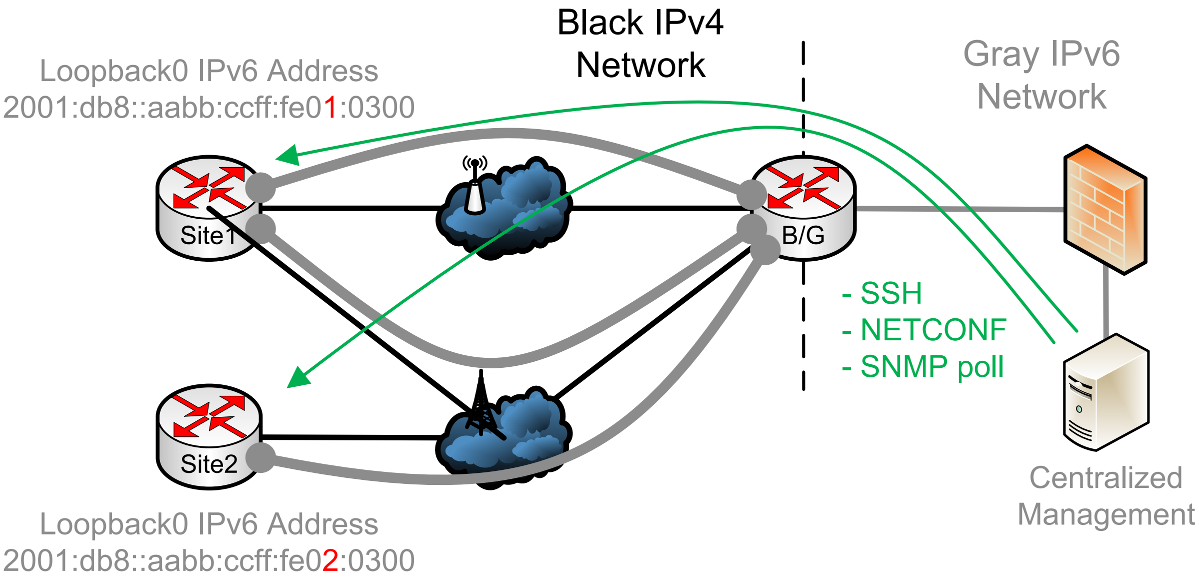
The Ansible inventory serves as an input to Ansible in general, providing a hierarchical collection of hosts and groups to which Ansible can connect for remote device management. Once populated, traditional top-down centralized management of remote sites is possible. The inventory file can be re-generated whenever new sites are added. Because the loopback's IPv6 management address never changes, this regeneration approach is safe for long-term usage.

If collecting MAC addresses manually is undesirable or infeasible, an alternate script can populate the Ansible inventory by examining the IPv6 BGP table on one of the hub routers. Given the overarching management /64 prefix that contains /128 management IPv6 addresses, the script performs the following steps:

1. Log into the router using SSH to access the CLI.
2. Capture the IPv6 BGP table output using a "show" command.
3. Extract all /128 management IPv6 prefixes using a parsing technique.
4. Write all /128 management IPv6 addresses to the Ansible inventory file.

The diagram below illustrates centralized Ansible management in action. This assumes that the initial EEM scripts have run, ultimately resulting in unique hostnames and unique IPv6 /128 prefixes for each device.

Figure 18 - Top-down Centralized Management to EUI-64 Loopbacks



4. Red Network Design

This section details the red network architecture. This is the network being protected by the dual-VPN design and contains sensitive user data. The primary goal is seamless and secure integration between the CSfC kit and the existing upstream network.

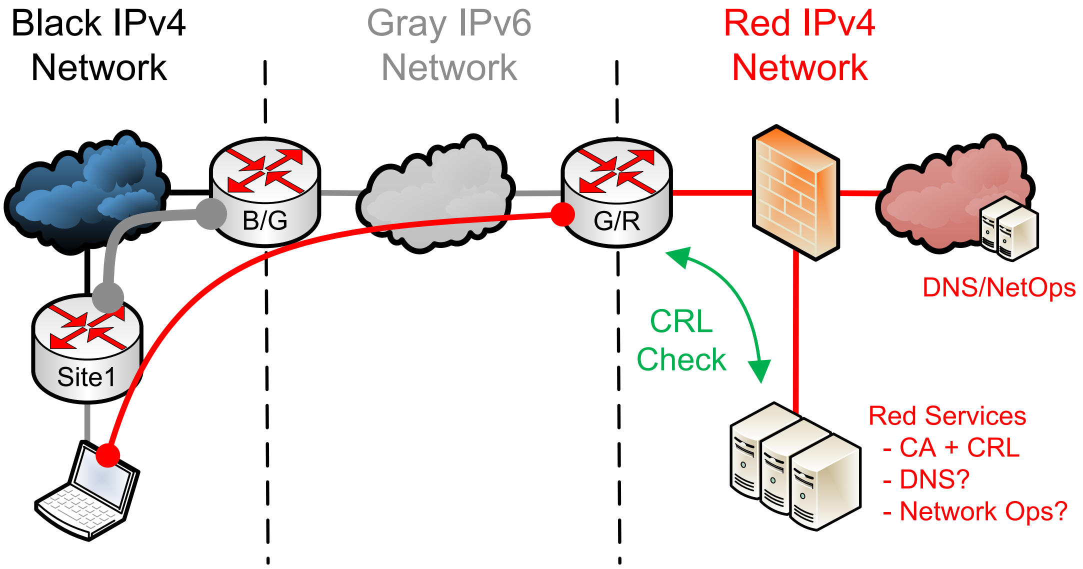
4.1. Required Headend Services

At a minimum, two critical services are required within the red network: a CA and a traffic-filtering firewall. The CA signs certificates for the gray/red VPN endpoints, which includes the remote-access VPN clients and the inner VPN headend. Just like with the black/gray router that terminates the outer VPN, the gray/red VPN headend (such as Cisco ASA) will evaluate remote-access clients against the CRL maintained by the CA. This ensures that compromised clients with revoked certificates cannot establish their inner VPNs to access the protected red network.

The presence of other basic services, such as DNS and network management, depend upon the overall size of the red network. In most cases, the red network is very large and encompasses the entire enterprise. As such, there are likely centralized DNS and network management services elsewhere, so including them in the CSfC kit's red network is duplicative.

The diagram below illustrates these services and the inner VPN formation at a high level.

Figure 19 - Red Network Services and VPN Formation



4.2. Gray/Red VPN Headend Discovery

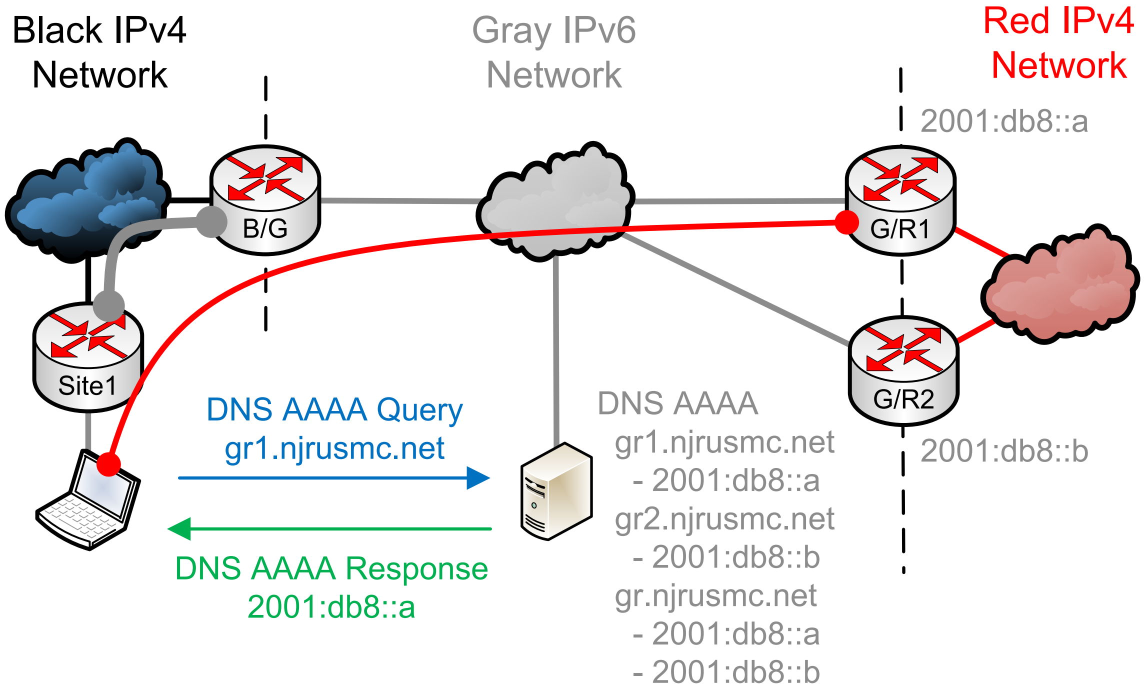
This section explains how the gray/red clients use DNS services (discussed briefly in previous gray networking sections) to discover the inner VPN headends.

A common approach when creating DNS records is to create both specific and generic records relating to a given service. The specific AAAA records would represent a single gray/red VPN headend device, such as an individual firewall. This record would contain a single IPv6 address so that clients resolving this specific hostname would always target a given device. These specific records are useful for clients that prefer specific VPN headends, perhaps due to performance reasons.

Additionally, at least one generic AAAA record should be created which contains all the gray/red VPN headend IPv6 addresses. Clients that don't care about which VPN headend is chosen can connect to this hostname, allowing the client operating system to select which IPv6 address to use. Additional generic AAAA records could be used to regionalize this process whereby each record contains a subset of the VPN headend IPv6 addresses. For example, one record could represent North American VPN headends while another represents Europeans VPN headends, and clients in each region should prefer VPN headends closest to them.

The diagram below illustrates a conceptual DNS design. The specific hostnames "gr1.njrusmc.net" and "gr2.njrusmc.net" map to the corresponding gray IPv6 addresses on G/R1 and G/R2, respectively. The generic hostname "gr.njrusmc.net" contains two IPv6 addresses, enumerating both gray/red gateways. Note that anycast designs whereby each gray/red VPN headend uses the same IPv6 address (and a single DNS AAAA) record are also supported.

Figure 20 - Gray/Red Headend Discovery using IPv6 DNS



4.3. Host-level Security Techniques

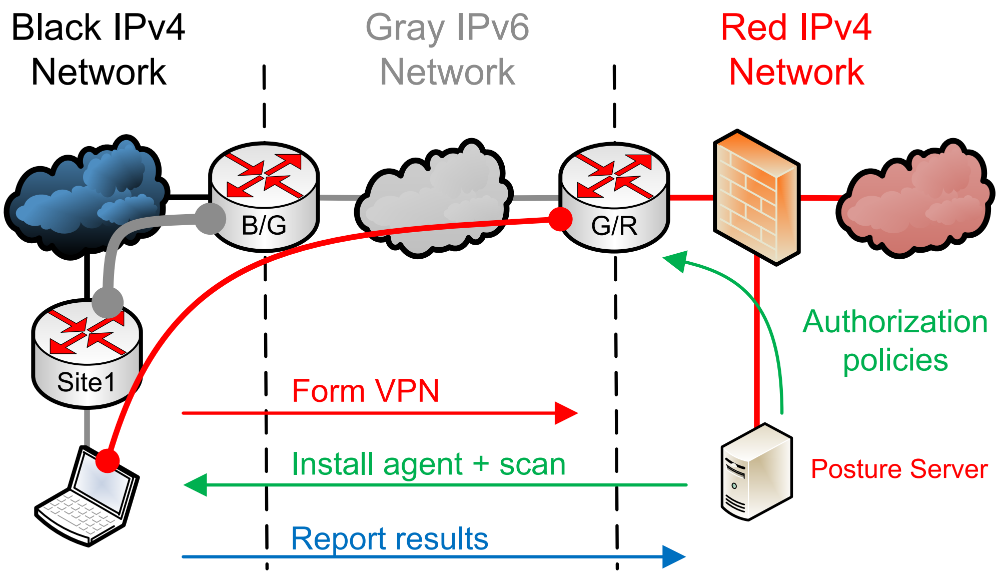
Clients tend to have more attack surfaces than network devices given their general-purpose nature and their regular usage by non-technical people. As such, the gray/red VPN headend should perform some additional security checks on the clients when they connect. Like all things in security, there exists a spectrum whereby complexity/cost typically increase proportionally with effectiveness.

For the sake of a realistic example, suppose the inner VPN headend is a Cisco ASA and the VPN client software is Cisco AnyConnect. At the time of this writing, both of these are CSfC-approved products. At one end of the spectrum, the formation of the VPN is the only security event that occurs with no additional checks performed on the host. This is simple and inexpensive but does not assess the client's suitability to connect to the network beyond presenting a valid CA-signed certificate.

A moderately secure solution would leverage Cisco ASA Dynamic Access Policies (DAP), also known as "host scan". This solution installs a software agent (in addition to Cisco AnyConnect) on each client that provides an interface to various operating system security features. For example, DAP can ensure that the operating system's firewall (such as Microsoft Defender, Linux iptables, etc.) is enabled. If the firewall is not enabled, the client's connection is rejected.

For greater security, Cisco Identity Services Engine (ISE) and comparable products can provide additional posture assessment services for connected clients. This is a more advanced version of DAP that can test for specific firewall rules, presence or absence of applications/services, and more. Naturally, this solution is the more expensive, both financially and operationally, but provides the highest degree of security for sensitive data. Common settings checked in a posture assessment include the presence and enablement of an OS firewall, antivirus software, software update service, and disk encryption. This process may also include authorization policies pushed to the VPN headend for additional, network-level security. The diagram below shows a high-level posture assessment in process.

Figure 21 - Centralized Host-level Security via Posture Assessment



4.4. Red Firewall Policy

Unlike the gray and black firewalls, the red firewall policy is highly variable between organizations. The red network is a data network, not a transport network, and so any traffic filtering policies will depend on the services offered within an organization. However, there are some generic rules that should be common in any design.

First, all traffic originating from the CSfC-connected clients on the outside of the firewall should be sourced from a valid VPN address pool. Any spoofed packets should be dropped. Next, any existing enterprise blacklists should be applied, matching existing security policies elsewhere in the network. The blacklist should be extended to drop IKE and IPsec related traffic if it isn't already specified. An unauthorized "triple VPN" being formed from the client's operating system to some other off-net red VPN concentrator could be a signal of a data exfiltration attack and should be blocked.

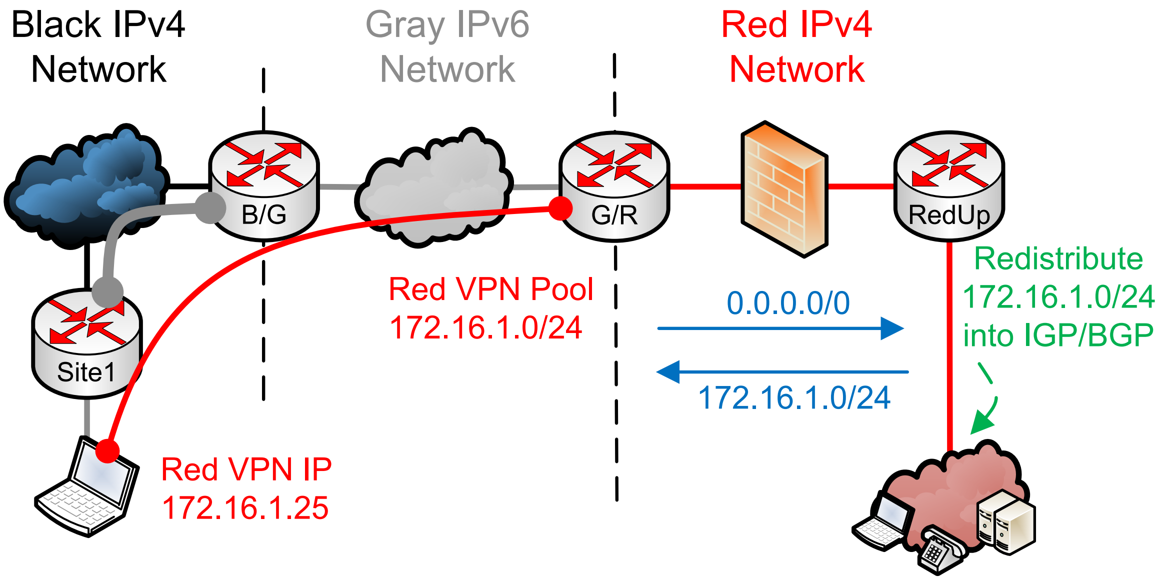
In terms of permitted flows, this document does not attempt to enumerate every application that organizations might use. Instead, consider this list of generic applications as a sanity check:

a. **Common services:** DNS, HTTP, FTP, and NTP/SNTP
b. **Collaborative applications:** voice over IP, chat, presence, and email
c. **System management:** Windows updates, Linux package management (yum/apt), anti-virus software updates, and other software distribution mechanisms

4.5. Upstream Red Network Integration

There are two general approaches for tying the CSfC solution into the rest of the red network. The first approach assumes no multi-tenancy and that all gray/red VPN clients can exist in the same subnet. This is common for traditional enterprise remote access VPN designs as the clients are all entering the network at the same point. It simplifies the routing and allows for a relatively basic exchange of routes as shown in the diagram below. The inner VPN gateway can advertise the VPN pool into IGP or BGP towards the upstream red router, or the devices can use static routing without much risk. The diagram below illustrates this simple design where only a single logical connection exists between the gray/red gateway and red upstream router. Note that the red firewall can run in layer-2 or layer-3 mode (and could be designed in an HA pair or cluster), but this example uses a layer-2 firewall for simplicity.

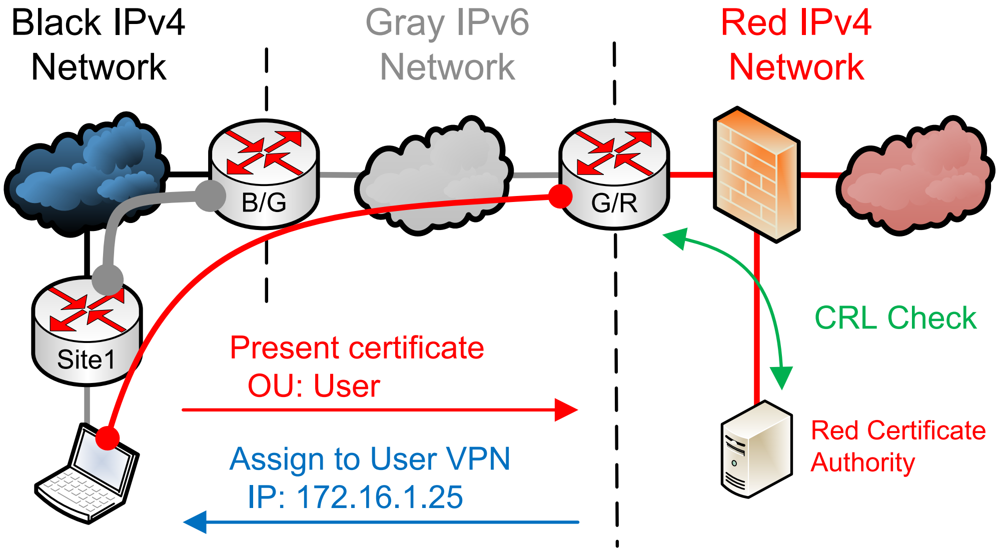
Figure 22 - Single-tenant Red Network Integration



More complex is designing multi-tenancy for remote clients. Clients must somehow signal their tenancy to the VPN headend. The simplest and most secure way to accomplish this is by using certificate metadata, such as the Organizational Unit (OU) field. This field can specify business departments, network VLANs, or other identifying information that the gray/red VPN headend can match for tenancy assignment. For example, an organization may want to separate users by functional area (finance, engineering, manufacturing, etc.) or by network privilege/permission (user, management, etc.) using this technique. Configuration-wise, this is more complex on any platform as it requires an enumeration of the different tenants, their subnets, their upstream VLAN identifiers, and more.

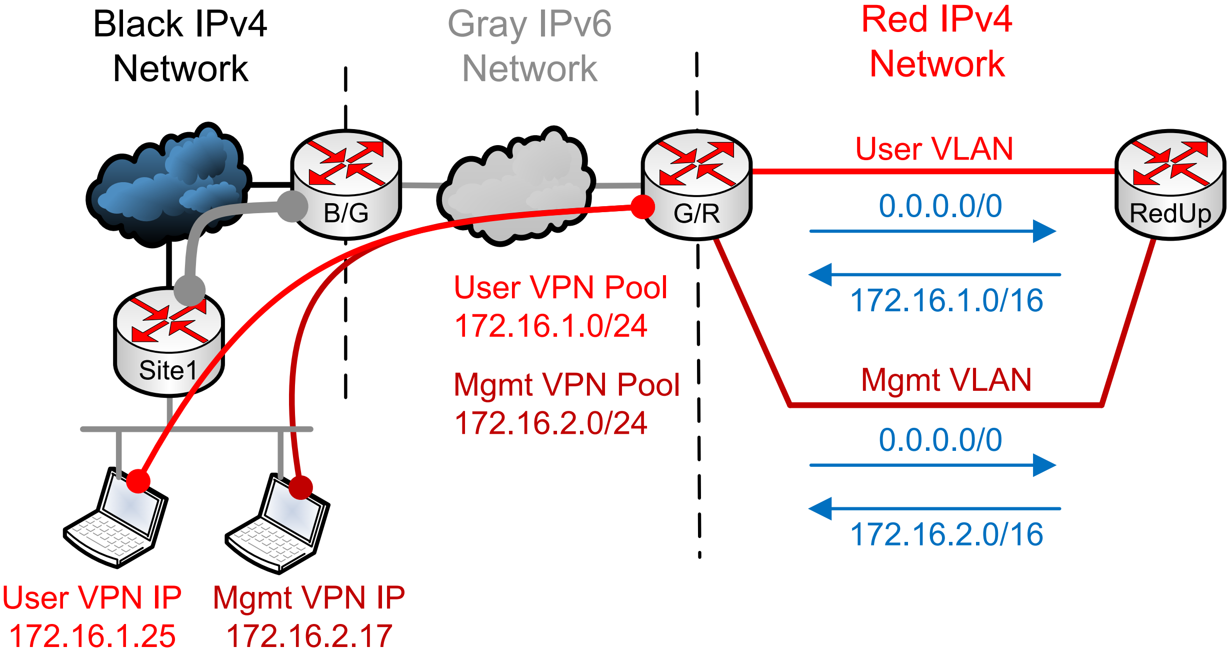
After presenting the certificate with a specific OU, the gray/red headend will perform the usual CRL check. Assuming the client certificate is valid and authentication succeeds, the headend will assign the proper tenancy to the client, issuing the correct IP and placing all received traffic into the proper logical routing instance. The diagram below illustrates this process at a high level.

Figure 23 - Using Certificate OU to Signal Tenancy



Additionally, for each tenant, there must be a separate routing exchange using either static or dynamic routing. The gray/red VPN gateway will likely need to use virtual routing instances or VRFs, allowing it to concurrently respect multiple default routes within each tenant routing instance. The diagram below illustrates a high-level multi-tenant VPN design. Some details, such as the red firewall, upstream red cloud, and static route redistribution are omitted for cleanliness. There is only a single physical uplink but multiple logical uplinks (using VLAN + VRF) with separate routing exchanges over each. This diagram depicts static routes, but individual VRF-aware routing processes can also work.

Figure 24 - Multi-tenant Red Network Integration



5. 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 different MACP solutions or comparing the various gray transport options described earlier.

5.1. 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.

One of the biggest advantages of this design is the concept of a single "golden configuration" for each remote device type. This scales in both constant time and constant space (i.e., one configuration services N remote sites), making it easy to provision new sites. The tunneling (DMVPN + IPsec) and DHCPv6 processes all scale linearly as there will typically be one or two entries per remote site. The exact number depends on the gray transport option selected, which impacts the slope of the graph when viewed on a two dimensional plane.

Put simply, the number of DMVPN hubs is directly proportional to amount of tunneling state as each hub must maintain copies of the same registration data from each spoke, along with unique IPsec tunnels (and possibly BGP peers). The number of transports is directly proportional to the amount of DHCPv6 PD state because remote sites receive a unique delegated prefix over each transport mesh. No matter which option is chosen, the table below enumerates these combinations.

Table 5 - Comparing State Retention across Gray Transport Options

	Tunneling state per site	**DHCPv6 PD state per site**
1 hub / 1 transport	1	1
N hubs / 1 transport	N	1
1 hub / N transports	1	N
N hubs / N transports	N	N

At a more fundamental level, the solution generally scales well. If the tunnel meshes leverage BGP for routing, the hubs can advertise aggregate routes (such as a default route or "core" network aggregate) to reduce churn over the mesh. If upstream failures occur that lead to components of an aggregate becoming unreachable, it becomes much less likely that this leads to BGP updates transmitted over the tunnel meshes. Across all gray options, the spokes always rely on some kind of gray IPv6 aggregate route, whether learned through ICMPv6 RA messages or through explicit BGP advertisements, allowing them to scale in constant time. The hubs scale linearly as they'll receive one or two DHCPv6 PD delegated prefixes from each remote site, plus one additional loopback prefix for top-down management.

5.2. 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, convergence time, and network overhead while maximizing utilization, manageability, and user experience.

The primary driver of the design described in this document is to maximize scale while minimizing deployment time and effort. The multi-hub and multi-transport options provide various degrees of high-availability, allowing for fast failover. When BGP is deployed, operations have fine-grained control over how traffic is forwarded by ranking transports from best to worst. In this regard, the design is highly optimal.

As is true with any design that relies on prefix aggregation, the possibility of suboptimal routing remains. In small, tactical-style environments where the gray network is tiny, this is operationally irrelevant. In a large, enterprise-wide gray network, remote sites in one region may be forced to route traffic through a geographically distant black/gray hub simply because the BGP policy demands it.

It is possible to configure the hubs to assign remote sites to different BGP peer-groups, allowing them to consume different BGP policies. This is very difficult; it requires trading off the scale advantage by creating additional "golden configurations" on a regional basis, likely using different tunnels over the same transports with different IPv6 address ranges. Another option is to leak longer-match routes from BGP without using prefix aggregation, relying on existing BGP policy attributes to implement the desired forwarding policy. Again, this trades off scale in the BGP control-plane and may lead to increase churn (and additional bandwidth consumption) between the DMVPN hubs and remote sites.

5.3. 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 relatively deep. Several completely different features must all work in concert, and in series, to deliver a successful outcome. These technologies are spread across many products and operating systems. To summarize:

a. Black/gray DMVPN/IPsec VPN formation using digital certificates.
b. DHCPv6 PD for prefix acquisition and subsequent assignment to gray LANs.
c. ICMPv6 RAs or dynamic eBGP neighbor formation for routing exchange.
d. ICMPv6 RAs for prefix signaling and SLAAC to gray/red clients.
e. Stateless DHCPv6 for "other-configuration" distribution to gray/red clients.
f. DNS lookups to resolve gray/red gateways using AAAA records.
g. Gray/red remote access VPN formation using digital certificates.

While the inner workings of the gray network have deep surface interactions, comparable interactions across networks of different colors/classifications (black to gray and gray to red) are minimal. This separation is strictly required by the CSfC framework and is enhanced by the design decisions described earlier in this document. For example, using front-door VRFs further separates the black underlay networks from the gray DMVPN overlay meshes. Using an exclusively IPv6 gray network creates a nearly impassable buffer between black and red networks, both of which are primarily using IPv4.

The surface interaction breadth of the solution is very wide. In a large-scale network, thousands of remote sites would be following the complex, sequential process outlined above. Having broad and deep surface interactions spread across a large, distributed network is considered highly complex. The best mitigation is to keep everything consistent and predictable. The design does not encourage one-off modifications, such as one mesh using BGP while another relies on default routes from ICMPv6 RA messages. These notional modifications may slightly improve optimization, but come at a high cost with respect to state and surface interaction.

Appendix A – Acronyms

Acronym	Definition
ACL	Access Control List
AES	Advanced Encryption Standard
AP	Access Point (WiFi)
APN	Access Point Name (LTE)
AS	Autonomous System (BGP)
ASA	Adaptive Security Appliance (Cisco)
ASN	AS Number (BGP)
BGP	Border Gateway Protocol
CA	Certificate Authority
CRL	Certificate Revocation List
CSfC	Commercial Solutions for Classified
DAP	Dynamic Access Policy (Cisco)
DAR	Data At Rest
DHCP	Dynamic Host Configuration Protocol
DMVPN	Dynamic Multipoint VPN
DMZ	De-Militarized Zone
DNS	Domain Name System
DUID	DHCPv6 Unique Identifier
EAP	Extensible Authentication Protocol
eBGP	External BGP
EEM	Embedded Event Manager (Cisco)

Acronym	Definition
eNodeB	Evolved Node B (LTE)
EPC	Evolved Packet core (LTE)
ESP	Encapsulating Security Payload
EUI-64	Extended Unique Identifier (IPv6 SLAAC)
GUA	Global Unicast Address (IPv6)
HA	High Availability
HAIPE	High Assurance IP Encryptor
HTTP	HyperText Transport Protocol
HTTPS	HTTP Secure
IAID	Identity Association Identifier (DHCPv6)
ICMP	Internet Control Message Protocol
IGP	Interior Gateway Protocol
IKE	Internet Key Exchange
IP	Internet Protocol
IR	Industrial Router (Cisco)
ISE	Identity Services Engine (ISE)
LAN	Local Area Network
LLA	Link Local Address (IPv6)
MAC	Media Access Control (Ethernet)
MACP	Mobile Access Capability Package
MED	Multi-Exit Discriminator (BGP)
MSCHAP	Microsoft Challenge Handshake Authentication Protocol
NAT	Network Address Translation

Acronym	Definition
NAT-T	NAT Traversal (IPsec)
NSA	National Security Agency (US Government)
NTP	Network Time Protocol
OS	Operating System
OSPF	Open Shortest Path First
OU	Organizational Unit (Certificate)
PD	Prefix Delegation (DHCPv6)
PEAP	Protected EAP
PSK	Pre-Shared Key
RA	Router Advertisement (ICMPv6)
RADIUS	Remote Authentication Dial-In User Service
RDP	Remote Desktop Protocol (Windows)
RS	Router Solicitation (ICMPv6)
SLAAC	StateLess Address AutoConfiguration
SNMP	Simple Network Management Protocol
SNTP	Simple NTP
SOS	State Optimization Surface
SSH	Secure Shell
SSID	Service Set ID (WiFi)
TLS	Transport Layer Security
UDP	User Datagram Protocol
UE	User Equipment (LTE)
ULA	Unique Local Address (IPv6)

Acronym	Definition
uWGB	Universal Workgroup Bridge
VLAN	Virtual LAN
VPN	Virtual Private Network
VRF	Virtual Routing and Forwarding
VRRP	Virtual Router Redundancy Protocol
WAN	Wide Area Network
WGB	Workgroup Bridge
WLAN	Wireless LAN
WLC	Wireless LAN Controller (WiFi)
WPA2	WiFi Protected Access 2

Appendix B – References

Automation Tools for IPv6 (GitHub)

BGP-4 (IETF RFC 4271)

DHCPv6 General/Aggregate (IETF RFC 8415)

DHCPv6 Prefix Delegation (IETF RFC 3633)

DHCPv6 Stateless (IETF RFC 3736)

ICMPv6 (IETF RFC 4443)

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

IP Security – IPsec (IETF RFC 4301)

WiFi CAPWAP (IETF RFC 5416)

Navigating Network Complexity (White and Tantsura)

NHRP (IETF RFC 2332)

NSA CSfC Resources

OSPF version 3 (IETF RFC 5340)