CHAPTER 3

Scalability Considerations

This chapter presents the following steps to selecting Cisco products for a VPN solution:

- Sizing the headend
- Choosing Cisco products that can be deployed for headend devices
- Product sizing and selection information for branch devices

General Scalability Considerations

This section provides general scalability considerations to assist with design requirements.

IPsec Encryption Throughput

The throughput capacity of the IPsec encryption engine in each platform (headend or branch) must be considered for scalable designs, because each packet that is encrypted must traverse through the encryption engine. Therefore, encryption throughput must consider bi-directional speeds. Several examples are shown in Table 3-1 and Table 3-2 for popular headend and branch connection speeds.

Table 3-1 Headend Connection Speeds

Connection Type	Speed (in Mbps)	Encryption Throughput Required (in Mbps)
T3/DS3	44.7	90.0
OC3	155.0	310.0
OC12	622.0	1250.0

Table 3-2 Branch Connection Speeds

Connection Type	Speed (in Mbps)	Encryption Throughput Required (in Mbps)
T1	1.5	3.0
2 x T1	3.0	6.0

Table 3-2 Branch Connection Speeds (continued)

T3/DS3	44.7	90.0
Broadband cable/DSL	384 Kbps uplink/ 2 Mbps downlink	2.4

In general, as throughput increases, the burden on router CPU also increases. However, with hardware-accelerated encryption available for all Cisco router products from the 871 through the 7600, impact to the main CPU is offloaded to the VPN hardware. However, main router CPU processing still occurs, so higher throughput typically results in higher CPU consumption.

Packets Per Second—Most Important Factor

Although bandwidth throughput capacity must be considered, the packet rate for the connection speeds being terminated or aggregated is more important.

In general, routers and encryption engines have upper boundaries for processing a given number of packets per second (pps). The size of packets used for testing and throughput evaluations can understate or overstate true performance. For example, if a router with a VPN module can handle 20 Kpps, 100-byte packets lead to 16 Mbps throughput while 1400-byte packets at the same packet rate lead to 224 Mbps.

Because of such a wide variance in throughput, pps is generally a better parameter to determine router forwarding potential than bits per second (bps). Scalability of the headend is the aggregate forwarding potential of all branches that terminate a tunnel to that headend. Therefore, the aggregate pps from all branches impacts the pps rate of that headend.

Tunnel Quantity Affects Throughput

Although throughput is highly dependent on platform architecture, as tunnel quantities are increased, the overall throughput generally tends to decrease. When a router receives a packet from a different peer than the peer whose packet was just decrypted, a lookup based on the security parameters index (SPI) of the new packet must be performed. The transform set information and negotiated session key of the new packet is then loaded into the hardware decryption engine for processing. Having traffic flowing on a larger numbers of SAs tends to negatively affect throughput performance.

Increasingly, platforms with hardware-accelerated IPsec encryption are designed to offload tunnel processing overhead as well, resulting in more linear performance regardless of the number of tunnels. For example, the VPN SPA blade for the Cisco 7600 has fairly linear throughput regardless of whether the traffic load is offered on a few tunnels or several thousand.

GRE Encapsulation Affects Throughput

Router encryption throughput is affected by the configuration of GRE. In addition to the headers that are added to the beginning of each packet, these headers also must be encrypted. The GRE encapsulation process, when not hardware-accelerated, increases total CPU utilization. Total throughput in a DMVPN design results in a lower throughput than that of an IPsec Direct Encapsulation design.

Routing Protocols Affect CPU Overhead

CPU overhead is affected by running a routing protocol. The processing of keepalives or hello packets and maintenance of a routing table uses a finite amount of CPU time. This amount varies with the number of routing peers and the size of the routing table. The network manager should design the routing protocol based on widely-known accepted practices for that particular routing protocol.

Scalable Dual DMVPN Cloud Topology—Hub-and-Spoke Deployment Model

This section discusses headend and branch scalability for the dual DMVPN cloud topology with the hub-and-spoke deployment model.

Headend Scalability

This section describes the various headend scalability factors to consider in a dual DMVPN cloud topology with the hub-and-spoke deployment model.

Tunnel Aggregation Scalability

The maximum number of IPsec tunnels that a headend can terminate must be considered. Tunnel scalability is a function of the number of branch routers that are terminated to the headend aggregation point. This number must include both the primary tunnels as well as any alternate tunnels that each headend may be responsible for in the event of a failover situation.

The number of IPsec tunnels that can be aggregated by a platform is used as the primary determining factor in recommending a platform. Equally or more important is the encryption pps rate.

Aggregation Scalability

Aside from the number of tunnels that a headend terminates, the aggregated pps must be considered. Requirements are influenced by several factors, including the following:

- Headend connection speed—What is the speed of the WAN link on which the IPsec tunnels of the branch routers are transported through at the headend (DS3, OC3, OC12, or other)?
- Branch connection speeds—What is the typical bandwidth at each branch office (fractional-T1, T1, T3, broadband DSL/cable, or other)?
- Expected utilization—What is the maximum utilization of the WAN bandwidth under normal operation (or perhaps peak, depending on customer requirements)?

The pps rate (traffic size and traffic mix) is the largest single factor in router scalability.

Customer Requirement Aggregation Scalability Case Studies

This section includes examples to illustrate headend scalability factors.

Customer Example—1000 Branches

Assume that a customer has the following requirements:

- Number of branch offices—1000
- Branch access speeds—384k/1.5M cable/DSL
- Headend access speed—OC12 (622 Mbps)
- Expected utilization—33 percent

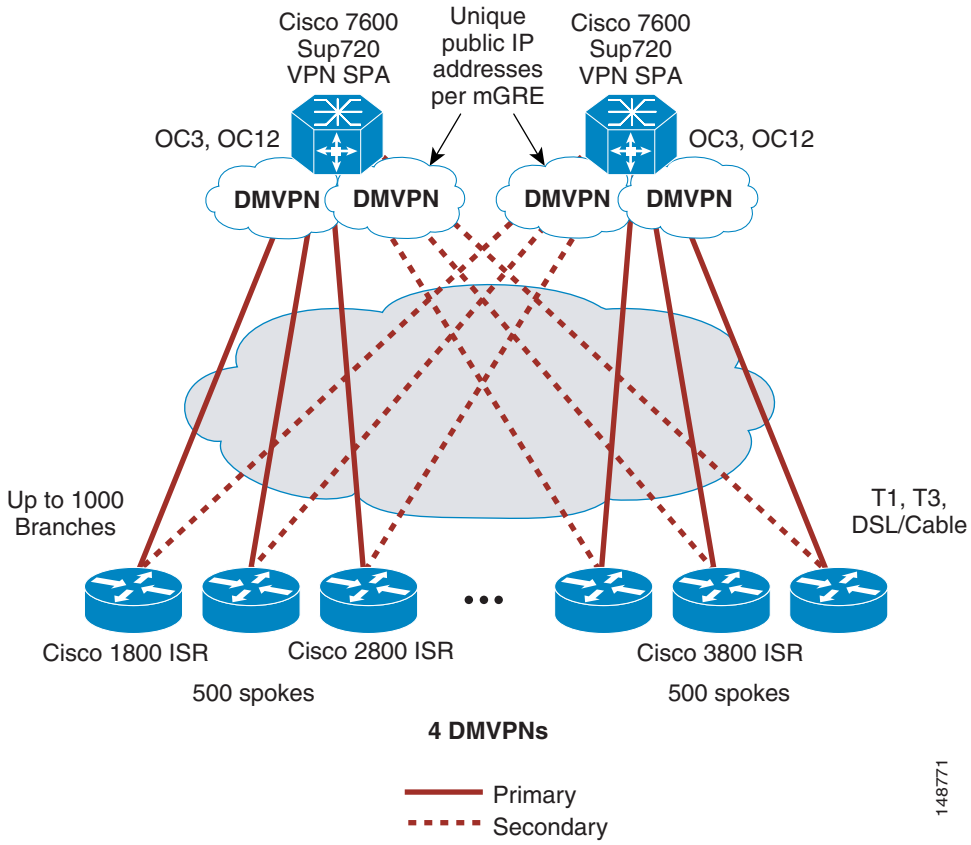
The calculation of aggregate bandwidth requirements is as follows:

- typical case—1000 x (384kbps + 1.5 Mbps) x 33 percent utilization = 628 Mbps
- worst case—1000 x (384kbps + 1.5 Mbps) x 100 percent utilization = 1.9 Gbps

Possible design options are to recommend a series of stacked Cisco 7200VXR platforms. At least four are required to aggregate 1000 tunnels; however, this does not provide the needed aggregate bandwidth of at least 628 Mbps.

A design alternative is to recommend a pair of Cisco 7600 routers, each with Sup720 and VPN SPA, as shown in Figure 3-1. The VPN SPAs each provide up to 1.2 Gbps of encryption performance, so the design can support up to OC12 connection speeds at each headend.

Figure 3-1 Cisco 7600-Based DMVPN Design



Although the VPN SPA can handle up to 5000 IPsec tunnels, as well as 1000 accelerated GRE tunnels, accelerated mGRE is not currently supported. This is primarily because of the lack of support for the mGRE tunnel key, which adds an extra four-byte field to the GRE packet header. However, there is a workaround to take advantage of the accelerated GRE functionality in the VPN SPA.

The mGRE tunnel key is used to distinguish to which DMVPN cloud the packet belongs when multiple DMVPN clouds are configured on the same router. Another way to differentiate DMVPN clouds is to allocate a unique public IP address to each mGRE interface. The remote routers can then connect to their assigned mGRE interface and use the IP address instead of the tunnel key to designate which DMVPN cloud is the destination on the headend. In this way, you can still take advantage of the IPsec acceleration on the VPN SPA, with the Sup720 processing the mGRE.

The design shown above can then support up to 1000 branch offices. Routing peers tend to be the limiting factor. On the Cisco 7200VXR platform, routing peers tend to be limited to 500–700. On the Cisco 7600 with Sup720, up to 1000 routing peers have been proven to work in the Cisco scalability test lab.

Customer Example—5000 Branches

Assume that a customer has the following requirements:

- Number of branch offices—5000
- Branch access speeds—128 Kpps/1 Mbps DSL
- Headend access speed—OC12 (622 Mbps)
- Expected utilization—25 percent

The calculation of aggregate bandwidth requirements is as follows:

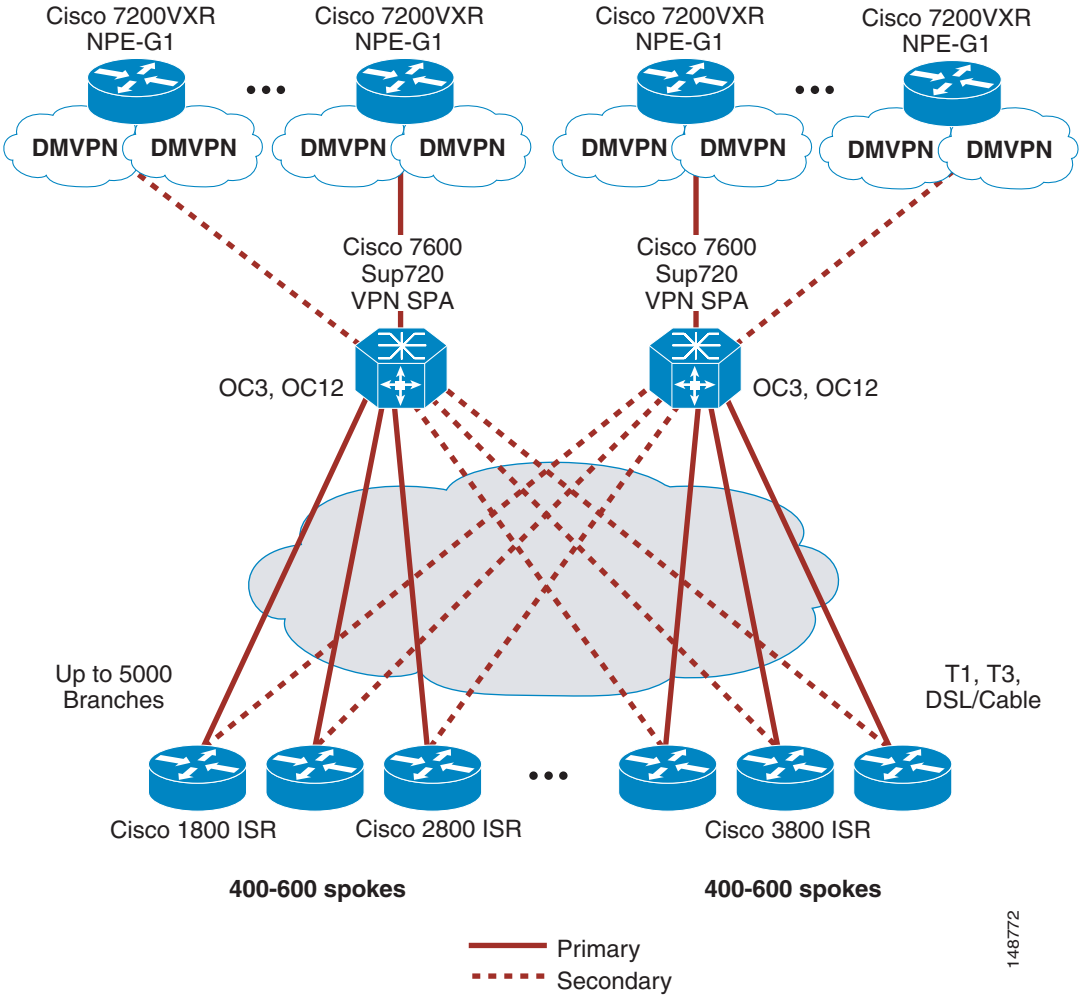
- typical case—1000 x (128kbps + 1 Mbps) x 25 percent utilization = 1.24 Gbps
- worst case—1000 x (128kbps + 1Mbps) x 100 percent utilization = 5.64 Gbps

Currently, no Cisco platform can aggregate 5000 DMVPN tunnels on a single box. Options for such large designs include the following:

- Duplicating a smaller scale design, such as either the Cisco 7200VXR-based design for 500–700 branch spokes, or the Cisco 7600-based design for 1000 spokes.
- Implementing a dual tier architecture using the Cisco 7200VXR platform to terminate the mGRE connections, and the Cisco 7600 platform for high-capacity IPsec encryption.

The dual tier architecture is shown in Figure 3-2.

Figure 3-2 DMVPN Dual Tier Headend Architecture



The Cisco 7200VXR platforms terminate the DMVPN clouds with mGRE interfaces. Because there are no IPsec encryption requirements in this tier of the design, no SA-VAM2+ is required. In addition, these platforms can typically handle more spokes than if the router is performing both mGRE and IPsec.

In the encryption tier of the design, the Cisco 7600 with Sup720 and VPN SPA performs IPsec encryption services. This enables a single Cisco 7600, providing up to OC12 encryption speed, to perform as the IPsec tunnel aggregation point for up to 5000 tunnels.

Two very important limitations of this design approach are the following:

- DMVPN spoke-to-spoke topologies are not supported in this design topology because the mGRE and IPsec terminate to separate IP addresses. For spoke-to-spoke functionality, the source and destination IP addresses must be the same.
- IP multicast limits the total number of tunnels that can be terminated through the VPN SPA. Too many tunnels create an instantaneous IP multicast fan-out packet replication burst that overwhelms the input queue of the VPN SPA. If IP multicast is a requirement, keep the number of total streams through the VPN SPA to less than 1000.

Branch Office Scalability

The branch routers are primarily responsible for the following:

- Terminating p2p GRE over IPsec or mGRE tunnels from the headend routers
- Running a routing protocol inside of the GRE tunnels to advertise internal routes

The most important factors to consider when choosing a product for the branch office include the following:

- Branch access speed and expected traffic throughput to the headend (for example, fractional T1, T1, T3, broadband cable/DSL)
- Other services provided by the branch router (for example, DHCP, NAT/PAT, VoIP, Cisco IOS firewall, IOS-IPS)

The pps rate (traffic size and traffic mix) is the largest single factor in branch router scalability.

The number of p2p GRE over IPsec tunnels does not play a large role in the branch sizing because each branch router must be able to terminate a single set of tunnels (primary and secondary) for this design in a hub-and-spoke model.

A primary concern is the amount of traffic throughput (pps and bps) along with the corresponding CPU utilization. Cisco recommends that branch routers be chosen so that CPU utilization does not exceed 65 percent under normal operational conditions. The branch router must have sufficient CPU cycles to service periodic events that require processing. Examples include ISAKMP and IPsec SA establishing and re-keying, SNMP, SYSLOG activities, as well as local CLI exec processing.

After initial deployment and testing, it may be possible to run branch routers at CPU utilization levels higher than 65 percent under normal operational conditions. However, this design guide conservatively recommends staying at or below 65 percent.

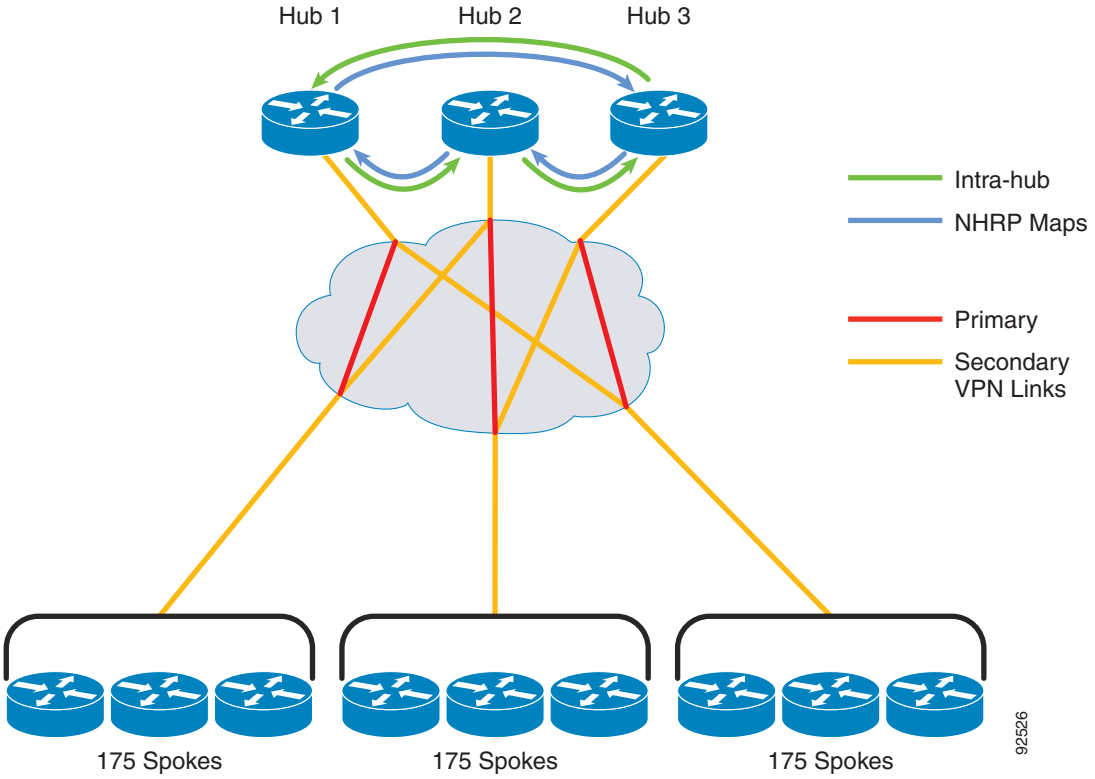
The Cisco Integrated Services Router (ISR) 1840, 2800, and 3800 Series of products have higher CPU performance than the products they replace. The ISR has an encryption module on the motherboard, and can be upgraded to an AIM series of encryption module for increased crypto performance.

Scalable Dual-DMVPN Cloud Topology—Spoke-to-Spoke Designs

Scalable spoke-to-spoke designs are a little more complicated to achieve. To begin with, to achieve resiliency in the design, Cisco recommends that each branch router have a link to two headends. In a dual DMVPN topology, each cloud has a single headend. Routing is used to determine the primary cloud. Headend load balancing can be achieved when a headend router serves two clouds; one being used as a primary for a set of branches and the other as a secondary for a different set of branches. This method allows additional headend routers to be added if more DMVPN clouds are needed to attach additional branch routers to the network.

The drawback of this approach is that spoke-to-spoke sessions are not allowed over different DMVPN clouds. This would require a single large DMVPN cloud for the primary connections and a second single DMVPN cloud that would be used if the primary failed. Building a large DMVPN cloud requires the headend routers to be daisy-chained. To build spoke-to-spoke tunnels between branches located on different headends, the headends must have NHRP maps to each other, just as the branches have NHRP maps to the headends. Consider the case where three headend routers are daisy-chained into a single DMVPN cloud, as shown in Figure 3-3.

Figure 3-3 Scaling the Single DMVPN Cloud



Note the following:

- Groups of spokes are mapped to different hubs as primary and secondary, as indicated by the orange and grey lines. Because each hub has only one mGRE interface to aggregate the primary and secondary tunnels from the spokes, the spoke fan-out is limited to 175 routers per group (totaling 350 spokes per hub).
- The hubs have NHRP maps to each other, as indicated by the blue and dark red lines.
- Intra-hub communications must flow over mGRE tunnels.
- Hubs must be routing peers of each other over the mGRE tunnels.

Note the hub configurations snippets below, with the relevant NHRP and NHS commands in italics:

- Hub1 router:

```
version 12.3
!
hostname Hub1
!
crypto ipsec transform-set ENTERPRISE esp-3des esp-sha-hmac
 mode transport
!
crypto ipsec profile VPN-DMVPN
 set transform-set ENTERPRISE
!
interface Tunnel0
 description mGRE Template Tunnel
 bandwidth 1000
 ip address 10.0.0.1 255.255.240.0
 ip mtu 1400
 no ip next-hop-self eigrp 1
```

```
 ip nhrp authentication cisco123
 ip nhrp map 10.0.0.2 172.16.0.5
 ip nhrp map multicast 172.16.0.5
 ip nhrp map 10.0.0.3 172.16.0.9
 ip nhrp map multicast 172.16.0.9
 ip nhrp map multicast dynamic
 ip nhrp network-id 100000
 ip nhrp holdtime 600
 ip nhrp nhs 10.0.0.2
 no ip split-horizon eigrp 1
 tunnel source FastEthernet0/0
 tunnel mode gre multipoint
 tunnel key 100000
 tunnel protection ipsec profile VPN-DMVPN
!
interface FastEthernet0/0
 description Outside Interface
 ip address 172.16.0.1 255.255.255.252
!
```

- Hub2 router:

```
version 12.3
!
hostname Hub2
!
crypto ipsec transform-set ENTERPRISE esp-3des esp-sha-hmac
 mode transport
!
crypto ipsec profile VPN-DMVPN
 set transform-set ENTERPRISE
!
interface Tunnel0
 description mGRE Template Tunnel
 bandwidth 1000
 ip address 10.0.0.2 255.255.240.0
 ip mtu 1400
 no ip next-hop-self eigrp 1
 ip nhrp authentication cisco123
 ip nhrp map 10.0.0.1 172.16.0.1
 ip nhrp map multicast 172.16.0.1
 ip nhrp map 10.0.0.3 172.16.0.9
 ip nhrp map multicast 172.16.0.9
 ip nhrp map multicast dynamic
 ip nhrp network-id 100000
 ip nhrp holdtime 600
 ip nhrp nhs 10.0.0.3
 no ip split-horizon eigrp 1
 tunnel source FastEthernet0/0
 tunnel mode gre multipoint
 tunnel key 100000
 tunnel protection ipsec profile VPN-DMVPN
!
interface FastEthernet0/0
 description Outside Interface
 ip address 172.16.0.5 255.255.255.252
!
```

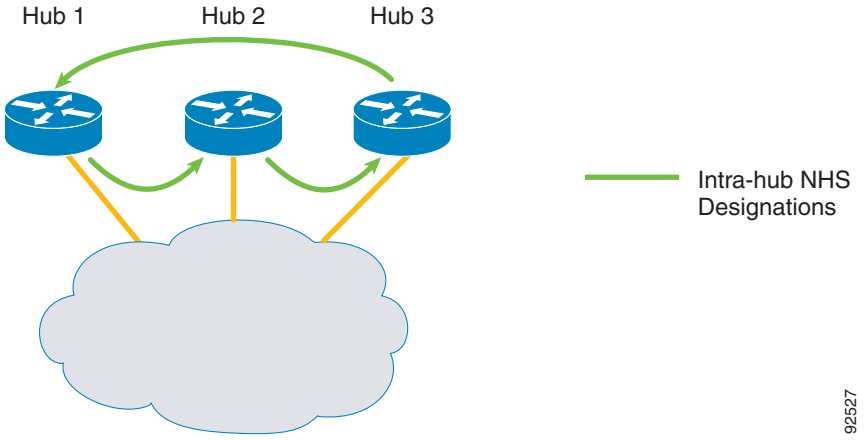
- Hub3 router:

```
version 12.3
!
hostname Hub3
!
```

```
crypto ipsec transform-set ENTERPRISE esp-3des esp-sha-hmac
 mode transport
!
crypto ipsec profile VPN-DMVPN
 set transform-set ENTERPRISE
!
interface Tunnel0
 description mGRE Template Tunnel
 bandwidth 1000
 ip address 10.0.0.3 255.255.240.0
 ip mtu 1400
 no ip next-hop-self eigrp 1
 ip nhrp authentication cisco123
 ip nhrp map 10.0.0.2 172.16.0.5
 ip nhrp map multicast 172.16.0.5
 ip nhrp map 10.0.0.1 172.16.0.1
 ip nhrp map multicast 172.16.0.1
 ip nhrp map multicast dynamic
 ip nhrp network-id 100000
 ip nhrp holdtime 600
 ip nhrp nhs 10.0.0.1
 no ip split-horizon eigrp 1
 tunnel source FastEthernet0/0
 tunnel mode gre multipoint
 tunnel key 100000
 tunnel protection ipsec profile VPN-DMVPN
!
interface FastEthernet0/0
 description Outside Interface
 ip address 172.16.0.9 255.255.255.252
!
```

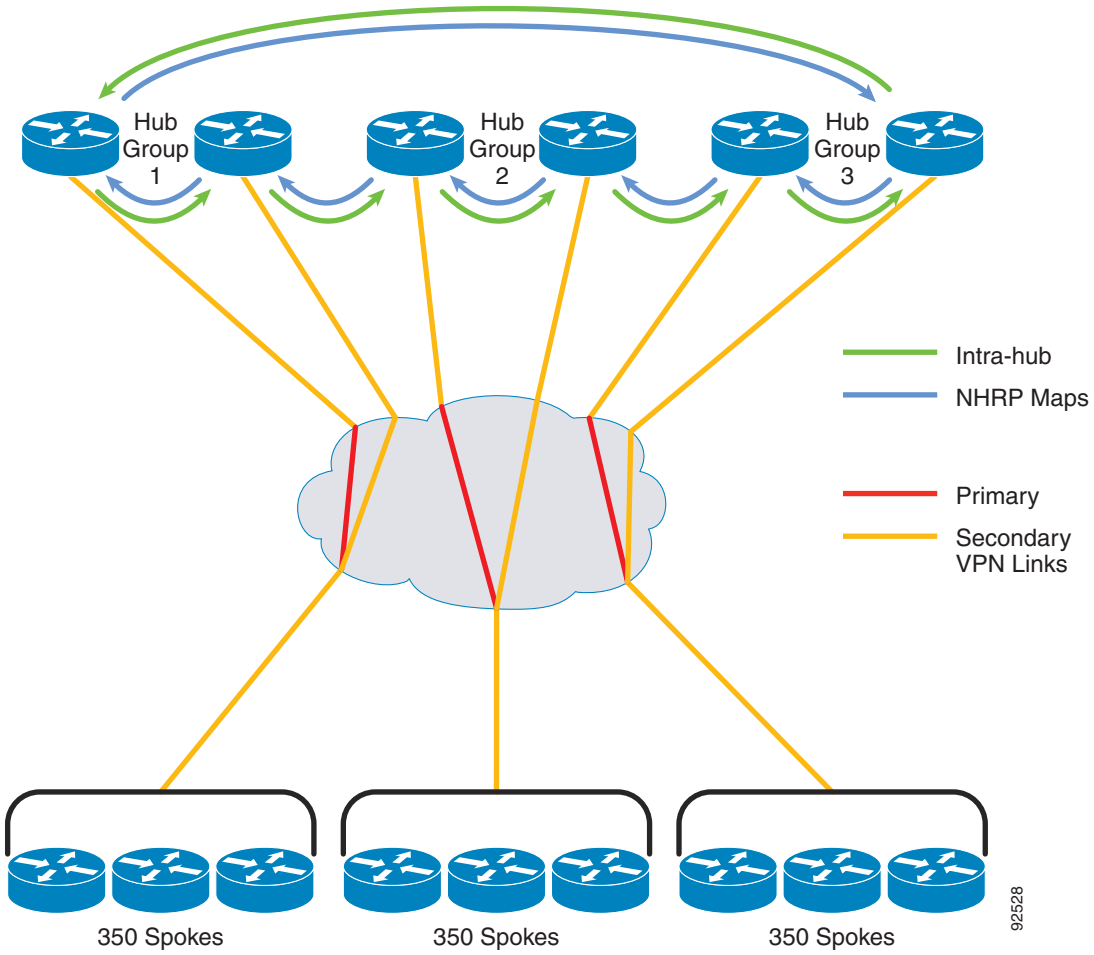
NHRP maps between hubs are bi-directional (1->2, 2->1, 2->3, 3->2, 3->1, 1->3), as shown in the configurations. Additionally, the hubs must point to each other as next-hop servers, which is done in a daisy-chain fashion (1->2, 2->3, 3->1), as shown in Figure 3-4.

Figure 3-4 Single DMVPN Cloud NHS Daisy Chain



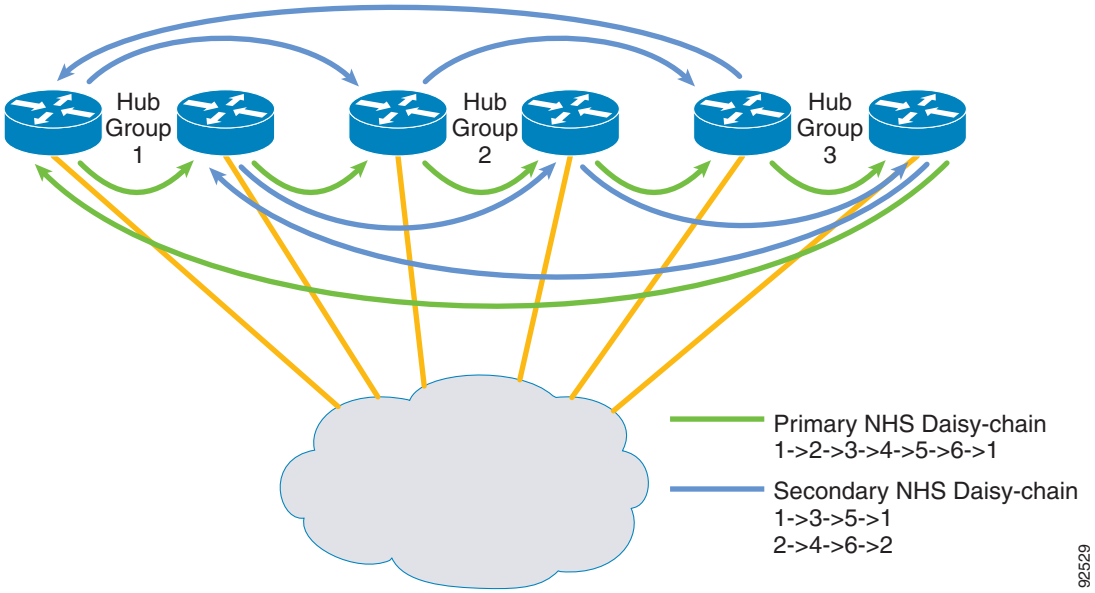
With an NHS daisy-chain deployed in this manner, a multi-hub design is subject to a single point of failure. If one hub is lost, no spoke-to-spoke tunnel setups between spokes connected to the surviving hubs are possible. A more resilient design is shown in Figure 3-5.

Figure 3-5 Single DMVPN Cloud with Primary/Secondary Hubs



In this design, hub routers are deployed in pairs, and operate in a dedicated primary or secondary role, allowing the spoke fan-out to be 350 spokes per hub. All hubs are NHRP-mapped to each other, creating bi-directional paths within the DMVPN for the routing updates to traverse and, as before, all hubs are routing peers or neighbors with each other. With regard to the next-hop server mappings, in this design, all routers belong to a primary daisy chain, with a secondary daisy chain connecting routers serving the primary or secondary role in each hub group, as shown in Figure 3-6.

Figure 3-6 Dual DMVPN with NHS Daisy Chain

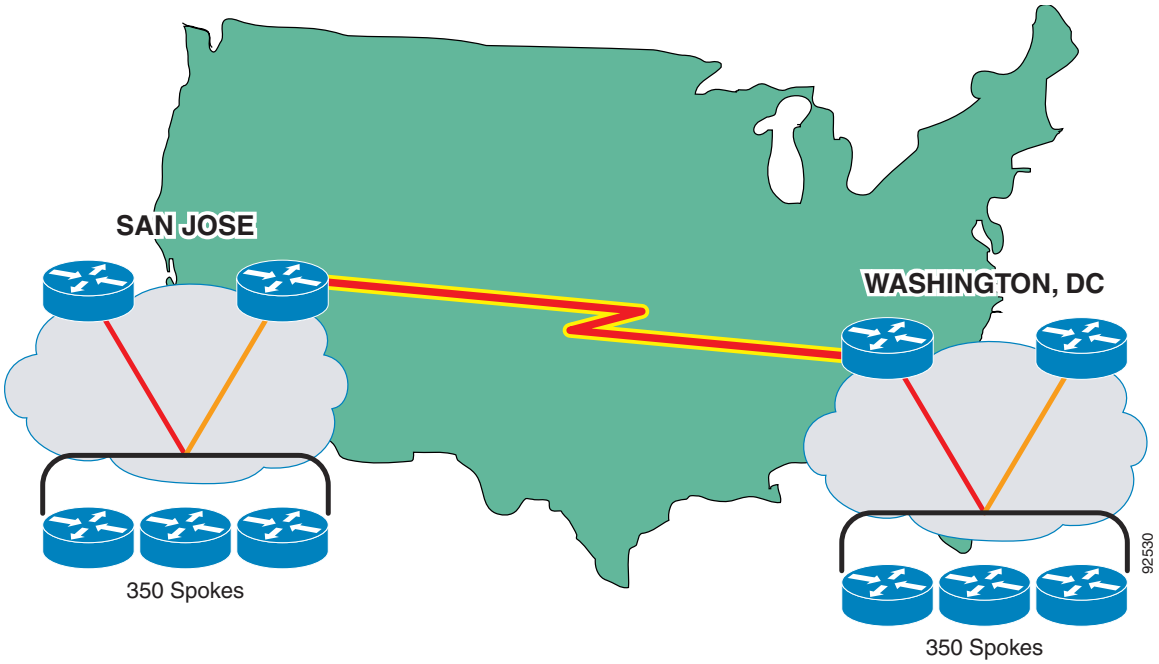


In this design, if a spoke must failover to its secondary hub router, it can still find a path through the secondary NHS daisy chain to open a spoke-to-spoke tunnel to a spoke homed in a different hub group.

Regional Spoke-to-Spoke Clusters

Another option for a very large DMVPN network with spoke-to-spoke requirements, especially one which covers a large geographic area, is to group local spokes into smaller, regional groups, with hubs connected by dedicated high-speed links, as shown in Figure 3-7.

Figure 3-7 Regional Spoke-to-Spoke Clusters



If spoke-to-spoke response time is important, it may be more advantageous to go spoke-hub-hub-spoke, when the two hubs are connected via a high-speed link, than to send traffic via a spoke-to-spoke connection over a long distance via the Internet. In this type of design, the hub devices in the different clusters are connected via any type of IP transport.

Additional Spoke-to-Spoke Design Considerations and Caveats

The ability to create dynamic spoke-to-spoke IPsec tunnels can create the potential for operational problems in the network. As mentioned before, the spoke-to-spoke tunnels do not create routing peers in the network, eliminating concerns about full mesh routing. Other problems can exist, however, which this section examines briefly.

Resiliency

Spoke-to-spoke tunnels are not as resilient to some forms of failure as spoke-to-hub tunnels. Because a routing protocol is not run through the tunnel, a spoke-to-spoke tunnel may fail, with no awareness by the endpoints, allowing traffic to be black-holed. Even ISAKMP keepalives (if configured) may succeed when the encrypted data tunnel is down, and it may take the endpoints an unacceptably long period of time to respond to the loss of the spoke-to-spoke path and resume use of the spoke-hub-spoke path.

Path Selection

The path that a spoke-to-spoke tunnel takes through a public infrastructure, such as the Internet, may actually be slower than a spoke-to-hub-to-spoke path. DMVPN has no way of measuring the delay incurred on the spoke-to-spoke path and adjusting its choice of path because of poor response time or other network quality degradation.

Overloading of Spoke Routers

There are no foolproof mechanisms to prevent a spoke from being overrun by incoming traffic from multiple remote spokes. Especially because spoke routers are likely to be the smaller routers (that is, less powerful CPU), it is possible that multiple tunnel setups can cause operational problems for the spoke device if too many other spokes attempt to create tunnels with it. Two Cisco IOS features (first introduced in IOS 12.3(8)T) help alleviate this situation: IKE Call Admission Control (CAC) and System CAC. The first limits the number of ISAKMP SAs the router can set up, based on an absolute value. The second limits the number of SAs based on the usage of system resources.

IKE CAC is configured as follows on the router:

```
Spoke1#crypto call admission limit ike sa 25
```

In this example, the number of ISAKMP SAs is limited to 25. The router rejects new SA requests after there are 25 active ISAKMP SAs. The state of IKE CAC can be monitored with the **show crypto call admission statistics** command.

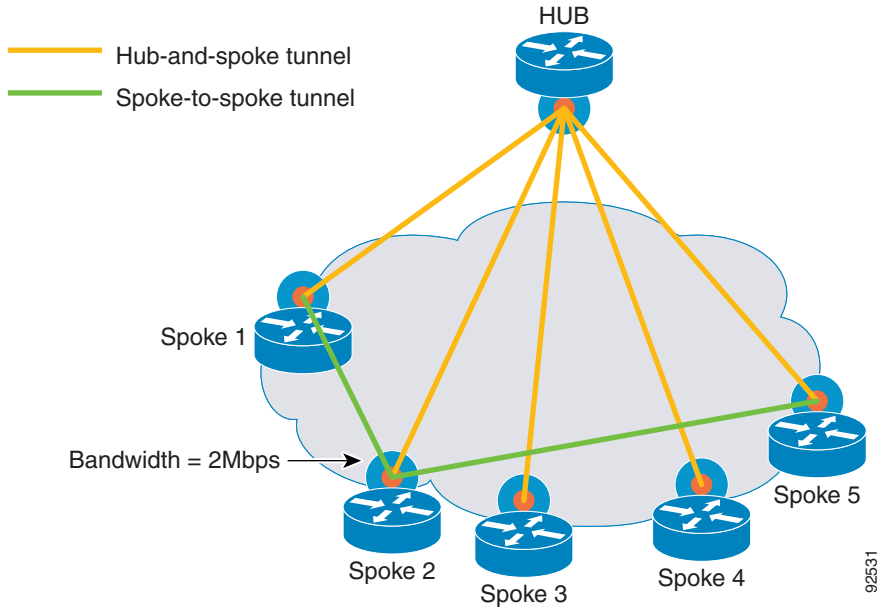
System CAC is configured as follows on the router:

```
Spoke1#call admission limit 80
```

In this example, the router drops new SA requests when 80 percent of system resources are being used. The state of System CAC can be monitored with the **show call admission statistics** command.

A further problem for spoke-to-spoke designs, not addressed by the CAC feature, is simply the overloading of spoke bandwidth by multiple concurrent spoke-to-spoke IPsec tunnels, which may occur even if the IKE authentication of the ISAKMP SA or system resource limits have not been reached. (See Figure 3-8.)

Figure 3-8 Spoke Router Bandwidth Overloading



In this example, Spoke 2 has a 2 Mbps connection to the Internet. It has existing spoke-to-spoke tunnels established with Spoke 1 and Spoke 5. It has not exceeded either its hard-configured IKE CAC or System CAC limits, but traffic on the existing tunnels with the other two spokes has completely consumed the 2 Mbps bandwidth. Spoke 3 attempts to set up a spoke-to-spoke tunnel with Spoke 2. There is enough

bandwidth available for the NHRP requests and ISAKMP and IPsec session establishment to occur, but after the tunnel is up, there is not enough bandwidth to send and receive application traffic. The problem here is that there simply is no way for Spoke 2 to tell Spokes 1 and 5 to throttle back their data flows to share the 2 Mbps access link fairly. Upper layer protocols such as UDP or TCP eventually adapt; however, RTP has no flow control mechanism. This is part of the greater QoS dilemma discussed earlier.

At this time, the only workarounds to the problems of bandwidth overloading are the following:

- Designing the network with adequate bandwidth for the anticipated application load
- Balancing the percentage of spoke-to-hub and spoke-to-spoke flows to a reasonable level; the design recommendation is 80 percent hub-to-spoke and 20 percent spoke-to-spoke
- Setting user expectations of the response time and availability of the link appropriately