IP Tunnels

In This Section

This section provides an overview of IP Security (IPSec) software features for the IPSec ISA.

Topics in this section include:

IP Tunnels Overview

This section discusses IP Security (IPSec), GRE tunneling, and IP-IP tunneling features supported by the MS-ISA. In these applications, the MS-ISA functions as a resource module for the system, providing encapsulation and (for IPSec) encryption functions. The IPSec encryption functions provided by the MS-ISA are applicable for many applications including: encrypted SDPs, video wholesale, site-to-site encrypted tunnel, and remote access VPN concentration.

Figure 26 shows an example of an IPSec deployment, and the way this would be supported inside a 7750. GRE and IP-IP tunnel deployments are very similar.

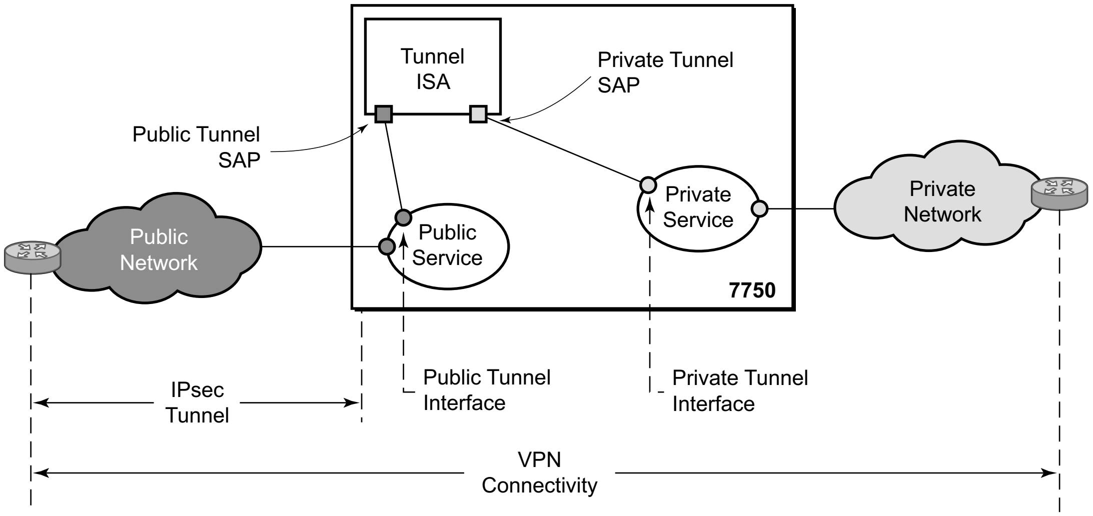

Figure 26: 7750 IPSec Implementation Architecture

Figure 26, the public network is typically an “insecure network” (for example, the public Internet) over which packets belonging to the private network in the diagram cannot be transmitted natively. Inside the 7750, a public service instance (IES or VPRN) connects to the public network and a private service instance (typically a VPRN) connects to the private network.

The public and private services are typically two different services, and the MS-ISA is the only “bridge” between the two. Traffic from the public network may need to be authenticated and encrypted inside an IPSec tunnel to reach the private network. In this way, the authenticity/confidentiality/integrity of accessing the private network can be enforced.If authentication and confidentiality are not important then access to the private network may alternatively be provided through GRE or IP-IP tunnels.

The MS-ISA provides a variety of encryption features required to establish bi-directional IPSec tunnels including:

Control Plane:

- Manual Keying
- Dynamic Keying: IKEv1/v2
- IKEv1 Mode: Main and Aggressive
- Authentication: Pre-Shared-Key /xauth with RADIUS support/X.509v3 Certificate
- Perfect Forward Secrecy (PFS)
- DPD
- NAT-Traversal
- Security Policy

Data Plane:

- ESP (with authentication) Tunnel mode
- Authentication Algorithm: MD5/SHA1/SHA256/SHA384/SHA512
- Encryption Algorithm: DES/3DES/AES128/AES192/AES256
- DH-Group: 1/2/5/14/15
- Anti-Replay Protection
- N:M IPSec ISA card redundancy

Note: SR OS will use a configured authentication algorithm in an ike-policy for Pseudorandom Function (PRF).

There are two types of tunnel interfaces and SAPs:

- Public tunnel interface: configured in the public service; outgoing tunnel packets have a source IP address in this subnet
- Public tunnel SAP: associated with the public tunnel interface; a logical access point to the MS-ISA card in the public service
- Private tunnel interface: configured in the private service; can be used to define the subnet for remote access IPSec clients.
- Private tunnel SAP: associated with the private tunnel interface, a logical access point to the MS-ISA card in the private service

Traffic flows to and through the MS-ISA card as follows:

- In the upstream direction, the encapsulated (and possibly encrypted) traffic is forwarded to a public tunnel interface if its destination address matches the local or gateway address of an IPSec tunnel or the source address of a GRE or IP-IP tunnel. Inside the MS-ISA card, encrypted traffic is decrypted, the tunnel header is removed, the payload IP packet is delivered to the private service, and from there, the traffic is forwarded again based on the destination address of the payload IP packet.

- In the downstream direction, unencapsulated/clear traffic belonging to the private service is forwarded into the tunnel by matching a route with the IPSec/GRE/IP-IP tunnel as next-hop. The route can be configured statically, learned by running OSPF on the private tunnel interface (GRE tunnels only), learned by running BGP over the tunnel (IPSec and GRE tunnels only), or learned dynamically during IKE negotiation (IPSec only). After clear traffic is forwarded to the MS-ISA card, it is encrypted if required, encapsulated per the tunnel type, delivered to the public service, and from there, the traffic is forwarded again based on the destination address of the tunnel header.

Tunnel ISAs

A tunnel-group is a collection of MS-ISAs (each having mda-type **isa-tunnel**) configured to handle the termination of one or more IPSec, GRE and/or IP-IP tunnels. An example tunnel-group configuration is shown below:

```
config isa
     tunnel-group 1 create
          primary 1/1
          backup 2/1
          no shutdown
          exit
```

A GRE, IP-IP, or IPSec tunnel belongs to only one tunnel group. There are two types of tunnel groups:

- A single-active tunnel-group can have one tunnel-ISA designated as primary and optionally one other tunnel-ISA designated as backup. If the primary ISA fails the affected failed tunnels are re-established on the backup (which is effectively a cold standby) if it is not already in use as a backup for another tunnel-group.
- A multi-active tunnel-group can have multiple tunnel-ISAs designated as primary. This is only supported on 7750 SR7/SR12/SR12E with chassis mode D or 7450 mixed mode with IOM3.

The show isa tunnel-group allows the operator to view information about all configured tunnelgroups.This command displays the following information for each tunnel-group: group ID, primary tunnel-ISAs, backup tunnel-ISAs, active tunnel-ISAs, admin state and oper state.

Public Tunnel SAPs

A VPRN or IES service (the delivery service) must have at least one IP interface associated with a public tunnel SAP to receive and process the following types of packets associated with GRE, IP-IP, and IPSec tunnels:

- GRE (IP protocol 47)
- IP-IP (IP protocol 4)
- IPSec ESP (IP protocol 50)
- IKE (UDP)

The public tunnel SAP type has the format tunnel-*tunnel-group*.public:*index*, as shown in the following CLI example.

```
config service ies 199 customer 1 create
     interface “public-1” create
```

```
          address 64.251.12.1/30
          sap tunnel-1.public:200 create
     exit all
```

Private Tunnel SAPs

The private service must have an IP interface to a GRE, IP-IP, or IPSec tunnel in order to forward IP packets into the tunnel, causing them to be encapsulated (and possibly encrypted) per the tunnel configuration and to receive IP packets from the tunnel after the encapsulation has been removed (and decryption). That IP interface is associated with a private tunnel SAP.

The private tunnel SAP has the format tunnel-*tunnel-group*.private:*index*, as shown in the following CLI example where a GRE tunnel is configured under the SAP.

```
config service vprn 1 customer 1 create
     interface “gre tunnel to ce1” tunnel create
          address 10.0.0.1/30
          ip-mtu 1476
          sap tunnel-1.private:210 create
               ip-tunnel “to ce1” create
                    gre-header
                    dest-ip 10.0.0.2
                    source 64.251.12.1
                    remote-ip 12.47.10.33
                    backup-remote-ip 12.47.51.7
                    delivery-service 199
                    dscp af11
                    no shutdown
                    exit
               ingress
               egress
               exit all
```

IP Interface Configuration

In the configuration example of the previous section the IP address 10.0.0.1 is the address of the GRE tunnel endpoint from the perspective of payload IP packets. This address belongs to the address space of the VPRN 1 service and will not be exposed to the public IP network carrying the GRE encapsulated packets. An IP interface associated with a private tunnel SAP does not support unnumbered operation.

It is possible to configure the IP MTU (M) of a private tunnel SAP interface. This sets the maximum payload IP packet size (including IP header) that can be sent into the tunnel – for example, it applies to the packet size before the tunnel encapsulation is added. When a payload IPv4 packet that needs to be forwarded to the tunnel is larger than M bytes:

- If the DF bit is clear, the payload packet is IP fragmented to the MTU size prior to tunnel encapsulation.
- If the DF bit is set, the payload packet is discarded and (if allowed by the ICMP setting of the sending interface) an ICMP type 3/code 4 is returned to the sender (with MTU of the private tunnel SAP interface in the payload).

GRE and IP-IP Tunnel Configuration

To bind an IP/GRE or IP-IP tunnel to a private tunnel SAP, the **ip-tunnel** command should be added under the SAP. To configure the tunnel as an IP/GRE tunnel, the **gre-header** command must be present in the configuration of the **ip-tunnel**. To configure the tunnel as an IP-IP tunnel, the **ip-tunnel** configuration should have the **no gre-header** command. When configuring a GRE or IP-IP tunnel, the **dest-ip** command is mandatory as this specifies the private IP address of the remote tunnel endpoint. If the dest-ip address is not within the subnet of the local private endpoint then the tunnel will not come up. In the CLI sub-tree under **ip-tunnel**, there are commands to configure the following:

- The source address of the GRE or IP-IP tunnel- This is the source IPv4 address of GRE or IP-IP encapsulated packets sent by the delivery service. It must be an address in the subnet of the associated public tunnel SAP interface.
- The remote IP address - If this address is reachable in the delivery service (there is a route) then this is the destination IPv4 address of GRE or IP-IP encapsulated packets sent by the delivery service.
- The backup remote IP address- If the remote IP address of the tunnel is not reachable then this is the destination IPv4 address of GRE or IP-IP encapsulated packets sent by the delivery service.
- The delivery service- This is the id or name of the IES or VPRN service where GRE or IP-IP encapsulated packets are injected and terminated. The delivery service can be the same service where the private tunnel SAP interface resides.

- The DSCP marking in the outer IP header of GRE encapsulated packets- If this is not configured then the default is to copy the DSCP from the inner IP header to the outer IP header.

A private tunnel SAP can have only one ip-tunnel sub-object (one GRE or IP-IP tunnel per SAP).

The show ip tunnel displays information about a specific IP tunnel or all configured IP tunnels. The following information is provided for each tunnel:

- service ID that owns the tunnel
- private tunnel SAP that owns the tunnel
- tunnel name, source address
- remote IP address
- backup remote IP address
- local (private) address
- destination (private) address
- delivery service
- dscp
- admin state
- oper state
- type (GRE or IP-IP)

IP Fragmentation and Reassembly for IP Tunnels

An IPSec, GRE or IP-IP tunnel packet that is larger than the IP MTU of some interface in the public network must either be discarded (if the Do Not Fragment bit is set in the outer IP header) or fragmented. If the tunnel packet is fragmented, it is then up to the destination tunnel endpoint to reassembly the tunnel packet from its fragments. Starting in R10, IP reassembly can be enabled for all the IPSec, GRE, and IP-IP tunnels belonging to a tunnel-group. For IP-IP and GRE tunnels, the reassembly option is also configurable on a per-tunnel basis so that some tunnels in the tunnel-group can have reassembly enabled, and others can have the extra processing disabled. When reassembly is disabled for a tunnel, all received fragments belonging to the tunnel are dropped.

To avoid public network fragmentation of IPSec, GRE, or IP-IP packets belonging to a particular tunnel, one possible strategy is to fragment IPv4 payload packets larger than a specified size M at entry into the tunnel (before encapsulation and encryption if applicable). The size M is configurable using the **ip-mtu** command under the **ip-tunnel** or **ipsec-tunnel** configuration.

If the payload IPv4 packets are all M bytes or less in length then it is guaranteed that all resulting tunnel packets will be less than M+N bytes in length, if N is the maximum overhead added by the tunneling protocol. If M+N is less than the smallest interface, IP MTU in the public network fragmentation will be avoided. In some cases, some of the IPv4 payload packets entering a tunnel may have their Don’t Fragment (DF) bit set. And if desired, SR-OS supports the option (also configurable on a per-tunnel basis) to clear the DF bit in these packets so that they can be fragmented.

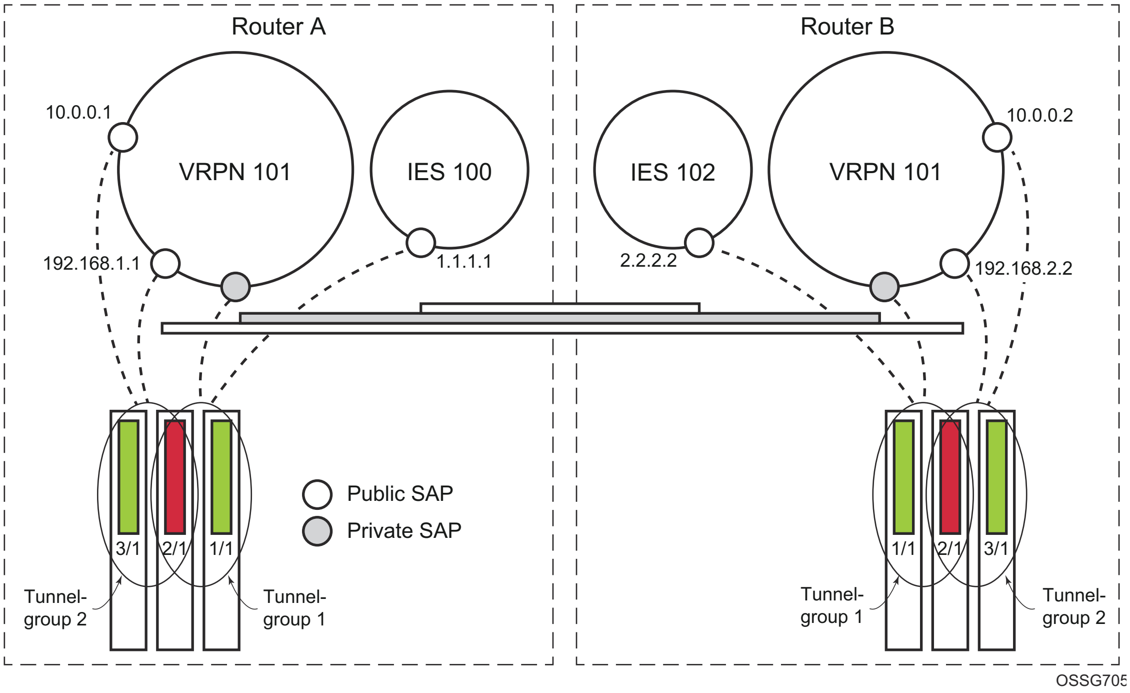

Figure 27: Example GRE over IPSec Tunnel Configuration

Configuration of Router A:

```
config isa
     tunnel-group 1 create
          primary 1/1
          backup 2/1
          no shutdown
          exit
     tunnel-group 2 create
          primary 3/1
          backup 2/1
          no shutdown
          exit
     exit all

config ipsec
     ike-policy 1 create
          auth-algorithm sha1
          dh-group 5
          encryption-algorithm aes128
          ike-mode main
          ike-version 1
          exit
     ipsec-transform 1
          esp-auth-algorithm sha1
          esp-encryption-algorithm aes256
          exit
     exit all

config service ies 100 customer 1 create
     interface “public-ipsec-1” create
          address 1.1.1.1/24
          sap tunnel-1.public:200 create
          exit all

config service vprn 101 customer 1 create
     ipsec
          security-policy 1 create
               entry 1 create
                    local-ip 192.168.1.0/24
                    remote-ip 192.168.2.0/24
                    exit
               exit
          exit
     interface “private-ipsec-1” tunnel create
          sap tunnel-1.private:201 create
               ipsec-tunnel “ipsec-tunnel-for-n-gre-tunnels” create
                    security-policy 1
                    local-gateway-address 1.1.1.2 peer 2.2.2.2 delivery-service 100
                    dynamic-keying
                         ike-policy 1
                         pre-shared-key “secret”
                         transform 1
                         exit
                    exit
               exit
          exit
     interface “public-gre-1” create
          address 192.168.1.1/24
          sap tunnel-2.public:200 create
          exit
     interface “private-gre-1” tunnel create
          address 10.0.0.1/30
          sap tunnel-2.private:201 create
               ip-tunnel “protected-gre-tunnel” create
                    gre-header
                    dest-ip 10.0.0.2
                    source 192.168.1.2
                    remote-ip 192.168.2.2
                    delivery-service 101
                    no shutdown
                    exit
               exit
          exit
static-route 192.168.2.0/24 next-hop ipsec-tunnel “ipsec-tunnel-for-n-gre-tunnels”
```

Operational Conditions

A tunnel group that is in use cannot be deleted. In single-active mode, changes to the primary ISA are allowed only in when the tunnel group is in a shutdown state. Change to the backup ISA (or the addition of a backup ISA) is allowed at any time unless the ISA is currently active for this tunnel group. When the backup module is active, changing the primary module is allowed without shutting down the tunnel group.

A shutdown of tunnel-group is required to do the following:

- Change the mode between multi-active and single-active.
- Change the primary-isa in single-active mode.
- Change the active-mda-number in multi-active mode.
- De-configure an active MS-ISA in multi-active mode.

In multi-active mode, if the active member ISA goes down, system will replace it with backup ISA; however, if there is no backup ISA, the tunnel-group will be "oper-down".

A change to the IPSec transform policy is allowed at any time. The change will not impact tunnels that have been established until they are renegotiated. If the change is required immediately the tunnel must be cleared (reset) for force renegotiation.

A change to the **ike-policy** is allowed at any time. The change will not impact tunnels that have been established until they are renegotiated. If the change is required immediately the tunnel must be cleared (reset) for force renegotiation.

The public interface address can be changed at any time (current behavior). If changed, tunnels that were configured to use it will require a configuration change. If the subnet changed the tunnels will be in an operationally down state until their configuration is corrected. The public service cannot be deleted while tunnels are configured to use it. A public service is the IES or VPRN service that hold the regular interface that connects the node to the public network. A private service connects to the private protected service.

A tunnel group ID or tag cannot be changed. To remove an tunnel group instance, it must be in a shutdown state (both front-door and back-door).

A change to the security policy is not allowed while a tunnel is active and using the policy.

The tunnel local-gateway-address, peer address, or delivery router parameters cannot be changed while the tunnel is operationally up (shutdown will make it both admin down and operationally down).

A tunnel security policy cannot be changed while the tunnel is operationally up. An IPSec transform policy or ike-policy assignments to a tunnel requires the tunnel to be shutdown.

QoS Interactions

The MS-ISA can interact with the queuing functions on the IOM through the ingress/egress QoS provisioning in the IES or IP VPN service where the IPSec session is bound. Multiple IPSec sessions can be assigned into a single IES or VPRN service. In this case, QoS defined at the IES or VPRN service level, is applied to the aggregate traffic coming out of or going into the set of sessions assigned to that service.

In order to keep marking relevant in the overall networking design, the ability to translate DSCP bit marking on packets into DSCP bit markings on the IPSec tunneled packets coming out of the tunnel is supported.

OAM Interactions

The MS-ISA is IP-addressed by an operator-controlled IP on the public side. That IP address can be used in Ping and Traceroute commands and the ISA can either respond or forward the packets to the CPM.

For static LAN-to-LAN tunnel, in multi-active mode, a ping requests to public tunnel address would not be answered if the source address is different from the remote address of the static tunnel.

The private side IP address is visible. The status of the interfaces and the tunnels can be viewed using show commands.

Traffic that ingresses or egresses an IES or VPRN service associated with certain IPSec tunnels can be mirrored like other traffic.

Mirroring is allowed per interface (public) or IPSec interface (private) side. A filter mirror is allowed for more specific mirroring.

Redundancy

In single-active mode, every tunnel group can be configured with primary and backup ISAs. An ISA can be used as a backup for multiple IPSec groups. The ISAs are cold standby such that upon failure of the primary the standby resumes operation after the tunnels re-negotiate state. While the backup ISA can be shared by multiple tunnel groups only one tunnel group can fail to a single ISA at one time (no double failure support).

In multi-active mode, system will select number of active-mda-number from all configured ISAs to be active ISA. The rest of ISAs will be standby ISA.The standby ISAs are cold standby.

IPSec also supports dead peer detection (DPD).

Note that BFD can be configured on the private tunnel interfaces associated with GRE tunnels and used by the OSPF, BGP or static routing that is configured inside the tunnel.

SR-OS also supports multi-chassis IPSec redundancy, which provides 1:1 stateful protection against MS-ISA failure or chassis failure

Statistics Collection

Input and output octets and packets per service queue are used for billing end customers who are on a metered service plan. Since multiple tunnels can be configured per interface the statistics can include multiple tunnels. These can be viewed in the CLI and SNMP.

Reporting (syslog, traps) for authentication failures and other IPSec errors are supported, including errors during IKE processing for session setup and errors during encryption or decryption.

A session log indicates the sort of SA setup when there is a possible negotiation. This includes the setup time, teardown time, and negotiated parameters (such as encryption algorithm) as well as identifying the service a particular session is mapped to, and the user associated with the session.

Security

The MS-ISA module provides security utilities for IPSec-related service entities that are assigned to interfaces and SAPs. These entities (such as card, isa-tunnel module, and IES or VPRN services) must be enabled in order for the security services to process. The module only listens to requests for security services from configured remote endpoints. In the case of a VPN concentrator application, these remote endpoints could come from anywhere on the Internet. In the cases where a point-to-point tunnel is configured, the module listens only to messages from that endpoint.

GRE Tunnel IPv4 Multicast Support

GRE tunnels only support unicast and multicast IPv4 packets as payload. From a multicast prospective, GRE tunnel IP interface (associated with a private tunnel SAP) can be configured as an IGMP interface and/or as a PIM interface. The following multicast features are supported:

- IGMP versions 1, 2 and 3
- IGMP import policies
- IGMP host tracking
- Static IGMP membership
- Configurable IGMP timers
- IGMP SSM translation
- Multicast CAC
- Per-interface, per-protocol (IGMP/PIM) multicast group limits
- MVPN support (draft-rosen)
- MVPN support (BGP-MPLS)
- PIM-SM and SSM operation
- PIM BFD support
- Configurable PIM timers
- Configurable PIM priority
- PIM tracking support
- PIM ECMP (bandwidth or hash-based)
- Static multicast route

IKEv2

IKEv2, defined in RFC 4306, *Internet Key Exchange (IKEv2) Protocol*, is the second version of the Internet Key Exchange Protocol. The main driver of IKEv2 is to simplify and optimize the IKEv1. An IKE_SA and a CHILD_SA could be created with only 4 IKEv2 messages exchanges. The 7750-SR supports IKEv2 with following features:

- Static lan-to-lan tunnel
- Dynamic lan-to-lan tunnel
- Pre-shared-key authentication, certificate authentication
- Liveness check
- IKE_SA rekey
- Child_SA rekey

SHA2 Support

According to RFC 4868, *Using HMAC-SHA-256, HMAC-SHA-384, and HMAC-SHA-512 with IPsec*, the following SHA2 variants are supported for authentication or pseudo-random functions:

Use HMAC-SHA-256+ algorithms for data origin authentication and integrity verification in IKEv1/2, ESP:

- AUTH_HMAC_SHA2_256_128
- AUTH_HMAC_SHA2_384_192
- AUTH_HMAC_SHA2_512_256

For use of HMAC-SHA-256+ as a PRF in IKEv1/2:

- PRF_HMAC_SHA2_256
- PRF_HMAC_SHA2_384
- PRF_HMAC_SHA2_512

X.509v3 Certificate Overview

X.509 is an ITU-T standard for a public key infrastructure (ing up) which allows entities to build trust relationships between each other based on their mutual trust of Certificate Authority (CA). The trusted CA issues certificate which includes the signed public key (by CA) of issued entity. Entities can trust the certificates because they trust the CA and can verify the CA's signature by using CA's root certificate.

SR OS's x.509 certificate management provides an infrastructure for x.509 certificate management including:

- Key generation
- CA profile management
- Certificate management
- CRL management

This feature can be used by other applications in the SR OS that require certificate authentication. The IKEv2 static and dynamic lan-to-lan tunnel are supported.

Key Generation

SR OS supports the generation of the following types of keys. They can be configured locally through the CLI.

- RSA
- DSA

With one of the following key sizes:

- 512
- 1024
- 2048

The generated key is stored in local CF card.

Note: The generated key file is a plain DER format file and must be imported before can be further used.

Formats and Local Storage

The following formats can be used by the SR OS directly:

- KEY file: encrypted file
- Certificate: DER
- Certificate-Request: PEM
- CRL: DER

The KEY/certificate/CRL files must be imported before they can be provisioned in a CA profile or tunnel configuration. SR OS stores imported files in the fixed directory **cf3:\system-pki**.

Import and Export

The process of import is to convert the format of input file to system’s format and save into fixed directory specified in section “Formats and Local Storage”; the process of export is to convert system’s format to one of specified format.

- Certificates can be import/export using following formats:
  - → PKCS#10
  - → PKCS#7 (DER and PEM)
  - → PEM
  - → DER

Note that certain formats of file could encapsulate multiple certificates, in this case, the SR-OS will only use the first certificate in the file.

- The Key pair can be import/export by using following formats:
  - → PKCS#12 (along with certificate)
  - → PEM
  - → DER
- The CRL can be import/export by using following formats:
  - → PKCS#7 (DER and PEM)
  - → PEM
  - → DER

Note that the PKCS#12 file may be encrypted with a password.

Certificate Enrollment

SR OS support X.509v3 certificate. Use the following steps to enroll a certificate:

1. Generate a key file.
2. Generate a certificate-request by using generated key file or an existing plain DER key file.
3. Send the certificate-request file to Certificate Authority (CA) via an out-of-band method.
4. CA signs the certificate-request and returns the signed certificate.
5. Import the signed certificate and the generated key.

CA Profile

The SR OS uses the ca-profile to manage Certificate Authority information. The ca-profile includes:

- CA name
- CA’s certificate
- Certificate Revocation List (CRLv2)

Use Certificate Authentication for IKEv2 Static/Dynamic LAN-to-LAN tunnel

The SR OS supports X.509v3 certificate authentication for IKEv2 LAN-to-LAN tunnel. SR OS also supports asymmetric authentication, which means the SR OS and the IKEv2 peer can use different methods to authenticate. For example, one side could use pre-shared-key and the other side could use x.509 certificate.

The SR OS supports certificate chain verification. For each static LAN-to-LAN tunnel or ipsec-gw, there will be a configurable trust-anchor, which specifies the expecting CA that should be present in the certificate chain before reaching the root or the self signed certificate.

The SR OS’s own key and certificate are also configurable per tunnel or ipsec-gw.

Note that when using certificate authentication, the SR OS will use the subject of the configured certificate as the its ID.

Remote Access VPN Concentrator Example

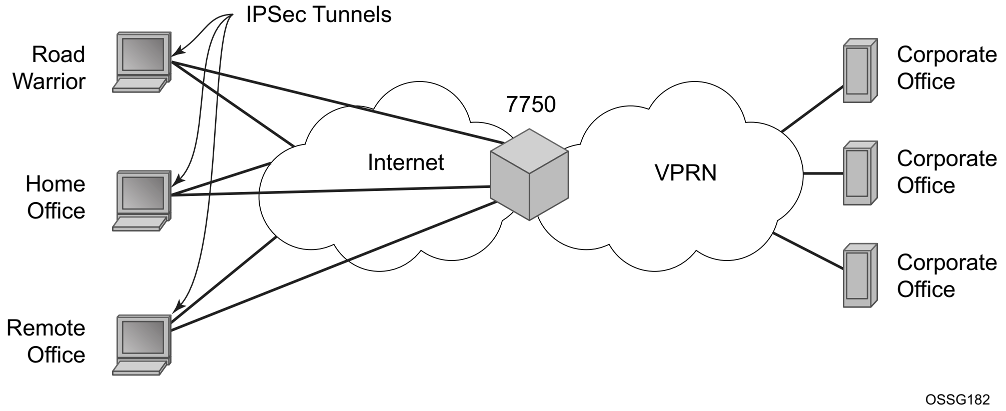

Figure 28: IPSec into VPRN Example

In this application (Figure 28), an IPSec client sets up encrypted tunnel across public network. The 7750 MS-ISA acts as a concentrator gathering, and terminating these IPSec tunnels into an IES or VPRN service. This mechanism allows as service provider to offer a global VPRN service even if node of the VPRN are on an uncontrolled or insecure portion of the network.

Video Wholesale Example

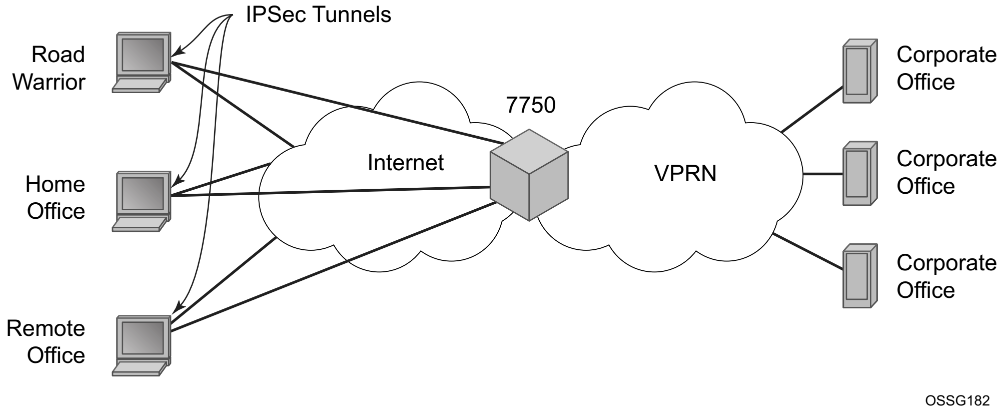

Figure 29: Video Wholesale Configuration

As satellite headend locations can be costly, many municipal and second tier operators cannot justify the investment in their own ground station in order to offer triple play features. However, it is possible for a larger provider or a cooperative of smaller providers to unite and provide a video headend. Each retail subscriber can purchase content from this single station, and receive it over IP. However, encryption is required so the signal cannot be understood if intercepted. A high speed encrypted tunnel is preferred over running two layers of double video protection which is cumbersome and computationally intensive.

Multi-Chassis IPSec Redundancy Overview

Multi-Chassis IPsec redundancy (MC-IPsec) provides a 1:1 IPsec stateful failover mechanism between two chassis.

- This feature provides protection against MS-ISA failure and chassis failure.
- IKEv2 static LAN-to-LAN is supported in SR-OS Release 10.0R5; IKEv2 dynamic LAN-to-LAN tunnel is supported in SR-OS Release 10.0R8.
- This feature is supported on following platforms:
  - → 7750 SR7, SR12 and SR12E
  - → IOM3 and chassis mode D
  - → 7450 mixed mode
  - → Multi-active tunnel-group only
- The granularity of failover is per tunnel-group, which means a specific tunnel-group could failover to standby chassis independent of other tunnel-groups on the master chassis.
- The following components are included in this feature:
  - → Master Election: MIMP (MC-IPsec Mastership Protocol) runs between chassis to elect master, MIMP run for each tunnel-group independently
  - → Synchronization: MCS (Multi-Chassis Synchronization) sync IPsec states between chassis
  - → Routing:
    - – MC-IPsec aware routing attract traffic to the master chassis
    - – Shunting support
    - – MC-IPsec aware VRRP (10.0R8)

Architecture

The overall MC-IPSec redundancy architecture is displayed in Figure 30:

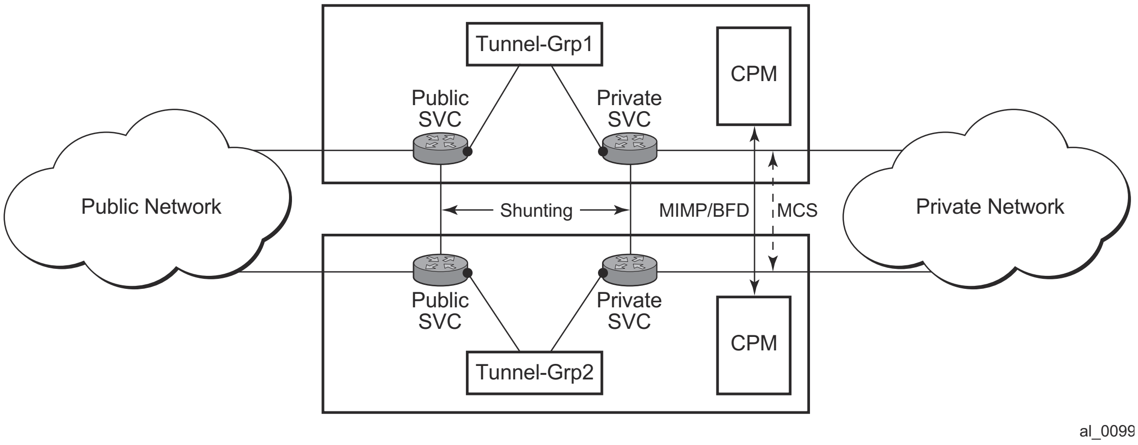

Figure 30: MC-IPSec Architecture

MC-IPSec Mastership Protocol (MIMP)

With MIMP enabled, there is a master chassis and a backup chassis. The state of the master or standby is per tunnel-group. For example (Table 12), chassis A and B, for tunnel-group 1, A is master, B is standby; for tunnel-group 2, A is standby, B is master.

Table 12: Master and Backup Chassis Example

	Master	Standby
Tunnel Group 1	A	B
Tunnel Group 2	B	A

All IKEv2 negotiation and ESP traffic encryption/decryption only occurs on the master chassis. If the backup chassis receives such traffic, if possible, it will shunt them to the master.

There will be a mastership election protocol (MIMP) running between the chassis to elect the master. This is an IP-based protocol to avoid any physical topology restrictions.

A central BFD session could be bound to MIMP to achieve fast chassis failure detection.

MIMP Protocol States

There are five MIMP states:

1. Discovery
   - Upon MC-IPSec is enabled for the tunnel-group, for example:
     - → System starts up.
     - → no shutdown MC-IPSec peer.
     - → no shutdown MC-IPSec tunnel-group.
   - Functionally, this means blackhole traffic to the MS-ISA and no shunting.
   - If the peer is reached before the discovery-interval (configurable) has expired, then the state will be changed to whatever the MIMP decides
   - If the peer is not reached before the discovery-interval has expired, then the state will be changed to **eligible** or **notEligible** depending on the oper-status of the tunnel-group.
2. notEligible
   - The tunnel-group is operationally down.
   - Functionally, this means blackhole traffic to the MS-ISA and no shunting.
3. Eligible
   - The peer is not reachable or the associated BFD session is down but the tunnel-group is operationally up.

   - Functionally, this means the MS-ISA processes traffic.
4. Standby
   - Peer is reachable, elected standby.
   - Functionally, this means blackhole traffic to MS-ISA and shunting if possible.
5. Master
   - Peer is reachable, elected master.
   - Functionally, this means the MS-ISA processes traffic.

Election Logic

The following election logic is executed when MIMP packets are exchanged.

Calculate Master Eligibility:

1. Set masterEligible to TRUE if the local tunnel group is operationally up, otherwise FALSE.
2. Set peerMasterEligible to TRUE if the peer’s tunnel group is operationally up, otherwise FALSE.

First elect based on eligibility:

3. If masterEligible and not peerMasterEligible, elect self master -> DONE.
4. If not masterEligible and peerMasterEligible, elect peer master -> DONE.
5. If not masterEligible and not peerMasterEligible, no master -> DONE.

Then apply stickyness rules (mastership tends not to change)

6. If l was “acting master” and peer was not “acting master”, then elect self master -> DONE.
7. If 1 was not “acting master” and peer was “acting master”, then elect peer master -> DONE.

Note: An “acting master” is either in MIMP state “master” or “eligible”.

Then elect based on priority and number of active ISA:

8. If my priority is higher than peer, elect self master -> DONE.
9. If peer priority is higher than mine, elect peer master -> DONE.
10. If I have more active ISA than peer, elect self master -> DONE.
11. If peer has more active ISA than me, elect peer master -> DONE.

The tie breaker:

12. If the local chassis's MIMP source address is higher than the peer's, elect self master -> DONE.
13. Elect peer master -> DONE.

Protection Status

Each MC-IPSec-enabled tunnel-group has a "protection status", which could be one of following:

- notReady — The tunnel-group is not ready for a switchover due to reasons such as no elected standby to takeover or there are pending IPSec states which need to be synced. If switchover occurs with this status, then there could be a significant traffic impact.
- nominal — The tunnel-group is in a better situation to switchover than notReady. However, traffic still may be impacted.

Protection status serves as an indication for the operator to decide the optimal time to perform a controlled switchover.

The **show redundancy multi-chassis mc-ipsec peer** <*ip-address*> **tunnel-group** <*tunnel-group-id*>" command can be used to check current protection status.

Other Details

- Mastership election is per tunnel-group.
- MIMP is running in the base routing instance.
- MIMP will use the configured value of the **config>redundancy>multi-chassis>peer>source-address** command as the source address. If not configured, then system address will be used.
- The priority range is from 0 to 255.
- When an mc-ipsec tunnel-group enters standby from acting master, the tunnel-group will be restarted.
- When a tunnel-group enters an admin shutdown state under the mc-ipsec configuration (add a tunnel-group to mc-ipsec configuration, or upon admin shutdown of an mc-ipsec enabled tunnel group):
  - → All tunnels in the tunnel-group will be deleted/reinstalled to the MS-ISAs.
  - → All IKE states associated with those tunnels are locally purged from the MS-ISAs.
  - → No IKE messages are sent to the IKE peer.

  These behaviors occur regardless of the presence of a redundant chassis or the state of a redundant chassis.

- With MC-IPSec enabled:
  - → auto-establish is blocked.
  - → For DPD configuration, only **no dpd** and **dpd** configurations with **reply-only** are allowed.

Routing

Routing in Public Service

A /32 route of the local tunnel address is created automatically for all tunnels on the MC-IPSec enabled tunnel-group.

This /32 route can be exported to a routing protocol by a route policy. The protocol type in route-policy is IPSec.

To attract traffic to the master chassis, a route metric of these /32 routes could be set according to the MIMP state, a metric from the master chassis is better than a metric from the standby chassis. There are three available states that can be used in the **from state** command in the route policy entry configuration:

- IPSec-master-with-peer
  → Corresponding MIMP states: master
- IPSec-master-without-peer
  → Corresponding MIMP states: eligible
- IPSec-non-master
  → Corresponding MIMP states: discovery/standby

However, if the standby chassis receives IPSec traffic, the traffic will be shunt to the master chassis by forwarding to a redundant next-hop. The redundant next-hop is an IP next-hop in the public routing instance.

Routing in Private Services

For static LAN-to-LAN tunnels, the static route with the IPSec tunnel as the next-hop could be exported to a routing protocol by a route policy. The protocol type remains **static**. For dynamic LAN-to-LAN tunnels, the reverse-route could be exported to a routing protocol by a route policy. The protocol type is **ipsec**

Similar to routing in public services, the route metric of the above the routes could be set according to the MIMP state. Only a static route with an IPSec tunnel as the nexthop and reverse route has an MIMP state.

If the standby chassis receives IPSec traffic, the traffic will be shunt to the master chassis by forwarding to a redundant next-hop. The redundant next-hop is an IP next-hop in a private routing instance.

Other Details About Shunting

Shunting only works when tunnel-group is operationally up.

Shunting is not supported over auto-bind tunnels.

MC-IPSec Aware VRRP

In many cases, the public side is a Layer 2 network and VRRP is used to provide link or node protection. However, VRRP and MC-IPSec are two independent processes, each has its own mastership state, which means the VRRP master could be different from MC-IPSec master. This will result unnecessary shunting traffic.

To address this issue, MC-IPSec aware VRRP is introduced in SR-OS Release 10.0R8, which add a new priority event in vrrp-policy: mc-ipsec-non-forwarding. If the configured tunnel-group enters non-forwarding (non-master) state, then the priority of associated VRRP instance will be set to the configured value. Delta priority is not supported for this type of event.

Synchronization

In order to achieve stateful failover, IPSec states are synced between chassis by using the MCS protocol.

- Only successfully created SA after a completed INITIAL EXCHANGES or CREATE_CILD_SA EXECHANGES is synced.
- Upon switchover, the new standby chassis will reboot the tunnel-group.
- The ESP sequence number is not synced.
- The CLI configuration is not synced.

The time must be the same on both chassis (using NTP/SNTP to sync to the same server is an option).

Automatic CHILD_SA Rekey

Because the ESP sequence number is not synced, a CHILD_SA rekey for each tunnel will be initiated by the new master to reset the sequence number upon switchover.

Responder Only

With MC-IPSec, it is preferable that only the 7750 SR acts as the IKEv2 responder (except for the automatic CHILD_SA rekey upon switchover). This is naturally the case for dynamic tunnels; however for static tunnels, there is a CLI command in tunnel-group to enable this behavior.

```
config>isa>tunnel-grp# ?
     [no] ipsec-responder-only
```

Best Practice Suggestions

Following is a list of best practice suggestions. They are not mandatory, but recommended.

Overall IPSec

To avoid high CPU loads and some complex cases, the following are suggestions to configure IKEv2 lifetime:

- Both IKE_SA and CHILD_SA lifetime on one side should be approximately 2 or 3 times larger than the other side.
- With the previous rule, the lifetime of the side with smaller lifetime should NOT be too small:
  → IKE_SA: >= 86400 seconds
  → CHILD_SA: >= 3600 seconds
- With 1st rule, on the side with smaller lifetime, the IKE_SA lifetime should be at least 3 times larger than CHILD_SA lifetime.

The IKE protocol is the control plane of IPSec, thus, the IKE packet should be treated as high QoS priority in the end-to-end path of public service.

- On a public interface, a sap-ingress QoS policy should be configured to ensure the IKE packet gets high QoS priority.

MC-IPSec Specific

MC-IPSec pair IKEv2 lifetime should be higher than the peer according to the suggestions in Overall IPSec on page 340.

Responder-only configuration for static tunnel if possible (meaning if the peer supports auto-establish)

DPD on peer side, **no dpd** on the 7750 SR side

Dedicate, direct physical link between chassis with enough bandwidth for MCS and shunting traffic, and proper QoS configuration to make sure the MIMP/MCS packet get the highest priority

Check and make sure the protection status is **nominal** on both chassis before a manual switchover is performed.

Wait at least five minutes between two consecutive switchovers, if possible.

In case of VRRP in public service, the VRRP/Layer 2 network should re-converge before the MC-IPSec switchover in case of chassis failure. One way to speed up VRRP switchover is to bind a BFD session to VRRP.