Transport Working Group
Internet-Draft
Expires: August 14, 2005

F. Baker
Cisco Systems
P. Bose
Lockheed Martin

February 10, 2005

QoS Signaling in a Nested Virtual Private Network
draft-baker-tsvwg-vpn-signaled-preemption-02

Status of this Memo

Copyright Notice

Abstract

Some networks require communication between an interior and exterior portion of a VPN, but have sensitivities about what information is communicated across the boundary. This note seeks to outline the issues and the nature of the proposed solutions.

Table of Contents

1. QoS in a nested VPN

More and more networks wish to guarantee secure transmission of IP traffic for across public LANs or WANs and therefore use Virtual Private Networks. Some networks require communication between an interior and exterior portion of a VPN, but have sensitivities about what information is communicated across the boundary. This note seeks to outline the issues and the nature of the proposed solutions. The outline of the QoS solution for real-time traffic has been described at a high level in [I-D.baker-tsvwg-mlpp-that-works] . The key characteristics of this proposal are that

- it uses standardized protocols,
- It includes reservation setup and teardown for guaranteed and controlled load services using the standardized protocols,
- it is independent of link delay, and therefore consistent with high delay*bandwidth networks as well as the more common variety,
- it has no single point of failure, such as a central reservation manager,
- It provides for the preemption of established data flows,
- In that preemption, it not only permits a policy-admitted data flow in, but selects a specific data flow to exclude based upon control input rather than simply accepting a loss of service dictated at the discretion of the network control function, and
- interoperates directly with SIP Proxies, H.323 Gatekeepers, or other call management subsystems to present the other services required in a preemptive or preferential telephone network.

The thrust of the memo surrounds VPNs that use encryption in some form, such as IPsec. As a result, we will discuss the VPN Router supporting "plaintext" and "ciphertext" interfaces. However, the concept extends readily to any form of aggregation, including the concept proposed in [RFC3175] of the IP traffic entering and leaving a network at identified points, and the use of other kinds of tunnels including GRE, IP/IP, MPLS, and so on.

A note on the use of the words "priority" and "precedence" in this document is in order. The term "priority" has been used in this context with a variety of meanings, resulting in a great deal of confusion. The term "priority" is used in this document to identify one of several possible datagram scheduling algorithms. A scheduler is used when deciding which datagram will be sent next on a computer

interface; a priority scheduler always chooses a datagram from the highest priority class (queue) that is occupied, shielding one class of traffic from jitter by passing jitter it would otherwise have experienced to another class. [RFC3181] applies the term to a reservation, in a sense that this document will refer to as "precedence". The term "precedence" is used in the sense implied in the phrase "Multi-Level Precedence and Preemption"[ITU.MLPP.1990] ; some classes of sessions take precedence over others, which may result in bandwidth being admitted that might not otherwise have been or may result in the prejudicial termination of a lower precedence session under a stated set of circumstances. For the purposes of the present discussion, "priority" is a set of algorithms applied to datagrams, where "precedence" is a policy attribute of sessions. The techniques of priority comparisons are used in a router or a policy decision point to implement precedence, but they are not the same thing.

Along the same lines, it is important for the reader to understand the difference between QoS policies and policies based on the "precedence" or "importance" of data to the person or function using it. Voice, regardless of the precedence level of the call, is impeded by high levels of variation in network-induced delay. As a result, voice is often serviced using a priority queue, transferring jitter from that application's traffic to other applications. This is as true of voice for routine calls as it is for flash traffic. There are classes of application traffic that require bounded delay. That is a different concept than "no jitter"; they can accept jitter within stated bounds. Routing protocols such as OSPF or BGP are critical to the correct functioning of network infrastructure. While they are designed to work well with moderate loss levels, they are not helped by them, and even a short period of high loss can result in dramatic network events. Variation in delay, however, is not at all an issue if it is within reasonable bounds. As a result, it is common for routers to treat routing protocol datagrams in a way that limits the probability of loss, accepting relatively high delay in some cases, even though the traffic is absolutely critical to the network. Telephone call setup exchanges have this characteristic as well: faced with a choice between loss and delay, protocols like SIP and H.323 far prefer the latter, as the call setup time is far less than it would be if datagrams had to be retransmitted, and this is true regardless of whether the call is routine or of high precedence. As such, QoS markings tell us how to provide good service to an application independent of how "important" it may be at the current time, while "importance" can be conveyed separately in many cases.

1.1 Nested VPNs

One could describe such a network in terms of three network diagrams.

Figure 1 shows a simple network stretched across a VPN connection. The VPN router (where, following [RFC2460] a "router" is "a node that forwards packets not explicitly addressed to itself"), performs the following steps: it

- receives an IP datagram from a plain text interface,
- determines what remote enclave and therefore other VPN router to forward it to,
- ensures that it has a security association with that router,
- encloses the encrypted datagram within another VPN (e.g. IPsec) and IP header, and
- forwards the encapsulated datagram toward the remote VPN router.

The receiving VPN router reverses the steps: it

- determines what security association the datagram was received from,
- decrypts the interior datagram,
- forwards the now-decapsulated datagram on a plain text interface.

The use of IPsec in this manner is described as the tunnel mode of [RFC2401] and [RFC2406] .

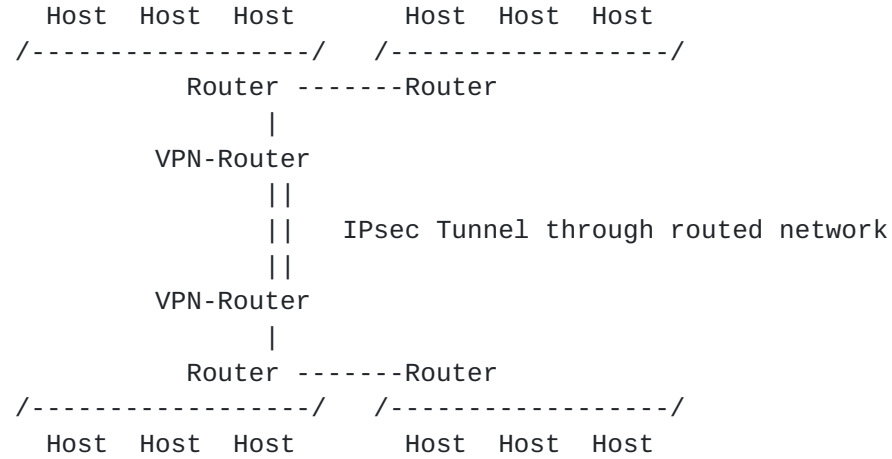

Figure 1: VPN-connected enclave

An important point to understand is that the VPN tunnel, like other features of the routed network, are invisible to the host. The host can infer that "something out there" is affecting the Path MTU, introducing delay, or otherwise affecting its data stream, but if

properly implemented it should be able to adapt to these. The words "if properly implemented" are the bane of every network manager, however; substandard implementations do demonstrably exist.

Outside of the enclave, the hosts are essentially invisible. The communicating enclaves look like a simple data exchange between peer hosts across a routed network, as shown in Figure 2 .
VPN-Router | Router | VPN-Router

Figure 2: VPN-connected enclave from exterior perspective

Such networks can be nested and re-used in a complex manner. As shown in Figure 3 a pair of enclaves might communicate across a ciphertext network which, for various reasons, is itself re-encrypted and transmitted across a larger ciphertext network. The reasons for doing this vary, but they relate to information-hiding in the wider network, different levels of security required for different enclosed enclaves, and so on.

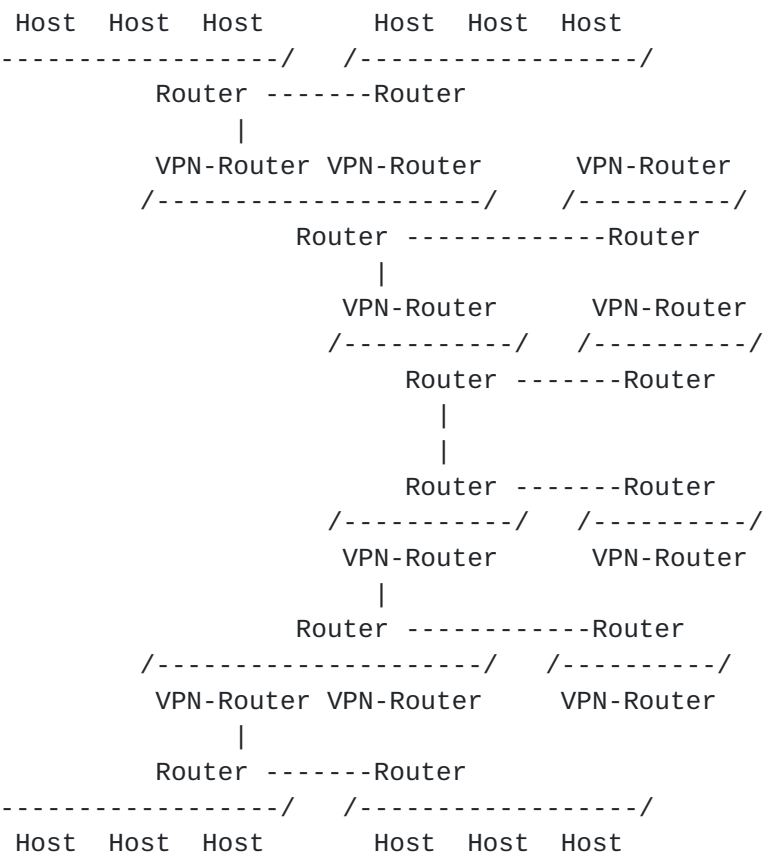

Figure 3: Nested VPN

The key question this document explores is "how do reservations, and preemption of reservations, work in such an environment?"

1.2 Signaled QoS technology

The Integrated Services model for networking was originally proposed in [RFC1633] . In short, it divides all applications into two broad classes: those that will adapt themselves to any available bandwidth, and those that will not or cannot. In its own words,

> One class of applications needs the data in each packet by a certain time and, if the data has not arrived by then, the data is essentially worthless; we call these "real-time" applications. Another class of applications will always wait for data to arrive; we call these "elastic" applications.

The Integrated Services model defines data flows supporting applications as either "real-time" or "elastic". It should be noted that "real-time" traffic is also referred to as "inelastic" traffic, and that elastic traffic is occasionally referred to as "non-real-time."

In this view, the key issue is the so-called "playback point": a real-time application is considered to have a certain point in time at which data describing the next sound, picture, or whatever to be delivered to a display device or forfeit its utility, while an elastic application has no such boundary. Another way to look at the difference is that real-time applications have an irreducible lower bound on their bandwidth requirements. For example, the typical G.711 payload is delivered in 160 byte samples (plus 40 bytes of IP/UDP/RTP headers) at 20 millisecond intervals. This will yield 80 KBPS of bandwidth, without silence suppression, and not accounting for the layer 2 overhead. To operate in real-time, a G.711 codec requires the network over which its data will be delivered to support communications at 80 KBPS at the IP layer with roughly constant end to end delay and nominal or no loss. If this is not possible (if there is significant loss or wide variations in delay), voice quality will suffer. On the other hand, if many megabits of capacity are available, the G.711 codec will not increase its bandwidth requirements either. Although adaptive codecs exist, (e.g., G.722.2 or G.726), the adaptive mechanism can either require greater or lesser bandwidth and can adapt only within a certain range of bandwidth requirements beyond which the quality of the data flow required is not met. Elastic applications, however, will generally adapt themselves to any network: if the bottleneck provides 9600 bits per second, a web transfer or electronic mail exchange will happen at 9600 bits per second, and if hundreds of megabits are available, the TCP (or SCTP) transport will increase their transfer rate in an attempt to reduce the time required to accomplish the transfer.

For real-time applications, those that require data to be delivered end to end with at least a certain rate and with delays varying between stated bounds, the Integrated Services architecture proposes the use of a signaling protocol that allows the communicating applications and the network to communicate about the application requirements and the network's capability to deliver them. Several such protocols have been developed or are under development, notably including RSVP and NSIS. The present discussion is limited to RSVP, although any protocol that delivers a similar set of capabilities could be considered.

1.3 The Resource Reservation Protocol (RSVP)

RSVP is initially defined in [RFC2205] with a set of datagram processing rules defined in [RFC2209] and datagram details for Integrated Services [RFC2210] . Conceptually, this protocol specifies a way to identify data flows from a source application to a destination application and request specific resources for them. The source may be a single machine or a set of machines listed explicitly or implied, whereas the destination may be a single machine or a

multicast group (and therefore all of the machines in it). Each application is specified by a transport protocol number in the IP protocol field, or may additionally include destination and perhaps source port numbers. The protocol is defined for both IPv4 [RFC0791] and IPv6 [RFC2460]. It was recognized immediately that it was also necessary to provide a means to perform the same function for various kinds of tunnels, which implies a relationship between what is inside and what is outside the tunnel. Definitions were therefore developed for IPsec [RFC2207] and [I-D.lefaucheur-rsvp-ipsec] and for more generic forms of tunnels [RFC2746]. With the later development of the Differentiated Services Architecture [RFC2475], definitions were added to specify the DSCP [RFC2474]to be used by a standard RSVP data flow in [RFC2996] and to use a pair of IP addresses and a DSCP as the identifying information for a data flow [RFC3175] .

In addition, the initial definition of the protocol included a placeholder for policy information, and for preemption of reservations. This placeholder was later specified in detail in [RFC2750] with a view to associating a policy [RFC2872] with an identity [RFC3182] and thereby enabling the network to provide a contracted service to an authenticated and authorized user. This was integrated with the Session Initiation Protocol [RFC3261] in [RFC3312] . Preemption of a reservation is specified in the context, in [RFC3181] which in essence specifies that a reservation installed in the network using an Preemption Priority and retained using a separate Defending Priority may be removed by the network via an RESV Error signal that removes the entire reservation. This has issues, however, in that the matter is often not quite so black and white. If the issue is that an existing reservation for 80 KBPS can no longer be sustained but a 60 KBPS reservation could, it is possible that a VoIP sender could change from a G.711 codec to a G.729 codec and achieve that. Or, if there are multiple sessions in a tunnel or other aggregate, one of the calls could be eliminated leaving capacity for the others. [I-D.polk-tsvwg-rsvp-bw-reduction] seeks to address this issue.

In a similar way, a capability was added to limit the possibility of control signals being spoofed or otherwise attacked [RFC2747][RFC3097] .

[RFC3175] describes several features that are unusual in RSVP, being specifically set up to handle aggregates in a service provider network. It describes three key components:

- The RFC 3175 session object, which identifies not the IP addresses of the packets that are identified, but the IP addresses of the ingress and egress devices in the network, and the DSCP that the traffic will use,

- The function of a reservation "aggregator", which operates in the ingress router and accepts individual reservations from the "customer" network which it aggregates into the ISP core in a tunnel, an MPLS LSP, or as a traffic stream that it known to leave at the deaggregator,

- The function of a reservation "deaggregator", which operates in the egress router and breaks the aggregate reservation and data streams back out into individual data streams that may be passed to other networks.

In retrospect, the Session Object specified by RFC 3175 is useful but not intrinsically necessary. If the ISP network uses tunnels, such as MPLS LSPs, IP/IP or GRE tunnels or enclosing IPsec Security Associations, the concepts of an aggregator and a deaggregator work in the same manner, although the reservation mechanism would be that of [RFC3473] and [RFC3474][RFC2207][I-D.lefaucheur-rsvp-ipsec] or [RFC2746] .

1.4 Logical structure of a VPN Router

The conceptual structure of a VPN Router is similar to that of any other router. In its simplest form, it is physically a two or more port device, similar to that shown in Figure 4 which has one or more interfaces to the protected enclave(s) and one or more interfaces to the outside world. On the latter, it structures some number of tunnels (in the case of an IPsec tunnel, having security associations) that it can treat as point to point interfaces from a routing perspective.

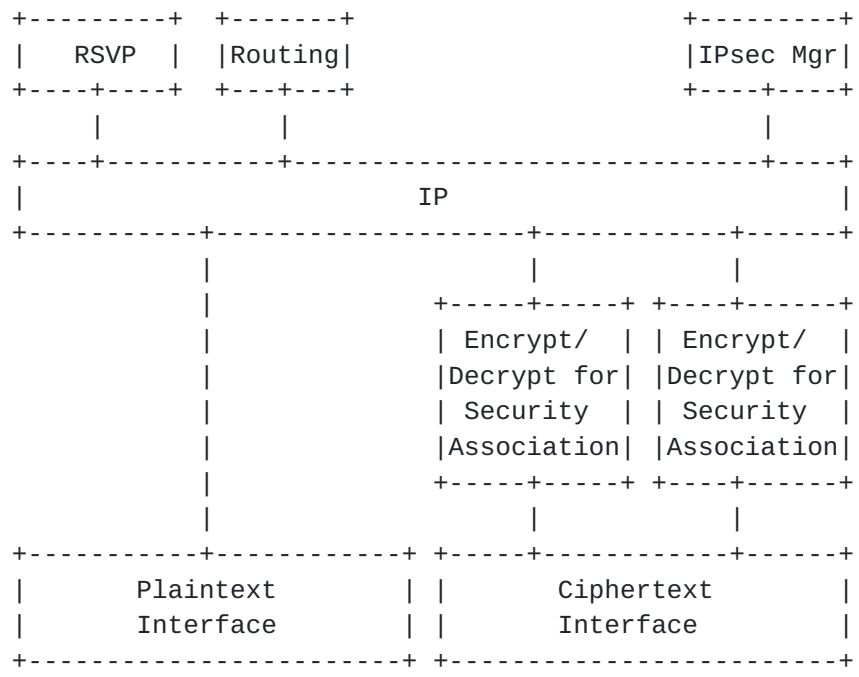

Figure 4: Logical structure of a VPN Router

In some environments, VPN Routers are used that are constructed of two half-routers with a private interface between them. This is show in Figure 5 . In such a design, one half-router is entirely within the enclave and one half-router is entirely within the VPN domain. They maintain separate routing tables for their various parts of the network: the enclave half-router knows of other enclave half-routers and the prefixes they offer, and the public half-router knows of other public half-routers. There is a private interface between them on which a very few datagrams are permitted to pass; these include

- IP datagrams,
- RSVP signaling between the half-routers,
- Control information coordinating security associations, and
- precious little else.

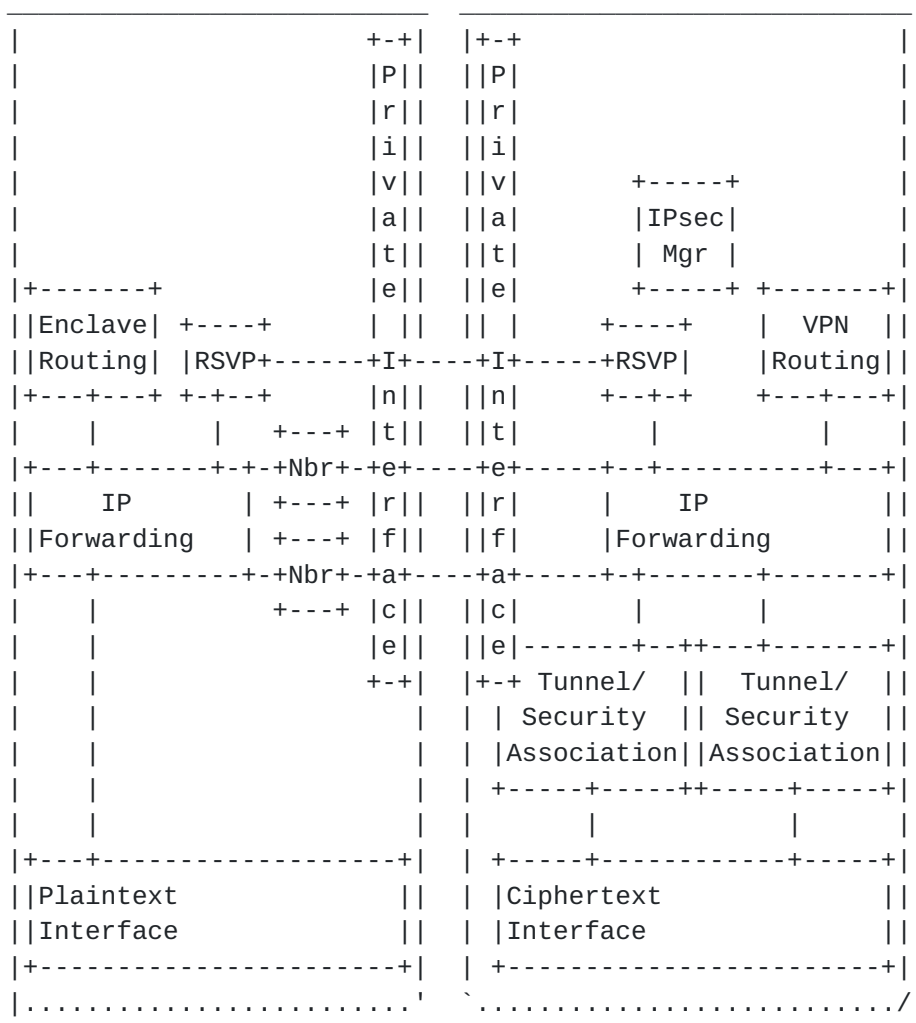

Figure 5: VPN Router shown as two half-routers

2. Reservation and Preemption in a nested VPN

```
                     /                           \
                    (       +--+   +--+   enclave )   ,---------.
      .----------.   \      |H2+---+R2|          / ,-'           `
       +--+   +--+`--.\     +--+   ++-+         / /   +--+   +--+
       |H1+---+R1|    \`.           |         ,' /    |R3+---+H3|
       +--+   +-++     ) '--.    +----++  _.-'  (     ++-+   +--+
                |     /    _.`---|VPN2||''-.     \     |
      enclave +----+-i.--''      +----++    `----.\ +----+ enclave
      --------|VPN1|'              |              ``|VPN3|       ,
             ,+----+               |                +----+-----'
           ,' --+-------+----------+------------------+---`.
         ,'            ++-+                                 `.
       ,'              |R7+--------+                          `.
      / interface      +--+        |                            \
        domain 1                 +-+--+                          \
                       _.--------|VPN7|--------.
               ,-----''          +--+-+         `------.         .
      `-.   ,-'                    |                   `-.   .-'
         `-:  inner domain        +-++                     `.'
         (                        |R9|                       )
         .'.                      ++-+                     ;-.
       .'   `-.                    |                   ,-'   `-.
      '        `------.          +-+--+         _.-----'         `
        interface      `---------|VPN8|-------''
        domain 2                 +-+--+                          /
      \                            |          +--+              /
       `.                          +----------+R8|            ,'
         `.                                   ++-+          ,'
           `. --+------------------+-----------+------+-- ,'
        ,-----+----+               |                +----+-----.
      ,'      |VPN6|.              |              _.|VPN4|       `
              +----+.`----.      +----+     _.--'' ,+----+
               |     \     `--=.-|VPN5|---:'      /    |
       +--+   ++-+    :   ,-''   +----+    `--.  ;    ++-+   +--+
       |H6+---+R6|    | ,'          |          `.|    |R4+---+H4|
       +--+   +--+    ;/    +--+   ++-+          :    +--+   +--+
                     //     |H5+---+R5|           \
       enclave     ,'(      +--+   +--+            `.     enclave
      `.         ,'   \                 enclave   /  '-.         ,
        `-------'      \                         /      `-------'
```

Figure 6: Reservations in a nested VPN

Let us discuss how a resource reservation protocol, and specifically RSVP, might be used in a nested virtual private network.

2.1 Reservation in a nested VPN

A reservation in a nest VPN is very much like a reservation in any other network, with one exception: it is composed of multiple reservations that must be coordinated. These include a reservation within the originating and receiving enclaves and a reservation at each layer of the VPN, as shown in Figure 6 .

Thus, when a host in one enclave opens a reservation to a host in another enclave, a reservation of the appropriate type and size is created end to end. As it traverses the VPN Router leaving its enclave, the reservation information and the data are placed within the appropriate tunnel (e. g., the IPsec Security Association for its precedence level to the appropriate remote VPN Router). At the remote VPN Router, it is extracted from the tunnel and passed on its way to the target system. The data in the enclave will be marked with a DSCP appropriate to its application and (if there is a difference) precedence level, and the signaling datagrams (PATH and RESV) are marked with a DCLASS object indicating that value. RSVP signaling datagrams (PATH and RESV) are marked with a DCLASS object indicating the value used fo the corresponding data. The DSCP on the signaling datagrams, however, is a DSCP for signaling, and has the one proviso that if routing varies by DSCP then it must be a DSCP that is routed the same way as the relevant data. The [RFC2872] policy object specifies the applicable policy (e. g., "routine service for voice traffic") and asserts a [RFC3182] credential indicating its user (which may be a person or a class of persons). As specified in [RFC3181] it also specifies its Preemption Priority and its Defending Priority; these enable the Preemption Priority of a new session to be compared with the Defending Priority of previously admitted sessions.

On the ciphertext side of the VPN Router, no guarantees result unless the VPN Router likewise sets up a reservation to the peer VPN Router across the ciphertext domain. Thus, the VPN Router sets up an [RFC2207][I-D.lefaucheur-rsvp-ipsec] or [RFC3175] reservation to its peer. In the [RFC2207] or [I-D.lefaucheur-rsvp-ipsec] cases (which are essentially equivalent except that the latter clarifies the use with a DSCP), the same DSCP may be used for all traffic in the class ([RFC3246] 's EF might be used for voice regardless of precedence level, for example), but the session reservations fit in different security associations with different policy objects by precedence level. In the [RFC3175] case, since the DSCP must be used to identify both the reservation and the corresponding data stream, the aggregate reservations for different precedence levels require different DSCP values.

As such, the fundamental necessity is for one VPN Router to act as

what [RFC3175] calls the "aggregator" and another to act as the "deaggregator", and extend a VPN tunnel between them. If the VPN Tunnel is an IPsec Security Association between the VPN routers and the IP packet is entirely contained within (such as is used to cross a firewall), then the behavior of [RFC2746] is required of the tunnel. That bearer will have the following characteristics:

- it will have a DSCP corollary or the same as the DSCP for the data it carries,
- the reservations and data will be carried in security associations between the VPN Routers, and
- the specification for the reservation for the tunnel itself will not be less than the sum of the requirements of the aggregated reservations.

The following requirements relationships apply between the set of enclosed reservations and the tunnel reservation:

- The sum of the average rates of the contained reservations, having been adjusted for the additional IP headers, will be less than or equal to the average rate of the tunnel reservation.
- The sum of the peak rates of the contained reservations, having been adjusted for the additional IP headers, will be less than or equal to the peak rate of the tunnel reservation.
- The sum of the burst sizes of the contained reservations, having been adjusted for the additional IP headers, will be less than or equal to the burst size of the tunnel reservation.
- The Preemption Priority of a tunnel reservation is identical to that of the individual reservations it aggregates.
- The Defending Priority of a tunnel reservation is identical to that of the individual reservations it aggregates.

This would differ only in the case that measurement-based admission is in use in the tunnel but not in the end system. In that case, the tunnel's average bandwidth specification would be greater than or equal to the actual average offered traffic. Such systems are beyond the scope of this specification.

As a policy matter, it may be useful to note a quirk in the way Internet QoS works. If the policies for various precedence levels specify different thresholds (e. g., "to accept a new routine call, the total reserved bandwidth after admission may not exceed X; to

accept a higher precedence level call, the total reserved bandwidth after admission may not exceed X+Y, and this may be achieved by preempting a lower precedence level call"), the bandwidth Y effectively comes from the bandwidth in use by elastic traffic rather than forcing a preemption event.

2.2 Preemption in a nested VPN

As discussed in Section 1.3 preemption is specified in [RFC3181] and further addressed in [I-D.polk-tsvwg-rsvp-bw-reduction] . The issue is that in many cases the need is to reduce the bandwidth of a reservation due to a change in the network, not simply to remove the reservation. In the case of an end system originated reservation, the end system might be able to accommodate the need through a change of codec; in the case of an aggregate of some kind, it could reduce the bandwidth it is sending by dropping one or more reservations entirely.

In a nested VPN or other kind of aggregated reservation, this means that the deaggregator (the VPN Router initiating the RESV signal for the tunnel) must

- receive the RESV Error signaling it to reduce its bandwidth,
- re-issue its RESV accordingly,
- identify one or more of its aggregated reservations, enough to do the job, and
- signal them to reduce their bandwidth accordingly.

It is possible, of course, that it is signaling them to reduce their bandwidth to zero, which is functionally equivalent to removing the reservation as described in [RFC3181] .

In the routers in the core, an additional case arises. One could imagine that some enclave presents the VPN with a single session, and that session has a higher precedence level. If some interior link is congested (e. g., the reserved bandwidth will exceed policy if the call is admitted), a session between a different pair of VPN Routers must be preempted. More generally, in selecting a reservation to preempt, the core router must always select a reservation at the lowest available Defending Priority. This is the reason that various precedence levels must be kept separate in the core.

2.3 Working through an example

The network in Figure 6 shows three security layers: six plaintext

enclaves around the periphery, two ciphertext domains connecting them at one layer (referred to in the diagram as an "interface domain"), and a third ciphertext domain connecting the first two (referred to in the diagram as an "inner domain"). The following distribution of information exists:

- Each enclave has access to general routing information concerning other enclaves it is authorized to communicate with: systems in it can translate a DNS name for a remote host or domain and obtain the corresponding address or prefix.

- Each enclave router also has specific routing information regarding its own enclave.

- A default route is distributed within the enclave, pointing to its VPN Router.

- VPN Routers 1-6 are able to translate remote enclave prefixes to the appropriate remote enclave's VPN Router addresses.

- Each interface domain has access to general routing information concerning the other interface domains, but not the enclaves. Systems in an interface domain can translate a DNS name for a remote interface domain and obtain the corresponding address or prefix.

- Each interface domain router also has specific routing information regarding its own interface domain.

- A default route is distributed within the interface domain, pointing to the "inner" VPN Router.

- VPN Routers 7 and 8 are able to translate remote interface domain prefixes to remote VPN Router addresses.

- Routers in the inner domain have routing information for that domain only.

While the example shows three levels, there is nothing magic about the number three. The model can be extended to any number of concentric layers.

Note that this example places unidirectional reservations in the forward direction. In voice and video applications, one generally has a reservation in each direction. The reverse direction is not discussed, for the sake of clarity, but operates in the same way in the reverse direction and uses the same security associations.

2.3.1 Initial routine reservations - generating network state

Now let us install a set of reservations from H1 to H4, H2 to H5, and H3 to H6, and for the sake of argument let us presume that these are at the "routine" precedence. H1, H2, and H3 each initiate an PATH signal describing their traffic. For the sake of argument, let us presume that H1's reservation is for an [RFC2205] session, H2's reservation is for a session encrypted using IPsec, and therefore depends on [RFC2207] and H3 (which is a PSTN Gateway) sends an [RFC3175] reservation comprising a number of distinct sessions. Since these are going to H4, H5, and H6 respectively, the default route leads them to VPN1, VPN2, and VPN3 respectively.

The VPN Routers each ensure that they have an appropriate security association or tunnel open to the indicated remote VPN Router (VPN4, VPN5, or VPN6). This will be a security association or tunnel for the indicated application at the indicated precedence level. Having accomplished that, it will place the PATH signal into the security association and forward it. If such does not already exist, following [RFC3175] 's aggregation model, it will now open a reservation (send a PATH signal) for the tunnel/SA within the interface domain; if the reservation does exist, the VPN Router will increase the bandwidth indicated in the ADSPEC appropriately. In this example, these tunnel/SA reservations will follow the default route to VPN7.

VPN7 ensures that it has an appropriate security association or tunnel open to VPN8. This will be a security association or tunnel for the indicated application at the indicated precedence level. Having accomplished that, it will place the PATH signal into the security association and forward it. If such does not already exist, following [RFC3175] 's aggregation model, it will now open a reservation (send a PATH signal) for the tunnel/SA within the interface domain; if the reservation does exist, the VPN Router will increase the bandwidth indicated in the ADSPEC appropriately. In this example, this tunnel/SA reservation is forwarded to VPN8.

VPN8 acts as an [RFC3175] deaggregator for the inner domain. This means that it receives the PATH signal for the inner domain reservation and stores state, decrypts the data stream from VPN7, operates on the RSVP signals as an RSVP-configured router, and forwards the received IP datagrams (including the updated PATH signals) into its interface domain. The PATH signals originated by VPN1, VPN2, and VPN3 are therefore forwarded towards VPN4, VPN5, and VPN6 according to the routing of the interface domain.

VPN4, VPN5, and VPN6 each act as an [RFC3175] deaggregator for the interface domain. This means that it receives the PATH signal for

the interface domain reservation and stores state, decrypts the data stream from its peer, operates on the RSVP signals as an RSVP-configured router, and forwards the received IP datagrams (including the updated PATH signals) into its enclave. The PATH signals originated by H1, H2, and H3 are therefore forwarded towards H4, H5, and H6 according to the routing of the enclave.

H4, H5, and H6 now receive the original PATH signals and deliver them to their application.

2.3.2 Initial routine reservations - request reservation

The application in H4, H5, and H6 decides to install the indicated reservations, meaning that they now reply with RESV signals. These signals request the bandwidth reservation. Following the trail left by the PATH signals, the RESV signals traipse back to their respective sources. The state left by the PATH signals leads them to VPN4, VPN5, and VPN6 respectively. If the routers in the enclaves are configured for RSVP, this will be explicitly via R4, R5, or R6; if they are not, routing will lead them through those routers.

The various RSVP-configured routers en route in the enclave (including the VPN Router on the "enclave" side) will verify that there is sufficient bandwidth on their links and that any other stated policy is also met. Having accomplished that, each will update its reservation state and forward the RESV signal to the next. The VPN Routers will also each generate an RESV for the reservation within the interface domain, attempting to set or increase the bandwidth of the reservation appropriately.

The various RSVP-configured routers en route in the interface domain (including VPN8) will verify that there is sufficient bandwidth on their links and that any other stated policy is also met. Having accomplished that, each will update its reservation state and forward the RESV signal to the next. VPN8 will also generate an RESV for the reservation within the inner domain, attempting to set or increase the bandwidth of the reservation appropriately. This gets the reservation to the inner deaggregator, VPN8.

The various RSVP-configured routers en route in the inner domain (including VPN7) will verify that there is sufficient bandwidth on their links and that any other stated policy is also met. Having accomplished that, each will update its reservation state and forward the RESV signal to the next. This gets the signal to VPN7.

VPN7 acts as an [RFC3175] aggregator for the inner domain. This means that it receives the RESV signal for the inner domain reservation and stores state, decrypts the data stream from VPN8,

operates on the RSVP signals as an RSVP-configured router, and forwards the received IP datagrams (including the updated RESV signals) into its interface domain. The RESV signals originated by VPN4, VPN5, and VPN6 are therefore forwarded towards VPN1, VPN2, and VPN3 through the interface domain.

VPN1, VPN2, and VPN3 each act as an [RFC3175] aggregator for the interface domain. This means that it receives the RESV signal for the interface domain reservation and stores state, decrypts the data stream from its peer, operates on the RSVP signals as an RSVP-configured router, and forwards the received IP datagrams (including the updated RESV signals) into its enclave. The RESV signals originated by H4, H5, and H6 are therefore forwarded towards H1, H2, and H3 according to the routing of the enclave.

H1, H2, and H3 now receive the original RESV signals and deliver them to their application.

2.3.3 Installation of a reservation using precedence

Without going through the details called out in Section 2.3.1 and Section 2.3.2 if sufficient bandwidth exists to support them, reservations of other precedence levels or other applications may also be installed across this network. If the "routine" reservations already described are for voice, for example, and sufficient bandwidth is available under the relevant policy, a reservation for voice at the "priority" precedence level might be installed. Due to the mechanics of preemption, however, this would not expand the existing "routine" reservations in the interface and inner domains, as doing this causes loss of information - how much of the reservation is now "routine" and how much is "priority"? Rather, this new reservation will open up a separate set of tunnels or security associations for traffic of its application class at its precedence between that aggregator and deaggregator.

As a side note, there is an opportunity here that does not exist in the PSTN. In the PSTN, all circuits are potentially usable by any PSTN application under a certain set of rules (H channels, such as are used by video streams, must be contiguous and ordered). As such, if a channel is not made available to routine traffic but is made available to priority traffic, the operator is potentially losing revenue on the reserved bandwidth and deserves remuneration. However, in the IP Internet, some bandwidth must be kept for basic functions such as routing, and in general policies will not permit 100% of the bandwidth on an interface to be allocated to one application at one precedence. As a result, it may be acceptable to permit a certain portion (e. g. 50%) to be used by routine voice and a larger amount (e. g. 60%) to be used by voice at a higher

precedence level. Under such a policy, a higher precedence reservation for voice might not result in the preemption of a routine call, but rather impact elastic traffic, and might be accepted at a time that a new reservation of lower precedence might be denied.

In microwave networks, such as satellite or mobile ad hoc, one could also imagine network management intervention that could change the characteristics of the radio signal to increase the bandwidth under some appropriate policy.

2.3.4 Installation of a reservation using preemption

So we now have a number of reservations across the network described in Figure 6 including several reservations at "routine" precedence and one at "priority" precedence. For sake of argument, let us presume that the link from VPN7 to R9 is now fully utilized - all of the bandwidth allocated by policy to voice at the routine or priority level has been reserved. Let us further imagine that a new "priority" reservation is now placed from H3 to H6.

The process described in Section 2.3.1 is followed, resulting in PATH state across the network for the new reservation. This is installed even though it is not possible to install a new reservation on VPN7-R9, as it does not install any reservation and the network does not know whether H6 will ultimately require a reservation.

The process described in Section 2.3.2 is also followed. The application in H6 decides to install the indicated reservation, meaning that it now replies with an RESV signal. Following the trail left by the PATH signal, the RESV signal traipses back towards H3. VPN6 and (if RSVP was configured) R6 verify that there is sufficient bandwidth on their links and that any other stated policy is also met. Having accomplished that, each will update its reservation state and forward the RESV signal to the next. VPN6 also generates an RESV for the reservation within the interface domain, attempting to set or increase the bandwidth of the reservation appropriately.

VPN6, R8, and VPN8's "interface domain" side now verify that there is sufficient bandwidth on their links and that any other stated policy is also met. Having accomplished that, each will update its reservation state and forward the RESV signal to the next. VPN8 will also generate an RESV for the reservation within the inner domain, attempting to set or increase the bandwidth of the reservation appropriately. This gets the reservation to the inner deaggregator, VPN8.

VPN8's "inner domain" side and R9 now verify that there is sufficient bandwidth on their links and that any other stated policy is also

met. At R9, a problem is detected - there is not sufficient bandwidth under the relevant policy. In the absence of precedence, R9 would now return an RESV Error indicating that the reservation could not be increased or installed. In such a case, if a pre-existing reservation of lower bandwidth already existed, the previous reservation would remain in place but the new bandwidth would not be granted, and the originator (H6) would be informed. Let us clarify what it means to be at a stated precedence: it means that the POLICY_DATA object in the RESV contains a Preemption Priority and a Defending Priority with values specified in some memo. With precedence, [I-D.polk-tsvwg-rsvp-bw-reduction] 's algorithm would have the Preemption Priority of the new reservation compared to the Defending Priority of extant reservations in the router, of which there are two: one VPN7->VPN8 at "routine" precedence and one VPN7->VPN8 at "priority" precedence. Since the Defending Priority of routine reservation is less than the Preemption Priority of a "priority" reservation, the "routine" reservation is selected. R9 determines that it will accept the increase in its "priority" reservation VPN7->VPN8 and reduce the corresponding "routine" reservation. Two processes now occur in parallel:

- The routine reservation is reduced following the algorithms in [I-D.polk-tsvwg-rsvp-bw-reduction] and

- The priority reservation continues according to the usual rules.

R9 reduces its "routine" reservation by sending an RESV Error updating its internal state to reflect the reduced reservation and sending an RESV Error to VPN8 requesting that it reduce its reservation to a number less than or equal to the relevant threshold less the sum of the competing reservations. VPN8, acting as a de-aggregator, makes two changes. On the "inner domain" side, it marks its reservation down to the indicated rate (the most it is now permitted to reserve), so that if an RESV Refresh event happens it will request the specified rate. On the "interface domain" side it selects one or more of the relevant reservations by an algorithm of its choosing and requests that it likewise reduce its rate. For sake of argument, let us imagine that the selected reservation is the one to VPN5. The RESV Error now makes its way through R8 to VPN5, which similarly reduces its bandwidth request to the stated amount and passes a RESV Error signal on the "enclave" side requesting that the reservation be appropriately reduced.

H5 is now faced with a decision. If the request is to reduce its reservation to zero, that is equivalent to tearing down the reservation. In this simple case, it sends an RESV Tear to tear down the reservation entirely and advises its application to adjust its expectations of the session accordingly, which may mean shutting down

the session. If the request is to reduce it below a certain value, however, it may be possible for the application to do so and remain viable. For example, if a VoIP application using a G. 711 codec (80 KBPS) is asked to reduce its bandwidth below 70 KBPS, it may be possible to renegotiate the codec in use to G. 729 or some other codec. In such a case, the originating application should re-reserve at the stated bandwidth (in this case, 70 KBPS), initiate the application level change, and let the application change the reservation again (perhaps to 60 KBPS) when it has completed that process.

For the "priority" reservation, at the same time, R9 believes that it has sufficient bandwidth and that any other stated policy is also met, it forwards the RESV to VPN7. Each will update its reservation state and forward the RESV signal to the next. VPN7 now acts as an [RFC3175] aggregator for the inner domain. This means that it receives the RESV signal for the inner domain reservation and stores state, decrypts the data stream from VPN8, operates on the RSVP signals as an RSVP-configured router, and forwards the received IP datagrams (including the updated RESV signals) into its interface domain. The RESV signals originated by VPN4, VPN5, and VPN6 are therefore forwarded towards VPN1, VPN2, and VPN3 through the interface domain.

VPN3 now acts as an [RFC3175] aggregator for the interface domain. This means that it receives the RESV signal for the interface domain reservation and stores state, decrypts the data stream from its peer, operates on the RSVP signals as an RSVP-configured router, and forwards the received IP datagrams (including the updated RESV signals) into its enclave. The RESV signal originated by H6 is therefore forwarded towards H3 according to the routing of the enclave.

H3 now receives the original RESV signals and deliver it to the relevant application.

3. Data flows within a VPN Router

This section details the data flows within a VPN Router, in the context of sessions as described in Section 2 .

3.1 Plaintext to Ciphertext Data Flows

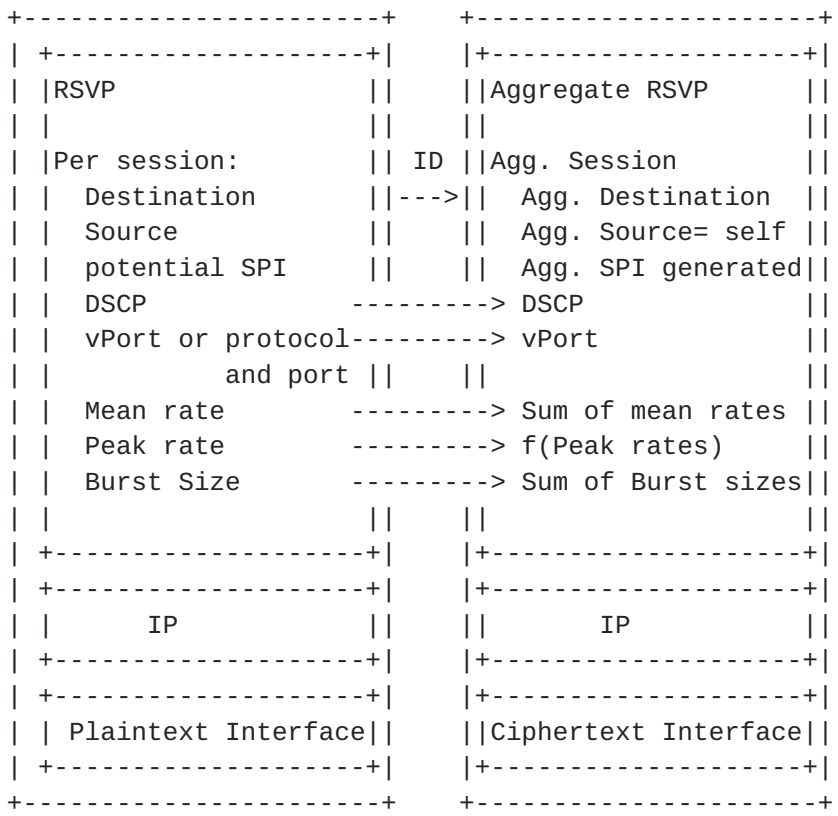

Figure 7: Data Flows in a VPN Router Outbound

Parameters on a reservation include:

Destination Address: On the plaintext side, the VPN Router participates in the end to end reservations being installed for plaintext sessions. These may include individual flows as described in [RFC2205] IPsec data flows [RFC2207] aggregate reservations [RFC3175] or other types. It passes an identifier for the ciphertext side of the deaggregator to its ciphertext unit.

DSCP: The DSCP of the plaintext data flow is provided to the ciphertext side.

Virtual Port: The virtual destination port is provided to the ciphertext side. This may be derived from an [RFC2207] session object or from policy information.

Mean Rate: The sum of the plaintext mean rates is provided to the ciphertext unit.

Peak Rate: A function of the plaintext peak rates is provided to the ciphertext unit. This function is less than or equal to the sum of the peak rates.

Burst Size: The sum of the burst sizes is provided to the ciphertext unit.

Messages include:

Path: The Plaintext PATH message is sent as encrypted data to the ciphertext unit. In parallel, a trigger needs to be sent to the ciphertext unit that results in it generating the corresponding aggregated PATH message for the ciphertext side.

Path Error: This indicates that a PATH message sent to the remote enclave was in error. In the error case, the message itself is sent on as encrypted data, but a signal is sent to the ciphertext side in case the error affects the ciphertext reservation (such as reomving or changing state).

Path Tear: The PATH Tear message is sent as encrypted data to the ciphertext unit. In parallel, a signal is sent to the ciphertext side which will trigger a Path Tear on its reservation in the event that this is the last aggregated session, or change the SENDER_TSPEC of the aggregated session.

RESV: The Plaintext RESV message is sent as encrypted data to the ciphertext unit. In parallel, a trigger needs to be sent to the ciphertext unit that results in it generating the corresponding aggregated RESV message for the ciphertext side.

RESV Error: This indicates that a RESV message recieved as data and forwarded into the enclave was in error or needed to be preempted as described in [RFC3181] or [I-D.polk-tsvwg-rsvp-bw-reduction] . In the error case, the message itself is sent on as encrypted data, but a signal is sent to the ciphertext side in case the error affects the ciphertext reservation (such as reomving or changing state).

RESV Tear: The RESV Tear message is sent as encrypted data to the ciphertext unit. In parallel, a signal is sent to the ciphertext side which will trigger a RESV Tear on its reservation in the event that this is the last aggregated session, or reduce the bandwidth of an existing reservation.

RESV Confirm: This indicates that a RESV message recieved as data and forwarded into the enclave, and is now being confirmed. This message is sent as encrypted data to the ciphertext side, and in parallel a signal is sent to potentially trigger an RESV Confirm on the aggregate reservation.

3.2 Ciphertext to Plaintext Data Flows

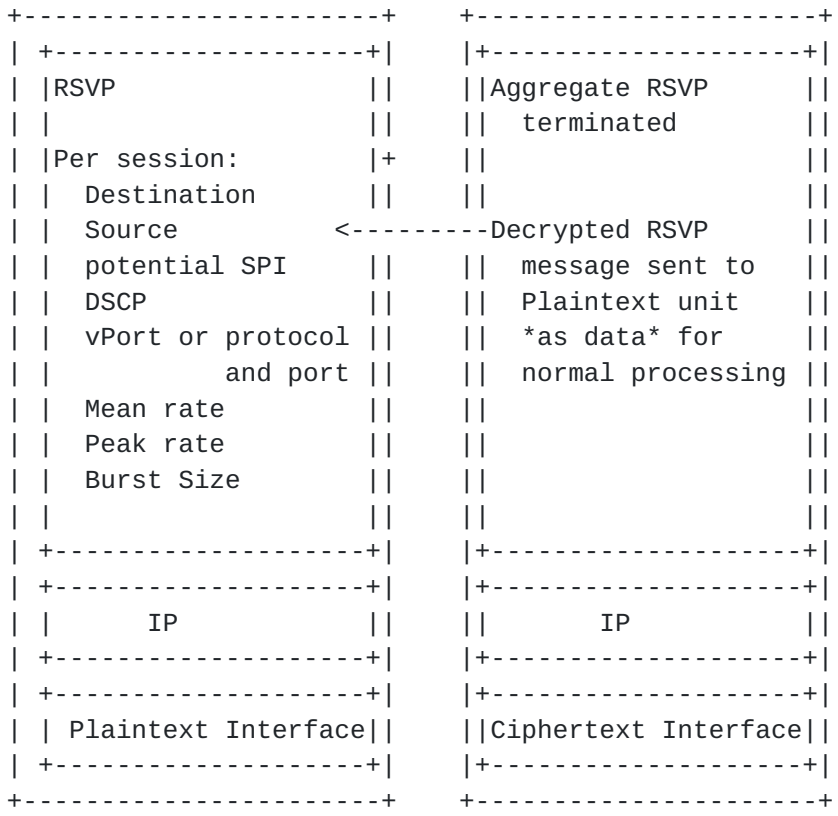

Figure 8: Data Flows in a VPN Router Inbound

The aggregate reservation is terminated by the ciphertext side of the VPN Router. The RSVP messages related to the subsidiary sessions are carried in the encrypted tunnel as data, and therefore arrive at the plaintext side with other data. As the plaintext side participates in these reservations, some information is returned to the ciphertext size to parameterize the aggregate reservation as described in Section 3.1 in the processing of the outbound messages.

4. IANA Considerations

This document makes no request of the IANA.

Note to RFC Editor: in the process assigning numbers and building IANA registries prior to publication, this section will have served its purpose. It may therefore be removed upon publication as an RFC.

5. Security Considerations

The typical security concerns of datagram integrity, node and user authentication are implicitly met by the security association that exists between the VPN routers. The secure data stream which flows between the VPN routers is also used for the reservation signaling datagrams flowing between VPN routers. Information that is contained in these signaling datagrams receives the same level of encryption that is received by the data streams.

One of the reasons cited for the nesting of VPN routes in Section 1.1 are the different levels of security across the nested VPN routers. If the security level decreases from one VPN router to the next VPN router in the nested path, the reservation signaling datagrams will by default receive the lower security level treatment. For most cases, the lower security treatment is acceptable. In certain networks, however, the reservation signaling across the entire nested path must receive the highest security level treatment (e. g. encryption, authentication of signaling nodes). For example the highest precedence level may only be signaled to VPN routers which can provide the highest security levels. If any VPN router in the nested path is incapable of providing the highest security level, it cannot participate in the reservation mechanism.

In the general case, the nested path may contain routers which are either incapable of participating in VPNs or providing required security levels. These routers can participate in the reservation only if the lower security level is acceptable (as configured by policy) for the signaling of reservation datagrams.

VPN routers encapsulate encrypted IP packets and prepend an extra header on each packet. These packets, whether used for signaling or data, should be identifiable, at a minimum by the IP addresses and DSCP value. The prepended header, therefore, should contain at a minimum the DSCP value corresponding to the signaled reservation in each packet. This may literally be the same DSCP as is used for the data (forcing control plane traffic to receive the same QoS treatment as its data), or a different DSCP that is routed identically (separating control and data plane traffic QoS but not routing).

6. Acknowledgements

Doug Marquis, James Polk, Mike Tibodeau, Pete Babendreier, Roger Levesque, and Subha Dhesikan gave early review comments.

Francois Le Faucheur, Bruce Davie, and Chris Christou (with Pratik Bose) added [I-D.lefaucheur-rsvp-ipsec], which clarified the interaction of this approach with the DSCP.

7. References

7.1 Normative References

[I-D.baker-tsvwg-mlpp-that-works]
Baker, F., "Implementing MLPP for Voice and Video in the Internet Protocol Suite", Internet-Draft draft-baker-tsvwg-mlpp-that-works-02, October 2004.

[I-D.lefaucheur-rsvp-ipsec]
Faucheur, F., Davie, B., Bose, P. and C. Christou, "Aggregate RSVP Reservations for IPsec Tunnels", Internet-Draft draft-lefaucheur-rsvp-ipsec-00, February 2005.

[I-D.polk-tsvwg-rsvp-bw-reduction]
Polk, J., "A Resource Reservation Extension for the Reduction of Bandwidth of a Reservation Flow", Internet-Draft draft-polk-tsvwg-rsvp-bw-reduction-00, October 2004.

[RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S. and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.

[RFC2207] Berger, L. and T. O'Malley, "RSVP Extensions for IPSEC Data Flows", RFC 2207, September 1997.

[RFC2746] Terzis, A., Krawczyk, J., Wroclawski, J. and L. Zhang, "RSVP Operation Over IP Tunnels", RFC 2746, January 2000.

[RFC2750] Herzog, S., "RSVP Extensions for Policy Control", RFC 2750, January 2000.

[RFC2996] Bernet, Y., "Format of the RSVP DCLASS Object", RFC 2996, November 2000.

[RFC3175] Baker, F., Iturralde, C., Le Faucheur, F. and B. Davie, "Aggregation of RSVP for IPv4 and IPv6 Reservations", RFC 3175, September 2001.

7.2 Informative References

[ANSI.MLPP.Spec]
American National Standards Institute, "Telecommunications - Integrated Services Digital Network (ISDN) - Multi-Level Precedence and Preemption (MLPP) Service Capability",

ANSI T1.619-1992 (R1999), 1992.

[ANSI.MLPP.Supplement]
American National Standards Institute, "MLPP Service Domain Cause Value Changes", ANSI ANSI T1.619a-1994 (R1999), 1990.

[ITU.MLPP.1990]
International Telecommunications Union, "Multilevel Precedence and Preemption Service (MLPP)", ITU-T Recommendation I.255.3, 1990.

[RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.

[RFC1633] Braden, B., Clark, D. and S. Shenker, "Integrated Services in the Internet Architecture: an Overview", RFC 1633, June 1994.

[RFC2209] Braden, B. and L. Zhang, "Resource ReSerVation Protocol (RSVP) -- Version 1 Message Processing Rules", RFC 2209, September 1997.

[RFC2210] Wroclawski, J., "The Use of RSVP with IETF Integrated Services", RFC 2210, September 1997.

[RFC2401] Kent, S. and R. Atkinson, "Security Architecture for the Internet Protocol", RFC 2401, November 1998.

[RFC2406] Kent, S. and R. Atkinson, "IP Encapsulating Security Payload (ESP)", RFC 2406, November 1998.

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

[RFC2474] Nichols, K., Blake, S., Baker, F. and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z. and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[RFC2747] Baker, F., Lindell, B. and M. Talwar, "RSVP Cryptographic Authentication", RFC 2747, January 2000.

[RFC2872] Bernet, Y. and R. Pabbati, "Application and Sub

Application Identity Policy Element for Use with RSVP", RFC 2872, June 2000.

[RFC3097] Braden, R. and L. Zhang, "RSVP Cryptographic Authentication -- Updated Message Type Value", RFC 3097, April 2001.

[RFC3181] Herzog, S., "Signaled Preemption Priority Policy Element", RFC 3181, October 2001.

[RFC3182] Yadav, S., Yavatkar, R., Pabbati, R., Ford, P., Moore, T., Herzog, S. and R. Hess, "Identity Representation for RSVP", RFC 3182, October 2001.

[RFC3246] Davie, B., Charny, A., Bennet, J., Benson, K., Le Boudec, J., Courtney, W., Davari, S., Firoiu, V. and D. Stiliadis, "An Expedited Forwarding PHB (Per-Hop Behavior)", RFC 3246, March 2002.

[RFC3261] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M. and E. Schooler, "SIP: Session Initiation Protocol", RFC 3261, June 2002.

[RFC3312] Camarillo, G., Marshall, W. and J. Rosenberg, "Integration of Resource Management and Session Initiation Protocol (SIP)", RFC 3312, October 2002.

[RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.

[RFC3474] Lin, Z. and D. Pendarakis, "Documentation of IANA assignments for Generalized MultiProtocol Label Switching (GMPLS) Resource Reservation Protocol - Traffic Engineering (RSVP-TE) Usage and Extensions for Automatically Switched Optical Network (ASON)", RFC 3474, March 2003.

Authors' Addresses

Fred Baker
Cisco Systems
1121 Via Del Rey
Santa Barbara, California 93117
USA

Phone: +1-408-526-4257
Fax: +1-413-473-2403
Email: fred@cisco.com

Pratik Bose
Lockheed Martin
22300 Comsat Drive
Clarksburg, Maryland 20871
USA

Phone: +1-301-428-4215
Fax: +1-301-428-5414
Email: pratik.bose@lmco.com

Intellectual Property Statement

Disclaimer of Validity

Copyright Statement

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.