Resolución fragmentación de IP, problemas MTU, MSS, y PMTUD con el GRE y el IPSEC

Contenido

Introducción

El documento describe cómo trabajo fragmentación de IP y del Path Maximum Transmission Unit Discovery (PMTUD) y también discute algunos escenarios que impliquen el comportamiento del PMTUD cuando están combinados con diversas combinaciones de túneles IP. El uso extenso actual de los túneles IP en Internet ha traído los problemas que implican fragmentación de IP y PMTUD a la vanguardia.

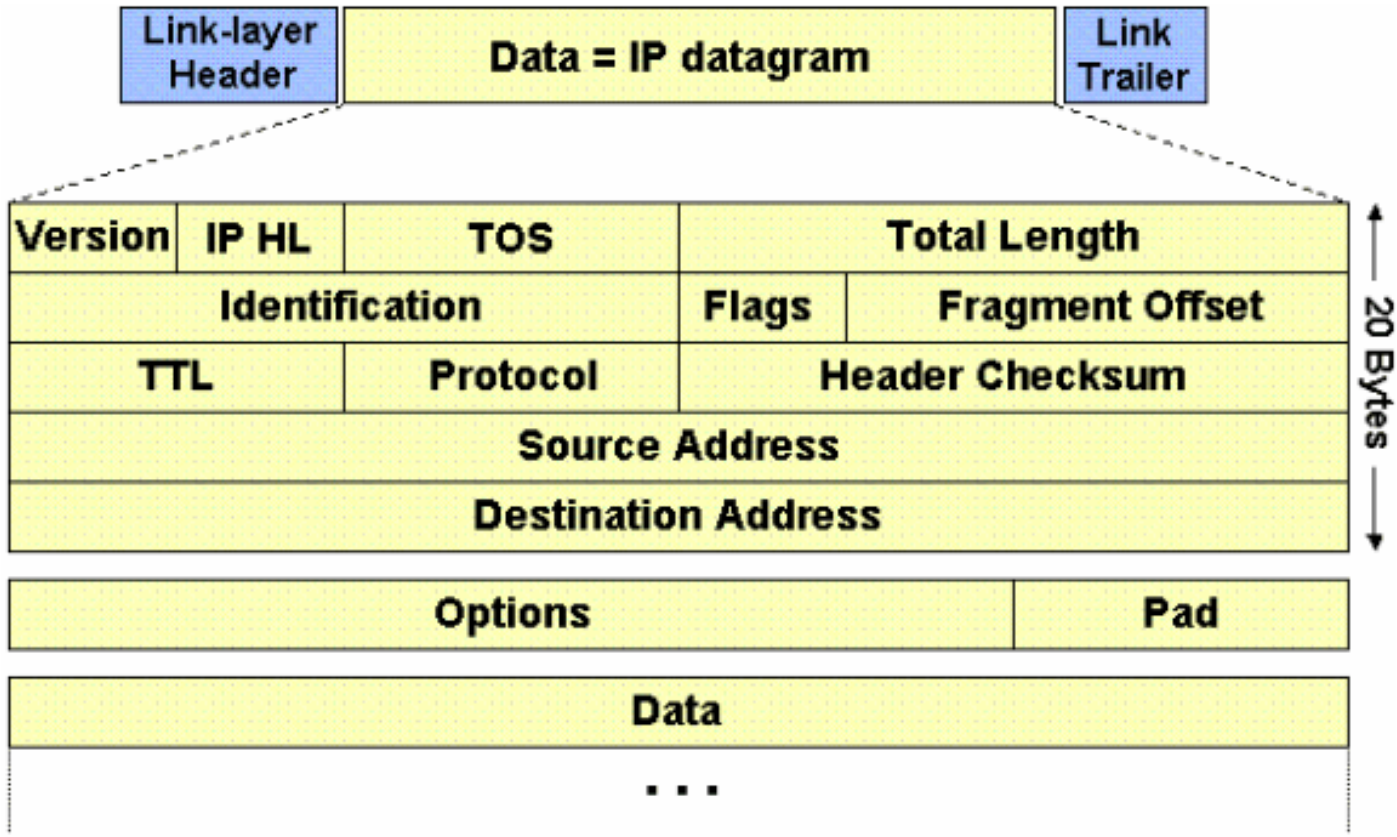
Fragmentación IP y Reensamblado

El protocolo IP fue diseñado para su uso en una amplia variedad de links de transmisión. Aunque el Largo máximo de un IP datagram sea 65535, la mayoría de los links de transmisión aplican un límite más pequeño de la longitud máxima de paquetes, llamado un MTU. El valor de MTU depende del tipo de link de transmisión. El diseño de IP acomoda las diferencias de MTU puesto

que permite que el Routers haga fragmentos de los datagramas IP cuanto sea necesario. La estación receptora es responsable del nuevo ensamble de los fragmentos nuevamente dentro del IP datagram del mismo tamaño original.

La fragmentación IP implica dividir un datagrama en varias partes que luego se puedan reensamblar. Los campos IP source (origen IP), destination (destino), identification (identificación), total length (longitud total) y fragment offset (desplazamiento de fragmentos), junto con los indicadores “more fragments” (más fragmentos) y “don’t fragment” (no fragmentar) en el encabezado IP se utilizan para la fragmentación y la reagrupación IP. Para más información sobre los mecánicos de la fragmentación y reconstrucción de IP, vea el RFC 791 .

Esta imagen representa la disposición de un encabezado IP.



La identificación es 16 bits y es un valor asignado por el remitente de un IP datagram para ayudar en el nuevo ensamble de los fragmentos de un datagrama.

El desplazamiento de fragmentos es de 13 bits e indica a dónde pertenece el fragmento en el datagrama IP original. Este valor es un múltiplo de ocho bytes.

En el campo de los indicadores del encabezado IP, hay tres bits para los indicadores de control. Es importante observar que el bit "don't fragment" (DF) tiene una función central en la PMTUD porque determina si se permite o no que se fragmente un paquete.

0 mordido es reservado, y se fija siempre a 0. mordió 1 es el bit DF (0 = “puede hacer fragmentos,” 1 = “no hace fragmento”). El Bit 2 es el bit "more fragments" (MF) (0 = "último fragmento", 1 = "más fragmentos").

Valor	Bit 0 Reservado	Bit 1 DF	Bit 2 MF
0	0	Mayo	Último
1	0	No se puede	Más

El gráfico siguiente muestra un ejemplo de fragmentación. Si usted suma todas las longitudes de los fragmentos IP, el valor excede la longitud del datagrama IP original por 60. La razón por la cual la longitud total aumenta en 60 es que se crearon tres encabezados IP adicionales, uno para cada fragmento después del primero.

El primer fragmento tiene un desplazamiento de 0 y la longitud de este fragmento es 1500; esto incluye 20 bytes para el encabezado IP original levemente modificado.

El segundo fragmento tiene un desplazamiento de 185 (185 x 8 = 1480), lo cual significa que la porción de datos de este fragmento se inicia a los 1480 bytes del datagrama IP original. La longitud de este fragmento es 1500; esto incluye el encabezado IP adicional creado para este fragmento.

El tercer fragmento tiene un desplazamiento de 370 (370 x 8 = 2960), lo cual significa que la porción de datos de este fragmento se inicia a los 2960 bytes del datagrama IP original. La longitud de este fragmento es 1500; esto incluye el encabezado IP adicional creado para este fragmento.

El cuarto fragmento tiene un desplazamiento de 555 (555 x 8 = 4440), lo cual significa que la porción de datos de este fragmento comienza 4440 bytes en el datagrama IP original. La longitud de este fragmento es 700 bytes; esto incluye el encabezado IP adicional creado para este fragmento.

Solo cuando se recibe el último fragmento, se puede determinar el tamaño del datagrama IP original.

El desplazamiento de fragmentos en el último fragmento (555) da un desplazamiento de datos de 4440 bytes en el datagrama IP original. Si usted luego suma los bytes de datos del último fragmento (680 = 700 - 20), obtiene como resultado 5120 bytes, que es la porción de datos del datagrama IP original. Luego, al agregar 20 bytes para un encabezado IP, se iguala el tamaño del datagrama IP original (4440 + 680 + 20 = 5140).

Original IP Datagram

Sequence	Identifier	Total Length	DF May / Don't	MF Last / More	Fragment Offset
0	345	5140	0	0	0

IP Fragments (Ethernet)

Sequence	Identifier	Total Length	DF May / Don't	MF Last / More	Fragment Offset
0-0	345	1500	0	1	0
0-1	345	1500	0	1	185
0-2	345	1500	0	1	370
0-3	345	700	0	0	555

Problemas con la Fragmentación IP

Hay varios problemas que hacen que la fragmentación IP no sea aconsejable. Hay un pequeño aumento en la sobrecarga del CPU y de la memoria para fragmentar un datagrama IP. Esto es válido tanto para el remitente como para un router en la trayectoria entre un remitente y un receptor. La creación de fragmentos implica simplemente crear encabezados de fragmento y copiar el datagrama original en los fragmentos. Esto se puede realizar de forma bastante eficaz, ya que toda la información necesaria para crear los fragmentos está disponible inmediatamente.

La fragmentación genera mayor sobrecarga para el receptor al reensamblar los fragmentos porque el receptor debe asignar memoria para los fragmentos que llegan y unirlos nuevamente en un datagrama una vez recibidos todos los fragmentos. El reensamblado en un host no se considera un problema porque el host tiene el tiempo y los recursos de memoria para dedicarlos a esta tarea.

Pero, el reensamblado es muy ineficaz en un router cuya tarea primaria sea reenviar los paquetes tan rápido como sea posible. Un router no está diseñado para conservar los paquetes durante un período de tiempo. También un router que hace el nuevo ensamble elige el buffer más grande disponible (18K) con el cual para trabajar porque no tiene ninguna manera de conocer el tamaño del paquete del IP original hasta que se reciba el fragmento más reciente.

Otro problema de fragmentación implica cómo se manejan los fragmentos caídos. Si se descarta un fragmento de un datagrama IP, se deberá reenviar el datagrama IP original entero y también se fragmentará. Puede verse un ejemplo de esto con Network File System (NFS). El NFS, por abandono, tiene un tamaño del bloque de lectura y escritura de 8192, así que un datagrama NFS IP/UDP será aproximadamente 8500 bytes (que incluye el NFS, el UDP, y los encabezados IP). Una estación remitente conectada a Ethernet (MTU 1500) deberá fragmentar el datagrama de 8500 bytes en seis partes; cinco fragmentos de 1500 bytes y un fragmento de 1100 bytes. Si es un de los seis fragmentos se caen debido a un link congestionado, el datagrama original completo tendrá que ser retransmitido, así que significa que seis más fragmentos tendrán que ser creados. Si este link descarta uno de seis paquetes, las probabilidades son bajas de que se puedan transferir datos NFS a través de este link, ya que por lo menos un fragmento IP se descartaría de cada datagrama IP original de 8500 bytes NFS.

Los Firewall que filtran o manipulan los paquetes basados en la capa 4 (L4) con la información de la capa 7 (L7) en el paquete pudieron tener problema que procesaba los fragmentos IP correctamente. Si los fragmentos IP están fuera de servicio, un Firewall pudo bloquear los fragmentos no iniciales porque no llevan la información que haría juego el filtro de paquete. Esto significaría que el datagrama IP original no podría ser reensamblado por el host receptor. Si el firewall está configurado para permitir que los fragmentos no iniciales con información insuficiente coincidan correctamente con el filtro, podría ocurrir un ataque de fragmentos no iniciales a través del firewall. También, los paquetes directos de algunos dispositivos de red (tales como motores de switch de contenido) basados en el L4 con la información L7, y si un paquete atraviesa los fragmentos múltiples, después el dispositivo pudieron tener problema que aplicaba sus directivas.

Evite fragmentación de IP: Qué hace TCP MSS y cómo funciona

El Tamaño máximo de segmento de TCP (MSS) define la cantidad máxima de datos que un host desea aceptar en un datagrama simple de TCP/IP. Este datagrama TCP/IP se pudo hacer fragmentos en la capa IP. El valor de MSS se envía como una opción de encabezado TCP solamente en los segmentos SYN de TCP. Cada lado de una conexión TCP informa su valor de MSS al otro lado. Contrariamente a la creencia popular, el valor de MSS no se negocia entre los

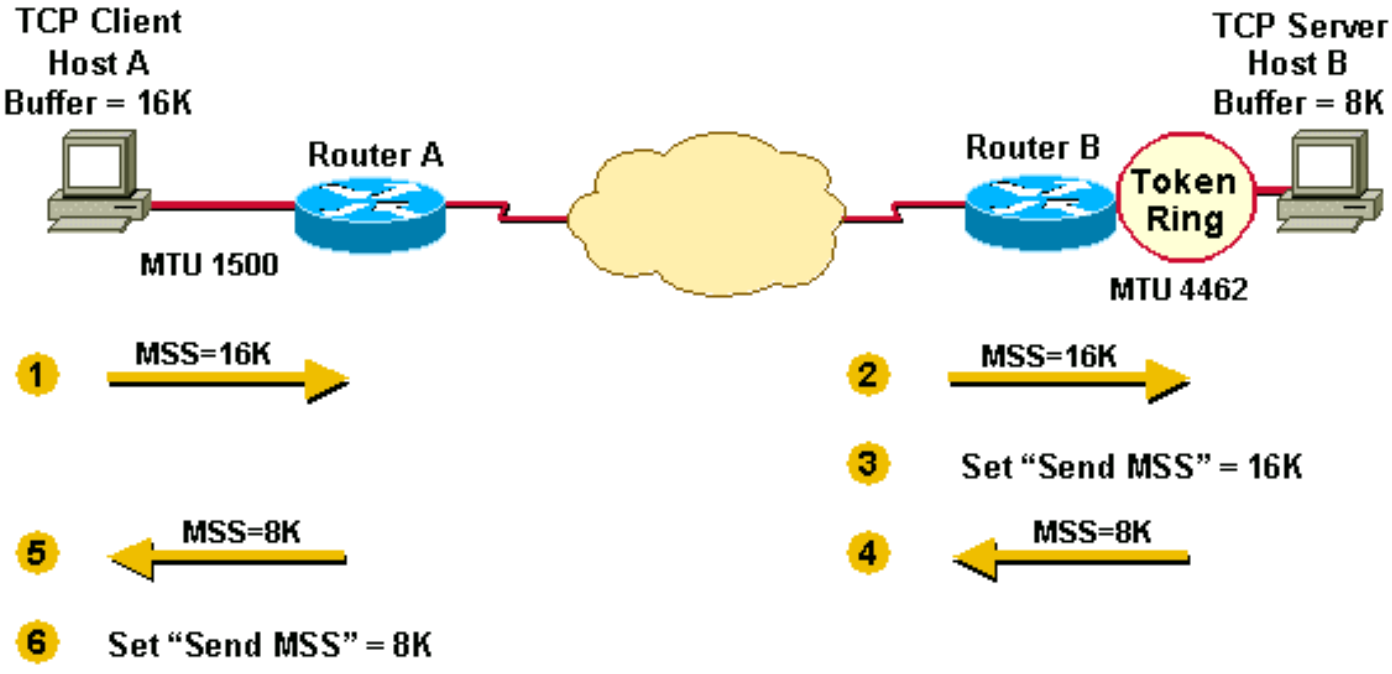
hosts. Se requiere que el host remitente limite el tamaño de los datos en un solo segmento TCP a un valor inferior o igual al valor de MSS informado por el host receptor.

Originalmente, el valor de MSS indicaba el tamaño de un buffer (mayor o igual que 65.496K) que estaba ubicado en una estación receptora para poder almacenar los datos TCP incluidos dentro de un solo datagrama IP. El valor de MSS era el segmento máxima (tramo) de datos que el receptor TCP estaba dispuesto a aceptar. Este segmento TCP podría tener un tamaño de 64K (el máximo tamaño de datagrama IP) y podría fragmentarse en la capa IP con el propósito de transmitirse a través de la red hacia el host receptor. El host receptor reensamblaba el datagrama IP antes de entregarle el segmento TCP completo a la capa TCP.

Abajo están un par de escenarios que muestran cómo los valores MSS se fijan y se utilizan para limitar los tamaños del segmento TCP, y por lo tanto, los tamaños del IP datagram.

En la situación 1, se ilustra la manera en que se implementó primero el valor de MSS. El Host A tiene un buffer de 16K y el Host B tiene un buffer de 8K. Envían y reciben sus valores MSS y ajustan sus envíos MSS para enviar información entre ellos. Note que el host A y el host B tendrán que hacer fragmentos de los datagramas IP que son más grandes que el MTU de interfaz, pero aún menos que el envío MSS porque el stack TCP podría pasar los bytes de dato 16K o 8K abajo del stack al IP. En el caso del Host B, los paquetes podrían fragmentarse dos veces, una vez para llegar a la LAN Token Ring y nuevamente para llegar a la LAN Ethernet.

Escenario 1



1. El Host A envía su valor de MSS de 16K al Host B.
2. El Host B recibe el valor de MSS de 16K del Host A.
3. El Host B configura su valor de MSS de envío en 16K.
4. El Host B envía su valor de MSS de 8K al Host A.
5. El Host A recibe el valor de MSS de 8K del Host B.
6. El Host A configura su valor de MSS de envío en 8K.

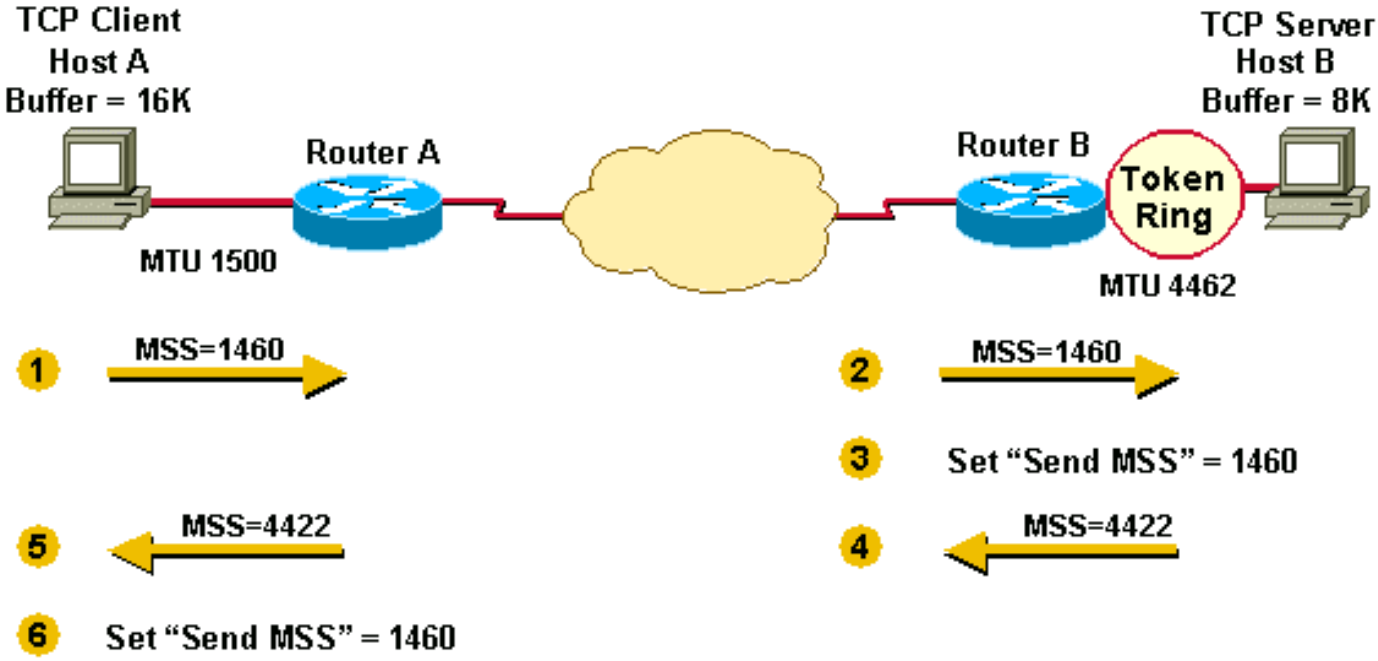
Para ayudar a evitar la fragmentación IP en los extremos de la conexión TCP, la selección del valor de MSS se cambió al tamaño mínimo de buffer y a la MTU de interfaz saliente (- 40). Los números de MSS son 40 bytes más pequeños que los números de MTU porque el valor de MSS

es apenas el tamaño de datos TCP, que no incluye el encabezado IP de 20 bytes ni el encabezado TCP de 20 bytes. El valor de MSS se basa en los tamaños de encabezado predeterminados; el stack del remitente debe restar los valores apropiados para el encabezado IP y el dependiente del encabezado TCP en qué TCP o las opciones IP se utilizan.

La forma en que funciona MSS ahora es que cada host primero comparará su MTU de interfaz saliente con su propio buffer y seleccionará el menor valor como el MSS que se enviará. Los hosts luego compararán el tamaño de MSS recibido con su propia MTU de interfaz y, de nuevo, elegirán el menor de los dos valores.

El escenario 2 ilustra este paso adicional tomado por el remitente para evitar la fragmentación en los cables locales y remotos. Observe cómo cada host tiene en cuenta la MTU de la interfaz saliente (antes de que los hosts se envíen entre sí sus valores de MSS) y cómo esto ayuda a evitar la fragmentación.

Escenario 2



1. El Host A compara su buffer de MSS (16K) y su MTU (1500 - 40 = 1460) y utiliza el valor más bajo como el valor de MSS (1460) para enviarlo al Host B.
2. El Host B recibe el valor de MSS de envío (1460) del Host A y lo compara con el valor de su MTU de interfaz saliente - 40 (4422).
3. El Host B configura el valor inferior (1460) como el valor de MSS para el envío de datagramas IP al Host A.
4. El Host B compara su buffer de MSS (8K) y su MTU (4462 - 40 = 4422) y utiliza 4422 como el valor de MSS para enviarlo al Host A.
5. El Host A recibe el valor de MSS de envío (4422) del Host B y lo compara con el valor de su MTU de interfaz saliente - 40 (1460).
6. El Host A configura el valor inferior (1460) como el valor de MSS para el envío de datagramas IP al Host B.

El valor elegido por ambos hosts como MSS de envío recíproco es 1460. A menudo, el valor de MSS de envío será el mismo en cada extremo de una conexión TCP.

En la situación 2, la fragmentación no ocurre en los extremos de una conexión TCP porque

ambas MTU de interfaz saliente son tenidas en cuenta por los hosts. Los paquetes aún pueden fragmentarse en la red entre el Router A y el Router B si encuentran un link con una MTU inferior a las de interfaz saliente de cualquiera de esos hosts.

¿Qué es PMTUD?

El TCP MSS como anterior descrito toma el cuidado de la fragmentación en los dos puntos finales de una conexión TCP, pero no maneja el caso donde hay un link más pequeño MTU en el centro entre estos dos puntos finales. El PMTUD fue desarrollado para evitar la fragmentación en la trayectoria entre los puntos finales. Se utiliza para determinar dinámicamente la MTU más baja a lo largo de la trayectoria del origen de un paquete a su destino.

> Nota: El PMTUD es soportado solamente por el TCP y el UDP. Otros protocolos no lo soportan. Si el PMTUD se habilita en un host, y está casi siempre, todo el TCP/IP o paquetes UDP del host tendrá el conjunto de bits DF.

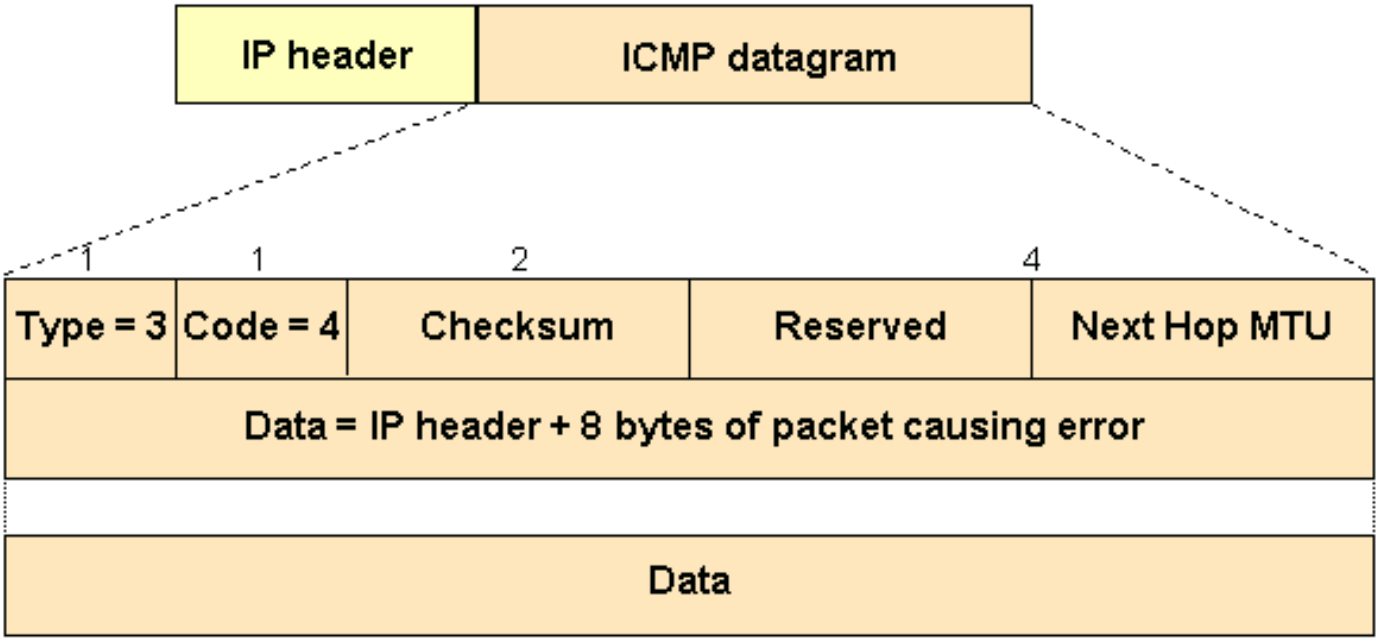
Cuando un host envía un paquete de datos lleno MSS con el conjunto de bits DF, el PMTUD reduce el valor del envío MSS para la conexión si recibe la información que el paquete requeriría la fragmentación. Un host “recuerda generalmente” el valor MTU para un destino puesto que crea una /32) entrada del “host” (en su tabla de ruteo con este valor MTU.

Si un router intenta remitir un IP datagram, con el conjunto de bits DF, sobre un link que tenga un MTU inferior que el tamaño del paquete, el router caerá el paquete y volverá un mensaje “Destino inalcanzable” del Internet Control Message Protocol (ICMP) a la fuente de este IP datagram, con el código que indica la “fragmentación necesaria y el DF para fijar” (el tipo 3, el código 4). Al recibir el mensaje de ICMP, la estación de origen disminuirá el MSS de envío y, cuando el TCP vuelve a transmitir el segmento, usará el tamaño menor de segmento.

Aquí está un ejemplo de un mensaje “fragmentación requerida y DF configurado” ICMP que usted puede ser que vea en un router después de que giren al **comando debug ip icmp**:

```
ICMP: dst (10.10.10.10) frag. needed and DF set
unreachable sent to 10.1.1.1
```

Este diagrama muestra el formato del encabezado ICMP de una “fragmentación necesaria y del DF para fijar” el mensaje “Destino inalcanzable”.



Por el RFC 1191 , un router que vuelve un mensaje ICMP que indica “fragmentación necesitó y el DF fijar” debe incluir el MTU de esa red del salto siguiente en los 16 bits de orden inferior del campo del encabezado adicional ICMP que se etiqueta “inusitado” en el RFC 792 de la especificación ICMP .

Las primeras implementaciones de RFC 1191 no suministraban la información de MTU de salto siguiente. Incluso cuando esta información se suministraba, algunos hosts la ignoraban. En este caso, RFC 1191 también contiene una tabla que enumera los valores sugeridos por los que se debería disminuir a la MTU durante PMTUD. Es utilizado por los host para llegar más rápidamente un valor razonable para el envío MSS.

Plateau	MTU	Comments	Reference
	65535	Official maximum MTU	RFC 791
	65535	Hyperchannel	RFC 1044
65535			
32000		Just in case	
	17914	16Mb IBM Token Ring	ref. [6]
17914			
	8166	IEEE 802.4	RFC 1042
8166			
	4464	IEEE 802.5 (4Mb max)	RFC 1042
	4352	FDDI (Revised)	RFC 1188
4352 (1%)			
	2048	Wideband Network	RFC 907
	2002	IEEE 802.5 (4Mb recommended)	RFC 1042
2002 (2%)			
	1536	Exp. Ethernet Nets	RFC 895
	1500	Ethernet Networks	RFC 894
	1500	Point-to-Point (default)	RFC 1134
	1492	IEEE 802.3	RFC 1042
1492 (3%)			
	1006	SLIP	RFC 1055
	1006	ARPANET	BBN 1822
1006			
	576	X.25 Networks	RFC 877
	544	DEC IP Portal	ref. [10]
	512	NETBIOS	RFC 1088
	508	IEEE 802/Source-Rt Bridge	RFC 1042
	508	ARCNET	RFC 1051
508 (13%)			
	296	Point-to-Point (low delay)	RFC 1144
296			
68		Official minimum MTU	RFC 791

El mecanismo PMTUD se lleva a cabo en todos los paquetes, ya que el trayecto entre el remitente y el receptor puede cambiar en forma dinámica. Cada vez que un remitente reciba mensajes de ICMP de "no se puede fragmentar", actualizará la información de ruteo (donde se almacena la PMTUD).

Durante PMTDU, pueden ocurrir dos cosas:

- El paquete puede llegar finalmente al receptor sin haber sido fragmentado.Nota: Para que un router proteja el CPU contra ataques de DOS, regula el número de mensajes de destino inalcanzable de ICMP que enviaría a dos por segundo. Por lo tanto, en este contexto, si usted tiene un escenario de red en el cual usted cuente con que el router necesite responder con más de dos mensajes ICMP (tipo = 3, código = 4) por segundo (pueden ser diversos host), usted querría inhabilitar estrangular de los mensajes ICMP con el **ningún comando interface inalcanzable del [df] del tarifa-límite ICMP del IP**.

- El emisor puede obtener mensajes ICMP "Can’t Fragment" (Imposible realizar la fragmentación) desde cualquier (o todos) los saltos a lo largo del trayecto hacia el receptor.

PMTUD se realiza independientemente para ambas direcciones de un flujo de TCP. Pudo haber los casos donde el PMTUD en una dirección de un flujo acciona una de las estaciones terminales para bajar el envío MSS y la estación del otro extremo guarda la original para enviar el MSS porque nunca envió un IP datagram bastante grande para accionar el PMTUD.

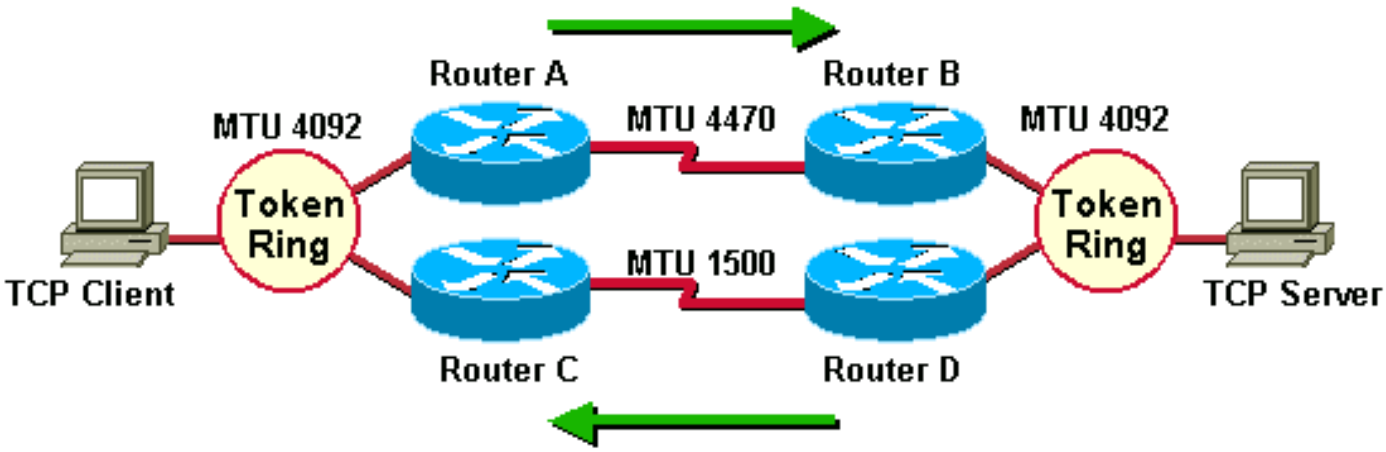
Un buen ejemplo de esto es la conexión HTTP representada a continuación en la situación 3. El cliente TCP envía los pequeños paquetes y el servidor envía los paquetes grandes. En este caso, solamente los paquetes grandes del servidor (mayor de 576 bytes) accionarán el PMTUD. Los paquetes del cliente son pequeños (menores que 576 bytes) y no activarán la PMTUD porque no requieren fragmentación para atravesar el link de MTU de 576.

Escenario 3

En la situación 4, se muestra un ejemplo de ruteo asimétrico donde una de las trayectorias tiene una MTU mínima más pequeña que la otra. El Asymmetric Routing ocurre cuando diversas trayectorias se toman para enviar y para recibir los datos entre dos puntos finales. En esta situación, la PMTUD accionará la disminución del valor de MSS de envío solamente en una dirección de un flujo de TCP. El tráfico del cliente TCP al servidor atraviesa el router A y al router B, mientras que el tráfico de retorno que viene del servidor al cliente atraviesa el router D y el C del router. Cuando el servidor TCP envíe paquetes al cliente, PMTUD hará que el servidor disminuya la velocidad del envío MSS porque el Router D debe fragmentar los paquetes de 4092 bytes antes de enviarlos al Router C.

El cliente, por otra parte, nunca recibirá un mensaje “Destino inalcanzable” ICMP con el código que indica la “fragmentación necesaria y el DF para fijar” porque no lo hace tuvo que el router A los paquetes de fragmento cuando los envía al servidor a través del router B.

Situación 4



Nota: El comando ip tcp path-mtu-discovery se utiliza para habilitar la detección de trayectoria de MTU de TCP para conexiones TCP iniciadas por routers (por ejemplo, BGP y Telnet).

Problemas con PMTUD

Existen tres eventos que pueden interrumpir una PMTUD, dos de ellos no son comunes y uno sí lo es.

- Un router puede descartar un paquete y no enviar un mensaje de ICMP. (Poco común)
- Un router puede generar y enviar un mensaje ICMP, pero el mensaje ICMP consigue bloqueado por un router o un Firewall entre este router y el remitente. (Campo común)
- Un router puede generar y enviar un mensaje de ICMP, pero el remitente ignora el mensaje. (Poco común)

La primera y la última de las tres viñetas anteriores son poco comunes y, por lo general, son producto de un error, pero la viñeta del medio describe un problema habitual. Los usuarios que implementan filtros de paquetes ICMP tienden a bloquear todos los tipos de mensajes de ICMP, en lugar de solo bloquear determinados tipos de mensajes de ICMP. Un filtro de paquete puede bloquear todos los tipos de mensajes de ICMP, *excepto* los que indiquen "destino inalcanzable" o "tiempo excedido". El éxito o el fracaso de PMTUD depende de los mensajes de "destino inalcanzable" de ICMP que lleguen a través del remitente de un paquete TCP/IP. Los mensajes de "tiempo excedido" de ICMP son importantes para otros problemas de IP. Un ejemplo de tal filtro de paquete, implementado en un router se muestra aquí.

```
access-list 101 permit icmp any any unreachable
access-list 101 permit icmp any any time-exceeded
access-list 101 deny icmp any any
access-list 101 permit ip any any
```

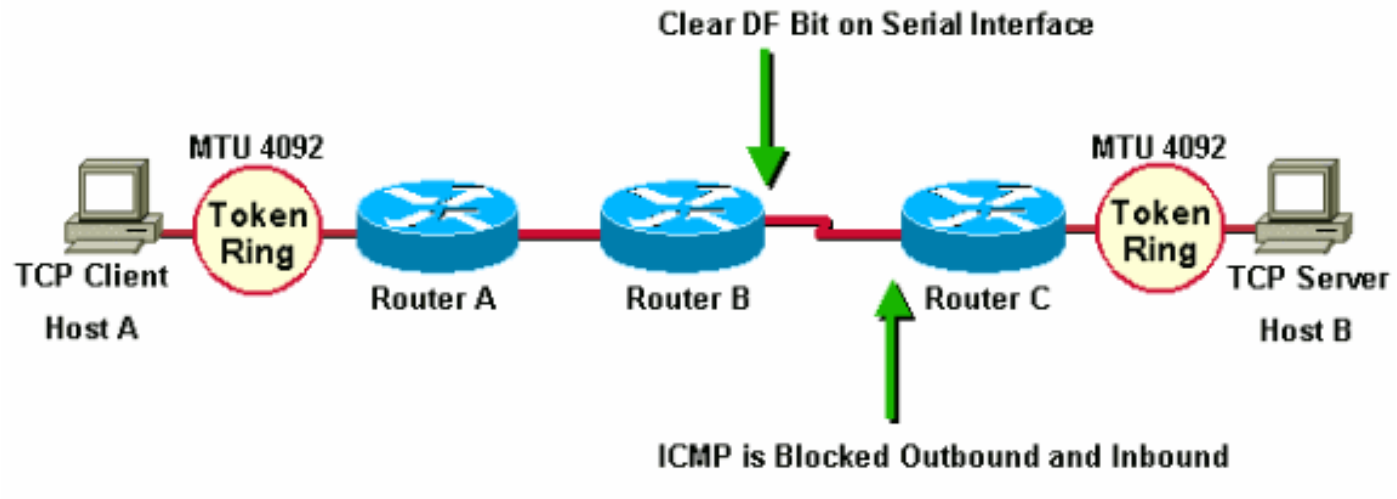
Existen otras técnicas que pueden utilizarse para ayudar a mitigar el problema asociado con el ICMP totalmente bloqueado.

- Borre el bit DF en el router y permita la fragmentación de todos modos (esto no pudo ser una buena idea, aunque. Consulte Problemas con la Fragmentación IP para obtener más información.
- Manipule el valor de opción MSS TCP MSS con el comando interface que el **IP tcp ajusta-mss <500-1460>**.

En el escenario siguiente, el router A y el router B están en el mismo dominio administrativo. El C del router es inaccesible y bloquea el ICMP, así que el PMTUD está quebrado. Una solución alternativa para esta situación está borrar el bit DF en las ambas direcciones en el router B para permitir la fragmentación. Esto se puede hacer con el Policy Routing. La sintaxis para borrar el bit DF está disponible en el Cisco IOS® Software, versión 12.1(6) y posteriores.

```
interface serial0
...
ip policy route-map clear-df-bit
route-map clear-df-bit permit 10
match ip address 111
set ip df 0

access-list 111 permit tcp any any
```



Otra opción es cambiar el valor de la opción MSS de TCP en los paquetes SYN que atraviesen el router (disponible en el Cisco IOS Software, versión 12.2(4)T y posteriores). Esto reduce el valor de opción MSS en paquete TCP Syn de modo que sea más pequeño que el valor (1460) en el **comando ip tcp adjust-mss**. El resultado es que el emisor TCP enviará segmentos no mayores a este valor. El tamaño del paquete IP será 40 bytes más grande (1500) que el valor MSS (1460 bytes) para explicar el encabezado TCP (20 bytes) y el encabezado IP (20 bytes).

Puede ajustar el MSS de los paquetes SYN de TCP con el **comando ip tcp adjust-mss**.
Este sintaxis reducirá el valor MSS en los segmentos TCP a 1460. Este comando afecta el tráfico tanto entrante como saliente en la interfaz serial0.

```
int s0
ip tcp adjust-mss 1460
```

Los problemas de la fragmentación IP se han generalizado debido a que los túneles IP se han implementado más ampliamente. La razón que los túneles causan más fragmentación es porque la encapsulación de túnel agrega los “gastos indirectos” al tamaño de un paquete. Por ejemplo, la adición de (GRE) del Generic Router Encapsulation agrega 24 bytes a un paquete, y después de que este aumento que el paquete pudo necesitar para ser hecho fragmentos porque es más grande que el MTU saliente. En una sección posterior de este documento, verá ejemplos de las clases de problemas que pueden presentarse con los túneles y la fragmentación IP.

Topologías de Red Comunes que Necesitan PMTUD

PMTUD se necesita en situaciones de red en las que los links intermedios tienen MTU más pequeñas que la MTU de los links extremos. Algunos motivos comunes para la existencia de estos links MTU más pequeños son:

- Token Ring (o FDDI): hosts extremos conectados con una conexión de Ethernet entre ellos. El Token Ring (o el FDDI) MTU en los extremos es mayor que los Ethernetes MTU en el centro.
- PPPoE (a menudo utilizado con ADSL) necesita un encabezado de 8 bytes. Esto reduce la MTU efectiva de Ethernet a 1492 (1500 - 8).

Los protocolos de tunelización como GRE, IPSec y L2TP también necesitan espacio para sus encabezados y colas correspondientes. Esto también reduce la MTU efectiva de la interfaz saliente.

En las siguientes secciones, el impacto del PMTUD donde un Tunneling Protocol se utiliza en

alguna parte entre los host del dos extremos se estudia. De los tres casos anteriores, este caso es el más complejo y cubre todos los problemas que usted puede ser que vea en los otros casos.

¿Qué es un Túnel?

Un túnel es una interfaz lógica en un router de Cisco que proporciona una manera de encapsular los paquetes pasajeros dentro de un protocolo de transporte. Es una arquitectura diseñada para proporcionar los servicios necesarios para implementar un esquema de encapsulación punto a punto. El Tunelización tiene estos tres componentes primarios:

- Protocolo pasajero (AppleTalk, Banyan VINES, CLNS, DECnet, IP o IPX).
- Protocolo de la portadora - Uno de estos protocolos de la encapsulación:GRE - El protocolo de la portadora de protocolo múltiple de Cisco. Consulte RFC 2784 y RFC 1701 para obtener más información.Túneles IP en IP: consulte RFC 2003 para obtener más información.
- Protocolo de transporte - El protocolo utilizado para llevar el protocolo encapsulado

Los paquetes mostrados en esta sección ilustran los conceptos del Tunelización IP donde está el Encapsulation Protocol el GRE y el IP es el Transport Protocol. El protocolo pasajero también es IP. En este caso, IP es tanto el protocolo de transporte como el protocolo pasajero.

Paquete Normal

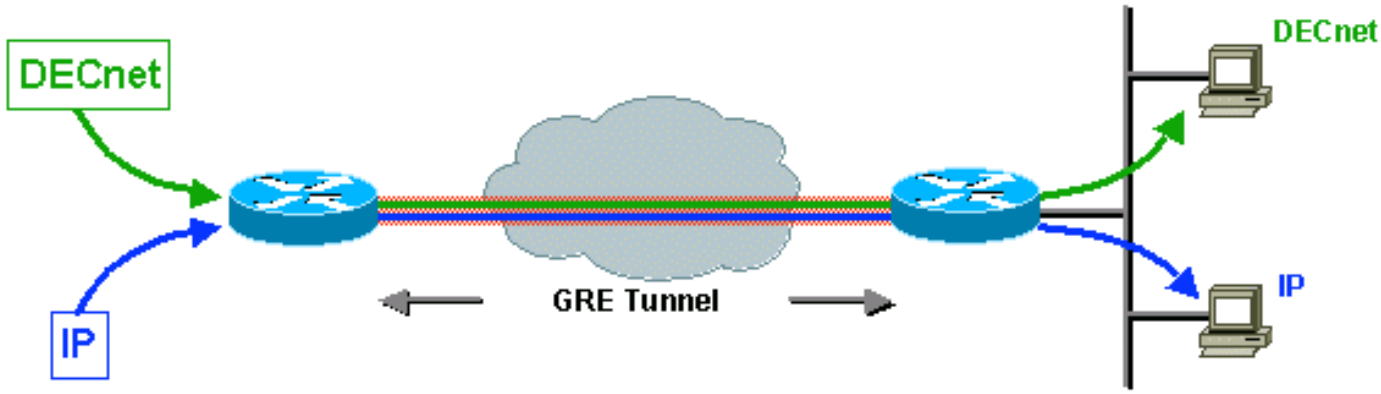
IP TCP Telnet

Paquete de Túnel

IP GRE IP TCP Telnet

- IP es el protocolo de transporte.
- GRE es el protocolo de encapsulación.
- IP es el protocolo pasajero.

En el siguiente ejemplo, se muestra la encapsulación de IP y DECnet como protocolos pasajeros con GRE funcionando como protocolo portador. Esto ilustra el hecho de que el protocolo de la portadora puede encapsular varios protocolos pasajeros.



Un administrador de la red puedo considerar la tunelización en una situación donde haya dos redes discontinuas que no sean IP separadas por una backbone IP. Si las redes no contiguas ejecutan el DECNet, el administrador no pudo querer conectarlas juntas configurando el DECNet en la estructura básica. El administrador no pudo querer permitir que el ruteo DECnet consuma el ancho de banda de estructura básica porque éste podría interferir con el funcionamiento de la red del IP.

Una alternativa viable es usar un túnel para DECnet a través de la backbone IP. El Tunelización encapsula los paquetes del DECNet dentro del IP, y los envía a través de la estructura básica al punto final del túnel donde se quita la encapsulación y los paquetes del DECNet se pueden rutear a su destino vía el DECNet.

El encapsulado del tráfico dentro de otro protocolo proporciona estas ventajas:

- Los puntos finales utilizan a las direcciones privadas (RFC 1918) y la estructura básica no soporta rutear estos direccionamientos.
- Permita redes privadas virtuales (VPN) a través de WAN o Internet.
- Una las redes discontinuas de varios protocolos a través de una backbone de un solo protocolo.
- Cifre el tráfico a través de la backbone o Internet.

Para el resto del documento, el IP se utiliza como el protocolo pasajero y el IP como el Transport Protocol.

Consideraciones Sobre las Interfaces de Túnel

Éstas son consideraciones al hacer un túnel.

- El fast switching de los túneles GRE se introdujo en el Cisco IOS, versión 11.1, y el CEF switching se introdujo en la versión 12.0. El CEF switching para los túneles GRE multipunto se introdujo en la versión 12.2(8)T. La encapsulación y el decapsulation en los puntos finales del túnel eran operaciones lentas en las versiones anteriores del Cisco IOS cuando solamente el process switching fue soportado.
- Hay problemas de topología y seguridad cuando se realiza la tunelización de paquetes. Los túneles pueden saltear listas de control de acceso (ACL) y firewalls. Si usted usa un túnel a través de un firewall, básicamente saltea el firewall para cualquier protocolo pasajero para el que use el túnel. Por lo tanto, se recomienda incluir la funcionalidad de firewall en los extremos de un túnel a fin de hacer cumplir cualquier política en los protocolos pasajeros.
- El Tunelización pudo crear los problemas con los protocolos de transporte que han limitado los temporizadores (por ejemplo, DECNet) debido a la mayor latencia.
- Haciendo un túnel a través de los entornos con diversos links de la velocidad, como los anillos FDDI rápidos y a través de las líneas telefónicas lentas 9600-bps, pudo introducir el paquete que reordenaba los problemas. Algunos protocolos pasajeros funcionan mal en redes de medios combinadas.
- Los túneles punto a punto pueden usar todo el ancho de banda en un link físico. Si usted funciona con los Routing Protocol sobre el punto múltiple para señalar los túneles, tenga presente que cada interfaz del túnel tiene un ancho de banda y que la interfaz física sobre la cual el túnel se ejecuta tiene un ancho de banda. Por ejemplo, desearía configurar el ancho de banda de túnel en 100 KB si hubiera 100 túneles en ejecución a través de un link de 10 MB. El ancho de banda predeterminado para un túnel es 9 Kb.
- Los Routing Protocol pudieron preferir un túnel sobre un link “real” porque el túnel pudo aparecer engañoso ser un link del uno-salto con la trayectoria más barata, aunque implique más saltos y sea realmente realmente más costoso que otra trayectoria. Esto se puede mitigar con una correcta configuración del protocolo de ruteo. Tal vez quiera considerar la ejecución de un protocolo de ruteo diferente sobre una interfaz de túnel, en lugar de ejecutar un protocolo de ruteo en la interfaz física.

- Los problemas de ruteo recurrente pueden evitarse al configurar rutas estáticas apropiadas al destino de túnel. Una ruta recurrente es cuando la mejor trayectoria al "destino de túnel" es a través del mismo túnel. Esta situación hace la interfaz del túnel despedir hacia arriba y hacia abajo. Usted verá este error cuando hay un problema de ruteo recursivo.`%TUN-RECURDOWN Interface Tunnel 0 temporarily disabled due to recursive routing`

El Router como un Participante de PMTUD en el Extremo de un Túnel

Cuando es el extremo de un túnel, el router tiene que realizar dos funciones de PMTUD diferentes.

- En la primera función, el router es el router de reenvío de un paquete de host. Para el procesamiento de PMTUD, el router debe verificar el bit DF y el tamaño de paquete del paquete de datos original, y realizar la acción apropiada, cuando sea necesario.
- La segunda función aparece después de que el router ha encapsulado el paquete IP original dentro del paquete de túnel. En esta etapa, el router actúa más bién un host en cuanto al PMTUD y con respecto al paquete del IP del túnel.

Deja el comienzo mirando qué sucede cuando el router actúa en el primer papel, un router que adelante IP del host los paquetes, en cuanto al PMTUD. Esta función aparece antes de que el router encapsula el paquete IP de host dentro del paquete de túnel.

Si participa el router pues el promotor de un paquete del host él completará estas acciones:

- Verificar si el bit DF está configurado.
- Verificar a qué tamaño de paquete puede adaptarse el túnel.
- Fragmentar (si el paquete es demasiado grande y el bit DF no está configurado), encapsular los fragmentos y enviar. o
- Descartar el paquete (si el paquete es demasiado grande y el bit DF está configurado) y enviar un mensaje de ICMP al remitente.
- Encapsular (si el paquete no es demasiado grande) y enviar.

Genéricamente, hay una elección de encapsulación y entonces una fragmentación (envíe dos fragmentos de la encapsulación) o una fragmentación y entonces una encapsulación (envíe dos fragmentos encapsulados).

Algunos ejemplos que describen a los mecánicos de la encapsulación y fragmentación del paquete del IP y dos escenarios que muestran la interacción del PMTUD y los paquetes que las redes de muestra transversales se detallan en esta sección.

El primer ejemplo muestra qué sucede a un paquete cuando el router (en el origen de túnel) actúa en el papel del router de reenvío. Recuerde que para procesar el PMTUD, el router necesita marcar el bit DF y el tamaño de paquetes del paquete de datos original y tomar la acción apropiada. Este ejemplo utiliza una encapsulación GRE para el túnel. Como puede ser visto, el GRE hace la fragmentación antes de la encapsulación. En los ejemplos posteriores, se muestran situaciones donde se realiza la fragmentación después de la encapsulación.

En el ejemplo 1, el bit DF no está configurado (DF = 0) y la MTU IP de túnel GRE es 1476 (1500 - 24).

Ejemplo 1

1. El router de reenvío (en el origen de túnel) recibe un datagrama de 1500 bytes con el bit DF borrado (DF = 0) del host remitente. Este datagrama está compuesto por un encabezado IP de 20 bytes más una carga útil TCP de 1480 bytes.
2. Debido a que el paquete será demasiado grande para la MTU IP después de agregar la sobrecarga de GRE (24 bytes), el router de reenvío divide el datagrama en dos fragmentos de 1476 (encabezado IP de 20 bytes + contenido IP de 1456 bytes) y 44 bytes (encabezado IP de 20 bytes + contenido IP de 24 bytes); por lo tanto, después de agregar la encapsulación de GRE, el paquete no será más grande que la MTU de interfaz física saliente.
3. El router de reenvío agrega la encapsulación de GRE, que incluye un encabezado GRE de 4 bytes más un encabezado IP de 20 bytes, a cada fragmento del datagrama IP original. Estos dos datagramas IP ahora tienen una longitud de 1500 y 68 bytes y estos datagramas se ven como IP individual datagramas, no como fragmentos.
4. El router de destino del túnel quita la encapsulación GRE de cada fragmento del datagrama original, que sale de dos fragmentos IP de las longitudes 1476 y 24 bytes. Estos fragmentos de datagrama IP reenviarán separadamente por este router al host receptor.
5. El host receptor reensamblará estos dos fragmentos en el datagrama original.

El escenario 5 describe la función del router de reenvío en el contexto de la topología de una red.

En este ejemplo el router actúa en el mismo papel del router de reenvío, pero este vez el bit DF se fija (DF = 1).

Ejemplo 2

1. El router de reenvío en el origen de túnel recibe un datagrama de 1500 bytes con DF = 1 del host remitente.
2. Dado que el bit DF está configurado y que el tamaño del datagrama (1500 bytes) es mayor que la MTU IP de túnel GRE (1476), el router descartará el datagrama y enviará un mensaje de "se necesita fragmentación de ICMP, pero el bit DF está configurado" al origen del datagrama. El mensaje de ICMP alertará al remitente que la MTU es 1476.
3. El host de envío recibe el mensaje ICMP, y cuando vuelve a enviar las informaciones originales él utilizará un IP datagram 1476-byte.
4. Esta longitud de datagrama IP (1476 bytes) es ahora igual en valor a la MTU IP de túnel GRE; por lo tanto, el router agrega la encapsulación de GRE al datagrama IP.
5. El router receptor (en el túnel de destino) elimina la encapsulación de GRE del datagrama IP y lo envía al host receptor.

Ahora podemos mirar qué sucede cuando el router actúa en el segundo papel como host de envío en cuanto al PMTUD y con respecto al paquete del IP del túnel. Recuerde que esta función aparece después de que el router ha encapsulado el paquete IP original dentro del paquete de túnel.

> Nota: Por abandono un router no hace el PMTUD en los paquetes de túnel GRE que genera. Se puede utilizar el **comando tunnel path-mtu-discovery** para activar PMTUD para los paquetes de túnel IP GRE.

El ejemplo 3 muestra qué sucede cuando el host envía los datagramas IP que son bastante pequeños caber dentro del IP MTU en la interfaz de túnel GRE. El bit DF en este caso puede configurarse o borrarse (1 o 0). La interfaz de túnel GRE no tiene el **comando tunnel path-mtu-**

discovery configuró así que el router no hará el PMTUD en el paquete GRE-IP.

Ejemplo 3

1. El router de reenvío en el origen de túnel recibe un datagrama de 1476 bytes del host remitente.
2. Este router encapsula el datagrama IP de 1476 bytes dentro de GRE para obtener un datagrama IP GRE de 1500 bytes. El bit DF en el encabezado IP GRE se borrará (DF = 0). Luego, este router reenvía este paquete al destino de túnel.
3. Suponga que hay un router entre el origen y el destino de túnel con una MTU de link de 1400. Este router fragmentará el paquete de túnel, ya que se borró el bit DF (DF = 0). Recuerde que este ejemplo fragmenta el IP externo; por lo tanto, los encabezados TCP, IP interno y GRE solo aparecerán en el primer fragmento.
4. El router de destino de túnel debe reensamblar el paquete de túnel GRE.
5. Una vez que se haya reensamblado el paquete de túnel GRE, el router quita el encabezado IP GRE y envía el datagrama IP original.

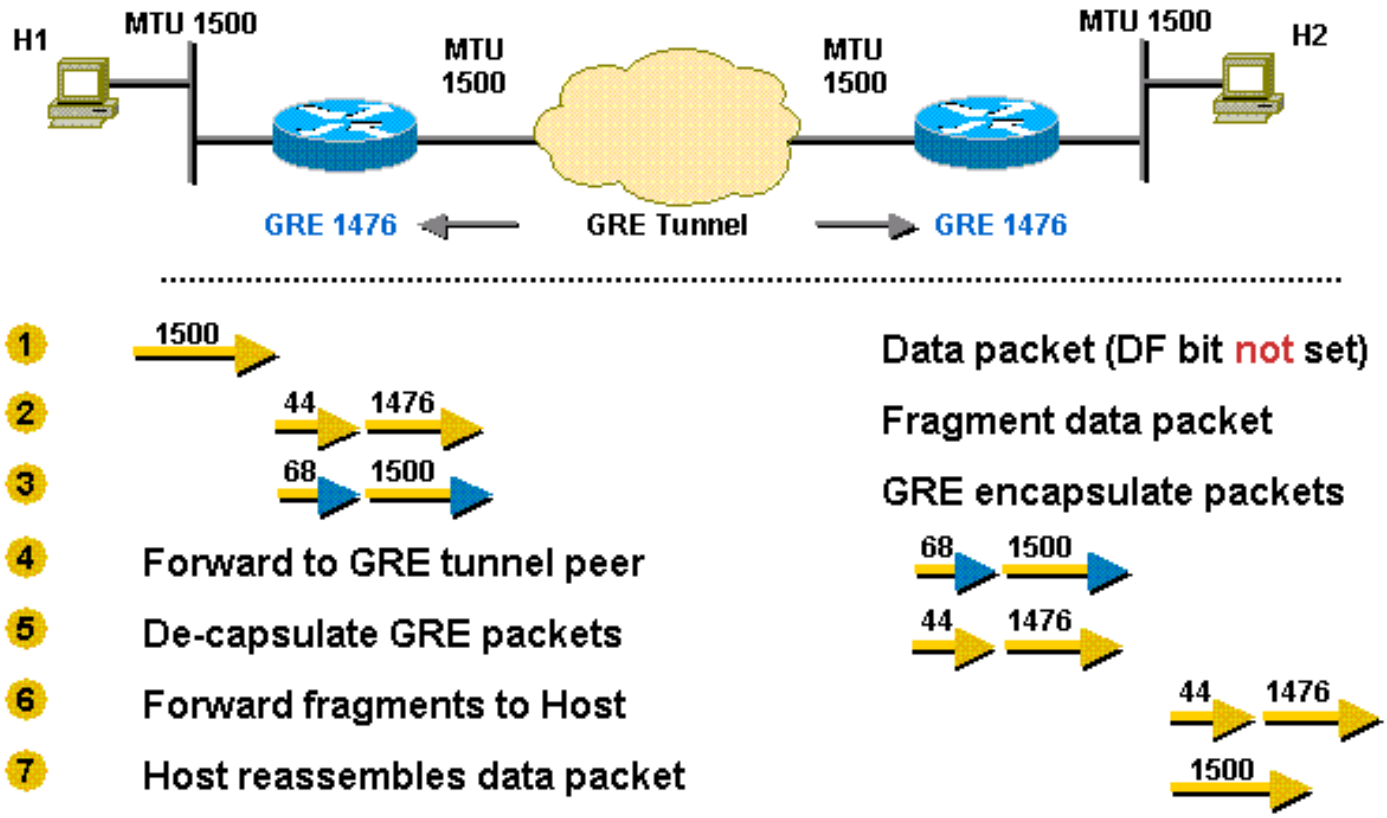
El próximo ejemplo muestra qué sucede cuando el router actúa en el papel de un host de envío en cuanto al PMTUD y con respecto al paquete del IP del túnel. Esta vez el bit DF se fija (se ha configurado el DF = 1) en el encabezado IP original y el **comando tunnel path-mtu-discovery** de modo que el bit DF sea copiado del encabezado IP interno (GRE +IP) a la encabezado externa.

Ejemplo 4

1. El router de reenvío en el origen de túnel recibe un datagrama de 1476 bytes con DF = 1 del host remitente.
2. Este router encapsula el datagrama IP de 1476 bytes dentro de GRE para obtener un datagrama IP GRE de 1500 bytes. Este encabezado IP GRE tendrá el bit DF configurado (DF = 1), ya que el datagrama IP original tenía el bit DF configurado. Luego, este router reenvía este paquete al destino de túnel.
3. Una vez más, suponga que hay un router entre el origen y el destino de túnel con una MTU de link de 1400. Este router no fragmentará el paquete de túnel porque el bit DF está configurado (DF = 1). Este router debe descartar el paquete y enviar un mensaje de error de ICMP al router de origen de túnel, ya que esa es la dirección IP de origen en el paquete.
4. El router de reenvío en el origen del túnel recibe este mensaje de error ICMP y reducirá el túnel IP MTU de GRE a 1376 (1400 - 24). La próxima vez que el host remitente retransmita los datos en un paquete IP de 1476 bytes, este paquete será muy grande y este router enviará un mensaje de error de ICMP al remitente con un valor de MTU de 1376. Cuando el host remitente retransmita los datos, los enviará en un paquete IP de 1376 bytes y este paquete pasará a través del túnel GRE al host receptor.

Situación 5

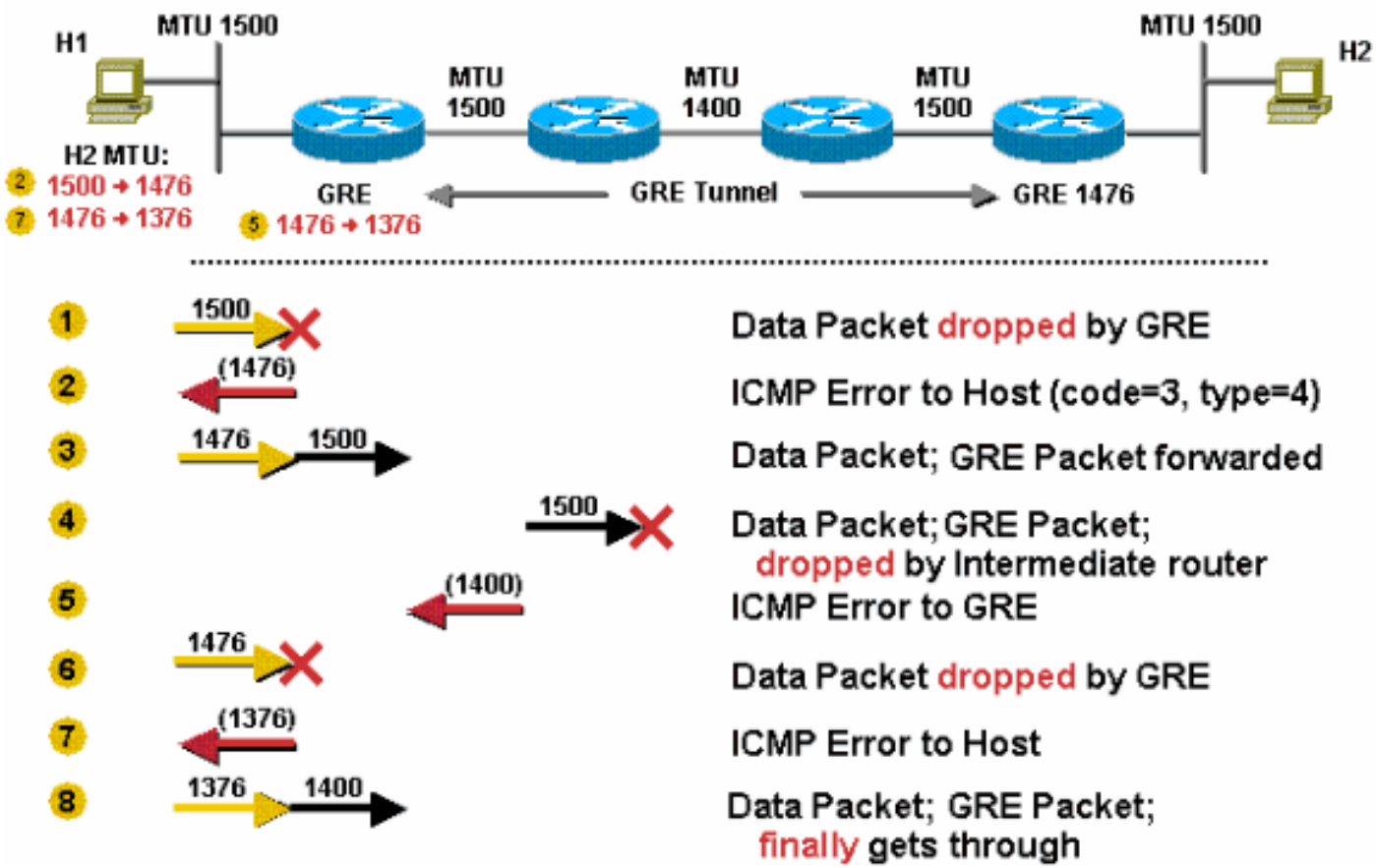
Este escenario ilustra la fragmentación de GRE. Recuerde que usted fragmenta antes de la encapsulación para GRE, después realiza la PMTUD para el paquete de datos, y el bit DF no se copia cuando el paquete IP es encapsulado por GRE. En esta situación, el bit DF no está configurado. Por defecto, la MTU de IP de la interfaz de túnel GRE es 24 bytes menos que la MTU de IP de la interfaz física, por lo que la MTU de IP de la interfaz GRE es 1476.



1. El remitente envía un paquete 1500-byte (byte ip encabezado 20 + 1480 bytes de carga útil de TCP).
2. Como la MTU del túnel GRE es 1476, el paquete de 1500 bytes se divide en dos fragmentos IP de 1476 y 44 bytes, cada uno anticipándose a los 24 bytes adicionales del encabezado GRE.
3. Los 24 bytes del encabezado GRE se agregan a cada fragmento IP. Ahora, los fragmentos son de 1500 (1476 + 24) y 68 (44 + 24) bytes cada uno.
4. Los paquetes GRE +IP que contienen los dos fragmentos IP se remiten al router del par del túnel GRE.
5. El router de peer de túnel GRE quita los encabezados GRE de los dos paquetes.
6. Este router reenvía ambos paquetes al host de destino.
7. El host de destino reensambla los fragmentos IP nuevamente en el datagrama IP original.

Situación 6

Este escenario es similar al escenario 5, pero este vez el bit DF se fija. En la situación 6, el router está configurado para realizar la PMTUD en los paquetes de túnel IP + GRE con el **comando tunnel path-mtu-discovery** y el bit DF se copia del encabezado IP original al encabezado IP GRE. Si el router recibe un error de ICMP para el paquete IP + GRE, reduce la MTU IP en la interfaz de túnel GRE. Una vez más, recuerde que la MTU IP de túnel GRE está configurada en 24 bytes menos que la MTU de interfaz física de forma predeterminada, así que la MTU IP GRE aquí es 1476. También observe que hay un link de MTU de 1400 en la trayectoria de túnel GRE.



1. El router recibe un paquete de 1500 bytes (encabezado IP de 20 bytes + contenido TCP de 1480 bytes) y descarta el paquete. El router cae el paquete porque es más grande que el IP MTU (1476) en la interfaz de túnel GRE.
2. El router envía un error de ICMP al remitente comunicándole que la MTU de salto siguiente es 1476. El host registrará esta información, generalmente como una ruta de host para el destino, en su tabla de ruteo.
3. El host remitente utiliza un tamaño de paquete de 1476 bytes cuando reenvía los datos. El router GRE agrega 24 bytes de encapsulación GRE y envía un paquete de 1500 bytes.
4. El paquete de 1500 bytes no puede atravesar el link de 1400 bytes y, por eso, será descartado por el router intermedio.
5. El router intermedio envía un ICMP (tipo = 3, código = 4) al router GRE con un Next-Hop MTU de 1400. El router GRE reduce esto a 1376 (1400 - 24) y configura un valor de MTU IP interno en la interfaz GRE. Este cambio solo se puede ver usando el **comando debug tunnel**; no se puede ver en el resultado del **comando show ip interface tunnel<#>**.
6. La próxima vez el host vuelve a enviar el paquete 1476-byte, el router GRE caerá el paquete, puesto que es más grande que el IP actual MTU (1376) en la interfaz de túnel GRE.
7. El router GRE enviará otro ICMP (el tipo = 3, código = 4) al remitente con un Next-Hop MTU de 1376 y el host pondrá al día su información actual con el nuevo valor.
8. El host reenvía nuevamente los datos, pero ahora en un paquete menor de 1376 bytes; GRE agregará 24 bytes de encapsulación y lo reenviará. Esta vez el paquete lo hará al par del túnel GRE, donde estará decapsulated y enviado el paquete a la computadora principal de destino.Nota: Si no se configurara el **comando tunnel path-mtu-discovery** en el router de reenvío en esta situación y el bit DF estuviera configurado en los paquetes reenviados a través del túnel GRE, el Host 1 podría enviar igualmente los paquetes TCP/IP al Host 2, pero no podrían fragmentarse en el medio en el link de MTU de 1400. Además, el peer de túnel

GRE debería reensamblarlos para poder desencapsularlos y reenviarlos.

Modo de Túnel IPSec "Puro"

El protocolo de la seguridad IP (IPSec) es un método de estándares que proporciona la aislamiento, la integridad, y la autenticidad a la información transferida a través de las redes del IP. IPSec proporciona cifrado de capa de red IP. IPSec alarga el paquete IP agregando por lo menos un encabezado IP (modo de túnel). La encabezado agregada varía de largo al dependiente en el modo de la configuración IPSec pero ella no excede ~58 bytes (Encapsulating Security Payload (ESP) y autenticación ESP (ESPauth)) por paquete.

IPSec tiene dos modos: modo de túnel y modo de transporte.

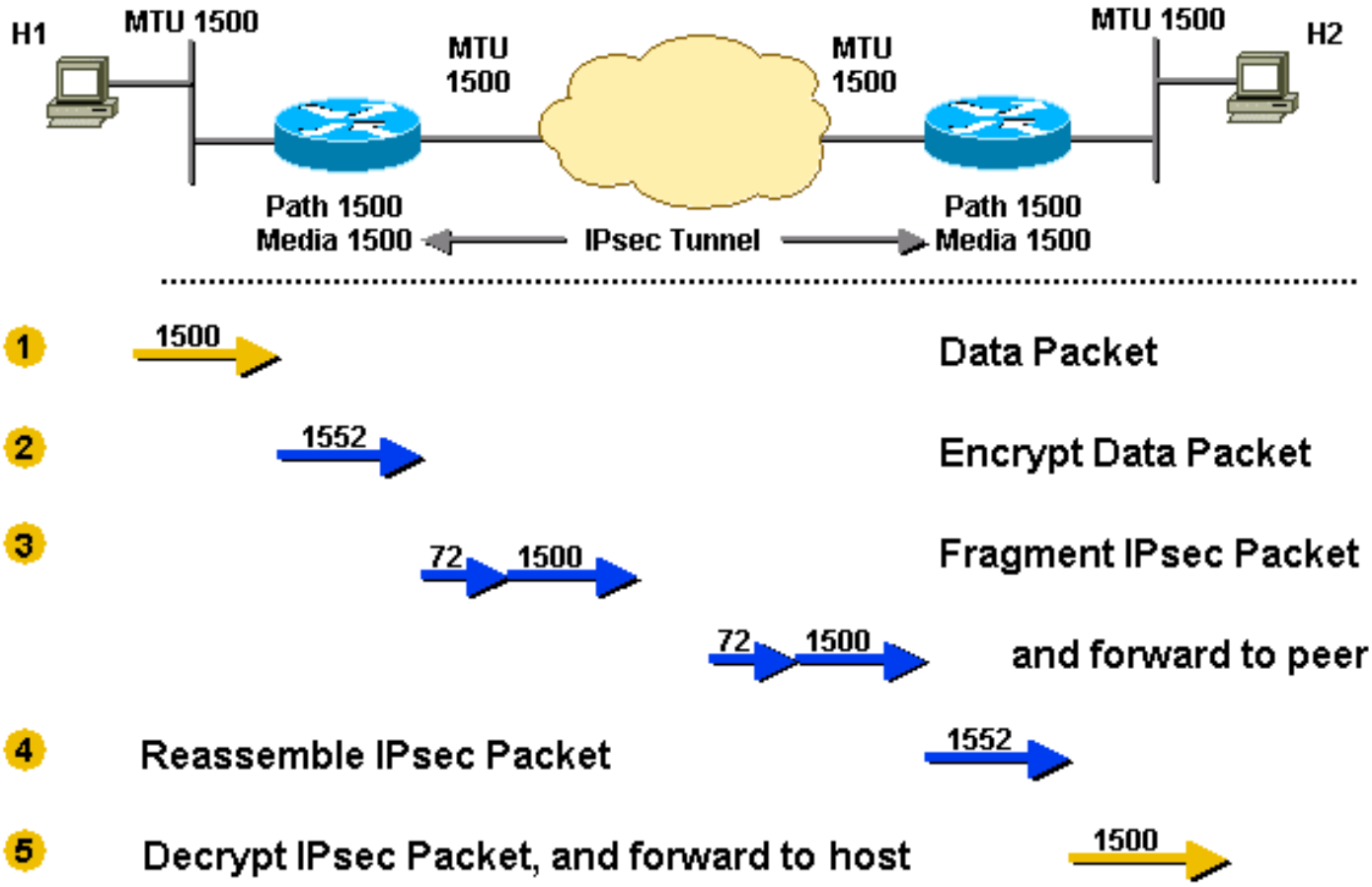
- El modo de túnel es el modo predeterminado. Con el modo de túnel, el paquete IP original entero es protegido (cifrado, autenticado o ambas opciones) y es encapsulado por los encabezados y las colas IPSec. Entonces un nuevo encabezado IP prepended al paquete, que los specifes los puntos finales de IPSec (pares) como la fuente y el destino. El modo de túnel puede ser usado con todo tipo de tráfico de unidifusión de IP y debe ser usado si IPsec está protegiendo al tráfico de los hosts ubicados detrás de los pares de IPsec. Por ejemplo, el modo de túnel se utiliza con las redes privadas virtuales (VPN) donde los hosts en una red protegida envían paquetes a hosts en una diferente red protegida vía un par de peers IPSec. Con VPN, el "túnel" IPsec protege el tráfico IP entre los hosts cifrándolo entre los routers de par IPsec.
- Con el modo transporte (configurado con el subcomando **mode transport** en la definición de transformación), sólo la carga útil del paquete IP original está protegida (cifrada, autenticada o ambas). Las colas y los encabezados IPSec encapsulan el contenido. Los encabezados IP originales permanecen intactos, salvo que el campo de protocolo IP se cambie a ESP (50), y el valor del protocolo original se guarda en la cola IPSec que se restablecerá cuando se descifre el paquete. El modo de transporte se utiliza solo cuando el tráfico IP que se desea proteger se encuentra entre los mismos peers IPSec; las direcciones IP de origen y de destino en el paquete son las mismas que las direcciones de peer IPSec. Normalmente, el modo de transporte IPSec solo se utiliza cuando otro protocolo de tunelización (como GRE) se utiliza para encapsular primero el paquete de datos IP y luego IPSec se utiliza para proteger los paquetes de túnel GRE.

IPSec siempre realiza la PMTUD para los paquetes de datos y para sus propios paquetes. Existen comandos de configuración de IPSec para modificar el procesamiento de PMTUD para los paquetes IP IPSec; IPSec puede borrar, configurar o copiar el bit DF del encabezado IP del paquete de datos al encabezado IP IPSec. Esta función se denomina "funcionalidad de invalidación del bit DF".

> Nota: Realmente se desea evitar la fragmentación después de la encapsulación cuando se realiza el cifrado del hardware con IPSec. El cifrado del hardware puede darle un rendimiento de 50 MB aproximadamente, según el hardware, pero si se fragmenta el paquete IPSec, pierde del 50 al 90 por ciento del rendimiento. Esta pérdida se produce debido a que los paquetes IPsec fragmentados son conmutados por proceso para su reensamblado y luego transferidos al motor de encripción de hardware para ser desencripcións. Esta pérdida de rendimiento puede bajar el rendimiento del cifrado del hardware al nivel de rendimiento del cifrado del software (2 - 10 MB).

Situación 7

En esta situación, se representa la fragmentación de IPSec en acción. En esta situación, la MTU junto con la trayectoria entera es 1500. En esta situación, el bit DF no está configurado.

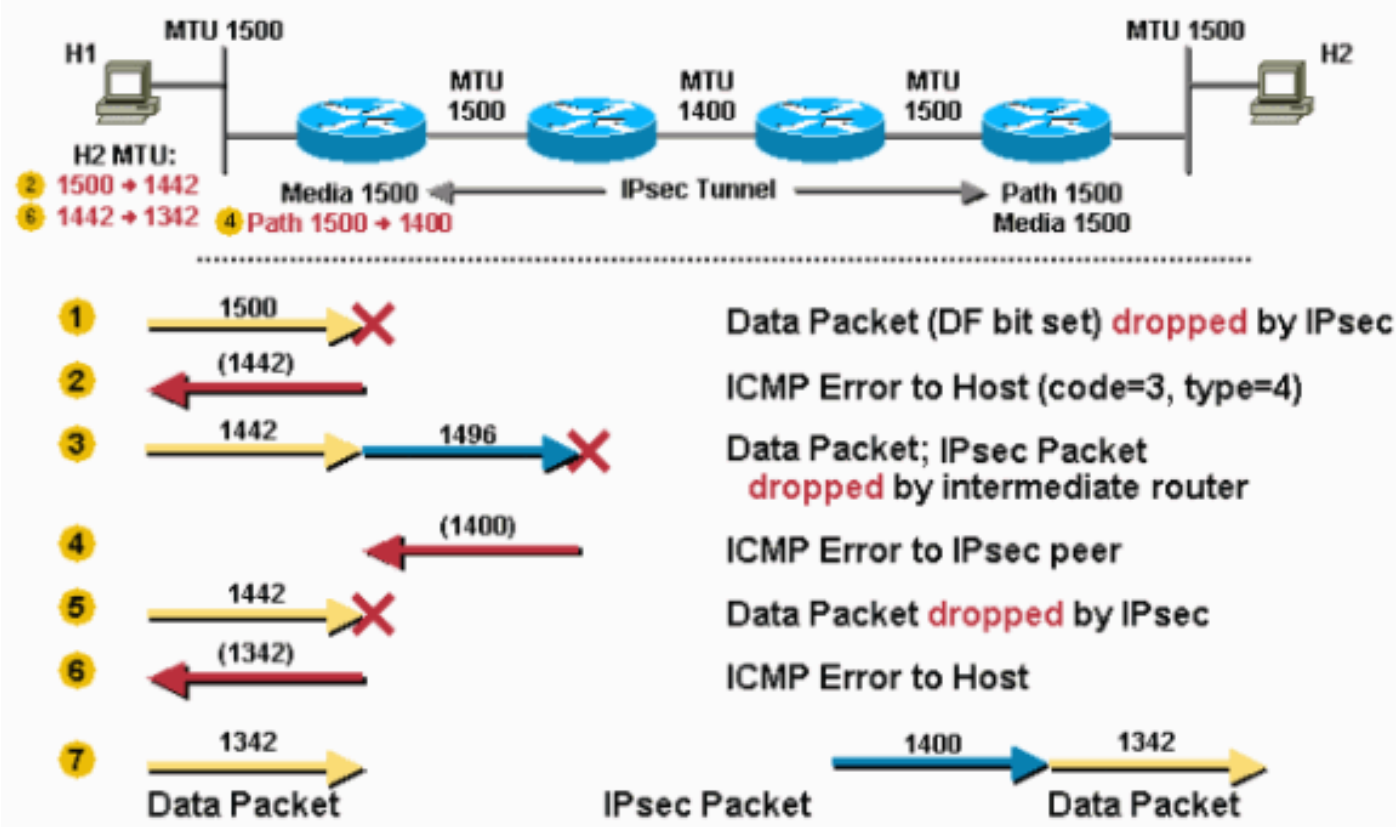


1. El router recibe un paquete de 1500 bytes (encabezado IP de 20 bytes + carga útil TCP de 1480 bytes) destinado al Host 2.
2. IPsec cifra el paquete de 1500 bytes y se agregan 52 bytes de sobrecarga (encabezado IPsec, cola y encabezado IP adicional). Ahora, IPSec necesita enviar un paquete de 1552 bytes. Dado que el MTU saliente es 1500, este paquete tendrá que ser fragmentado.
3. Dos fragmentos se crean a partir del paquete IPsec. Durante la fragmentación, se agrega un encabezado IP de 20 bytes adicional para el segundo fragmento, lo que da como resultado un fragmento de 1500 bytes y un fragmento IP de 72 bytes.
4. El router de peer de túnel IPSec recibe los fragmentos, quita el encabezado IP adicional y une los fragmentos IP nuevamente en el paquete IPSec original. Después, IPSec descifra este paquete.
5. Luego, el router reenvía el paquete de datos original de 1500 bytes al Host 2.

Situación 8

Esta situación es similar a la situación 6, salvo que en este caso el bit DF está configurado en el paquete de datos original y hay un link en la trayectoria entre los peers de túnel IPSec que tiene una MTU inferior en comparación con los otros links. En esta situación, se muestra cómo el router de peer IPSec realiza ambas funciones de PMTUD, como se describe en la sección El Router como un Participante de PMTUD en el Extremo de un Túnel.

Usted verá en esta situación cómo la PMTU IPSec cambia a un valor inferior como resultado de la necesidad de fragmentación. Recuerde que el bit DF se copia del encabezado IP interno al encabezado IP externo cuando IPsec cifra un paquete. Los valores de PMTU y MTU de medios se almacenan en la asociación de seguridad (SA) IPSec. La MTU de medios se basa en la MTU de interfaz de router saliente y la PMTU se basa en la MTU mínima vista en la trayectoria entre los peers IPSec. Recuerde que IPSec encapsula/cifra el paquete antes de intentar fragmentarlo.



1. El router recibe un paquete de 1500 bytes y lo descarta porque la sobrecarga de IPSec, cuando se agrega, hará que el paquete sea más grande que la PMTU (1500).
2. El router envía un mensaje de ICMP al Host 1 comunicándole que la MTU de salto siguiente es 1442 (1500 - 58 = 1442). Estos 58 bytes son la tasa máxima de IPsec cuando se utiliza IPsec ESP y ESPauth. El consumo de recursos de IPsec real puede ser tanto como 7 bytes menos que este valor. El Host 1 registra esta información, generalmente como una ruta de host para el destino (Host 2), en su tabla de ruteo.
3. El Host 1 disminuye su PMTU para el Host 2 a 1442, así que el Host 1 enviará paquetes más pequeños (de 1442 bytes) cuando retransmita los datos al Host 2. El router recibe el paquete de 1442 bytes e IPSec agrega 52 bytes de sobrecarga de cifrado; por lo tanto, el resultado es un paquete IPSec de 1496 bytes. Porque este paquete tiene el bit DF configurado en su encabezado, es descartado por el router del medio con el link de MTU de 1400 bytes.
4. El router del medio que descartó el paquete envía un mensaje de ICMP al remitente del paquete IPSec (el primer router) comunicándole que la MTU de salto siguiente es 1400 bytes. Este valor se registra en la PMTU SA IPSec.
5. La próxima vez que el Host 1 retransmita el paquete de 1442 bytes (si no recibió reconocimiento), IPSec descartará el paquete. El router caerá otra vez el paquete porque la tara Ipsec, cuando está agregada al paquete, lo hará más grande que el PMTU (1400).
6. El router envía un mensaje de ICMP al Host 1 comunicándole que la MTU de salto siguiente ahora es 1342. (1400 - 58 = 1342). El Host 1 registrará nuevamente esta información.

7. Cuando el Host 1 retransmita otra vez los datos, utilizará el paquete con el tamaño más pequeño (1342). Este paquete no requerirá fragmentación y pasará a través del túnel IPSec al Host 2.

GRE e IPSec Juntos

Más interacciones complejas para la fragmentación y PMTUD ocurren cuando el IPSec se utiliza para cifrar los túneles GRE. El IPSec y el GRE se combinan de este modo porque el IPSec no soporta los paquetes del Multicast IP, así que significa que usted no puede funcionar con un Dynamic Routing Protocol sobre la red del IPSec VPN. Los túneles GRE soportan multicast; por lo tanto, un túnel GRE se puede utilizar para primero encapsular el paquete de multicast de protocolo de ruteo dinámico en un paquete de unicast IP GRE, que luego puede ser cifrado por IPSec. Cuando se realiza esto, IPSec frecuentemente se implementa en el modo de transporte sobre GRE porque los peers IPSec y los extremos de túnel GRE (los routers) son los mismos y el modo de transporte guardará 20 bytes de sobrecarga de IPSec.

Un caso interesante ocurre cuando un paquete IP se divide en dos fragmentos y GRE lo encapsula. En este caso, IPSec verá dos paquetes independientes IP + GRE. A menudo, en una configuración predeterminada, uno de estos paquetes será muy grande y necesitará ser fragmentado después de su cifrado. El peer IPSec deberá reensamblar este paquete antes de descifrarlo. Esta "doble fragmentación" (una vez antes de GRE y otra vez después de IPSec) en el router remitente aumenta la latencia y baja el rendimiento. Además, el reensamblado se convierte en process-switched; por lo tanto, habrá un impacto en el CPU en el router receptor siempre que suceda esto.

Se puede evitar esta situación al configurar la "MTU IP" en la interfaz de túnel GRE lo suficientemente baja como para considerar la sobrecarga de GRE e IPSec (de forma predeterminada, la "MTU IP" de interfaz de túnel GRE se configura en la MTU de interfaz real saliente, los bytes de sobrecarga de GRE).

Esta tabla enumera los valores sugeridos MTU para cada túnel/combinación del modo que asumen que la interfaz de física saliente tenga un MTU de 1500.

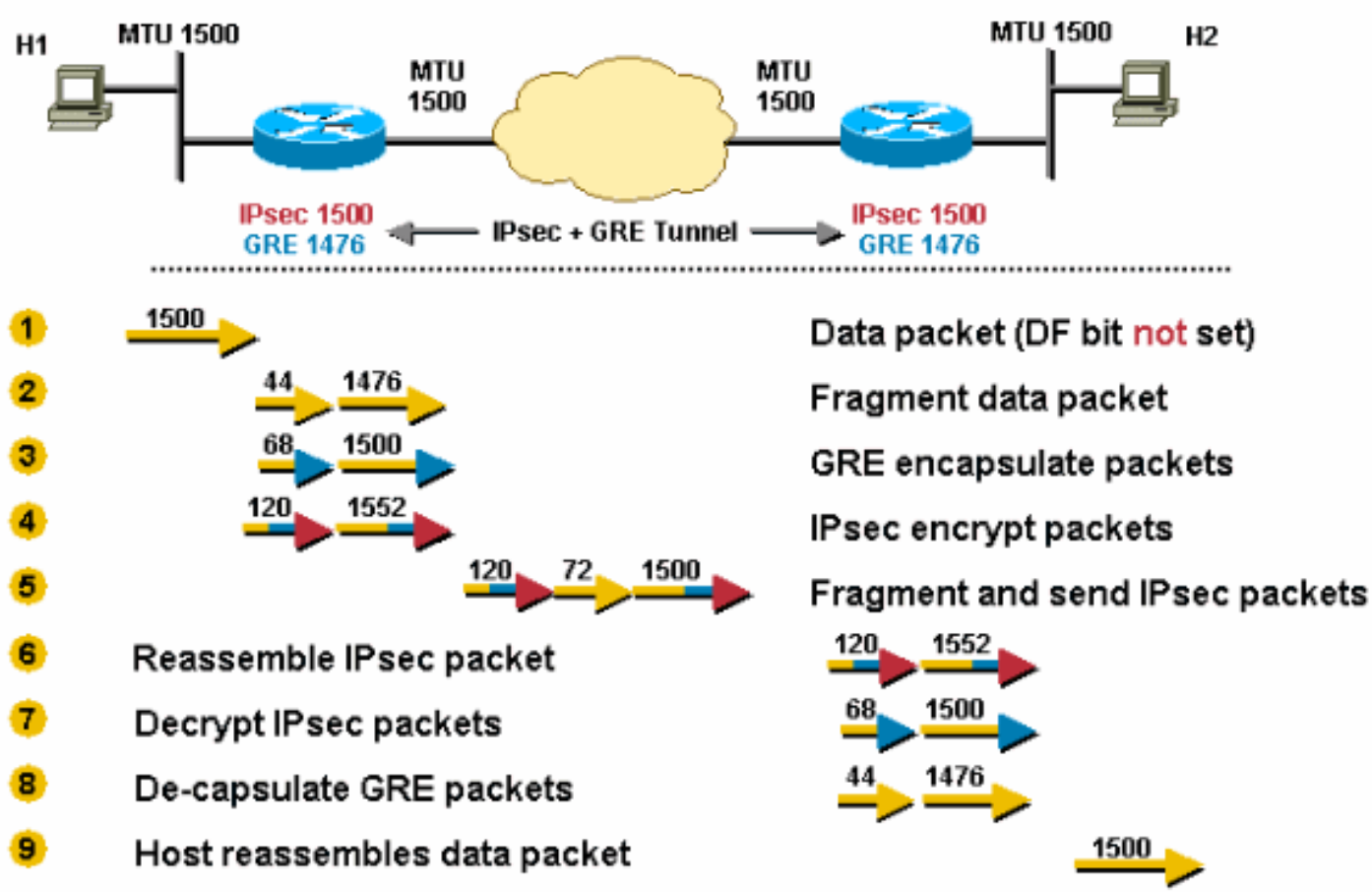
Combinación de Túneles	MTU Específica Necesaria	MTU Recomendada
GRE + IPSec (modo de transporte)	1440 bytes	1400 bytes
GRE + IPSec (modo de túnel)	1420 bytes	1400 bytes

> Nota: Se recomienda el valor de MTU de 1400 porque cubre las combinaciones de modos IPSec + GRE más comunes. Además, no hay desventaja notable en permitir una sobrecarga adicional de 20 o 40 bytes. Es más fácil recordar y configurar un valor y este valor cubre casi todas las situaciones.

Escenario 9

IPSec se implementa sobre GRE. La MTU física saliente es 1500, la PMTU IPSec es 1500 y la MTU IP GRE es 1476 (1500 - 24 = 1476). Debido a esto, los paquetes TCP/IP se fragmentarán dos veces, una vez antes de GRE y otra vez después de IPSec. El paquete se fragmentará antes de la encapsulación de GRE y uno de estos paquetes GRE se volverá a fragmentar después del cifrado IPSec.

La configuración de "MTU IP 1440" (modo de transporte IPSec) o "MTU IP 1420" (modo de túnel IPSec) en el túnel GRE quitaría la posibilidad de la doble fragmentación en esta situación.



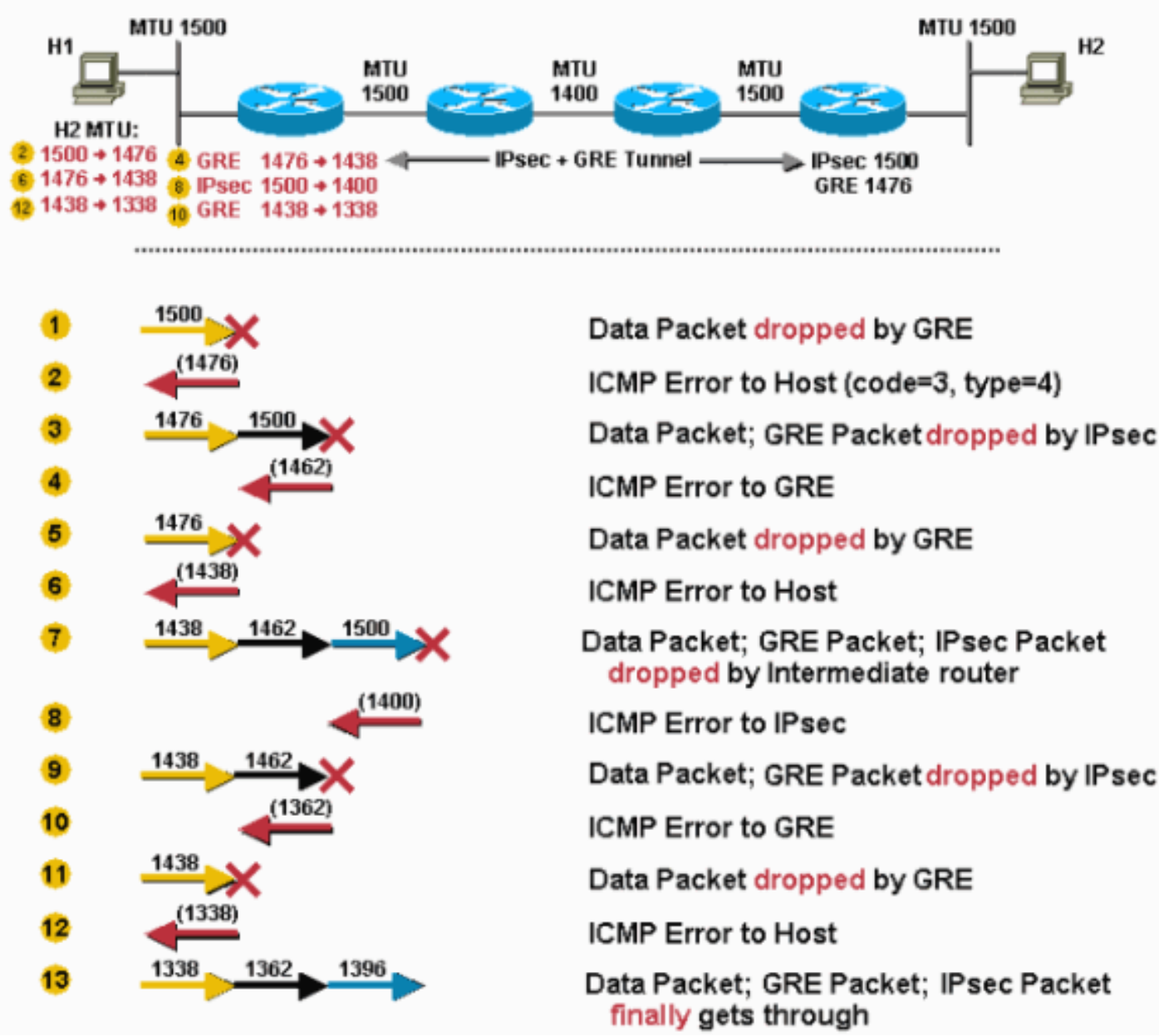
1. El router recibe un datagrama de 1500 bytes.
2. Antes de la encapsulación, GRE fragmenta el paquete de 1500 bytes en dos partes: uno de 1476 bytes (1500 - 24 = 1476) y otro de 44 bytes (24 bytes de datos + 20 bytes de encabezado IP).
3. GRE encapsula los fragmentos IP, lo que agrega 24 bytes a cada paquete. Como resultado, se obtienen dos paquetes IPSec + GRE de 1500 bytes (1476 + 24 = 1500) y 68 bytes (44 + 24).
4. El IPSec cifra los dos paquetes, agregando 52 byes (modo túnel del IPSec) de la tara de encapsulación a cada uno, para dar un 1552-byte y un paquete del 120-byte.
5. El paquete IPsec 1552-byte es hecho fragmentos por el router porque es más grande que el MTU saliente (1500). El paquete de 1552 bytes se divide en partes: un paquete de 1500 bytes y un paquete de 72 bytes (una "carga útil" de 52 bytes más un encabezado IP adicional de 20 bytes para el segundo fragmento). Los tres paquetes de 1500 bytes, 72 bytes y 120 bytes se reenvían al peer IPSec + GRE.
6. El router de recepción vuelve a montar los dos fragmentos del IPSec (1500 bytes y 72 bytes) para conseguir el paquete original del IPsec+GRE 1552-byte. No se debe realizar ninguna tarea para el paquete IPSec + GRE de 120 bytes.
7. El IPSec desencripta 1552-byte y los paquetes del IPsec+GRE del 120-byte para conseguir los Paquetes GRE 1500-byte y 68-byte.
8. Decapsulates GRE los Paquetes GRE 1500-byte y 68-byte para conseguir los fragmentos del paquete del IP 1476-byte y 44-byte. Estos fragmentos de paquete IP se reenvían al host de destino.
9. El host 2 vuelve a montar estos fragmentos IP para conseguir el IP datagram original 1500-

byte.

La situación 10 es similar a la situación 8, salvo que hay un link de MTU inferior en la trayectoria de túnel. Esta es la situación del "peor caso" para el primer paquete enviado del Host 1 al Host 2. Después del último paso en esta situación, el Host 1 configura la PMTU correcta para el Host 2 y todo funciona bien para las conexiones TCP entre el Host 1 y el Host 2. Los flujos de TCP entre el Host 1 y otros hosts (accesibles vía el túnel IPSec + GRE) solo tendrán que atravesar los últimos tres pasos de la situación 10.

En este escenario, configuran al **comando tunnel path-mtu-discovery** en el túnel GRE y el bit DF se fija en los paquetes TCP/IP que originan del host 1.

Situación 10



1. El router recibe un paquete de 1500 bytes. GRE descarta este paquete porque no puede fragmentar ni reenviar el paquete, ya que el bit DF está configurado y el tamaño del paquete excede la "MTU IP" de interfaz saliente una vez que se agrega la sobrecarga de GRE (24 bytes).
2. El router envía un mensaje ICMP para recibir 1 para dejarlo saber que el Next-Hop MTU es

1476 (1500 - 24 = 1476).
3. El Host 1 cambia su PMTU para el Host 2 a 1476 y envía el tamaño más pequeño cuando retransmite el paquete. GRE lo encapsula y entrega el paquete de 1500 bytes a IPSec. IPSec descarta el paquete porque GRE ha copiado el bit DF (configurado) del encabezado IP interno y, con la sobrecarga de IPSec (un máximo de 38 bytes), el paquete es demasiado grande para su reenvío a través de la interfaz física.
4. El IPSec envía un mensaje ICMP al GRE que indica que el Next-Hop MTU es 1462 bytes (puesto que un máximo 38 bytes será agregado para el cifrado y el IP por encima). GRE registra el valor 1438 (1462 - 24) como la "MTU IP" en la interfaz de túnel.Nota: Este cambio en el valor se almacena internamente y no se puede ver en el resultado del **comando show ip interface tunnel<#>**. Solo verá ese cambio si utiliza el **comando debug tunnel**.
5. La próxima vez que el Host 1 retransmita el paquete de 1476 bytes, GRE lo descartará.
6. El router envía un mensaje ICMP para recibir 1 que indique que 1438 es el Next-Hop MTU.
7. El Host 1 disminuye la PMTU para el Host 2 y retransmite un paquete de 1438 bytes. Esta vez, GRE acepta el paquete, lo encapsula y lo entrega a IPSec para su cifrado. El paquete IPsec se reenvía al router intermedio y deja de transmitirse porque tiene una interfaz saliente MTU de 1400.
8. El router intermedio envía un mensaje ICMP al IPSec que le diga que el Next-Hop MTU es 1400. Este valor es registrado por IPSec en el valor de PMTU de la SA IPSec asociada.
9. Cuando el host 1 retransmite el paquete de 1438 bytes, GRE lo encapsula y lo entrega al IPsec. IPSec descarta el paquete porque ha cambiado su propia PMTU a 1400.
10. El IPSec envía un error ICMP al GRE que indica que el Next-Hop MTU es 1362, y el GRE registra el valor 1338 internamente.
11. Cuando el Host 1 retransmite el paquete original (porque no recibió reconocimiento), GRE lo descarta.
12. El router envía un mensaje ICMP para recibir 1 que indique que el Next-Hop MTU es 1338 (1362 - 24 bytes). El Host 1 disminuye su PMTU para el Host 2 a 1338.
13. El Host 1 retransmite un paquete de 1338 bytes y esta vez puede pasarlo finalmente a través del Host 2.

Más Recomendaciones

La configuración del **comando tunnel path-mtu-discovery** en una interfaz de túnel puede ayudar con la interacción de IPSec y GRE cuando se configuran en el mismo router. Recuerde que sin el **comando tunnel path-mtu-discovery** configurado, el bit DF siempre se borraría en el encabezado IP GRE. Esto permite el paquete del IP GRE sea hecho fragmentos aunque el encabezado IP de los datos encapsulados tenía el conjunto de bits DF, que no permitiría normalmente que el paquete fuera hecho fragmentos.

Si configuran al **comando tunnel path-mtu-discovery** en la interfaz de túnel GRE, ésta sucederá.

1. GRE copiará el bit DF del encabezado IP de datos al encabezado IP GRE.
2. Si el bit DF está configurado en el encabezado IP GRE y el paquete será "demasiado grande" después del cifrado de IPSec para la MTU IP en la interfaz física saliente, IPSec descartará el paquete y notificará al túnel GRE que reduzca su tamaño de MTU IP.
3. El IPSec hace el PMTUD para sus propios paquetes y si el IPSec PMTU cambia (si se reduce), después IPSec no notifica inmediatamente el GRE, pero cuando viene otro paquete “demasiado grande” completo, después el proceso en el paso 2 ocurre.

4. La MTU IP de GRE ahora es más pequeña; por lo tanto, descartará cualquier paquete IP de datos con el bit DF configurado que sea demasiado grande y enviará un mensaje de ICMP al host remitente.

El **comando tunnel path-mtu-discovery** ayuda a que la interfaz GRE establezca su MTU IP dinámicamente, en lugar de estáticamente con el **comando ip mtu**. En realidad, se recomienda utilizar ambos comandos. El **comando ip mtu** se utiliza para proporcionar espacio para las sobrecargas de IPSec y de GRE relativas a la MTU IP de interfaz física saliente local. El **comando tunnel path-mtu-discovery** permite que se reduzca incluso más la MTU IP de túnel GRE si hay un link de MTU IP inferior en la trayectoria entre los peers IPSec.

A continuación, se incluyen algunas medidas que puede tomar si tiene problemas con la PMTUD en una red donde hay túneles GRE + IPSec configurados.

Esta lista comienza con la mayoría de la solución deseable.

- Repare el problema con el PMTUD que no trabaja, que es causado generalmente por un router o un Firewall que bloquee el ICMP.
- Utilice el **comando ip tcp adjust-mss** en las interfaces de túnel para que el router reduzca el valor de MSS de TCP en el paquete SYN de TCP. Esto ayudará a que los dos host extremos (el receptor y el remitente de TCP) utilicen paquetes lo suficientemente pequeños para que no se necesite PMTUD.
- Utilice el ruteo basado en políticas en la interfaz de ingreso del router y configure un mapa de ruta para borrar el bit DF en el encabezado IP de datos antes de que llegue a la interfaz de túnel GRE. Esto permitirá que el paquete IP de datos sea fragmentado antes de la encapsulación de GRE.
- Aumente la "MTU IP" en la interfaz de túnel GRE para que sea igual a la MTU de interfaz saliente. Esto permitirá que se aplique la encapsulación de GRE al paquete IP de datos sin la fragmentación primero. El paquete GRE será cifrado por IPSec y después será fragmentado para salir de la interfaz física saliente. En este caso, usted no configuraría el **comando tunnel path-mtu-discovery** en la interfaz de túnel GRE. Esto puede reducir de manera significativa el rendimiento porque el reensamblado de paquetes IP en el peer IPSec se realiza en el modo de process switching.

Información Relacionada

- Página de Soporte de IP Routing
- Página de Soporte de IPSec (Protocolo de Seguridad IP)
- Calculadora de la tara Ipsec (calcule el tamaño de paquetes con los protocolos de la encapsulación de IPSec)
- RFC 1191:Detección de MTU de Trayectoria
- RFC 1063: Opciones de Detección de MTU IP
- RFC 791 Internet Protocol
- RFC 793 Protocolo de Control de Transmisión
- RFC 879 Tamaño Máximo de Segmento de TCP y Temas Relacionados
- RFC 1701 Generic Routing Encapsulation (GRE)
- RFC 1241 Esquema de un Protocolo de Encapsulación de Internet
- RFC 2003 RFC 2003 Encapsulación de IP dentro de IP
- Soporte Técnico - Cisco Systems