DOCUMENTO TÉCNICO

Magic Transit de Cloudflare — Arquitectura de referencia

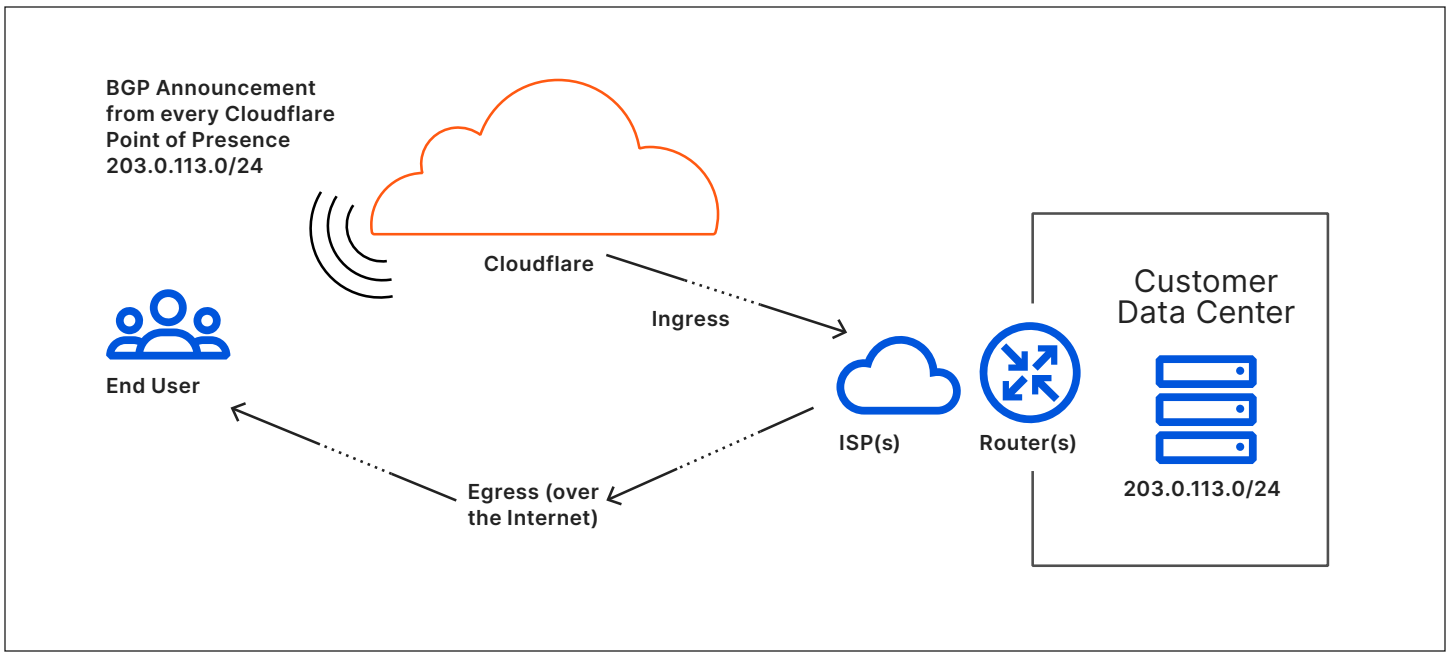
Magic Transit de Cloudflare proporciona protección frente a DDoS y aceleración del tráfico para redes locales, en la nube e híbridas. Con centros de datos que abarcan 200 ciudades y más de 35 Tbps en capacidad de mitigación, Magic Transit puede detectar y mitigar los ataques cerca de su origen entre 0 a 3 segundos (y menos de 10 segundos en promedio), todo mientras enruta el tráfico más rápido que la Internet pública.

En este documento, creamos una implementación de ejemplo y seguimos el recorrido de un paquete de un usuario a la red de un cliente de Magic Transit en Internet.

Situación:

El cliente Acme Corp. posee el prefijo IP 203.0.113.0/24, que utilizan para abordar un gabinete de hardware que ejecutan en su propio centro de datos físico. Actualmente, Acme anuncia rutas a Internet desde el equipo local del cliente (CPE, o un router en el perímetro de su centro de datos), que indica al mundo que 203.0.113.0/24 es accesible desde su número de sistema autónomo, AS64512.

Acme quiere conectarse a la red de Cloudflare para mejorar la seguridad y el rendimiento de su propia red. En especial, han sido objetivo de ataques de denegación de servicio distribuidos (DDoS).

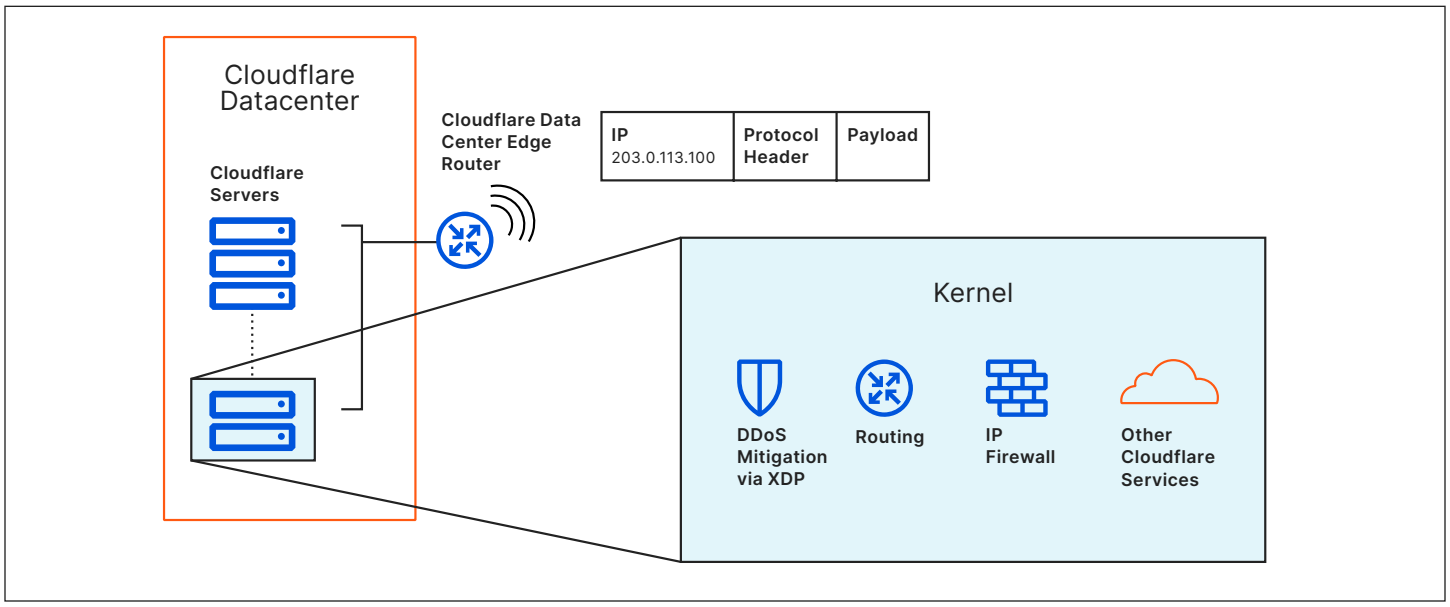


Cloudflare utiliza el protocolo de puerta de enlace de borde (BGP) para anunciar el prefijo de Acme desde el extremo de Cloudflare:

Cuando Acme trae su prefijo IP 203.0.113.0/24 a Cloudflare, empezamos a anunciar ese prefijo a nuestros proveedores de tránsito, a nuestros pares y a los intercambios de Internet en cada uno de nuestros centros de datos alrededor del mundo. Además, Acme deja de anunciar el prefijo a sus propios ISP. Esto significa que cualquier paquete de IP en Internet con una dirección de destino dentro del prefijo de Acme se entrega a un centro de datos Cloudflare cercano, y no al router de Acme.

Cuando un usuario final desea acceder, por ejemplo, al servidor FTP de Acme en 203.0.113.100, el paquete TCP SYN llega al centro de datos de Cloudflare más cercano (en términos de distancia de enrutamiento de Internet) al usuario final. El paquete llega al router del centro de datos de Cloudflare, que utiliza el enrutamiento ECMP (Equal Cost Multi-Path) para seleccionar qué servidor debe manejar el paquete. Este envía el paquete al servidor seleccionado.

Una vez en el servidor, el paquete fluye a través de las funciones de detección y de mitigación de DoS basadas en XDP e iptables de Cloudflare. Si se determinara que este paquete TCP SYN es parte de un ataque, sería eliminado y eso sería el final de este. Si el tráfico está limpio, entonces se le permite el paso.

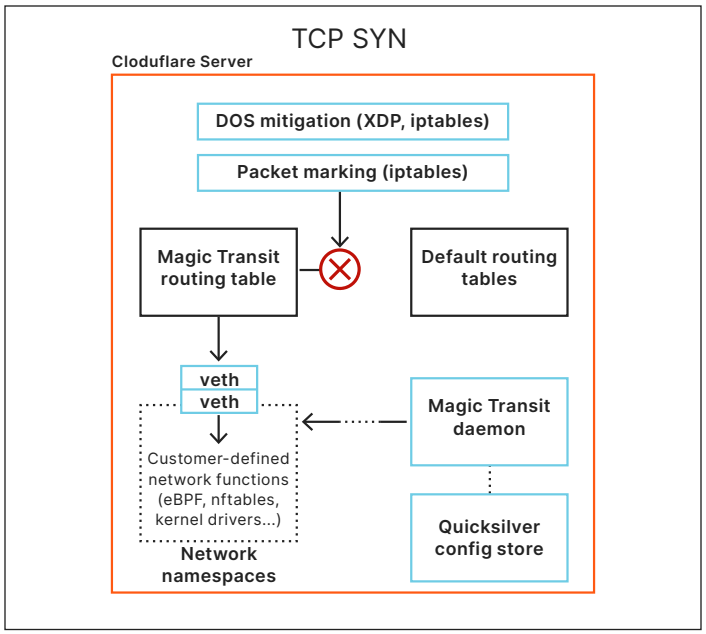


Espacios de nombres de red para el aislamiento y el control

El espacio de nombres es una colección de funciones del núcleo de Linux para crear instancias virtuales ligeras de recursos del sistema que pueden compartirse entre un grupo de procesos. Los espacios de nombres son un bloque de construcción fundamental para la contenedorización en Linux, en particular, Docker se desarrolla en el espacio de nombres de Linux. Un espacio de nombres de red es una instancia aislada de la pila de red de Linux, incluidas sus propias interfaces de red (con sus propios ganchos eBPF), tablas de enrutamiento, configuración de filtros de red, etc. Los espacios de nombres de red le dan a Cloudflare un mecanismo de bajo costo para aplicar con rapidez configuraciones de red definidas por el cliente de forma aislada, todo con funciones incorporadas en el núcleo de Linux, de manera que no hay aciertos en el rendimiento por el reenvío o la redirección de paquetes del espacio del usuario.

Cuando un nuevo cliente comienza a usar Magic Transit, Cloudflare crea un nuevo espacio de nombres de red para ese cliente en cada servidor del extremo de nuestra red. Lograr que el tráfico del cliente llegue a su espacio de nombres de red requiere una pequeña configuración de enrutamiento en el espacio de nombres de red predeterminado. Cuando se crea un espacio de nombres de red, también se crea un par de interfaces de Ethernet virtuales (veth): una en el espacio de nombres predeterminado y otra en el espacio de nombres recién creado. Este par de interfaces crea un "cable virtual" para entregar el tráfico de red dentro y fuera del nuevo espacio de nombres de red. En el espacio de nombres de la red predeterminado, mantenemos una tabla de enrutamiento que reenvía los prefijos IP de los clientes de Magic Transit a los veths que corresponden a los espacios de nombres de esos clientes. Utilizamos iptables para marcar los paquetes destinados a los prefijos de clientes de Magic Transit y tenemos una regla de enrutamiento que especifica que estos paquetes especialmente marcados deben utilizar la tabla de enrutamiento de Magic Transit.

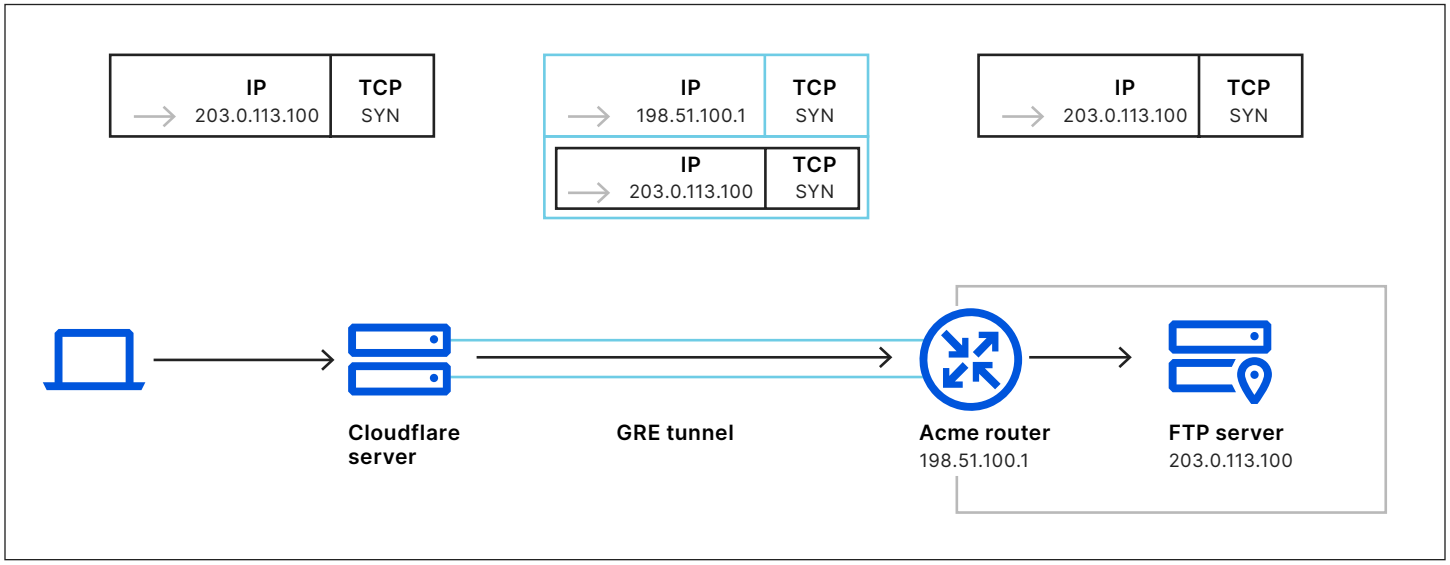
Los espacios de nombres de red proporcionan un entorno ligero en el que un cliente de Magic Transit puede ejecutar y gestionar funciones de red de forma aislada, lo que proporciona un control total al cliente.



Cloudflare emite paquetes con destino a Acme por los túneles de GRE

Después de pasar por las funciones del extremo de la red, el paquete TCP SYN está listo para ser entregado de nuevo a la infraestructura de red del cliente. Debido a que Acme Corp. no tiene una red en una instalación de colocación con Cloudflare, Cloudflare necesita entregar su tráfico de red a través de la Internet pública. Una manera en la que Cloudflare hace esto es a través de túneles.

La tunelización es un método para transportar el tráfico desde una red a través de otra red. En este caso, se trata de encapsular los paquetes IP de Acme dentro de paquetes de IP que pueden ser entregados al router de Acme a través de Internet. Existen varios protocolos de tunelización comunes, pero la Encapsulación de enrutamiento genérico (GRE) se utiliza a menudo por su simplicidad y el amplio apoyo de los proveedores. Los puntos de conexión del túnel GRE están configurados tanto en los servidores de Cloudflare (dentro del espacio de nombres de la red de Acme) como en el router de Acme. Los servidores de Cloudflare luego encapsulan los paquetes de IP destinados a 203.0.113.0/24 dentro de los paquetes IP destinados a una dirección IP públicamente redirigida para el router de Acme, que desencapsula los paquetes y los emite a la red interna de Acme.



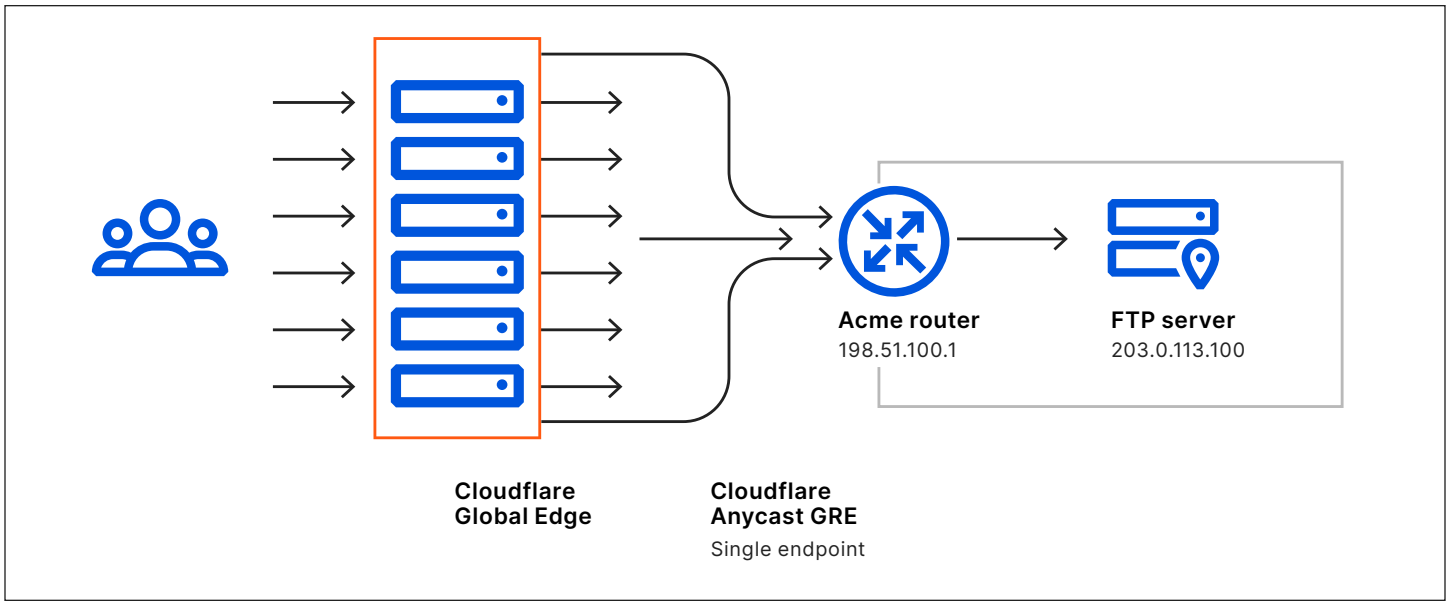
Tunelización Anycast GRE

Cloudflare utiliza las direcciones IP de Anycast para nuestros puntos de conexión del túnel GRE, lo que significa que cualquier servidor en cualquier centro de datos es capaz de encapsular y desencapsular paquetes para el mismo túnel GRE. En el contexto de Anycast, el término “túnel" es engañoso, ya que implica un vínculo entre dos puntos fijos. El protocolo GRE no tiene estado, cada paquete se procesa de forma independiente y sin necesidad de negociar ni coordinar entre los puntos de conexión del túnel. Aunque el túnel está vinculado técnicamente a una dirección IP, no es necesario vincularlo a un dispositivo específico. Cualquier dispositivo que pueda quitar los encabezados externos y luego enrutar el paquete interno puede manejar cualquier paquete GRE enviado por el túnel.

Con el Anycast GRE de Cloudflare, un "túnel" único proporciona a los clientes un conducto a cada servidor en cada centro de datos en el extremo global de Cloudflare. Una consecuencia muy poderosa del Anycast GRE es que elimina los puntos de falla únicos. Tradicionalmente, GRE a través de Internet puede ser problemático porque una interrupción de Internet entre los dos extremos GRE interrumpe por completo el túnel. Esto significa que la entrega de datos confiable requiere pasar por el dolor de cabeza de establecer y mantener túneles GRE redundantes que terminen en diferentes sitios físicos y redirigir el tráfico cuando uno de los túneles se bloquea.

Pero debido a que Cloudflare encapsula y entrega el tráfico del cliente desde cada servidor en cada centro de datos, no hay un “túnel" único que bloquear. Esto significa que los clientes de Magic Transit pueden disfrutar de la redundancia y confiabilidad de la terminación de túneles en varios sitios físicos, a la vez que solo configuran y mantienen un terminal GRE único.

MAGIC TRANSIT DE CLOUDFLARE



Funciones de la red a escala

Magic Transit es una forma nueva y potente de implementar la funciones de red a escala. Magic Transit toma los dispositivos de hardware que los clientes suelen instalar en su red local y los distribuye a través de cada servidor en cada centro de datos en la red de Cloudflare.