Cloudflare Magic Transit – Architecture de référence

Cloudflare Magic Transit offre une protection contre les attaques DDoS et une accélération du trafic pour les réseaux sur site, hybrides et dans le Cloud. Avec des datacenters couvrant 200 villes et une capacité de mitigation supérieure à 37 Tb/s, Magic Transit peut détecter et mitiger les attaques à proximité de leur origine dans un délai de 0 à 3 secondes (et en moins de 10 secondes, en moyenne), tout en acheminant le trafic plus rapidement que l'Internet public.

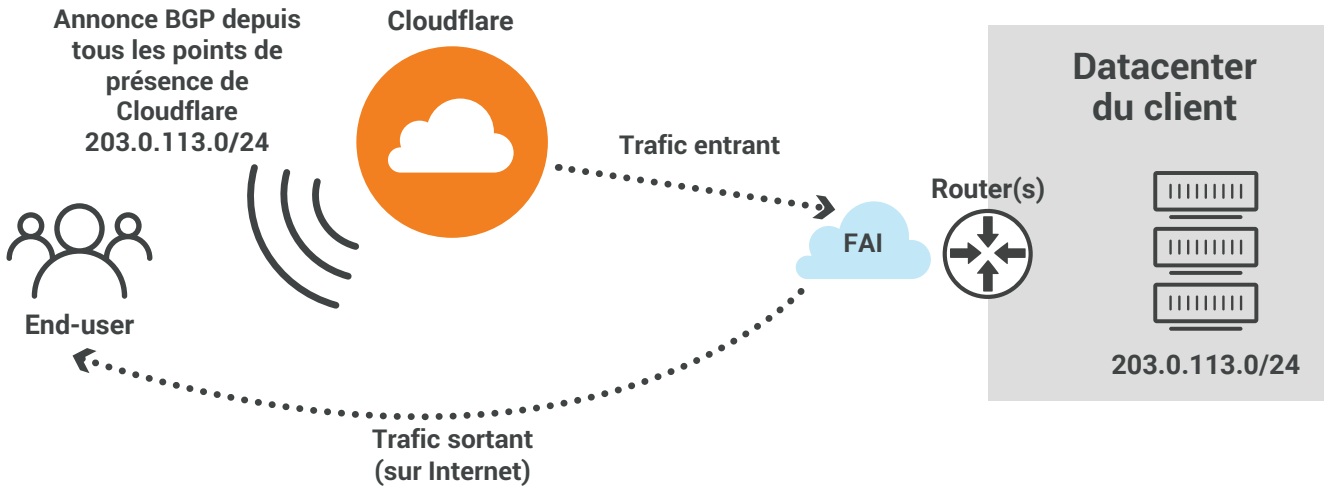
Dans ce document, nous créons un exemple de déploiement et suivons le parcours d'un paquet depuis un utilisateur d'Internet jusqu'au réseau d'un client de Magic Transit.

+33 75 7 90 52 73 | enterprise@cloudflare.com | www.cloudflare.com/fr-fr/

Situation :

L'entreprise Client Acme Corp. possède le préfixe IP 203.0.113.0/24, qu'elle utilise pour l'adressage d'un rack de matériel mis en œuvre dans son datacenter physique. Acme annonce actuellement les itinéraires vers Internet depuis ses équipements locaux d'abonné (CPE, c'est-à-dire un routeur situé sur le périmètre de son datacenter), en informant le monde que le 203.0.113.0/24 est accessible depuis son numéro de système autonome, AS64512.

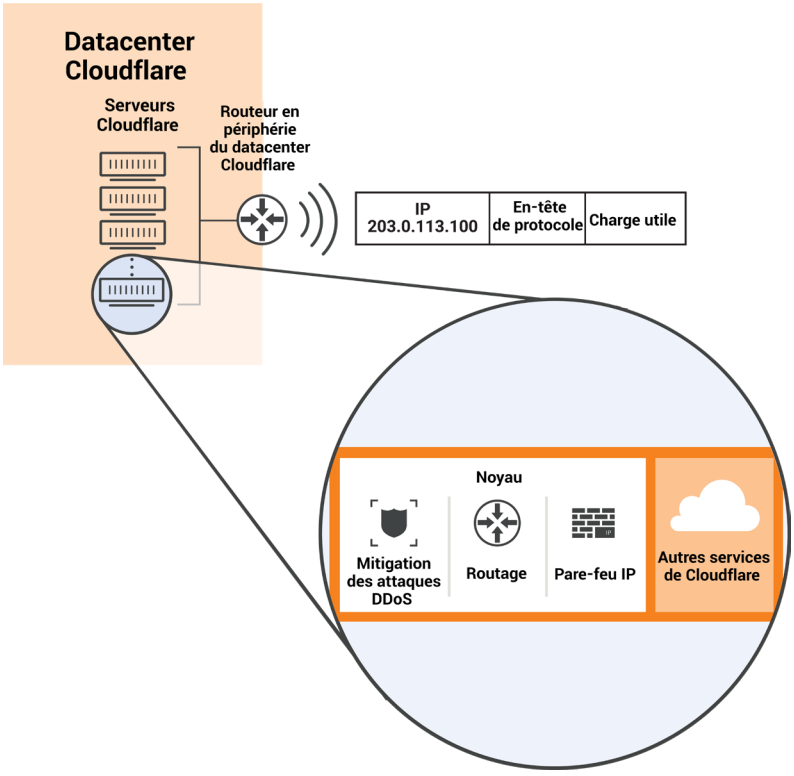
Acme souhaite se connecter au réseau Cloudflare pour améliorer la sécurité et la performance de son propre réseau. Plus spécifiquement, l'entreprise a été la cible d'attaques par déni de service distribué (DDoS).



Cloudflare utilise le protocole BGP (Border Gateway Protocol) pour annoncer le préfixe d'Acme depuis la périphérie du réseau Cloudflare :

Lorsque Acme fournit son préfixe IP 203.0.113.0/24 à Cloudflare, nous commençons à annoncer ce préfixe à nos fournisseurs de transit, à nos pairs et aux points d'interconnexion Internet dans chacun de nos datacenters, partout dans le monde. Par ailleurs, Acme cesse d'annoncer le préfixe à ses propres FAI. Ceci signifie que tout paquet IP sur Internet comportant une adresse de destination correspondant au préfixe d'Acme est acheminé vers un datacenter Cloudflare à proximité, et non vers le routeur d'Acme.

Si un utilisateur final souhaite accéder, par exemple, au serveur FTP d'Acme à l'adresse 203.0.113.100, le paquet TCP SYN atteint le datacenter Cloudflare le plus proche (en termes de distance de routage sur Internet) de l'utilisateur final. Le paquet atteint le routeur du datacenter de Cloudflare, qui utilise le routage ECMP (Equal Cost Multi-Path) pour sélectionner le serveur devant traiter le paquet. Il transfère le paquet vers le serveur sélectionné.



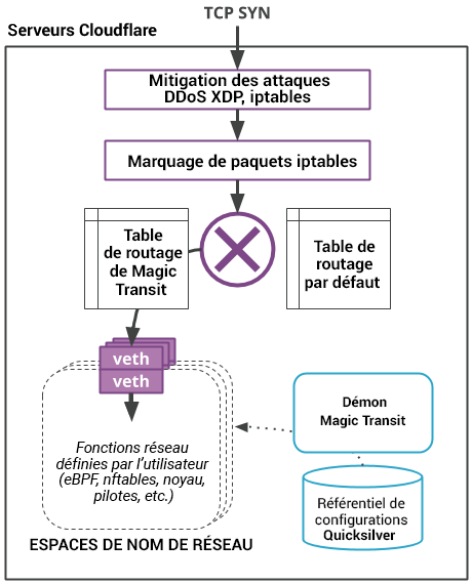
Une fois sur le serveur, le paquet transite par les fonctions de détection et de mitigation d'attaques DoS basées sur XDP et iptables de Cloudflare. Si ce paquet TCP SYN est identifié comme faisant partie d'une attaque, il est définitivement abandonné. Si le trafic dont il fait partie est légitime, il est autorisé à passer.

Espaces de noms de réseaux pour l'isolement et le contrôle

Les espaces de noms sont une collection de fonctionnalités du noyau Linux permettant de créer des instances virtuelles légères des ressources système, qui peuvent être partagées au sein d'un groupe de processus. Les espaces de noms sont une composante fondamentale de la conteneurisation sous Linux. Notamment, Docker est construit sur la base d'espaces de noms Linux. Un espace de noms de réseau est une instance isolée de la pile réseau de Linux, qui comprend ses propres interfaces réseau (avec leurs propres crochets eBPF), tables de routage, configuration Netfilter, etc. Les espaces de noms de réseau fournissent à Cloudflare un mécanisme économique permettant d'appliquer rapidement, en isolement, des configurations réseau définies par le client avec les fonctionnalités intégrées du noyau Linux. Ainsi, le transfert de paquets dans l'espace utilisateur ou l'utilisation d'un proxy ne provoque aucune baisse de performance.

Lorsqu'un nouveau client commence à utiliser Magic Transit, Cloudflare crée un tout nouvel espace de noms de réseau pour ce client sur chaque serveur sur l'ensemble de notre réseau de périphérie.

L'acheminement du trafic du client vers son espace de noms de réseau nécessite de configurer le routage dans l'espace de noms de réseau par défaut. Lorsqu'un espace de noms de réseau est créé, une paire d'interfaces virtual Ethernet (veth) est également créée : une dans l'espace de noms par défaut et une autre dans le nouvel espace de noms créé. Cette paire d'interfaces crée un « fil virtuel » permettant d'acheminer le trafic du réseau vers et depuis le nouvel espace de noms du réseau. Dans l'espace de noms du réseau par défaut, nous gérons une table de routage qui transmet les préfixes IP des clients de Magic Transit aux interfaces veth correspondant aux espaces de noms de ces clients. Nous utilisons iptables pour marquer les paquets destinés aux préfixes des clients de Magic Transit, et nous appliquons une règle de routage qui spécifie que ces paquets spécialement marqués doivent utiliser la table de routage de Magic Transit.



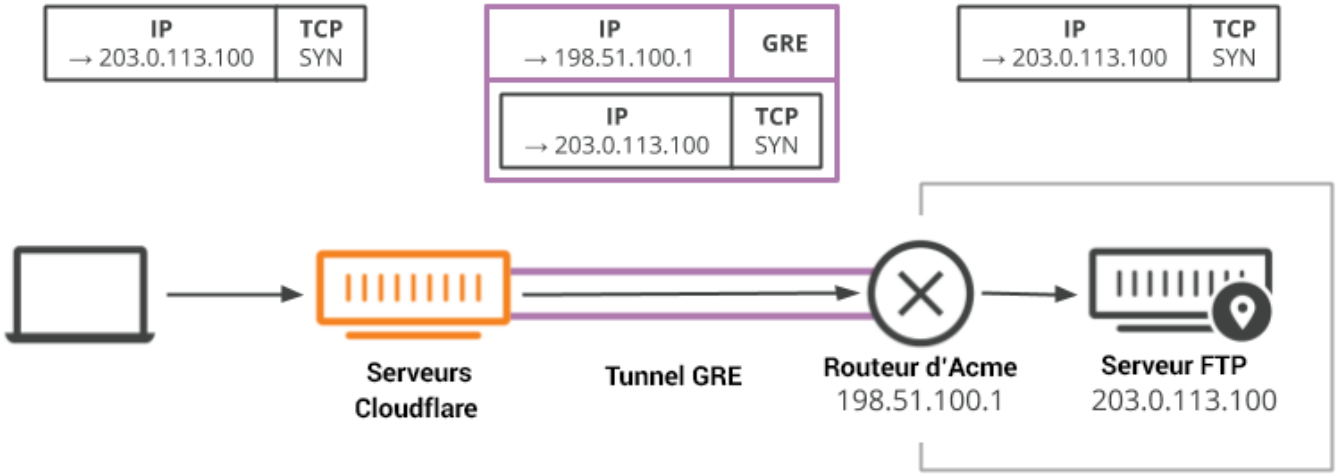
Les espaces de noms de réseau offrent un environnement léger dans lequel un client de Magic Transit peut exécuter et gérer les fonctions réseau en isolement, offrant ainsi un contrôle total au client.

Cloudflare extrait les paquets destinés à Acme via des tunnels GRE

Après avoir transité par les fonctions en périphérie de réseau, le paquet TCP SYN est prêt à être renvoyé à l'infrastructure réseau du client. Acme Corp. ne dispose pas d'une emprise réseau dans une installation en colocation avec Cloudflare. Cloudflare doit donc acheminer son trafic réseau sur l'Internet public. À cette fin, une des approches utilisées par Cloudflare est la tunnellisation.

La tunnellisation est une méthode consistant à transférer le trafic depuis un réseau sur un autre réseau. Dans ce cas, il s'agit d'encapsuler les paquets IP d'Acme dans des paquets IP pouvant être acheminés via Internet jusqu'au routeur d'Acme. Il existe plusieurs protocoles de tunnellisation répandus, toutefois, le protocole GRE (Generic Routing Encapsulation) est souvent utilisé pour sa simplicité et sa prise en charge par de nombreux fournisseurs.

Les terminaisons du tunnel GRE sont configurées à la fois sur les serveurs de Cloudflare (dans l'espace de noms d'Acme) et sur le routeur d'Acme. Les serveurs Cloudflare encapsulent ensuite les paquets IP destinés à 203.0.113.0/24 dans des paquets IP destinés à une adresse IP à routage public pour le routeur d'Acme, qui désencapsule les paquets et les diffuse ensuite sur le réseau interne d'Acme.

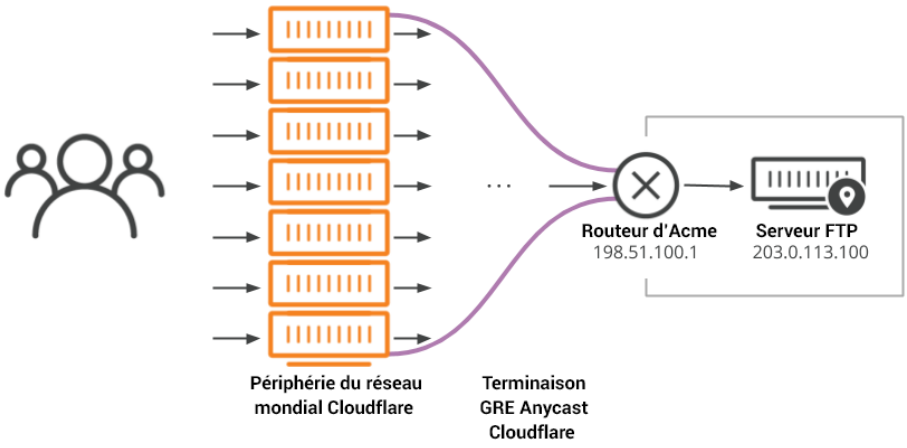


Tunnellisation GRE Anycast

Cloudflare utilise les adresses IP Anycast pour ses terminaisons de tunnel GRE, ce qui signifie que tout serveur dans tout datacenter est capable d'encapsuler et de désencapsuler des paquets pour un même tunnel GRE.

D'ailleurs, dans le contexte d'Anycast, le terme « tunnel » est trompeur, car il implique une liaison entre deux points fixes. Le protocole GRE est statique : chaque paquet est traité indépendamment et ne nécessite ni négociation, ni coordination entre les points d'extrémité du tunnel. Bien que le tunnel soit techniquement lié à une adresse IP, il n'est pas nécessaire qu'il soit lié à un appareil spécifique. Tout appareil capable d'extraire les en-têtes extérieurs, puis d'acheminer le paquet interne peut traiter un paquet GRE transmis via le tunnel.

Avec le protocole Anycast GRE de Cloudflare, un même « tunnel » permet aux clients d'accéder à tous les serveurs dans tous les datacenters à la périphérie du réseau mondial de Cloudflare.

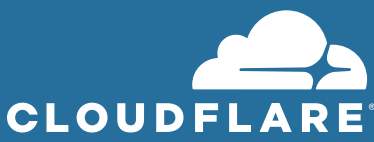


Un autre effet déterminant d'Anycast GRE est sa capacité à éliminer les points de défaillance uniques. Traditionnellement, l'utilisation du protocole GRE sur Internet peut être problématique, car toute défaillance d'Internet entre les deux terminaisons du protocole GRE interrompt complètement le tunnel. Cela signifie que l'acheminement fiable des données nécessite d'assurer la configuration et la maintenance complexes de tunnels GRE redondants, dont les terminaisons se trouvent sur différents sites physiques, et de réacheminer le trafic en cas de défaillance d'un des tunnels.

Cependant, parce que Cloudflare encapsule et achemine le trafic des clients depuis chaque serveur dans chaque datacenter, il n'existe aucun « tunnel » unique pouvant être mis hors service. Ceci signifie que les utilisateurs de Magic Transit peuvent bénéficier de la redondance et de la fiabilité de tunnels comportant des terminaisons sur plusieurs sites physiques, tout en gérant la configuration et la maintenance d'une seule terminaison GRE.

Fonctionnalités réseau à grande échelle

Magic Transit offre une nouvelle solution puissante pour déployer des fonctions réseau à grande échelle. Magic Transit répartit les équipements physiques que les clients connectent habituellement à leur réseau sur site et les répartit sur chaque serveur dans chaque datacenter du réseau de Cloudflare.



+33 75 7 90 52 73 | enterprise@cloudflare.com | www.cloudflare.com/fr-fr/

RÉV : 200505