Cloudflare Magic Transit — Referenz-Architektur

Cloudflare Magic Transit bietet Schutz vor DDoS-Angriffen und eine Beschleunigung des Traffic für lokale, cloudbasierte und hybride Netzwerke. Mit Rechenzentren in 200 Städten und mehr als 37 Tbps Abwehrkapazität kann Magic Transit teilweise innerhalb von 0 bis 3 Sekunden (und im Schnitt in weniger als 10 Sekunden) Angriffe in der Nähe ihres Ursprungs erkennen und neutralisieren. Zugleich wird der Traffic schneller geroutet, als dies über das öffentliche Internet der Fall wäre.

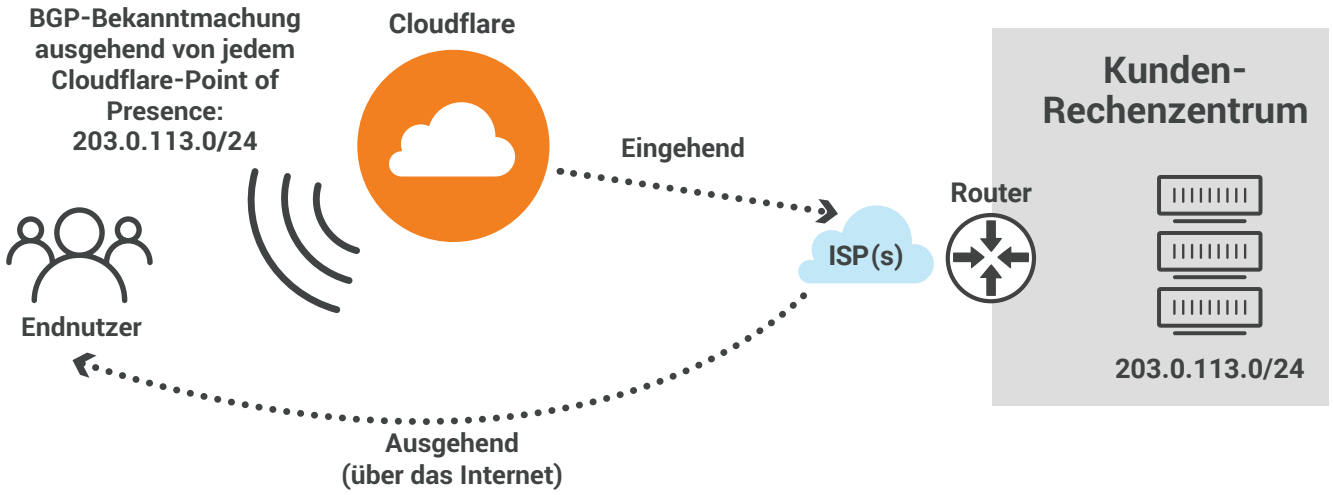
In diesem Paper erstellen wir eine Beispiel-Implementierung und folgen dem Weg eines Pakets von einem Nutzer im Internet zum Netzwerk eines Magic Transit-Kunden.

+49 89 2555 2276 | enterprise@cloudflare.com | www.cloudflare.com/de-de/

Die Ausgangslage:

Dem Kunden Acme Corp. gehört das IP-Präfix 203.0.113.0/24, das er für die Adressierung von Hardware in seinem eigenen physischen Rechenzentrum nutzt. Zurzeit veröffentlicht Acme Routen vom Internet zum eigenen Customer Premise Equipment (CPE; bzw. zu einem Router am Perimeter des Acme-Rechenzentrums). Die Firma teilt der Welt also mit, dass 203.0.113.0/24 über ihre AS-Nummer AS64512 erreichbar ist.

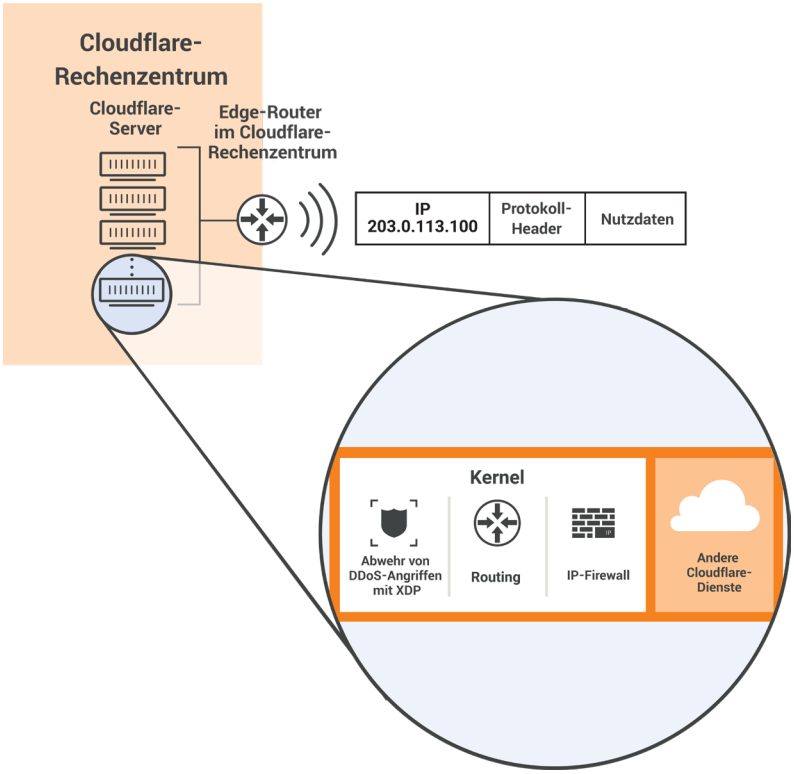
Acme möchte sich mit dem Cloudflare-Netzwerk verbinden, um die Sicherheit und Performance des eigenen Netzwerks zu erhöhen. Genauer gesagt ist das Unternehmen zum Ziel von Distributed-Denial-of-Service-(DDoS)-Angriffen geworden.



Cloudflare verwendet das Border Gateway Protocol (BGP), um das Präfix von Acme vom Rand des Cloudlare-Netzwerks zu senden:

Wenn Acme das eigene IP-Präfix 203.0.113.0/24 an Cloudflare übermittelt, beginnen wir, dieses Präfix an unsere Transit-Provider, unsere Peers und Internetknoten in jedem unserer Rechenzentren rund um den Globus zu senden. Außerdem hört Acme auf, den eigenen Internetdienstanbietern (ISPs) das Präfix mitzuteilen. Somit wird jedes IP-Paket im Internet mit einer Zieladresse innerhalb des Acme-Präfixes nicht an den Acme-Router, sondern an ein nahe gelegenes Cloudflare-Rechenzentrum geschickt.

Wenn ein Endnutzer beispielsweise auf den FTP-Server von Acme unter 203.0.113.100 zugreifen möchte, geht das TCP-SYN-Paket an das Cloudflare-Rechenzentrum, das (in Bezug auf die Internet-Routing-Entfernung) dem Endnutzer am nächsten ist. Das Paket erreicht den Router des Cloudflare-Rechenzentrums. Dieser nutzt ECMP (Equal Cost Multi-Path)-Routing, um den Server auszuwählen, der das Paket empfangen soll. Der Router sendet das Paket an den ausgewählten Server.



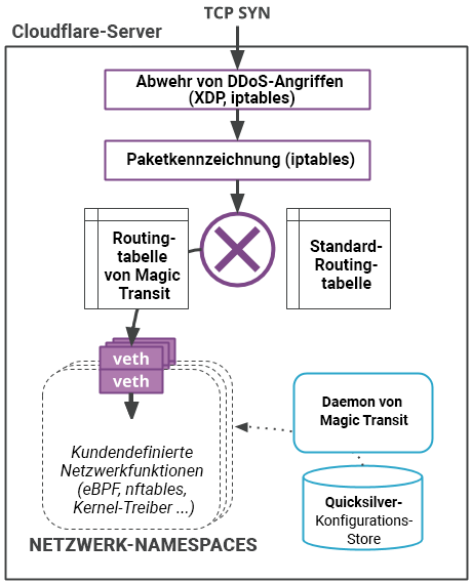
Wenn es sein Ziel erreicht hat, wird das Paket durch Cloudflare-Funktionen zur Erkennung und Abwehr von DoS-Angriffen geleitet, die auf XDP oder iptables aufbauen. Stellt sich dabei heraus, dass dieses TCP-SYN-Paket Teil eines Angriffs ist, wird es verworfen. Damit wäre die Sache dann erledigt. Ist der Traffic „sauber", darf er passieren.

Netzwerk-Namespaces zur Isolation und Kontrolle

Bei Namespaces handelt es sich um eine Sammlung von Linux-Kernel-Funktionen zum Erstellen einfacher virtueller Instanzen von Systemressourcen, die von einer Gruppe von Prozessen genutzt werden können. Namespaces sind ein wesentlicher Baustein der Linux-Containerisierung – insbesondere Docker baut auf Linux-Namespaces auf. Ein Netzwerk-Namespace ist eine isolierte Instanz des Netzwerk-Stacks unter Linux, die unter anderem über eigene Netzwerk-Schnittstellen (mit eigenen eBPF-Hooks), Routingtabellen und Netzfilterkonfiguration verfügt. Mit Netzwerk-Namespaces steht Cloudflare ein kostengünstiger Mechanismus zur Verfügung, um kundendefinierte Netzwerkkonfigurationen schnell und isoliert anzuwenden – jeweils mit integrierten Linux-Kernel-Funktionen, sodass Userspace-Paketweiterleitungen oder Proxying die Performance nicht beeinträchtigen.

Wenn ein Neukunde beginnt, Magic Transit zu verwenden, erstellt Cloudflare für diesen einen brandneuen Netzwerk-Namespace auf jedem Server unseres Netzwerkrands.

Um den Datenverkehr des Kunden zu seinem Netzwerk-Namespace befördern zu können, muss ein wenig Konfigurationsarbeit am Routing des Standard-Netzwerk-Namespace vorgenommen werden. Bei der Erstellung eines Netzwerk-Namespace werden auch zwei virtuelle Ethernet (veth)-Schnittstellen geschaffen: eine im Standard-Namespace und eine in dem neu erstellten Namespace. Dieses Schnittstellenpaar schafft eine „virtuelle Leitung“ für den Netzwerkdatenverkehr in und aus dem neuen Netzwerk-Namespace. In dem Standard-Netzwerk-Namespace unterhalten wir eine Routingtabelle, anhand derer die IP-Präfixe von Magic Transit-Kunden an die veths weitergeleitet werden, die denen der Kunden-Namespaces entsprechen. Mithilfe von iptables werden die Pakete gekennzeichnet, die für Magic Transit-Kundenpräfixe bestimmt sind. Außerdem verfügen wir über eine Routingregel, die festlegt, dass für diese speziell markierten Pakete die Routingtabelle von Magic Transit verwendet werden soll.



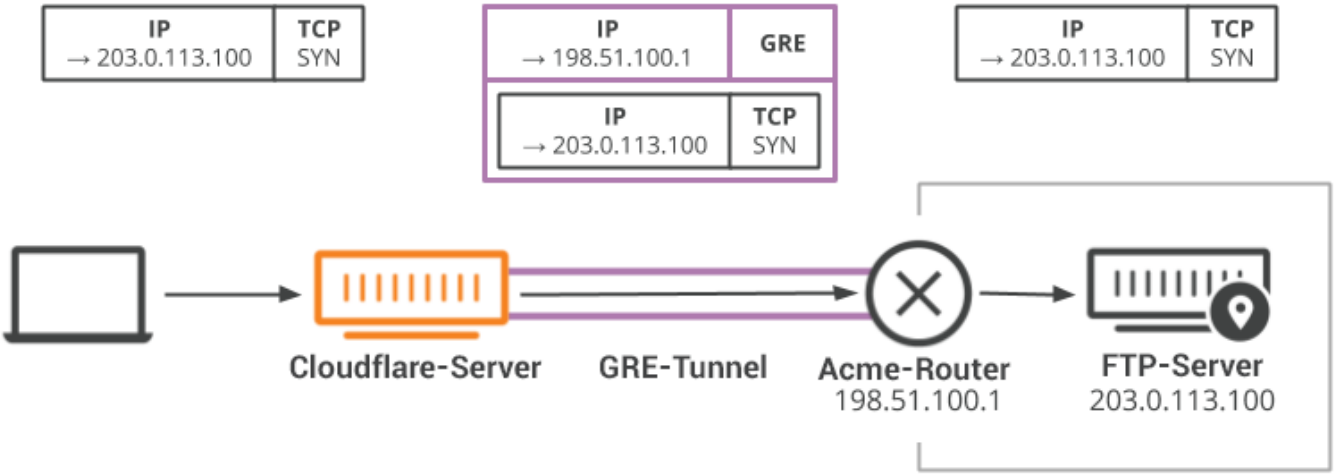
Netzwerk-Namespaces bieten eine einfache Umgebung, in der ein Magic Transit-Kunde Netzwerkfunktionen isoliert ausführen und verwalten kann, sodass er über die volle Kontrolle verfügt.

Cloudflare verschickt für Acme bestimmte Pakete über GRE-Tunnel

Nachdem das TCP-SYN-Paket die Funktionen am Netzwerkrand passiert hat, kann es wieder an die Netzwerkinfrastruktur des Kunden zurückgesendet werden. Weil Acme Corp. keine eigenen Netzwerk-Server in Rechenzentren von Cloudflare unterhält, müssen wir den Netzwerk-Traffic des Unternehmens über das öffentliche Internet leiten. Dafür nutzten wir unter anderem Tunneling.

Dabei handelt es sich um eine Methode, um Datenverkehr von einem Netzwerk über ein anderes zu befördern. In unserem Beispielfall werden die IP-Pakete von Acme in IP-Pakete verpackt, die über das Internet an den Acme-Router übermittelt werden können. Es existiert eine Reihe von gängigen Tunneling-Protokollen, aber wegen seiner Einfachheit und seiner Beliebtheit bei Anbietern wird häufig auf Generic Routing Encapsulation (GRE) zurückgegriffen.

Sowohl auf den Cloudflare-Servern (im Netzwerk-Namespace von Acme) als auch auf dem Acme-Router werden GRE-Tunnel-Endpunkte konfiguriert. Anschließend verpacken Cloudflare-Server IP-Pakete, die für 203.0.113.0/24 bestimmt sind, in IP-Pakete, die für eine öffentlich routingfähige IP-Adresse des Acme-Routers bestimmt sind. Dieser entpackt die Pakete wieder und sendet sie an das interne Acme-Netzwerk.

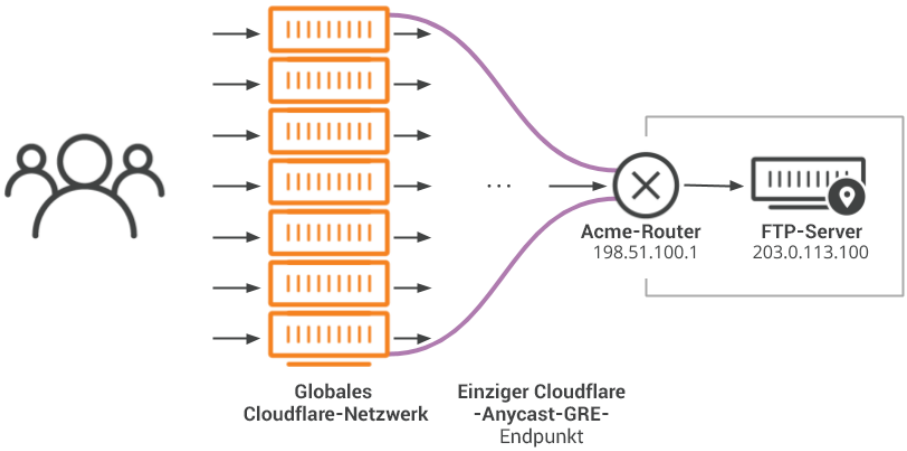


GRE-Tunneling mit Anycast

Wir verwenden Anycast-IP-Adressen für unsere GRE-Tunnel-Endpunkte. Somit ist jeder Server in jedem beliebigen Rechenzentrum in der Lage, Pakete für den gleichen GRE-Tunnel zu verpacken und zu entpacken.

Im Anycast-Kontext ist der Begriff „Tunnel“ irreführend, weil er nahelegt, das eine Verbindung zwischen zwei festen Punkten besteht. Das GRE-Protokoll ist zustandslos – jedes Paket wird unabhängig verarbeitet, ein Austausch oder eine Abstimmung zwischen Tunnelendpunkten ist nicht erforderlich. Der Tunnel ist zwar technisch an eine IP-Adresse gebunden, jedoch nicht zwangsläufig an ein bestimmtes Gerät. Jedes Gerät, das die äußeren Header abstreifen und dann das im Inneren enthaltene Paket weiterleiten kann, ist auch in der Lage, jedes beliebige über den Tunnel gesendete GRE-Paket zu verarbeiten.

Mit dem Anycast GRE von Cloudflare steht dem Kunden mit einem einzigen „Tunnel“ eine Leitung zu jedem Server in jedem beliebigen Rechenzentrum des globalen Netzwerkrands von Cloudflare zur Verfügung.

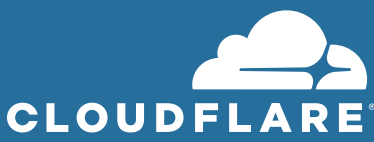


Ein sehr wichtiger Aspekt von Anycast GRE besteht darin, dass Single Points of Failure beseitigt werden. GRE über das Internet gilt traditionell als heikel, weil bei einem Internetausfall zwischen den beiden GRE-Endpunkten der Tunnel vollständig zusammenbricht. Deshalb müssen für eine zuverlässige Datenübermittlung redundante GRE-Tunnel mit Endpunkten an unterschiedlichen physischen Standorten mühsam eingerichtet und unterhalten werden. Außerdem muss der der Datenverkehr umgeleitet werden, wenn einer der Tunnel zusammenbricht.

Da Cloudflare jedoch Kundendatenverkehr von jedem Server in jedem Rechenzentrum verpackt und sendet, gibt es keinen einzelnen Tunnel, der zusammenbrechen kann. Somit profitieren Magic Transit-Kunden von der Redundanz und Zuverlässigkeit von Endtunneln an mehreren physischen Standorten, müssen selbst jedoch nur einen einzigen GRE-Endpunkt einrichten und unterhalten.

Netzwerkfunktionen in großem Maßstab

Mit Magic Transit gibt es jetzt eine neue und leistungsstarke Möglichkeit, Netzwerkfunktionen in großem Maßstab einzusetzen. Hardware-Appliances, die von Kunden normalerweise in ihrem lokalen Netzwerk betrieben würden, werden von Magic Transit auf jeden Server in jedem Rechenzentrum des Cloudflare-Netzwerks verteilt.



+49 89 2555 2276 | enterprise@cloudflare.com | www.cloudflare.com/de-de/

REV: 200505