WHITEPAPER

CLOUDFLARE

Cloudflare Magic Transit — Reference Architecture

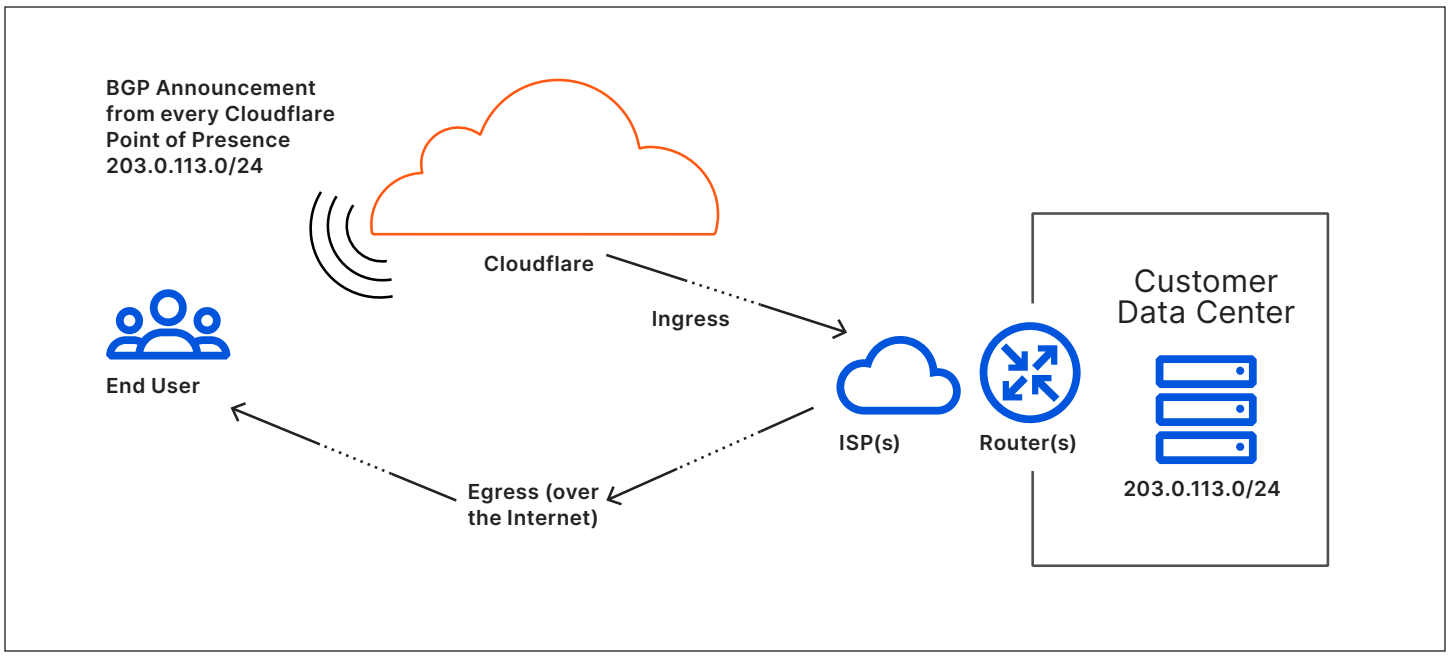
Cloudflare Magic Transit provides DDoS protection and traffic acceleration for on-premise, cloud, and hybrid networks. With data centers spanning 200 cities and over 35 Tbps in mitigation capacity, Magic Transit can detect and mitigate attacks close to their source of origin within 0-3 seconds (and less than 10 seconds on average) — all while routing traffic faster than the public Internet.

In this paper, we build an example deployment and follow the journey of a packet from a user on the Internet to a Magic Transit customer's network.

Situation:

Customer Acme Corp. owns the IP prefix 203.0.113.0/24, which they use to address a rack of hardware they run in their own physical data center. Acme currently announces routes to the Internet from their customer-premise equipment (CPE, or a router at the perimeter of their data center), telling the world 203.0.113.0/24 is reachable from their autonomous system number, AS64512.

Acme wants to connect to the Cloudflare network to improve the security and performance of their own network. Specifically, they've been the target of distributed denial-of-service (DDoS) attacks.

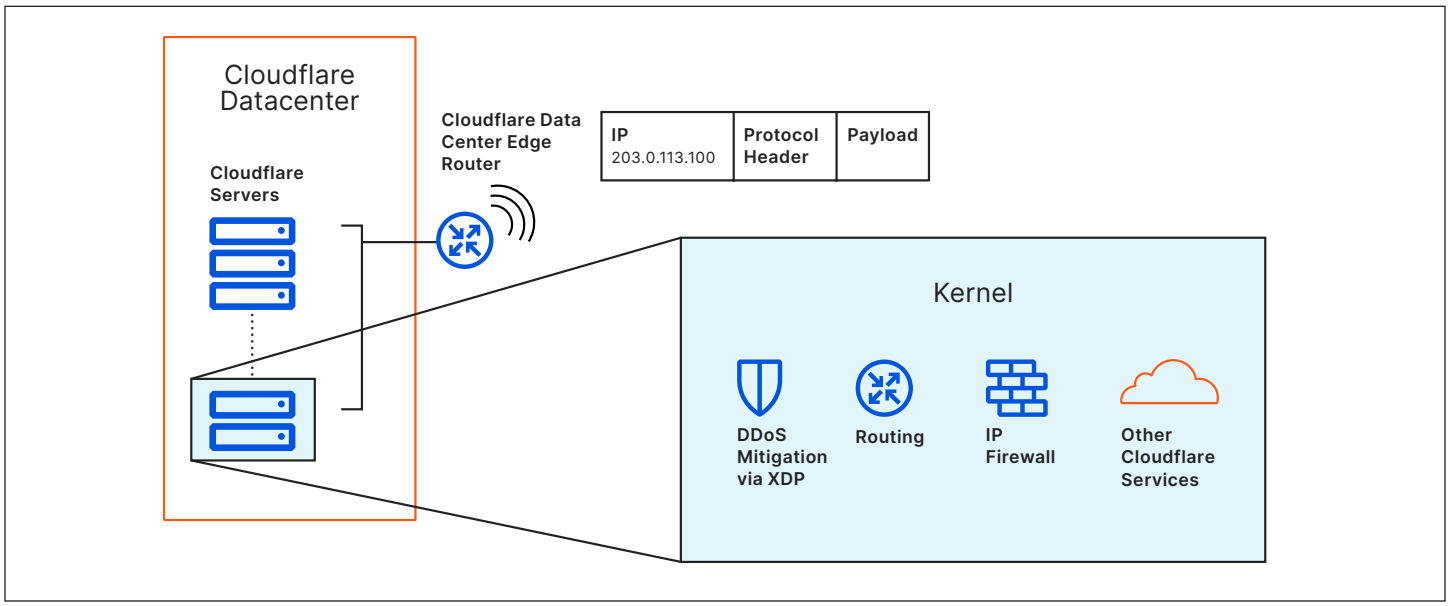


Cloudflare uses Border Gateway Protocol (BGP) to announce Acme's prefix from Cloudflare's edge:

When Acme brings their IP prefix 203.0.113.0/24 to Cloudflare, we start announcing that prefix to our transit providers, our peers, and Internet exchanges in each of our data centers around the globe. Additionally, Acme stops announcing the prefix to their own ISPs. This means that any IP packet on the Internet with a destination address within Acme's prefix is delivered to a nearby Cloudflare data center, not to Acme's router.

When an end user wants to access, for instance, Acme's FTP server on 203.0.113.100, the TCP SYN packet hits the Cloudflare data center closest (in terms of Internet routing distance) to the end user. The packet arrives on Cloudflare's data center's router, which uses ECMP (Equal Cost Multi-Path) routing to select which server should handle the packet. It dispatches the packet to the selected server.

Once at the server, the packet flows through Cloudflare's XDP- and iptables-based DoS detection and mitigation functions. If this TCP SYN packet were determined to be part of an attack, it would be dropped and that would be the end of it. If the traffic is clean, then it is permitted to pass.

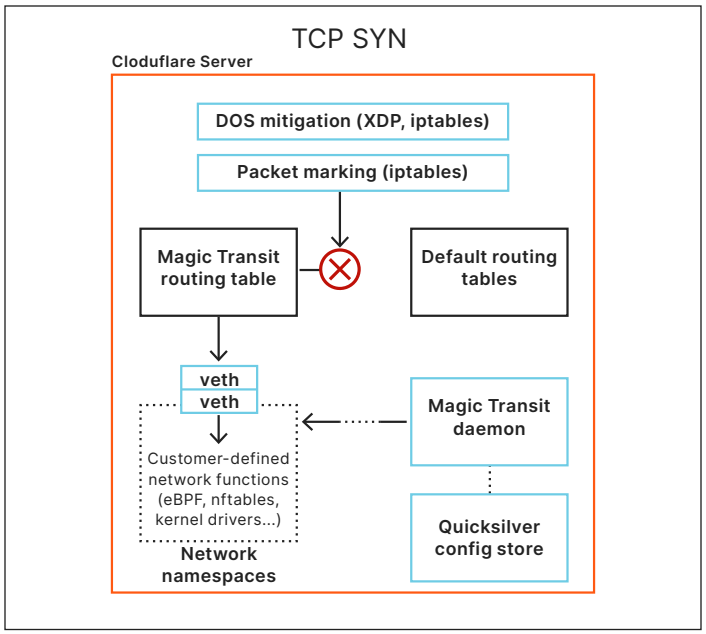


Network namespaces for isolation and control

Namespaces are a collection of Linux kernel features for creating lightweight virtual instances of system resources that can be shared among a group of processes. Namespaces are a fundamental building block for containerization in Linux — notably, Docker is built on Linux namespaces. A network namespace is an isolated instance of the Linux network stack, including its own network interfaces (with their own eBPF hooks), routing tables, netfilter configuration, and so on. Network namespaces give Cloudflare a low-cost mechanism to rapidly apply customer-defined network configurations in isolation, all with built-in Linux kernel features so there's no performance hit from userspace packet forwarding or proxying.

When a new customer starts using Magic Transit, Cloudflare creates a brand new network namespace for that customer on every server across our edge network. Getting the customer's traffic to their network namespace requires a little routing configuration in the default network namespace. When a network namespace is created, a pair of virtual Ethernet (veth) interfaces is also created: one in the default namespace and one in the newly created namespace. This interface pair creates a "virtual wire" for delivering network traffic into and out of the new network namespace. In the default network namespace, we maintain a routing table that forwards Magic Transit customer IP prefixes to the veths corresponding to those customers' namespaces.
We use iptables to mark the packets that are destined for Magic Transit customer prefixes, and we have a routing rule that specifies that these specially marked packets should use the Magic Transit routing table.

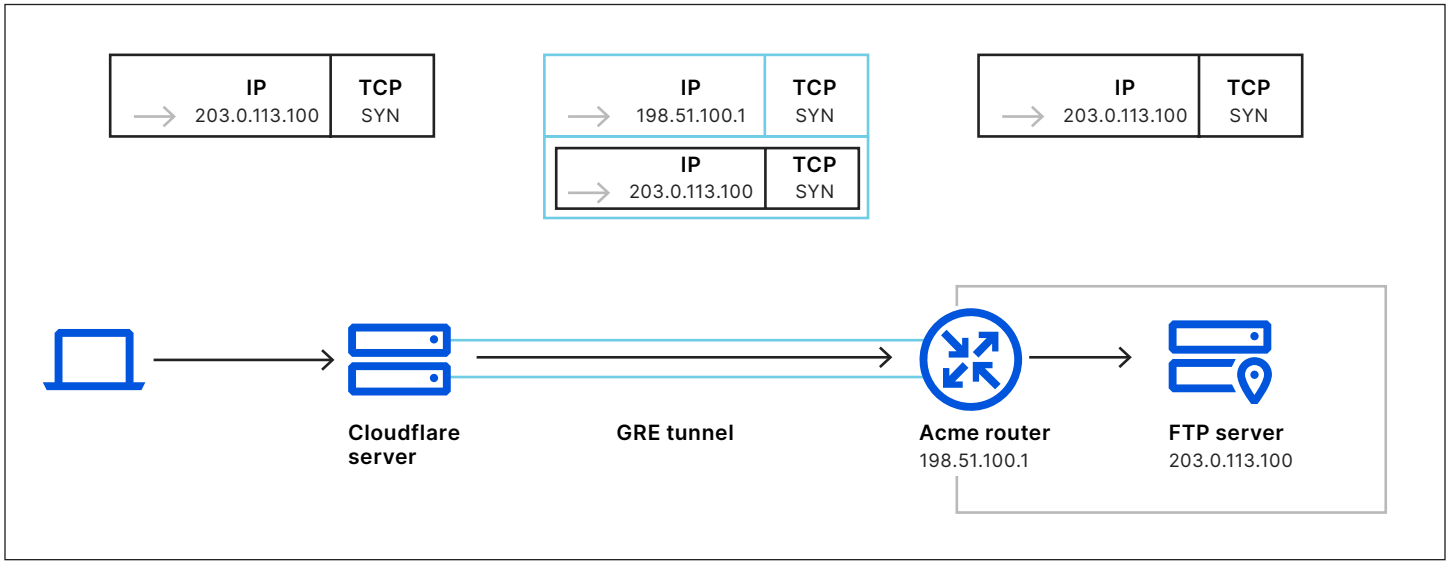
Network namespaces provide a lightweight environment where a Magic Transit customer can run and manage network functions in isolation, delivering full control to the customer.



Cloudflare egresses packets bound for Acme over GRE tunnels

After passing through the edge network functions, the TCP SYN packet is ready to be delivered back to the customer's network infrastructure. Because Acme Corp. does not have a network footprint in a colocation facility with Cloudflare, Cloudflare needs to deliver their network traffic over the public Internet. One of the ways Cloudflare does this is via tunneling.

Tunneling is a method of carrying traffic from one network over another network. In this case, it involves encapsulating Acme's IP packets inside of IP packets that can be delivered to Acme's router over the Internet. There are a number of common tunneling protocols, but Generic Routing Encapsulation (GRE) is often used for its simplicity and widespread vendor support. GRE tunnel endpoints are configured both on Cloudflare's servers (inside of Acme's network namespace) and on Acme's router. Cloudflare servers then encapsulate IP packets destined for 203.0.113.0/24 inside of IP packets destined for a publicly routable IP address for Acme's router, which decapsulates the packets and emits them into Acme's internal network.



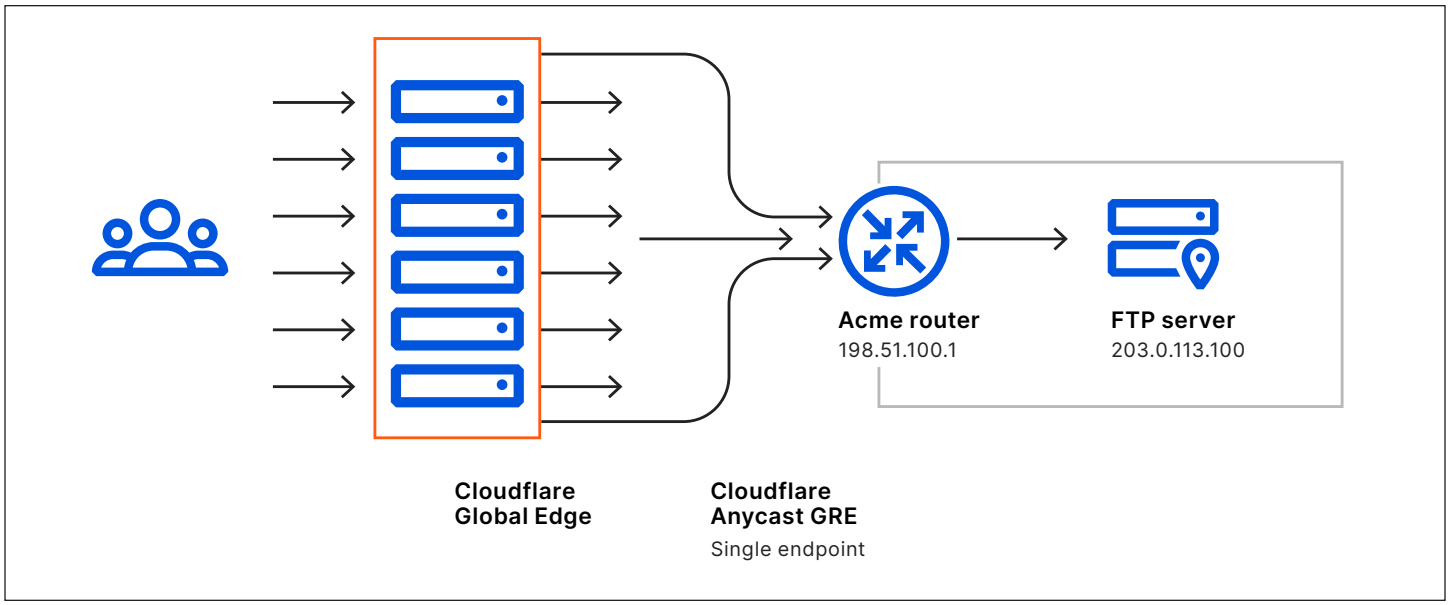
Anycast GRE Tunneling

Cloudflare uses Anycast IP addresses for our GRE tunnel endpoints, meaning that any server in any data center is capable of encapsulating and decapsulating packets for the same GRE tunnel. In the context of anycast, the term "tunnel" is misleading since it implies a link between two fixed points. The GRE protocol is stateless — each packet is processed independently and without requiring any negotiation or coordination between tunnel endpoints. While the tunnel is technically bound to an IP address, it need not be bound to a specific device. Any device that can strip off the outer headers and then route the inner packet can handle any GRE packet sent over the tunnel.

With Cloudflare's Anycast GRE, a single "tunnel" gives customers a conduit to every server in every data center on Cloudflare's global edge. A very powerful consequence of Anycast GRE is that it eliminates single points of failure. Traditionally, GRE-over-Internet can be problematic because an Internet outage between the two GRE endpoints fully breaks the tunnel. This means reliable data delivery requires going through the headache of setting up and maintaining redundant GRE tunnels terminating at different physical sites and rerouting traffic when one of the tunnels breaks.

But because Cloudflare is encapsulating and delivering customer traffic from every server in every data center, there is no single "tunnel" to break — this means Magic Transit customers can enjoy the redundancy and reliability of terminating tunnels at multiple physical sites while only setting up and maintaining a single GRE endpoint.

CLOUDFLARE MAGIC TRANSIT



Network functions at scale

Magic Transit is a powerful new way to deploy network functions at scale. Magic Transit takes the hardware appliances customers would typically rack in their on-premise network and distributes them across every server in every data center in Cloudflare's network.