A peering perspective from a global CDN

Marty Strong

GORE15 - 18th May 2015 - Madrid, Spain

Agenda

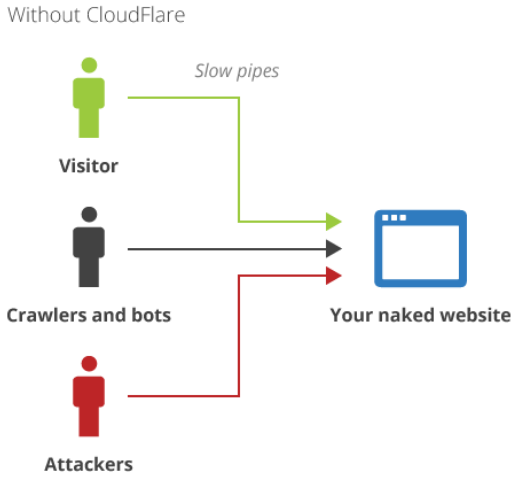
- What is CloudFlare?
- Why do we peer?
- Where do we peer?
- Why Madrid?
  - The EspanIX experience
- What would we like to see done differently?
- Any questions?

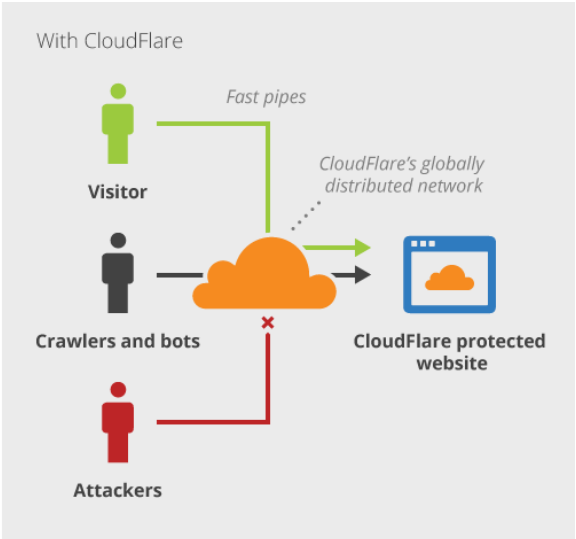
What is CloudFlare?

What is CloudFlare?

CloudFlare makes websites faster and safer using our globally distributed network to deliver essential services to any website

- Performance
- Content
- Optimisation
- Security
- 3rd party services
- Analytics





CloudFlare has customers globally

Nearly two million websites

How does CloudFlare work?

CloudFlare works at the network level

- Once a website is part of the CloudFlare community, its web traffic is routed through our global network of 30+ data centres.
- At each edge node, CloudFlare manages DNS, caching, bot filtering, web content optimisation and third party app installations.

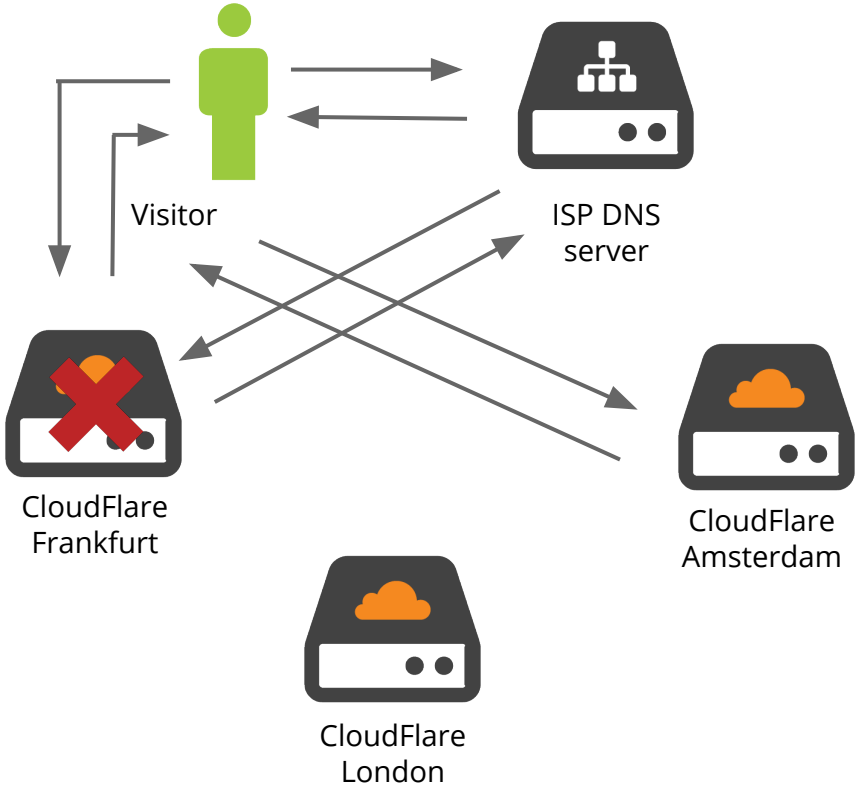
How does CloudFlare work?

How does it work?

- DNS Query - to anycast DNS address
- DNS result returned with Anycast IP
- Client makes connection to returned IP
- CloudFlare replies, session established

What happens in the event of an outage?

- Anycast prefixes are withdrawn from problematic PoP
- Traffic re-routes to next closest PoP
  - TCP session resets at this point



Anycast CDN – equally IPv4 and IPv6

Anycast prefixes

- Same IP prefixes (IPv4 & IPv6) advertised in each of the 30+ sites around the world
- Unicast (from separate site-specific prefixes) used to pull traffic from “origin” web sources

Traffic Control

- “Eyeball” ISPs (should) route to closest node, resulting in very low latency to our services from everywhere in the world
- If ISP A routes to CloudFlare in Germany then traffic will be served from Frankfurt or Düsseldorf
- If ISP B routes to CloudFlare in New Zealand then traffic will be served from Auckland

This results in a reasonable distribution of attack traffic between our sites

- Easier to mitigate 30 sites receiving a ~50Gbit DDoS than 1 site receiving 1,500Gbit DDoS

Unicast traceroutes

London

```
core1.lon2.he.net> traceroute 194.176.119.250
 traceroute to 194.176.119.250 (194.176.119.250), 30 hops max, 60 byte packets
 1  172.20.4.101  0.656 ms  0.757 ms  8.145 ms
 2  linx-1.init7.net (195.66.224.175)  6.531 ms  6.512 ms  6.600 ms
 3  r1par1.core.init7.net (77.109.128.213)  8.363 ms  8.366 ms  8.361 ms
 4  r1mad3.core.init7.net (82.197.164.248)  39.405 ms  28.315 ms  28.328 ms
 5  r1mad1.core.init7.net (77.109.140.241)  28.364 ms  28.591 ms  28.481 ms
 6  gw-customer.init7.net (77.109.135.50)  28.350 ms  28.346 ms  28.256 ms
 7  10.128.0.11  30.269 ms  30.517 ms  30.730 ms
 8  smtp1.bondis.org (194.176.119.250)  28.878 ms  28.663 ms  28.658 ms
```

New York

```
core1.nyc4.he.net> traceroute 194.176.119.250
 traceroute to 194.176.119.250 (194.176.119.250), 30 hops max, 60 byte packets
 1  ge3-8.core1.nyc4.he.net (209.51.161.13)  0.169 ms  0.192 ms  0.246 ms
 2  PAIX-NYC.init7.net (198.32.118.38)  0.362 ms  0.421 ms  0.461 ms
 3  r1lon2.core.init7.net (77.109.128.69)  72.958 ms  73.203 ms  73.308 ms
 4  r1par1.core.init7.net (77.109.128.213)  74.159 ms  74.312 ms  74.293 ms
 5  r1mad3.core.init7.net (82.197.164.248)  94.019 ms  103.161 ms  94.104 ms
 6  r1mad1.core.init7.net (77.109.140.241)  94.146 ms  94.244 ms  94.312 ms
 7  gw-customer.init7.net (77.109.135.50)  94.120 ms  94.038 ms  94.135 ms
 8  10.128.0.12  94.608 ms  96.503 ms  96.586 ms
 9  smtp1.bondis.org (194.176.119.250)  94.488 ms  94.436 ms  94.248 ms
```

Tokyo

```
core1.tyo1.he.net> traceroute 194.176.119.250
 traceroute to 194.176.119.250 (194.176.119.250), 30 hops max, 60 byte packets
 1  ge2-3.core1.tyo1.he.net (74.82.46.5)  0.227 ms  0.250 ms  0.326 ms
 2  10ge1-13.core1.lax2.he.net (184.105.223.105)  120.309 ms  120.338 ms  120.567 ms
 3  100ge2-1.core1.lax1.he.net (72.52.92.121)  97.686 ms  97.674 ms  97.696 ms
 4  100ge9-2.core1.ash1.he.net (184.105.80.201)  159.624 ms  159.612 ms  159.595 ms
 5  100ge5-1.core1.nyc4.he.net (184.105.223.166)  171.146 ms  171.176 ms  171.161 ms
 6  PAIX-NYC.init7.net (198.32.118.38)  164.897 ms  164.922 ms  165.001 ms
 7  r1lon2.core.init7.net (77.109.128.69)  237.044 ms  237.054 ms  236.933 ms
 8  r1par1.core.init7.net (77.109.128.213)  237.898 ms  237.799 ms  237.862 ms
 9  r1mad3.core.init7.net (82.197.164.248)  264.259 ms  260.805 ms  260.886 ms
10  r1mad1.core.init7.net (77.109.140.241)  257.697 ms  259.683 ms  259.773 ms
11  gw-customer.init7.net (77.109.135.50)  257.691 ms  257.680 ms  257.645 ms
12  10.128.0.11  260.181 ms  260.011 ms  258.721 ms
13  smtp1.bondis.org (194.176.119.250)  258.225 ms  257.977 ms  258.367 ms
```

Anycast traceroutes

London

```
core1.lon2.he.net> traceroute 198.41.208.141
 traceroute to 198.41.208.141 (198.41.208.141), 30 hops max, 60 byte packets
 1  172.20.4.101  0.095 ms  0.125 ms  0.176 ms
 2  linx-juniper.as13335.net (195.66.225.179)  0.643 ms  0.563 ms  0.639 ms
 3  198.41.208.141  0.463 ms  0.510 ms  0.507 ms
```

New York

```
core1.nyc4.he.net> traceroute 198.41.208.141
 traceroute to 198.41.208.141 (198.41.208.141), 30 hops max, 60 byte packets
 1  ge3-8.core1.nyc4.he.net (209.51.161.13)  2.143 ms  2.389 ms  2.375 ms
 2  xe-0-0-0.edge01.ewr01.as13335.net(198.32.118.206)  0.692 ms  0.696 ms  0.751 ms
 3  198.41.208.141  0.626 ms  0.617 ms  0.621 ms
```

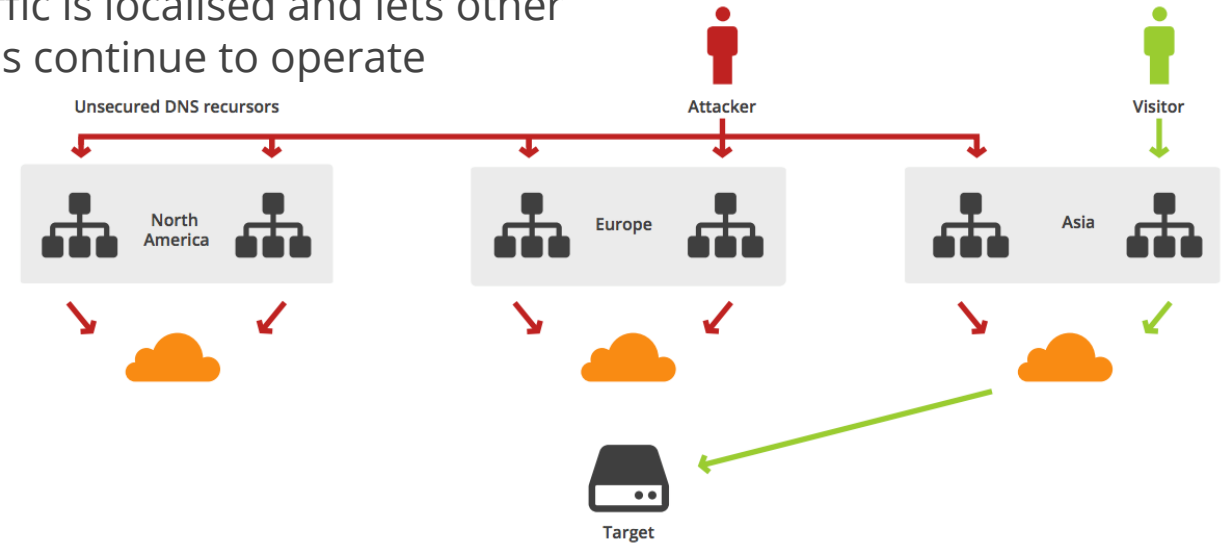
Tokyo

```
core1.tyo1.he.net> traceroute 198.41.208.141
 traceroute to 198.41.208.141 (198.41.208.141), 30 hops max, 60 byte packets
 1  ge2-3.core1.tyo1.he.net (74.82.46.5  0.192) ms  0.260 ms  0.288 ms
 2  13335.tyo.equinix.com (203.190.230.72)  0.467 ms  0.455 ms  0.479 ms
 3  198.41.208.141  0.465 ms  1.189 ms  0.527 ms
```

CloudFlare works globally

CloudFlare protects globally

- DDoS attack traffic is localised and lets other geographic areas continue to operate



Why do we peer?

Why do we peer?

"In computer networking, peering is a voluntary interconnection of administratively separate Internet networks for the purpose of exchanging traffic between the users of each network."

- To improve performance (reduce hop count, etc...)
- To reduce costs
- To ensure anycast traffic lands locally
- To gain more control over routing
- To gain more control of DDoS traffic

Where do we peer?

Where do we peer?

- AKL-IX (Auckland)
- AMS-IX (Amsterdam)
- APE (Auckland)
- CABASE-BUE (Buenos Aires)
- DE-CIX (Frankfurt)
- ECIX-DUS (Düsseldorf)
- ECIX-FRA (Frankfurt)
- ESPANIX (Madrid)
- Equinix (Ashburn, Atlanta, Chicago, Dallas, Hong Kong, Los Angeles, New York, Paris, San Jose, Seattle, Singapore, Sydney, Tokyo)
- FL-IX (Miami)
- France-IX (Paris)
- HKIX (Hong Kong)
- LINX (London)
- LONAP (London)
- MIX-IT (Milan)
- Megaport (Sydney, Auckland)
- NIX (Prague)
- NL-IX (Amsterdam)
- NOTA (Miami)
- Netnod (Stockholm)
- PIPE (Sydney)
- PLIX (Warsaw)
- PTT-SP (São Paulo)
- Peering.cz (Prague
- SIX (Seattle)
- STHIX (Stockholm)
- Telx (Atlanta)
- TorIX (Toronto)
- VIX (Vienna)

Why Madrid?

Why Madrid?

- 6th Largest city by population in Europe
- Many other smaller cities already covered
- It just makes sense!

The EspanIX experience

[redacted]

13 August 2014 08:15

CL

To: Marty Strong

Reply-To: [redacted]

Re: CloudFlare at EspanIX

Hi Marty,

It is Interxion who will perform your instructions there, we will just let them know your port ID. The freeze period ends in September, but we still have to receive our optics for you.

Regards,

[redacted]

> El 12 de agosto de 2014 a las 18:37 Marty Strong <marty@cloudflare.com> escribió:

See More from Marty Strong

The EspanIX experience

“Internet Exchange Points play a critical role in the way the internet works, not only through keeping traffic local, but enabling people and organizations to come together as a community.

If we continue to knock on the door and nobody answers, then a very fundamental piece is missing and we need to wonder what's going on.”

Christian Koch - Microsoft

What would we like to see done differently?

What would we like to see done differently?

- No more freezes, the internet doesn't sleep!
- A working management portal (e.g. https://github.com/inex/IXP-Manager)
- Machine readable member list (e.g. https://github.com/euro-ix/json-schemas)

Thank you!

Questions?

AS13335

http://as13335.peeringdb.com/

Marty Strong, Network Engineer
@martystronguk / @cloudflare
marty@cloudflare.com
https://www.cloudflare.com/

Additional material

DDoS Mitigation - in the network

Null route and move on

When an attacker targets a website or a service, while they may want to take this website/service down, they target the IP address in order to do this.

First order of business can be to update the DNS A/AAAA record and move on.

If the attacker follows, keep doing this.

Easy to automate, requires an attacker to continually change the attack to follow.

Depends on DNS resolver operators honouring our TTLs

FlowSpec (RFC 5575)

Important to understand from the outset that ALL flowspec does is automate the provisioning of a backplane-wide firewall filter on multiple devices. Having said that, **it does this really well**.

Can use most “from” and “then” actions available in Juniper firewall filters in FlowSpec. While Juniper have been an early adopter, other vendors have struggled to get this into their code. Even Juniper has only recently implemented IPv6 support for FlowSpec.

Being able to match “TCP packets from this /24, to this /32, with SYN but no ACK and a packet length of 63 bytes” and “rate-limit to 5Mbit” per edge router is incredibly useful.

Being able to configure this in one place and have it push to the entire network is awesome!

Regional enforcement

Under certain circumstances, it makes sense to enforce regionally

- Seeing 300Gbit of traffic targeted at AMS, LHR, FRA, CDG for a website with 99% of legitimate traffic being served into HKG and SIN
  - Can implement strict flowspec enforcement in sites targeted, while no enforcement needed in sites traffic is legitimately needed in.
  - Take advantage of any opportunity presented

Regional null routing can also be worthwhile at times

- Want to move site to new IPs and move on.
  - Null route in only the regions that are being targeted.

Have your transit provider configure firewall filters in their network to filter certain packet types / lengths / src-IPs / dst-IPs / etc upstream in one region only, to help filter malicious traffic.

Dealing with attacks on infrastructure IPs

Relatively easy to mitigate attacks on Anycast IP space.

- Multiple hundred gig attack on an anycast IP
  - Distributed over 30+ sites
  - Multiple tens of gigs per site

Vs:

- Multiple hundred gig attack on an IP specific to a single router, link or DC
  - Very hard to mitigate
  - Multiple hundred gig attack traffic > 100Gbit link

How do we prevent this from happening?

What can we do about it? What gain do you get from exposing this?

Attacks on Infrastructure - obfuscation of IPs

Traceroutes that show you the full path are nice... but... at what expense?

- Reveals a lot of the IP addressing information of your infrastructure to the entire internet
  - Becomes easy to figure out what to attack.
  - Makes every linknet, loopback, and infrastructure IP a target

Worth at least considering obscuring some of your infrastructure

- Stop responding to ICMP and UDP ttl expired
- Avoid ICMP-Packet-Too-Big in IPv6
  - Killing this can cause serious problems.

Attacks on Infrastructure - kill routability to IPs

Can take the next step and kill reachability entirely.

Make your linknet IPs non-routable;

- Take all your linknet IPs from a /24 that is not advertised on the internet
- Use RFC1918 space
- Blackhole all your linknets
  - Don’t forget to blackhole the provider side also!

This can make debugging significantly harder!

A lot of work will need to be done in the pre-sales stage with transit providers to ensure that one of these options is possible.

Peering exchanges should not be reachable on the internet anyway

Scaling the network to respond

Scaling the network - capacity

Ultimately, this is all a capacity game.

If you are seeing attacks roughly equivalent to your transit capacity you'll struggle to mitigate it, however if your transit capacity is 10x or 100x the size of the attack you'll struggle a lot less

Attack traffic will often be very very small packets (<80bytes) so need to ensure that any routers, line cards, are specced for capacity based on those numbers.

- As an example, MPC4E cards perform at nowhere near line rate when challenged with hundreds of gigs of 64 byte packets
- How oversubscribed is your backplane?
- Measurements should always be in PPS rather than Gbit/sec

As you scale up your routers, you may discover that PPS bottlenecks simply move to your transit providers.

Peering vs Transit

Far more difficult to mitigate a DDoS coming in on an IX than a DDoS coming in via a transit provider.

- Can negotiate with transit provider for features such as RTBH, NOC implementing firewall filters for you, etc
- Peering exchanges generally don't have these features.

Peering exchanges are also surprisingly expensive to scale up for DDoS. Generally will be more expensive to order more 10Gbit ports at an IX vs additional handovers to a transit provider.

Often end up de-peering a network sourcing large amounts of attack traffic to force them onto a transit provider where you have more control.

This seems broken - surely there is a better way to ingest this traffic?

The solution - lots and lots of PNIs

PNI with networks sourcing DDoS traffic in multiple locations

- Limits the scope of impact to the network sourcing the attack, or their upstream who you are peering with

Easy to scale very cheaply

- Only costs are router interfaces, DC cross connects

Can easily filter traffic from that specific network on your router interfaces

Scaling using caches

A step beyond scaling up the sites. Ability to place caches in hundreds, if not thousands of locations within ISP networks.

Distribute attack traffic significantly.