Routing for an Anycast CDN

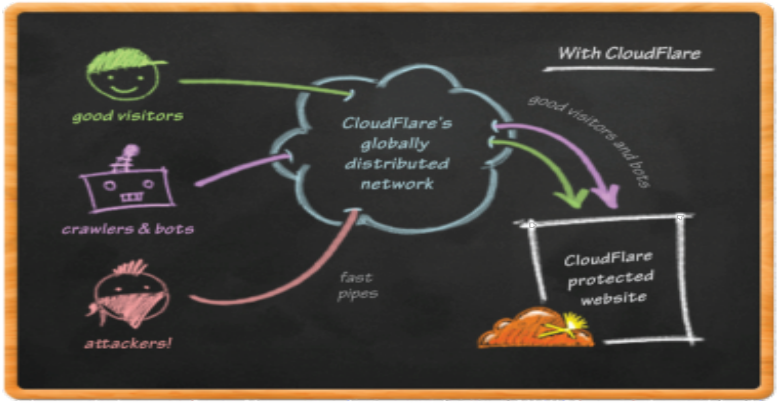
Martin J Levy @ CloudFlare

MENOG14 - Dubai - March 2014

What is CloudFlare?

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

  - Performance
  - Content Optimization
  - Security
  - Analytics
  - Third party services

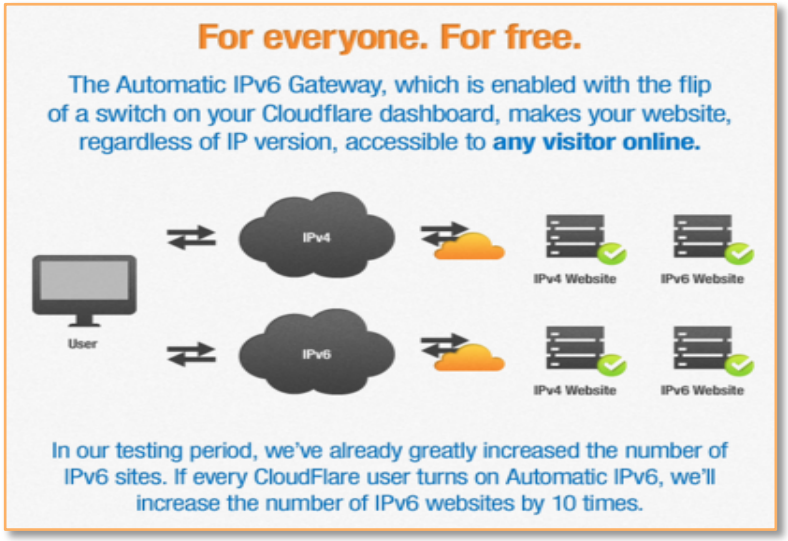


How does CloudFlare Work?

- CloudFlare works at the network level.
  - Once a website is part of the CloudFlare community, its web traffic is routed through CloudFlare's global network of 24 (and growing) data centers.
  - At each edge node, CloudFlare manages DNS, caching, bot filtering, web content optimization and third party app installations.

IPv4/IPv6 – automatically enabled

- With the Internet's explosive growth and the number of on-net devices closing in on IPv4's maximum capacity, CloudFlare now offers an automatic IPv6 gateway seamlessly bridging the IPv4 and IPv6 networks
  - For most businesses, upgrading to the IPv6 protocol is costly and time consuming
  - CloudFlare's solution requires NO hardware, software, or other infrastructure changes by the site owner or hosting provider
  - Enabled via the flip of a switch on the site owner's CloudFlare dashboard
  - Users can choose two options: (FULL) which will enable IPv6 on all subdomains that are CloudFlare Enabled, or (SAFE) which will automatically create specific IPv6-only subdomains (e.g. www.ipv6.yoursite.com)



Anycast CDN

Anycast CDN

- Anycast prefixes
  - Same IP prefixes (IPv4 & IPv6) advertised in each of the 24 sites around the world (and growing)
  - Unicast used to pull traffic from “origin” web source
- Traffic Control
  - How the eyeball ISP routes
  - If ISP A routes to CloudFlare in Germany then traffic will be served for ISP A from Frankfurt
  - If ISP B routes to CloudFlare in Central USA then traffic will be served for ISP B from Dallas or Chicago

Anycast CDN

- Traceroute from Hong Kong

```
traceroute to 173.245.61.248 (173.245.61.248), 30 hops max, 40 byte packets
 1  202-150-221-169.rev.ne.com.sg (202.150.221.169)     0.351 ms    0.406 ms    0.456 ms
 2  s4-6-r10.cyberway.com.sg (203.117.6.209)            0.610 ms    0.652 ms    0.692 ms
 3  anutsi10.starhub.net.sg (203.118.3.162)             2.579 ms    2.575 ms    2.562 ms
 4  six2utsi1.starhub.net.sg (203.118.3.189)            1.452 ms    1.633 ms    1.768 ms
 5  SH.gw5.sin1.asianetcom.net (203.192.169.41)         1.561 ms    1.620 ms    1.610 ms
 6  te0-0-2-0.wr1.sin0.asianetcom.net (61.14.157.109)   2.135 ms    1.921 ms    1.950 ms
 7  gi4-0-0.gw2.sin3.asianetcom.net (61.14.157.134)     1.909 ms    1.907 ms    1.882 ms
 8  CDF-0003.gw2.sin3.asianetcom.net (203.192.154.26)   1.417 ms    1.504 ms    1.493 ms
 9  cf-173-245-61-248.cloudflare.com (173.245.61.248)   1.470 ms    1.461 ms    1.520 ms
Traceroute Completed.
```

- Traceroute from Singapore

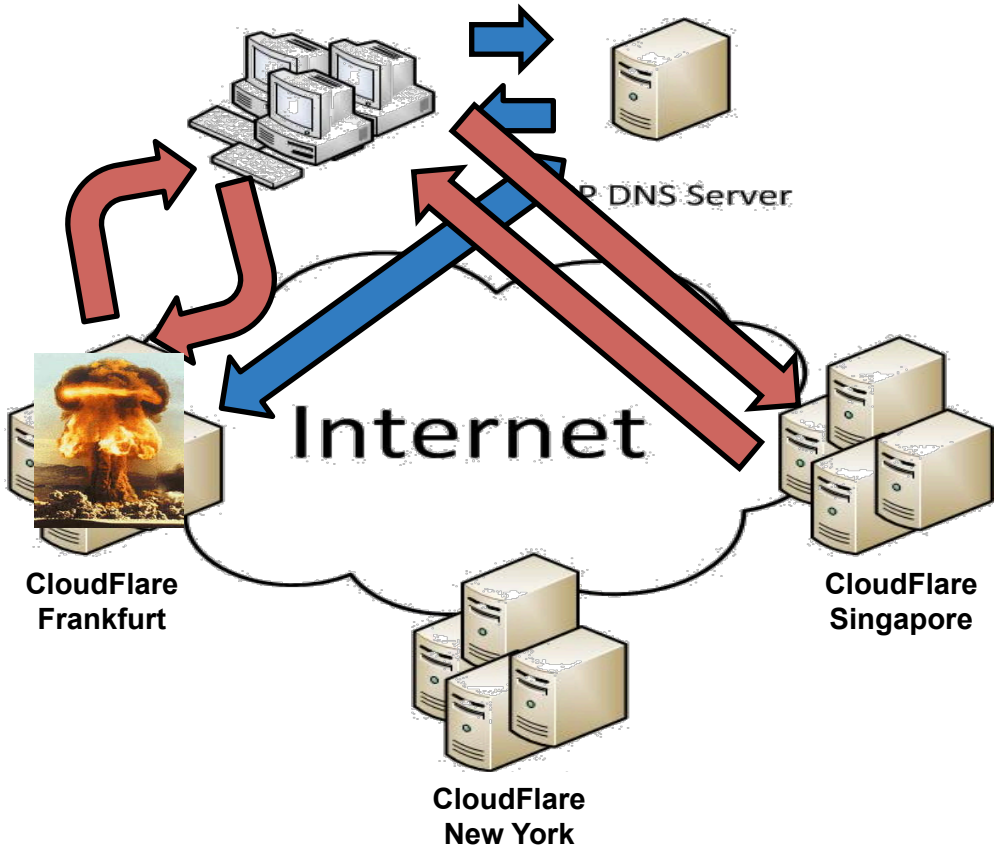
```
traceroute to 173.245.61.248 (173.245.61.248), 64 hops max, 44 byte packets
 1  bbs-1-250-0-210.on-nets.com (210.0.250.1)           0.423 ms    0.329 ms    0.320 ms
 2  10.2.193.17 (10.2.193.17)                           0.719 ms    0.661 ms    0.682 ms
 3  peer (218.189.96.62)                                0.569 ms    0.550 ms    0.545 ms
 4  cloudflare-RGE.hkix.net (202.40.160.246)            1.893 ms    2.419 ms    1.910 ms
 5  cf-173-245-61-248.cloudflare.com (173.245.61.248)   2.101 ms    1.973 ms    1.780 ms
Traceroute Completed.
```

- Traceroute from Dubai

```
traceroute to 173.245.61.248 (173.245.61.248), 64 hops max, 52 byte packets
 1  10.50.0.1 (10.50.0.1)                               7.423 ms   56.679 ms   23.776 ms
 2  94.200.91.194 (94.200.91.194)                      22.342 ms   35.165 ms    9.889 ms
 3  10.171.0.49 (10.171.0.49)                          20.604 ms   28.953 ms   24.390 ms
 4  10.128.144.29 (10.128.144.29)                      24.678 ms    6.069 ms    5.836 ms
 5  10.44.19.177 (10.44.19.177)                         9.389 ms    7.570 ms    8.151 ms
 6  10.44.247.89 (10.44.247.89)                         6.074 ms    6.196 ms    8.031 ms
 7  94.201.0.65 (94.201.0.65)                          21.275 ms   14.462 ms   11.681 ms
 8  10.44.24.58 (10.44.24.58)                         194.345 ms  162.052 ms  199.916 ms
 9  * * *
10  cf-173-245-61-248.cloudflare.com (173.245.61.248) 152.790 ms  167.038 ms  143.949 ms
11  Traceroute Completed.
```

Anycast CDN

- DNS Query
- DNS result returned with “Anycast” IP
- Client makes connection to closest server
- CloudFlare replies
- Outage Re-routes to next closest cluster



Transit

Transit

- Who?
  - Choice of Transit Provider is VERY important
  - We've chosen one provider per region – One in US/EU and one in Asia (or the same provider for both regions)
  - Single Provider makes routing easier
  - Transit provider should offer good routing controls
    - You need to be able to keep routes within a region
    - Prepend to specific peers
  - Transit Provider should make use of "Hot Potato" routing to their peers
    - i.e. Peer and exchange traffic in every mutual location

Transit

- Routing Controls?
  - Transit must be able to keep advertisements within region.
    - A customer of your European transit provider is likely to be a peer of your Asian transit provider
    - You don't want to serve traffic from Asia for Europe
  - A lot of work should be done in the presales stage to understand the providers network and how they peer.
- Example location to look for controls
  - Looking at AS1299's (Telia-Sonera) whois entry gives a good idea how they peer
  - Some routing controls listed at http://www.onesc.net/communities/

Transit

- Choices?
  - Many providers give you good coverage for common US/EU locations (LA, New York, London, Amsterdam, etc ...)
  - One provider can’t do it all in Asia
  - Asian networks are usually somewhat ‘disconnected’
    - Few peer with NTT in Asia & NTT, Pacnet and TATA are all disconnected from each other
  - Transit in the US/EU could be far cheaper for the provider than within Asia or Africa or Middle East
  - Supplement this with peering in all regions

Peering

Peering

- North America Peering
  - Is it economic to peer?
  - Transit is < $1
  - Eyeball networks probably *wont* peer with you
    - Comcast (not at any exchange)
    - ATT
  - South America?
    - Peering in Miami
    - Most networks open to peering
- EU Peering
  - Same argument as US, might be more costly to peer
  - Many networks open to peering however
  - Major providers / incumbents more difficult, probably wont peer:
    - DTAG, TeliaSonera
    - Telecom Italia Sparkle
    - Telefonica, France Telecom
  - IX's have good reach to surrounding regions.
    - AMS-IX, DE-CIX, NETNOD, LINX
- Asia Peering
  - Very economical; however large providers may not peer
  - HKIX and Hong Kong Equinix
    - No IX charges and HKIX will get you 100% of domestic Hong Kong.
    - Very good Vietnam and some Taiwan, Korea, Japan and China routes too
  - Singapore Equinix
    - Priced competitively and great coverage for South East Asia (Indonesia, Thailand, Malaysia, India)
  - Tokyo Japan
    - JPIX and JPNAP much more costly

Challenges

Challenges

- Challenges
  - Routing
  - Inefficient routing, optimizing.
  - Turning up peering, causing unexpected routing changes
  - Russian Network preferred our routes via HKIX instead of in Europe.
  - Keeping optimal routing to Eyeball Networks
- Deployments into new markets
  - China, South America, Africa, Middle East

Questions?

martin@cloudflare.com
AS13335