Building and operating a global DNS content delivery anycast network

APTLD 72

Gael Hernandez
Packet Clearing House
Tbilisi, 15 September 2017
Who are we?

- Packet Clearing House (PCH) is the global non-profit organisation providing operational support and security to critical Internet infrastructure, including Internet exchange points and the core of the DNS, since 1993.

- Funded by government grants, service-provision fees from the Internet operations industry and specialised consultancies on IXP construction/operation.

- Global human footprint with head office in San Francisco (US) and regional offices in Buenos Aires, Johannesburg, Dublin and Kathmandu.
An anycast cloud is a distributed cluster of identical instances of a server, each typically containing identical data, and capable of servicing requests identically.

Each instance has a regular unique globally routable IP address for management purposes, but... each instance also shares an IP address in common with all the others.

The Internet’s normal global routing system (BGP) routes every query to the instance of the anycast cloud that is closest in routing terms to the user who originated the query.
Anycast technology (ii)

DNS lookup of `ns.pch.net` produces a single answer

```
ns.pch.net       IN    A    206.220.231.1
```

Router 1 routing table

<table>
<thead>
<tr>
<th>Destination</th>
<th>Mask</th>
<th>Next-Hop</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>192.0.2.0</td>
<td>/29</td>
<td>127.0.0.1</td>
<td>0</td>
</tr>
<tr>
<td>206.220.231.1</td>
<td>/32</td>
<td>192.0.2.1</td>
<td>1</td>
</tr>
<tr>
<td>206.220.231.1</td>
<td>/32</td>
<td>192.0.2.2</td>
<td>2</td>
</tr>
<tr>
<td>206.220.231.1</td>
<td>/32</td>
<td>192.0.2.3</td>
<td>3</td>
</tr>
</tbody>
</table>
Anycast technology (iii)
Anycast for DNS

- PCH and its precursors have run production anycast services have been run since 1989.
- Bill Woodcock (PCH) and Mark Kosters (then at Verisign) first proposed the idea of anycasting authoritative root and TLD DNS at the Montreal IEPG in 1995.
- PCH began operating production anycast for ccTLDs and in-addr in 1997, and there’s been 100% up-time over more than twenty years.
- PCH first hosted an anycast production of a root name server in 2002. We operate services through IPv6 since 2000.
PCH’s Anycast Cloud (AS42)

- 118 anycast nodes in all five continents
  - 14 global nodes + 4 high traffic nodes
- 152 locations in five continents
  - 33 in ARIN region
  - 28 in RIPE region
  - 25 in AFRINIC region
  - 18 in APNIC region
  - 14 in LACNIC region
- 2,691 unique ASN peers
  - 150 route-servers ASN
- Secondary authoritative service to 400+ TLDs and two letters of the DNS root.
  - ~105 ccTLDs
  - ~120 million resource records
PCH’s Anycast Network (ii)

- By some estimates, the Internet duplicates in size every year.
- The growth rate of our locations follows very closely the creation rate of Internet exchanges, ~15/20 per year for the past 10 years.
- Our expansion rate has allowed us to increase the number of hosted TLDs without affecting the performance.
PCH’s 8th generation architecture

• Small, medium and full cluster installations.
• Routing vendor redundancy: Cisco and Quagga.
• Cisco servers with hardware specs based on site demand.
• VMware ESX clusters, supporting any x86 64-bit OS.
• Hosted servers fully integrated with BGP routing architecture.
• OS redundancy: Solaris and CentOS.
• Name server redundancy: Bind and NSD.

• Long-term strategic relationships with all involved vendors:
  • Cisco, AMD, Sun, VMware, ISC, and NLNet Labs.
Planning Anycast Nodes

• Anycast is a robust and well-proven technology
  • E-root is the fastest in the U.S., South Africa, Poland, Ireland, and Malaysia and D-root is the fastest in the U.K., Netherlands, Austria, and Thailand (Thousand Eyes, June 2017)

• Delivering content in some regions is challenging
  • Less developed interconnection market in emerging economies
  • Absence of open and neutral exchanges with public peering
  • Large networks won’t be peering at small exchanges

• Considerations when planning for new sites
  • Invitation from an IX operator to host a DNS node
  • Traffic levels, number of participants and prefixes at the IX
  • Availability of our transit providers
  • Relative location of other nodes
Operations

- Services run in separated virtual machines
  - Dedicated VMs for root servers, TLDs and monitoring services.

- A failure in the DNS service triggers the removing of the node from the routing table by stopping its BGP announcement

- Depending on the type of deployment (small/medium/large) and type of node (local/global), we announce via BGP a full or a partial set of services:
  - Small sites: anywhere in the world, local-only and partial service announcements.
  - Medium sites: medium to high-volume locations, local-only and partial service announcements.
  - Full sites: global nodes in high volume locations, with full service announcements via our transit providers NTT and Level3.
Monitoring

• Multiple layers of monitoring to proactively detect issues that could be leading to a degradation of the service
  • Hardware layer: CPU levels, temperature, RAM.
  • Interconnection layer: ports and traffic levels.
  • Routing layer: AS-PATH and prefix announcements.
  • Service layer: queries per second, replies per second.

• Passive monitoring tools
  • Nagios with custom plugins for DNS and DNSSEC
  • Netflow monitoring traffic levels

• Active monitoring of global performance using RIPE Atlas and RIPE DNSMon measurements on a regular basis
A day in PCH’s anycast network
A day in PCH’s anycast network (ii)

DNS queries processed by global and rest of nodes

DNS queries by region

- Rest of nodes
- Global nodes

0:00 2:00 4:00 6:00 8:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00

Time (UTC)

- A day in PCH’s anycast network (ii)
- DNS queries processed by global and rest of nodes
- DNS queries by region

0:00 2:00 4:00 6:00 8:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00

Time (UTC)
What keeps us busy?

- UDP spoofing and network operators not implementing BCP38
- Network operators doing too much traffic engineering
- Critical zero-day exploits affecting name servers and other critical software
- Automating the provisioning process to reduce the time to deploy new nodes
- Research lab work and benchmark of software alternatives, for instance Knot DNS by CZ.NIC.
Questions?
Thanks for your attention

Gael Hernandez
Senior Manager, Interconnection Policy and Regulatory Affairs
gael@pch.net