PCH
Packet Clearing House

Building and operating a global DNS content delivery anycast network

APRICOT 2018

Gael Hernandez

Packet Clearing House

Kathmandu, Nepal

Agenda

- Anycast introduction
- Overview of PCH’s anycast network
- A day in PCH’s network
- Planning Anycast Nodes
- Operation and Monitoring

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.

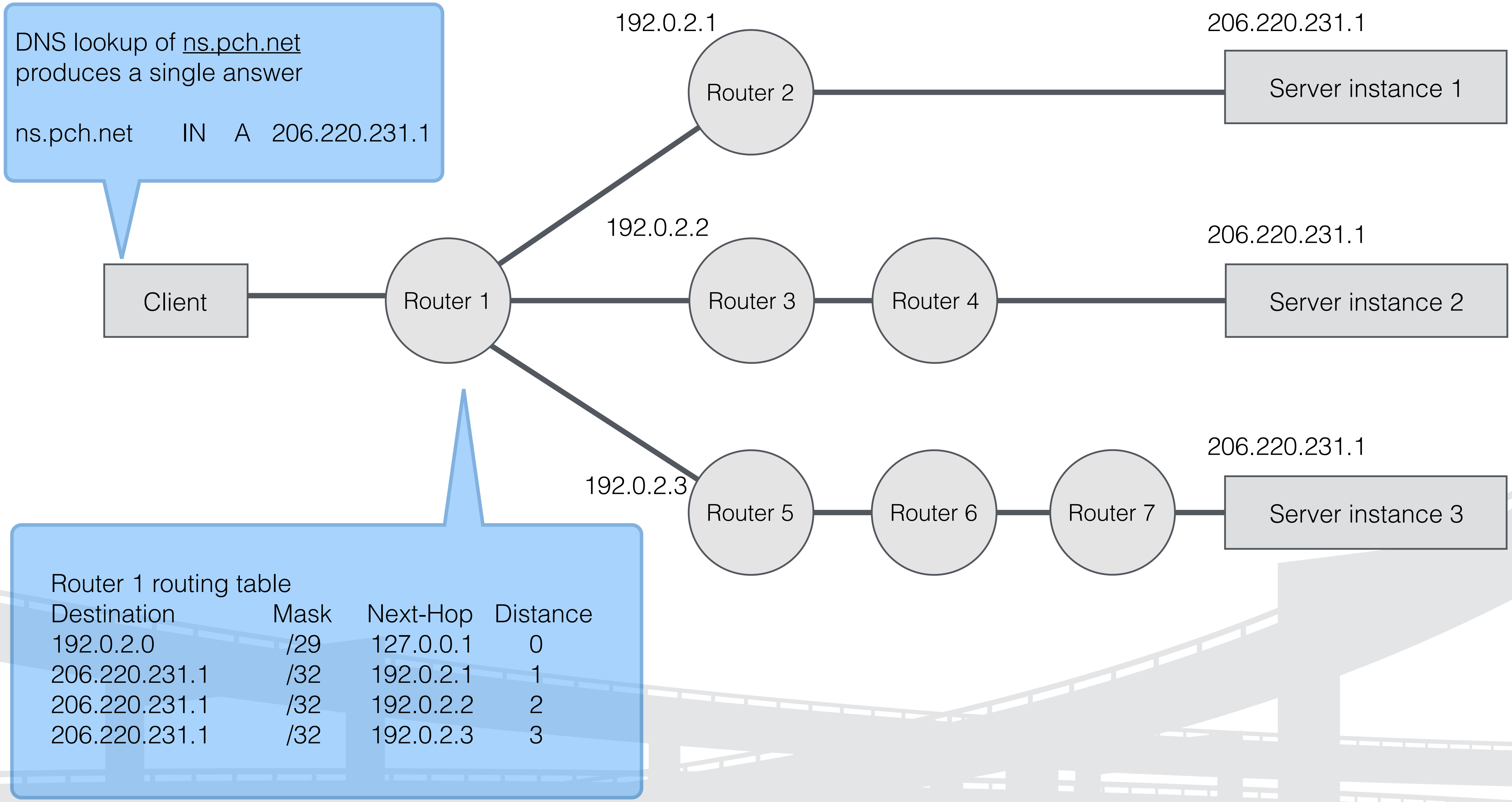
- Global footprint with head office in San Francisco (US) and regional offices in Buenos Aires, Johannesburg, Dublin and Kathmandu.

Packet Clearing House

Anycast technology

- Routing technology used among content delivery networks as it allows optimal routing to closer cluster.
- 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 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)



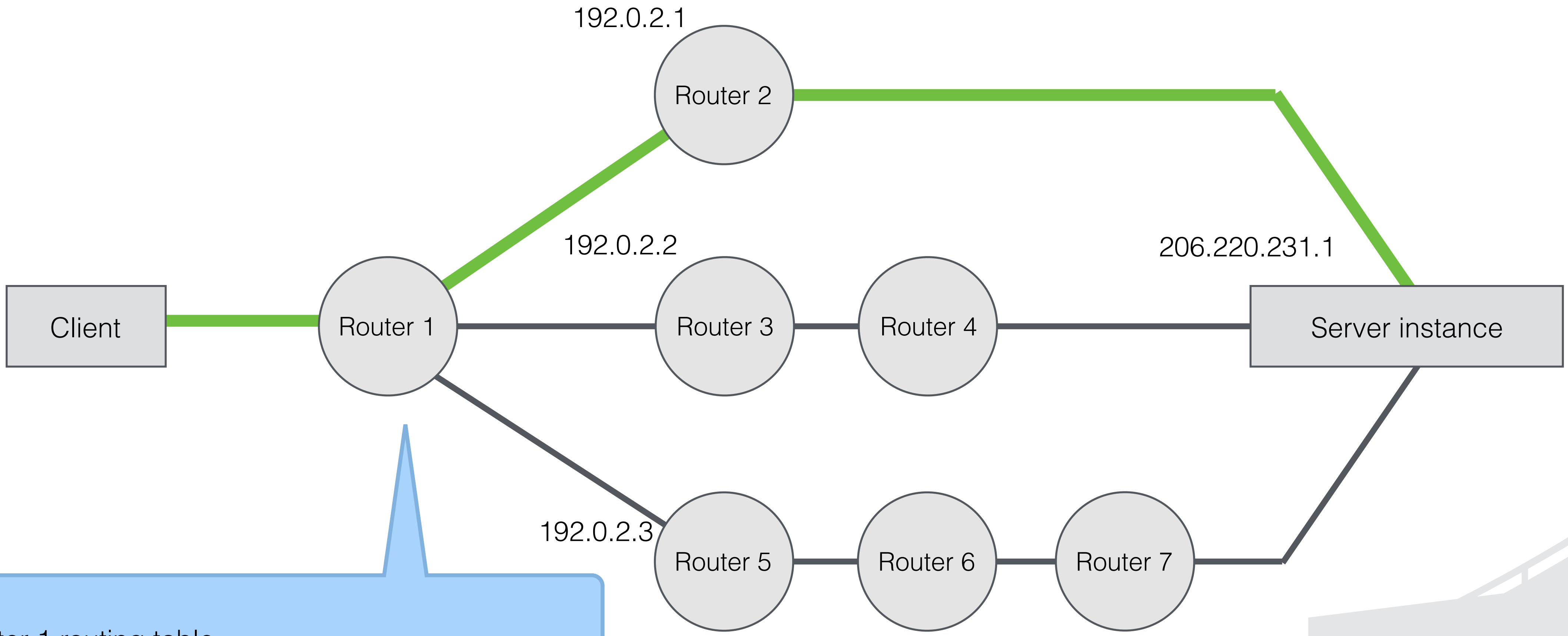
DNS lookup of ns.pch.net
produces a single answer

ns.pch.net IN A 206.220.231.1

Router 1 routing table

Destination	Mask	Next-Hop	Distance
192.0.2.0	/29	127.0.0.1	0
206.220.231.1	/32	192.0.2.1	1
206.220.231.1	/32	192.0.2.2	2
206.220.231.1	/32	192.0.2.3	3

Anycast technology (iii)



Router 1 routing table

Destination	Mask	Next-Hop	Distance
192.0.2.0	/29	127.0.0.1	0
206.220.231.1	**/32**	**192.0.2.1**	**1**
206.220.231.1	/32	192.0.2.2	2
206.220.231.1	/32	192.0.2.3	3

Anycast for DNS

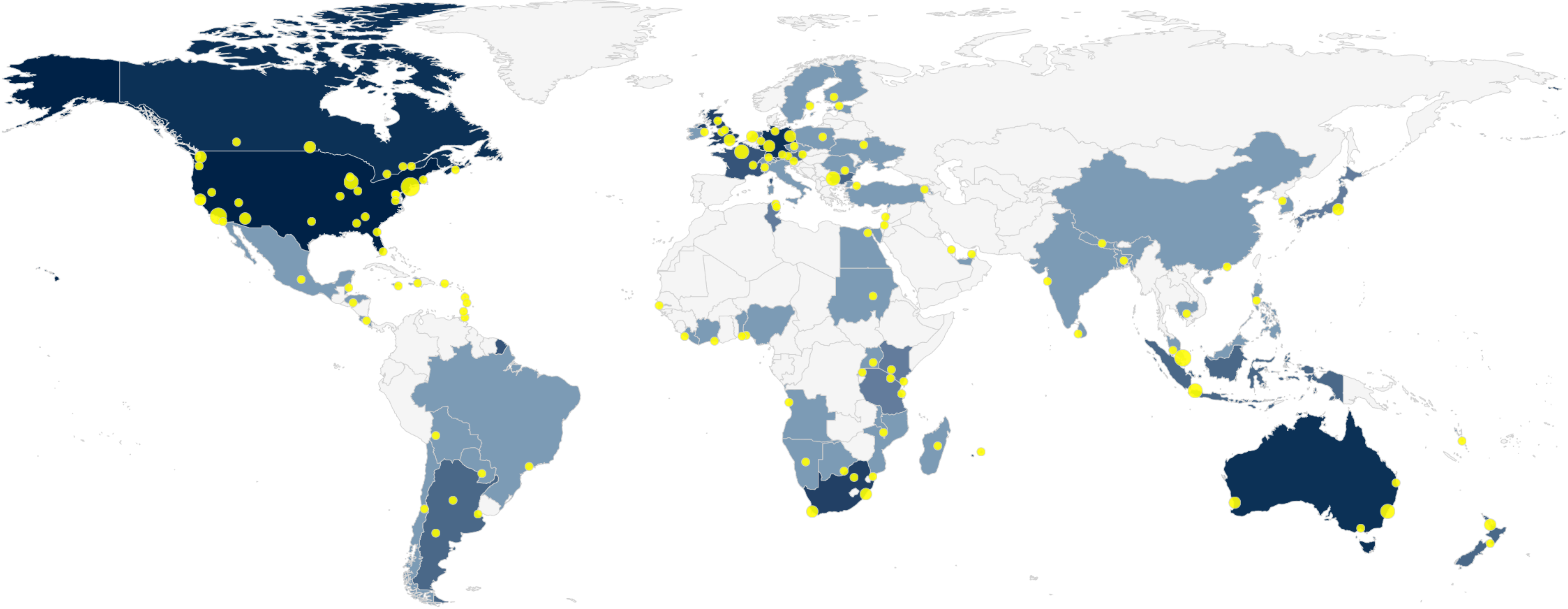
- PCH and its precursors have run production anycast services have been run since 1989.

- 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 Network AS42



120 nodes

14 global nodes

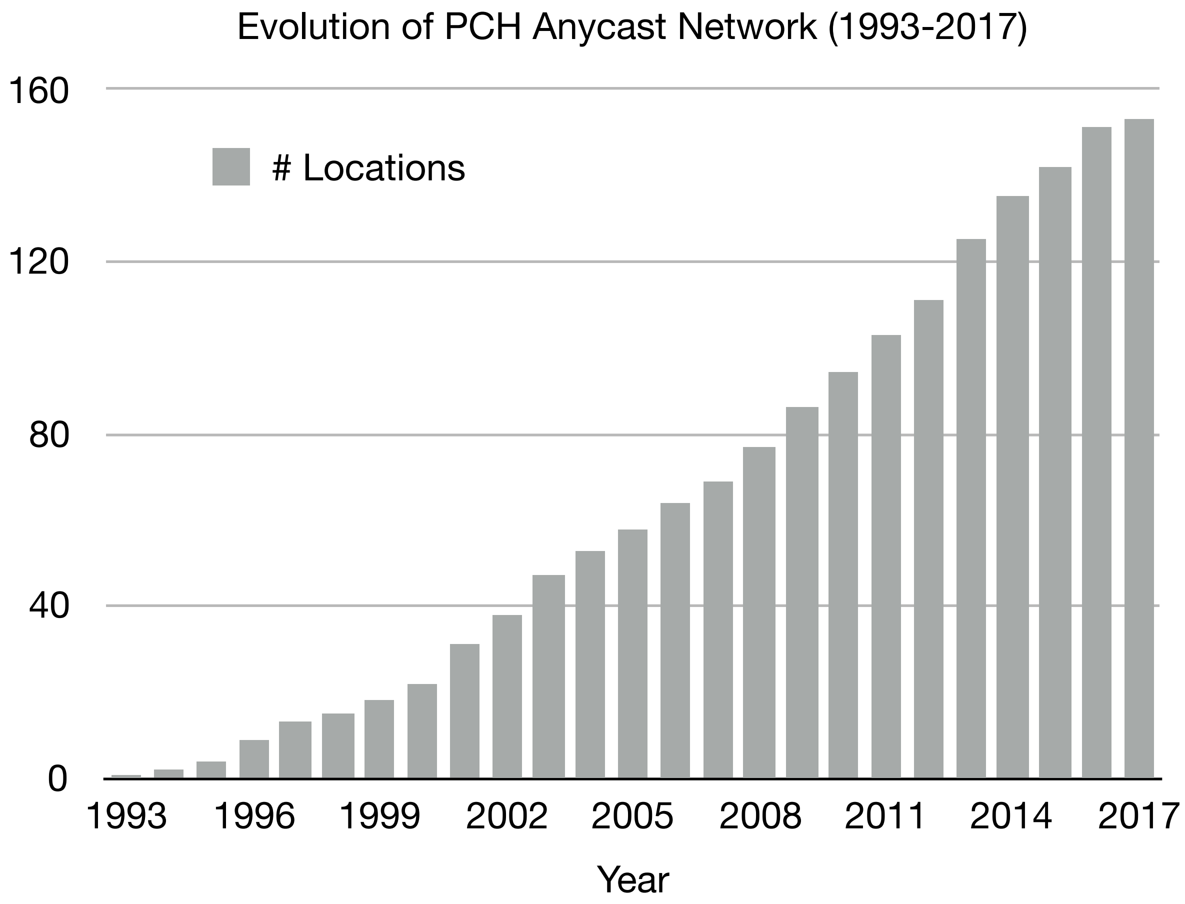
167 IXP locations

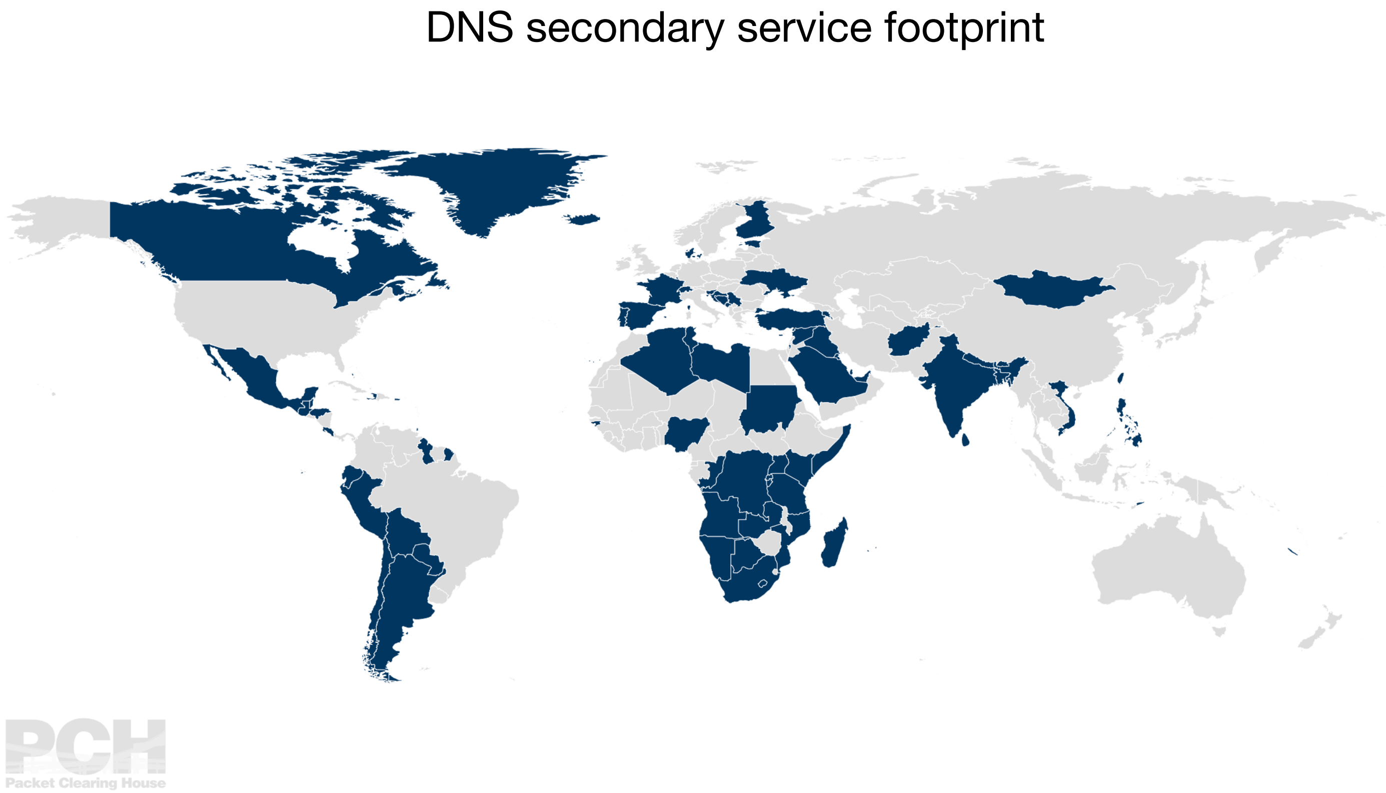
18 in APNIC region

+3000 unique ASN peers

150 route-servers ASN

PCH's Anycast Network AS42 (ii)





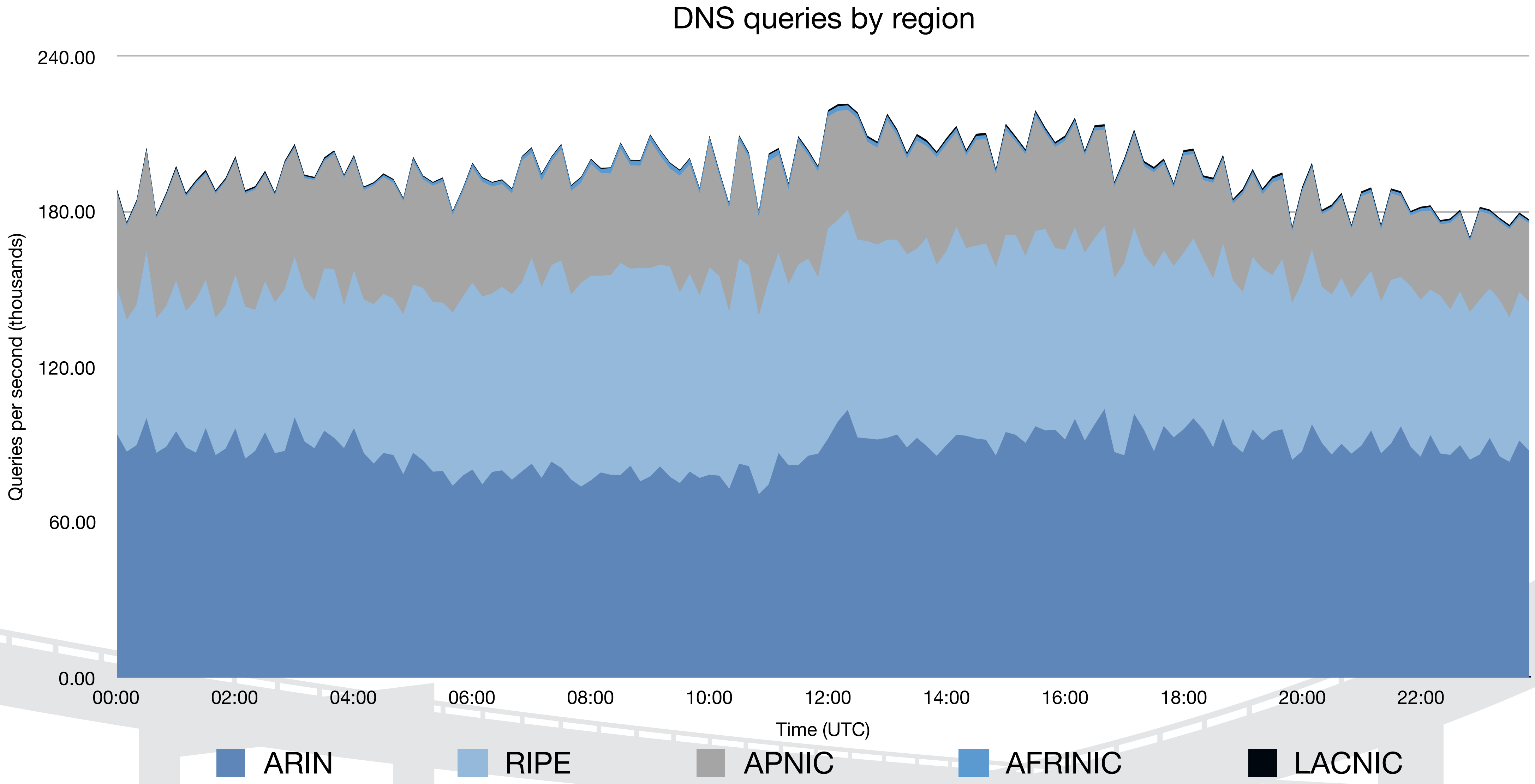
- By some estimates the Internet duplicates in size every 13 months.
- The growth rate of our locations follows very closely the creation rate of Internet exchanges, ~15/20 per year for the past 10 years.
- Our expansion rate has allowed us to increase the number of hosted TLDs without affecting the 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.

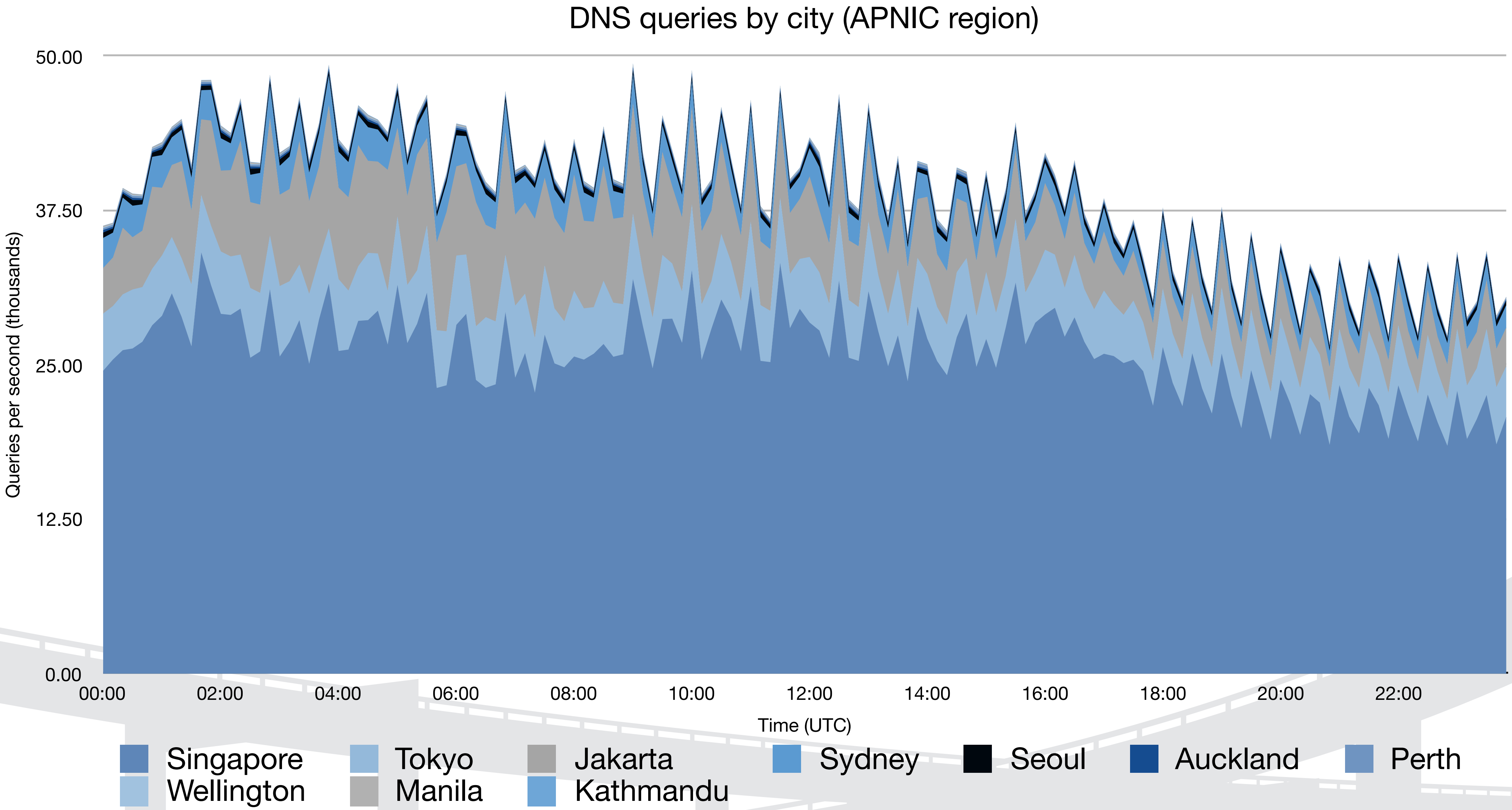
- Long-term strategic relationships with all involved vendors:
    - Cisco, AMD, Sun, VMware, ISC, and NLNet Labs.

A day in PCH’s anycast network

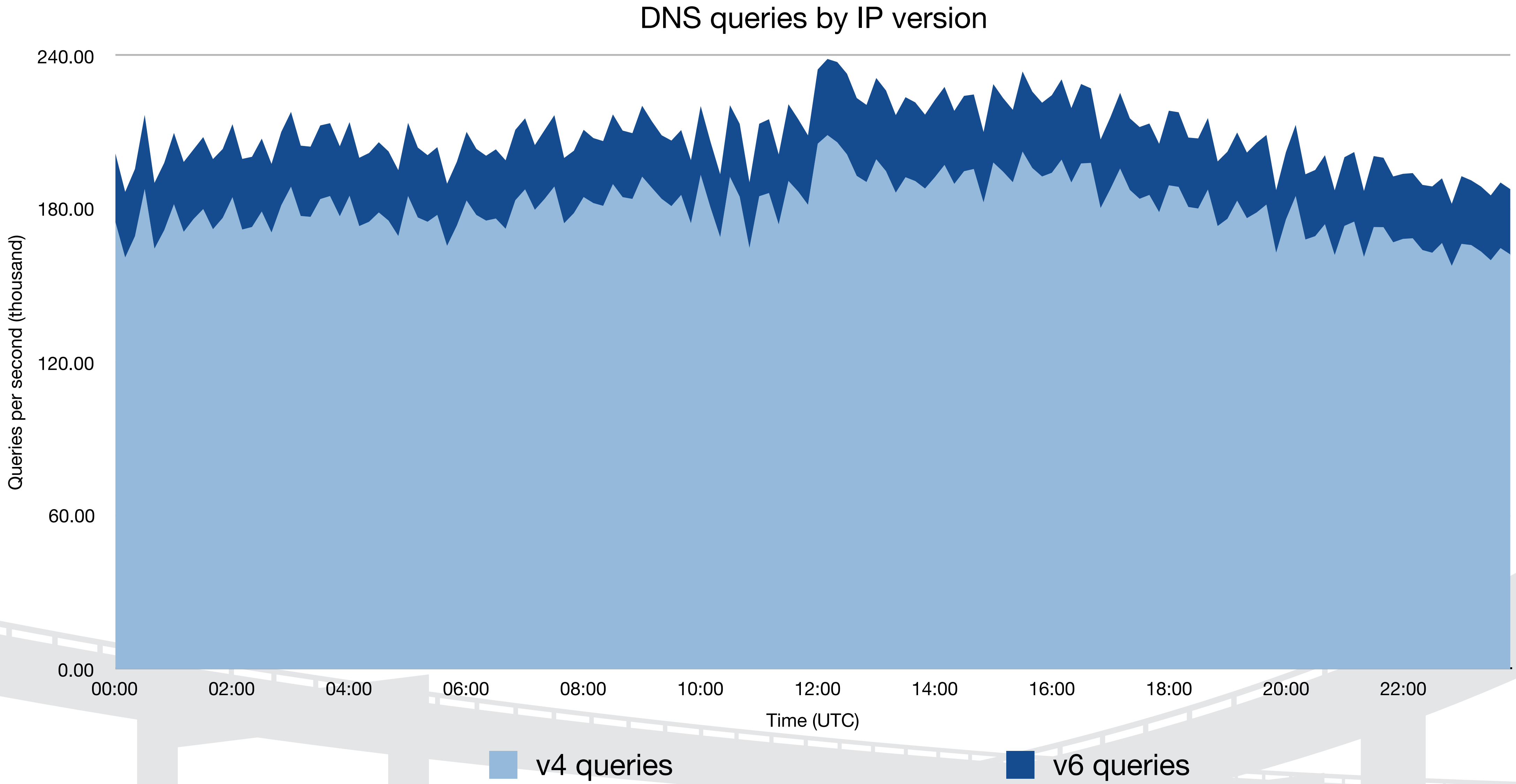


Packet Clearing House

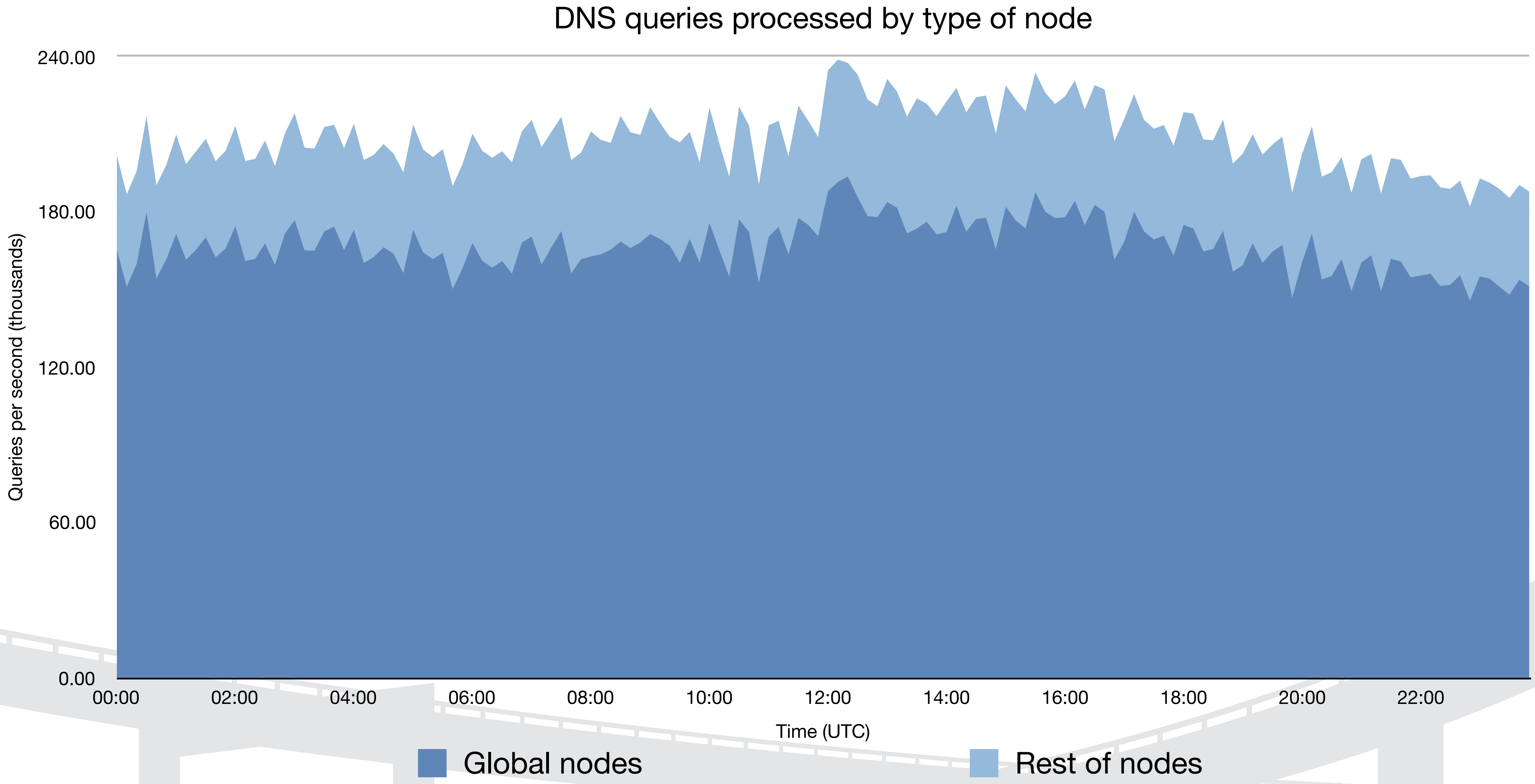
A day in PCH’s anycast network (ii)



A day in PCH’s anycast network (iii)



A day in PCH’s anycast network (vi)



Planning Anycast Nodes

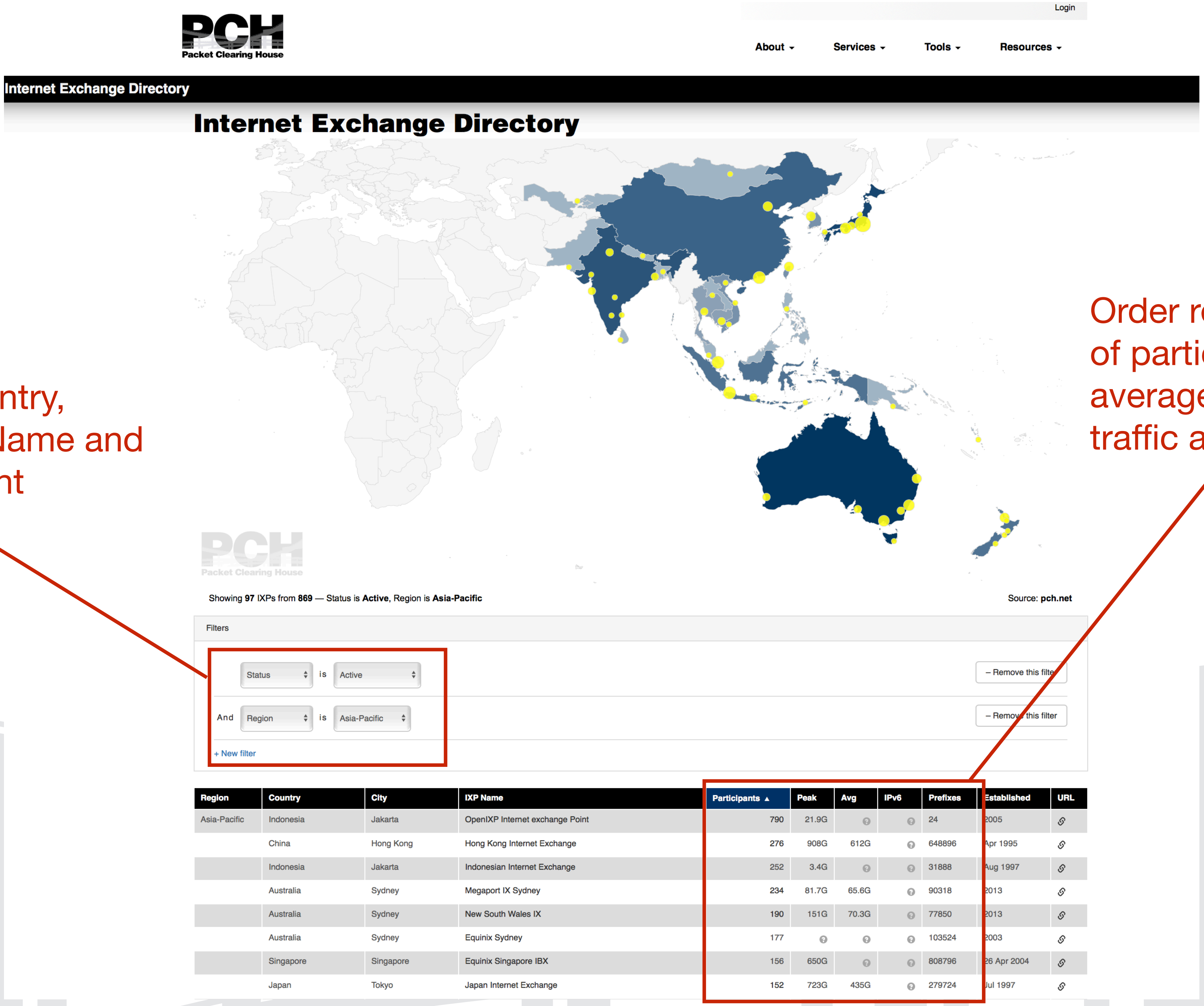
- Anycast is a robust and well-proven technology
  - **E-root** is the fastest in the U.S., South Africa, Poland, Ireland, and Malaysia and **D-root** is the fastest in the U.K., Netherlands, Austria, and Thailand (Thousand Eyes, June 2017)

- 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 (NTT and CenturyLink)
  - Relative location of neighbouring nodes

- 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

PCH
Packet Clearing House

Using IXPDirectory for planning purposes



Internet Exchange Directory

Showing **97** IXPs from **869** — Status is **Active**, Region is **Asia-Pacific**

Source: **pch.net**

Filters

Status is Active

And Region is Asia-Pacific

+ New filter

Region	Country	City	IXP Name	Participants ▲	Peak	Avg	IPv6	Prefixes	Established	URL
Asia-Pacific	Indonesia	Jakarta	OpenIXP Internet exchange Point	790	21.9G	?	?	24	2005	🔗
	China	Hong Kong	Hong Kong Internet Exchange	276	908G	612G	?	648896	Apr 1995	🔗
	Indonesia	Jakarta	Indonesian Internet Exchange	252	3.4G	?	?	31888	Aug 1997	🔗
	Australia	Sydney	Megaport IX Sydney	234	81.7G	65.6G	?	90318	2013	🔗
	Australia	Sydney	New South Wales IX	190	151G	70.3G	?	77850	2013	🔗
	Australia	Sydney	Equinix Sydney	177	?	?	?	103524	2003	🔗
	Singapore	Singapore	Equinix Singapore IBX	156	650G	?	?	808796	26 Apr 2004	🔗
	Japan	Tokyo	Japan Internet Exchange	152	723G	435G	?	279724	Jul 1997	🔗

Filter by: Country, Region, IXP Name and PCH is present

Order results by number of participants, peak and average traffic, % of v6 traffic and prefixes

Operations

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

- Depending on the type of deployment (small/medium/large) and type of node (local/global), we BGP-announce 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.

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

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
  - Cacti/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

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