

Operating a Global DNS anycast network

SANOG - 43

Dibya Khatiwada

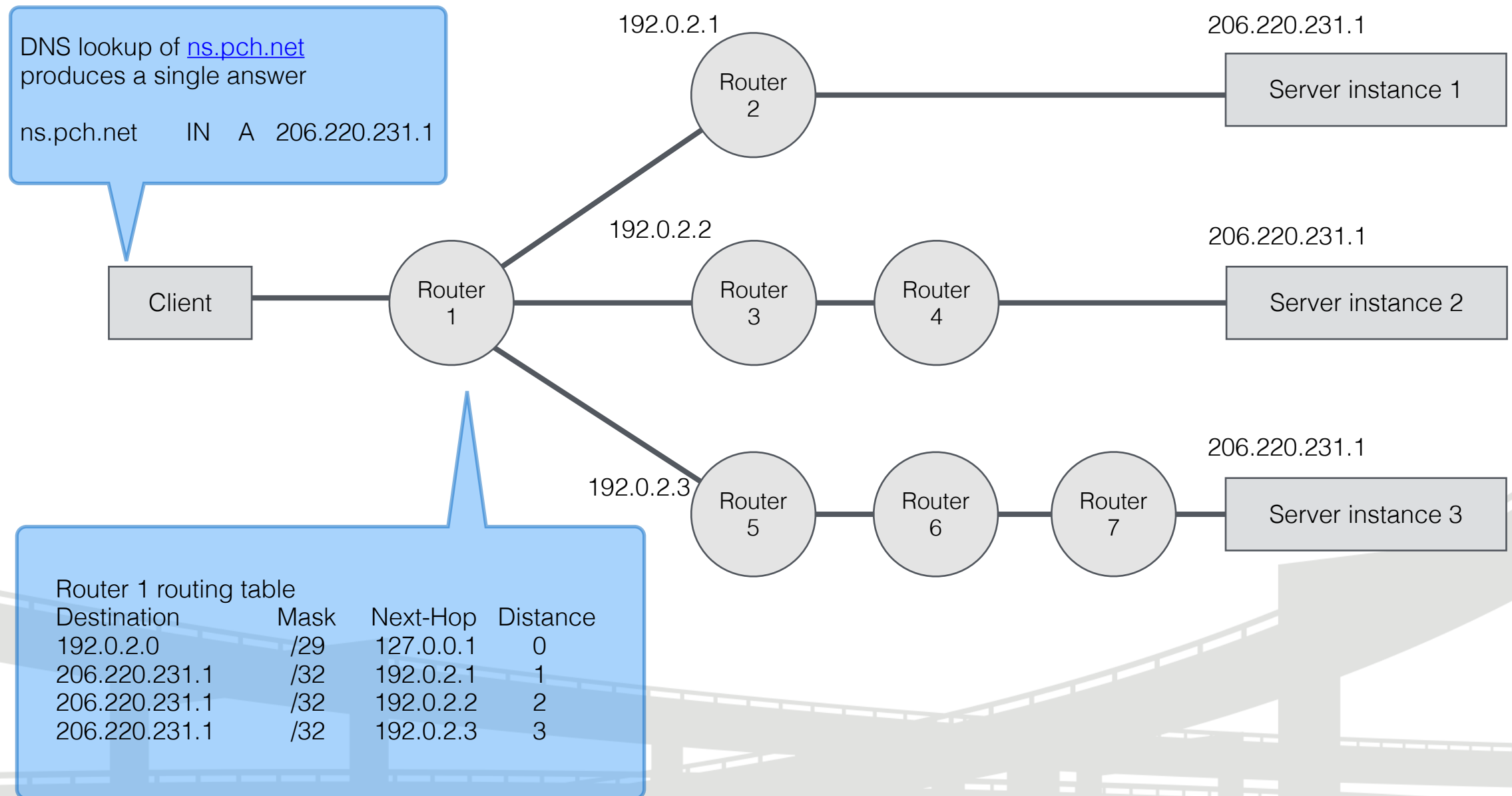
dibya@pch.net

Packet Clearing House (PCH)

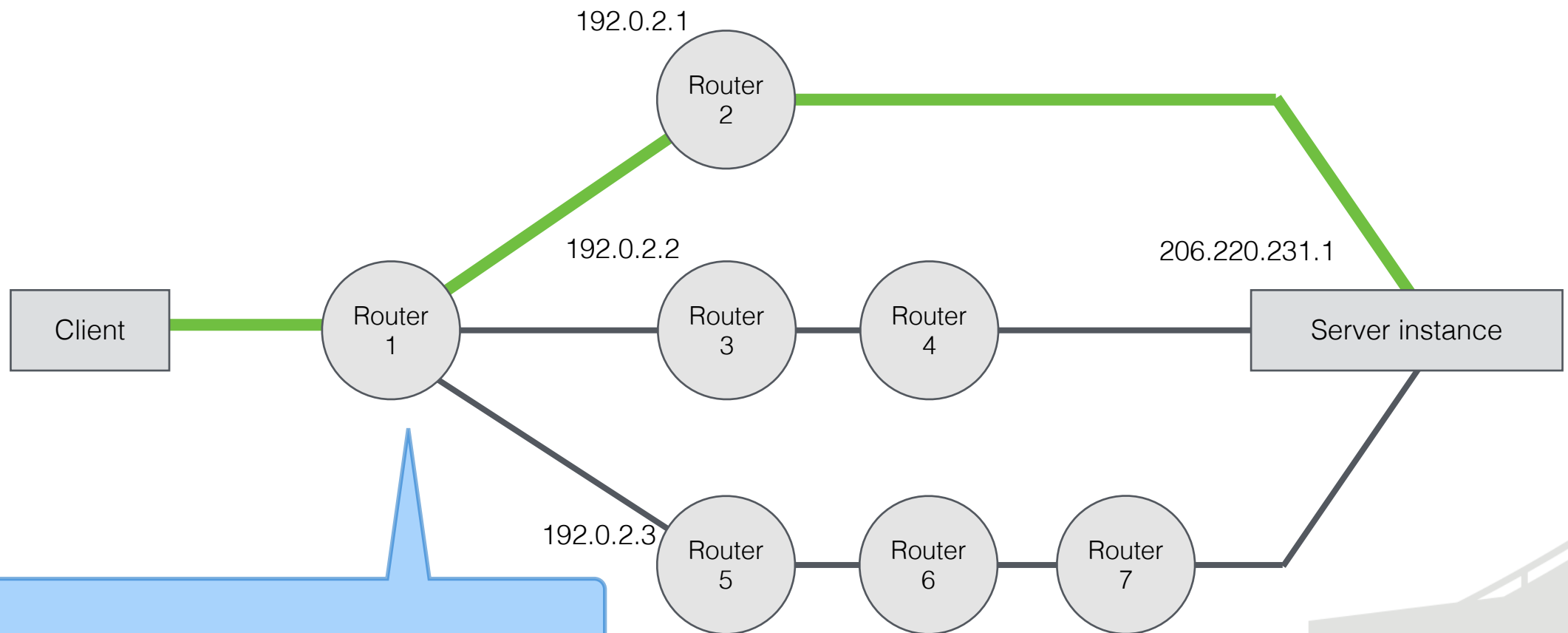
Anycast Technology

- 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)



Anycast technology (iii)



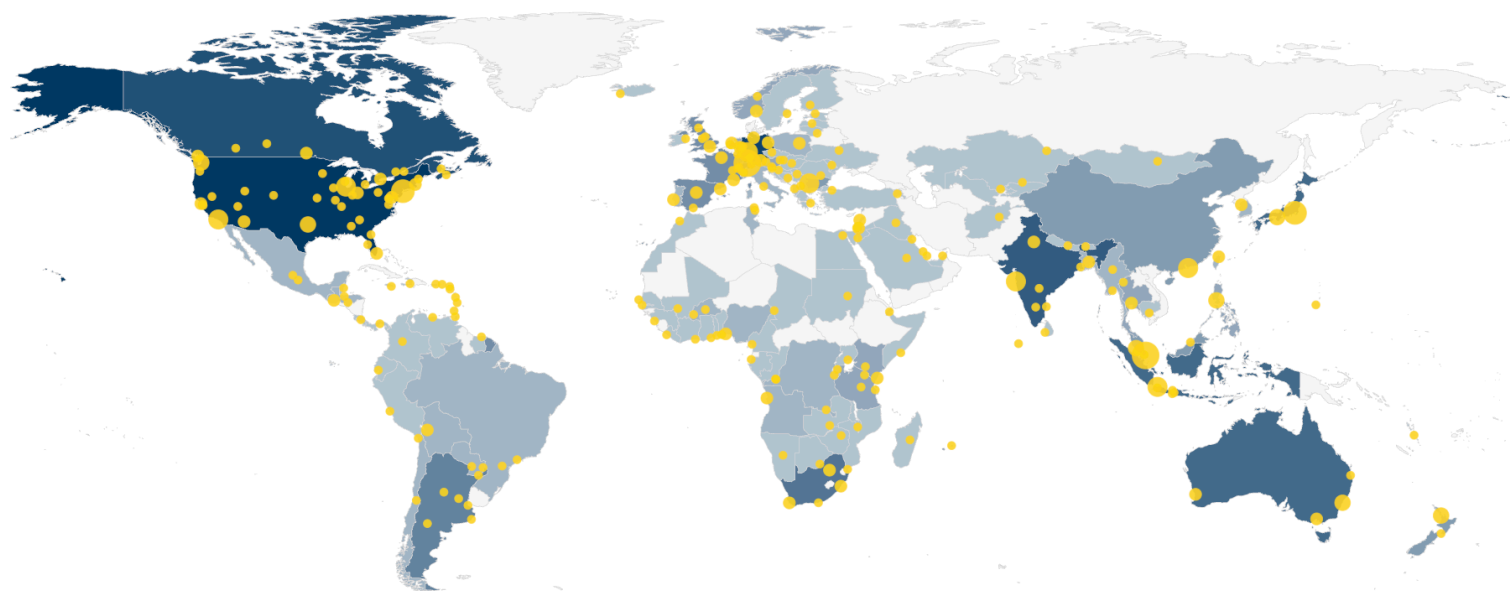
Router 1 routing table

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

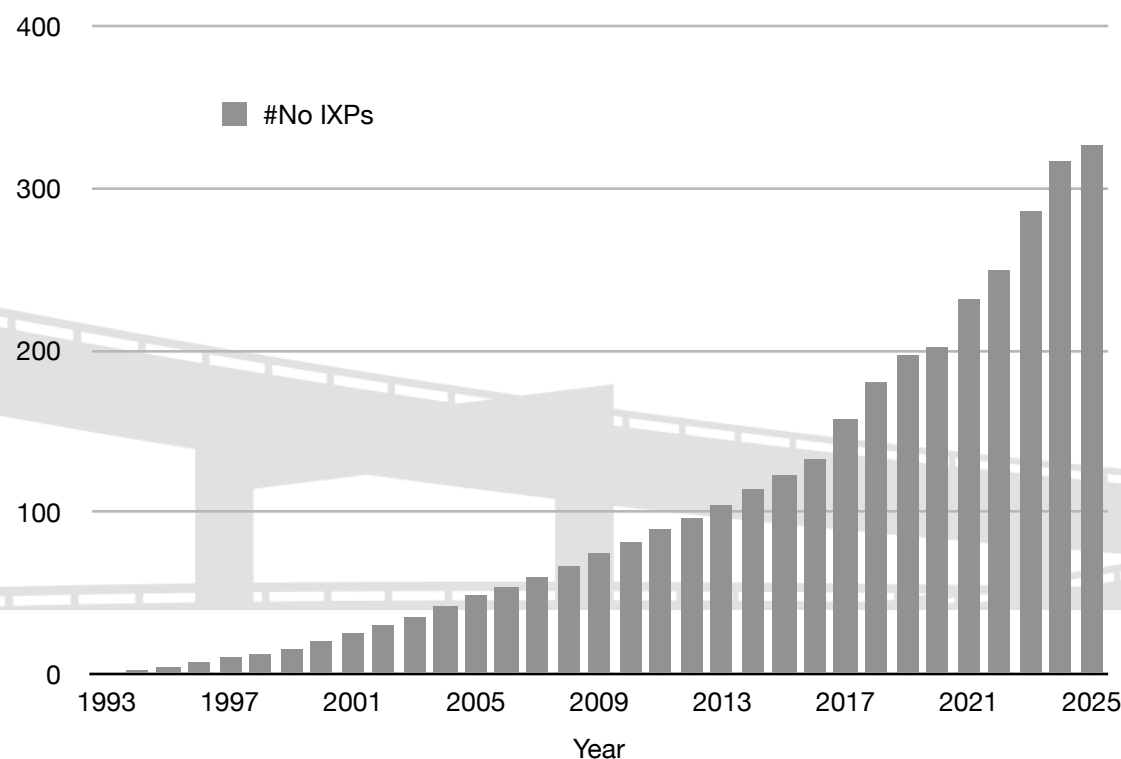
Anycast for DNS

- PCH and its precursors have run production anycast services since 1989.
- Bill Woodcock (PCH) and Mark Koster (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 zones in 1997.
- 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)



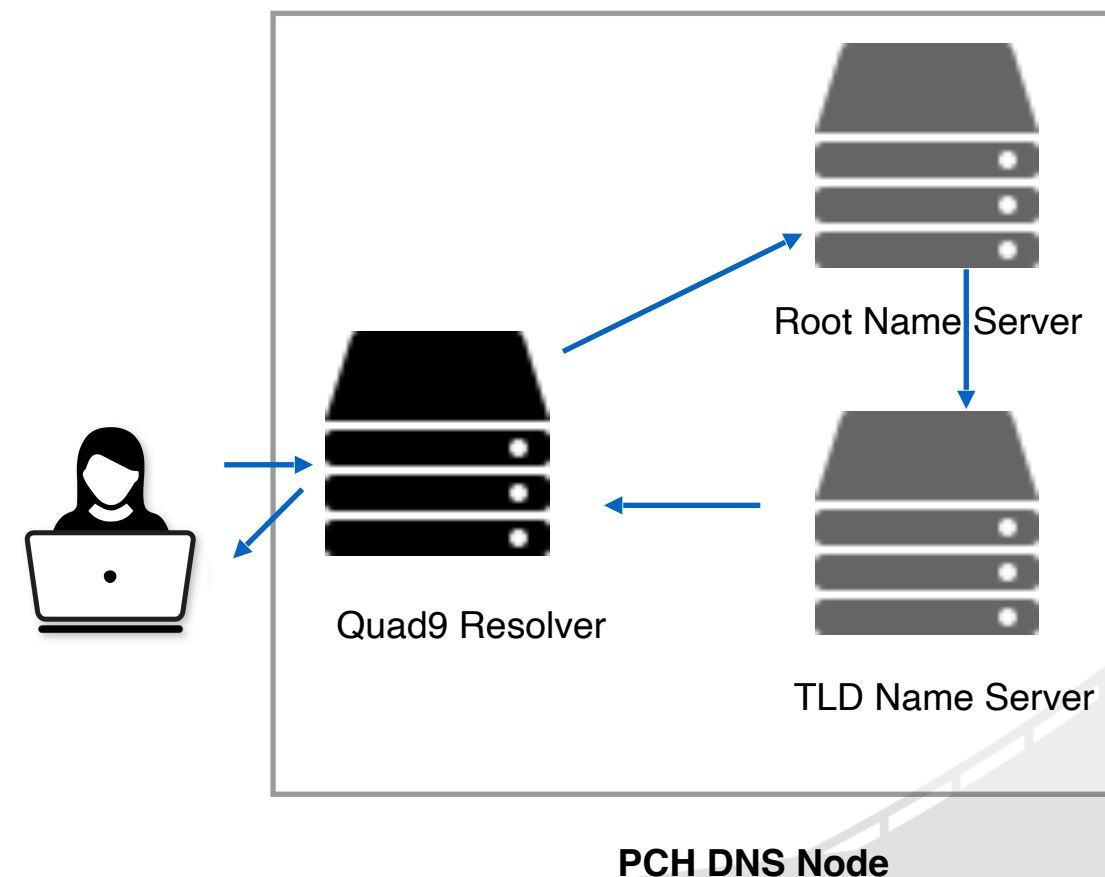
Evolution of PCH's Anycast Network (1993 - Present)



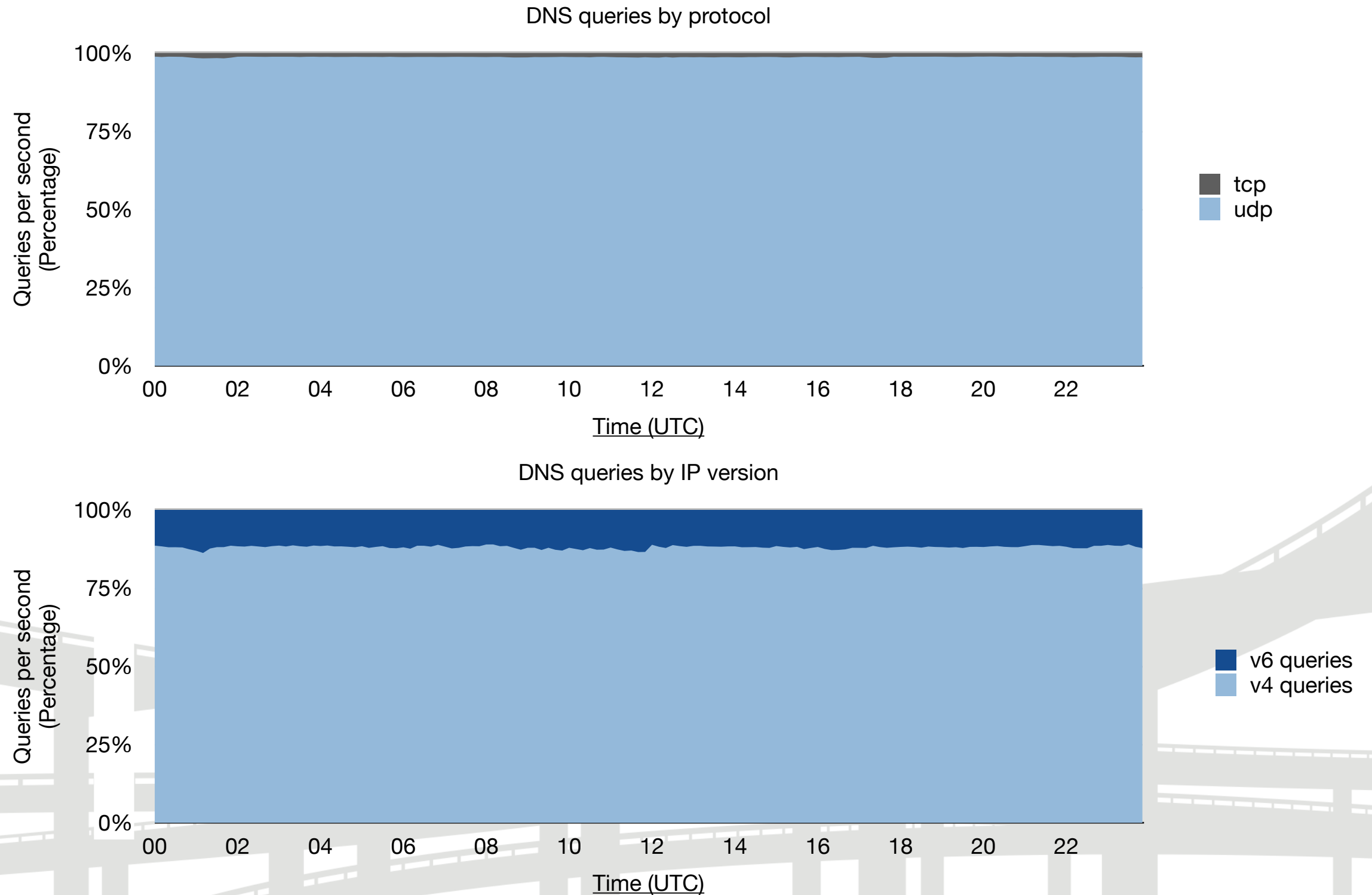
- 260 unique anycast nodes in all five continents
 - 28 global nodes (high traffic nodes)
- 160 cities in five continents
 - 38 in APNIC region
 - 38 in RIPE region
 - 30 in AFRINIC region
 - 30 in ARIN region
 - 20 in LACNIC region
- 5000+ unique peering ASNs
- Secondary authoritative service to 400+ TLDs and two letters of the DNS root.
 - ~128 ccTLDs
 - ~200 million resource records

PCH's 9th Generation Architecture

- Small, medium and full cluster installations (regional)
- Routing Vendor redundancy: Cisco, **FRR** ↑
- Servers (Cisco, Dell) with hardware specs based on site demand.
- VMware ESXi clusters, supporting any x86 64-bits OS.
- OS redundancy: Centos and Debian
- Orchestration: Ansible, Salt stack
- Name server redundancy: Bind, Unbound, NSD, Knot DNS.
- Long term strategic relationship with all involved vendors:
 - Cisco, AMD, VMware, ISC and NLNet Labs.

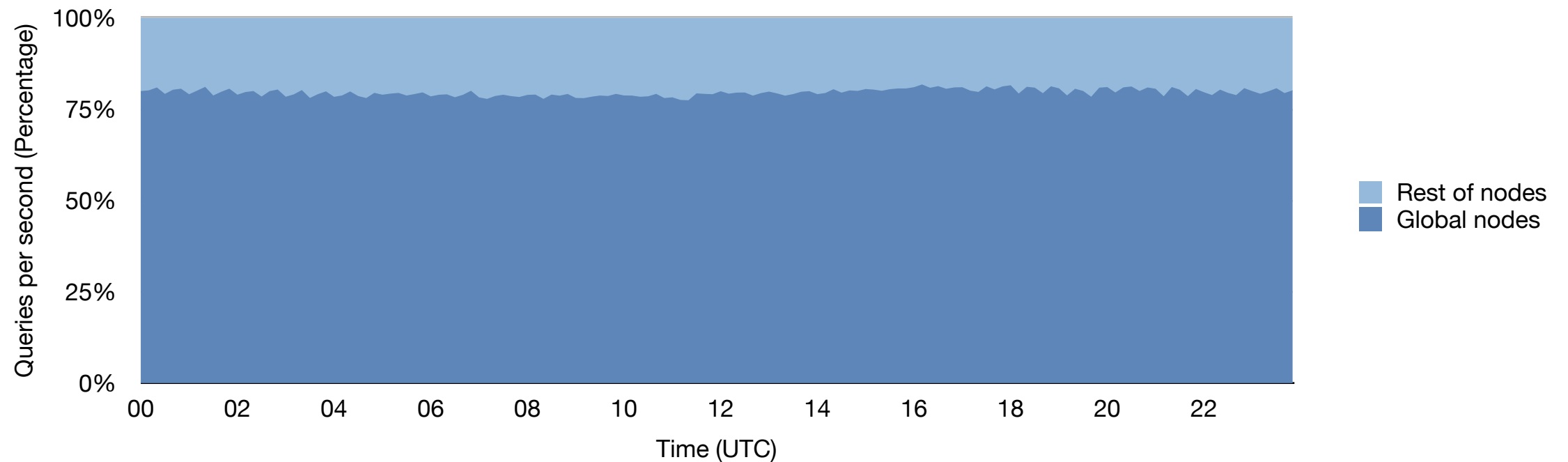


A day in PCH's anycast network

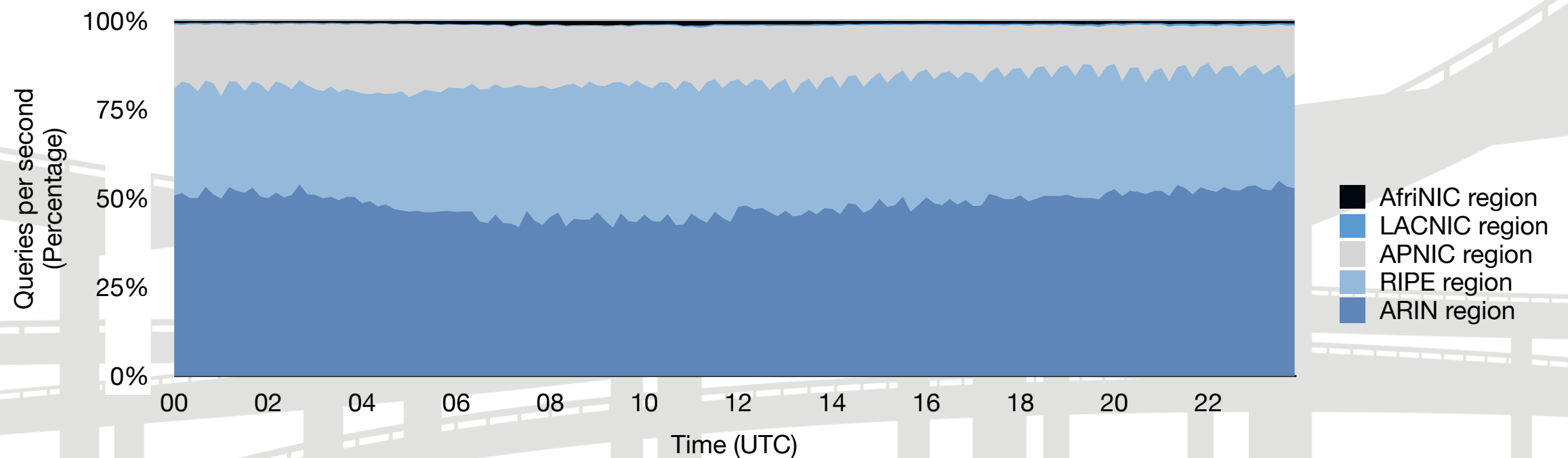


A day in PCH's anycast network (ii)

DNS queries processed by global and rest of nodes



DNS queries by region



Planning Anycast Nodes

- **Requirements:**

- Colocation (1ru) and power (200-300 watts)
- Stable OOB access (50 - 100 mbps) (static route or a full BGP table)
- 2x IX peering ports (10G at-least)

- **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
- Relative location of other nodes (correlates to measurements)
- Availability of our transit providers

- **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

Operations

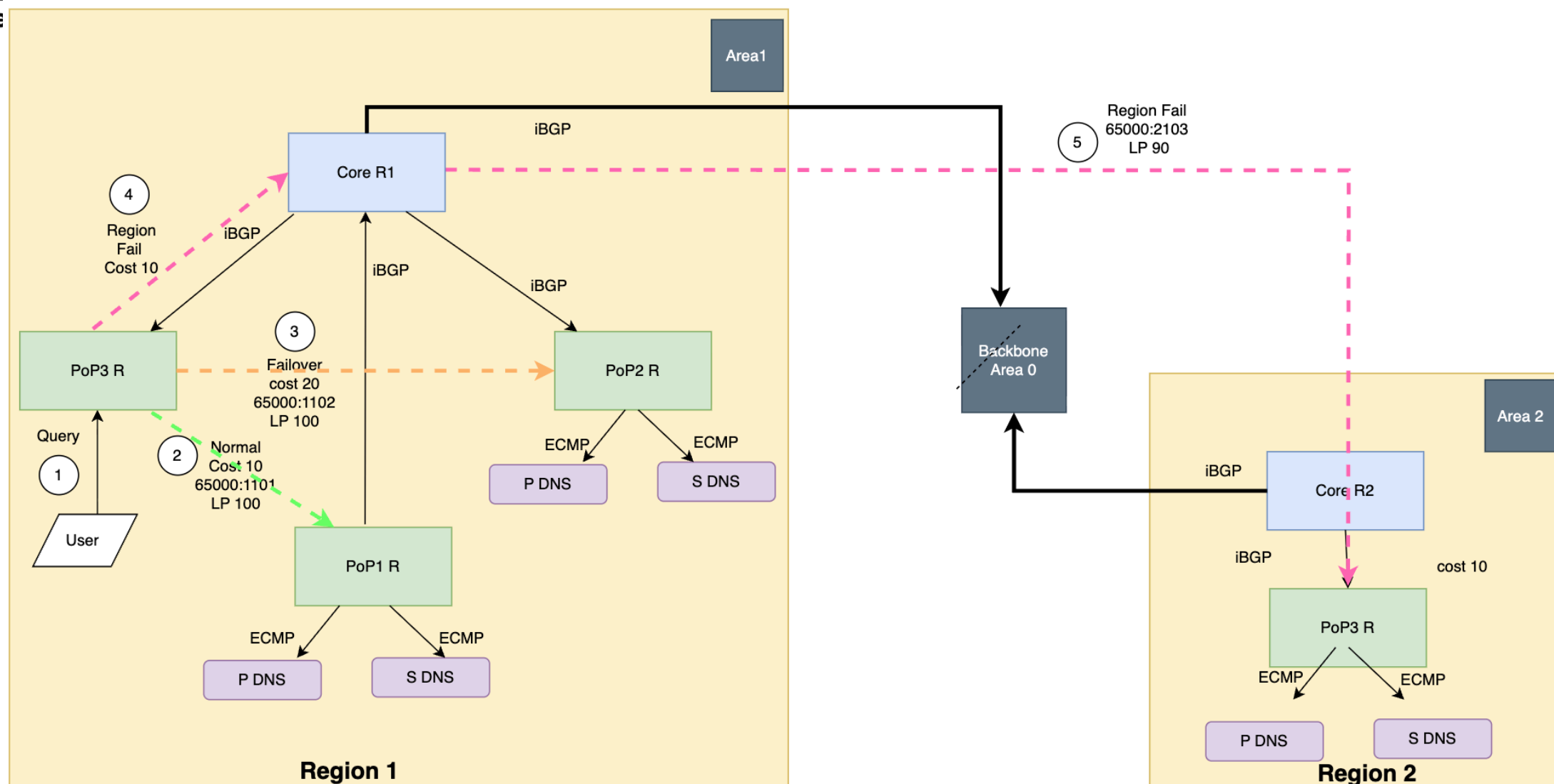
- Services run in separated virtual machines
 - Dedicated VMs for root servers, TLDs and monitoring services.
- Depending on the type of deployment (small/medium/large) and type of node (local/global), we announce via BGP a full or a partial set of services:
 - Small sites: anywhere in the world, local-only 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.
- A **failure** in the DNS service triggers the removing of the node from the routing table by withdrawing its BGP announcement automatically

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

Planning Your Anycast

- **Dynamic Failover:** Use of BGP communities and local preferences according to regions and PoPs:
- **65000:XYZZ, where X=action, Y=region, ZZ=PoP**
 - **65000:1101/65000:1102:** Same Region, PoP1/PoP2 (Primary/Backup, LP 100).
 - **65000:2103:** Nearby Region (Region 2, PoP3, LP 90).
 - **65000:5105:** Distant Region (e.g., Region 5, LP 80).
 - **65000:6666:** Withdraw route for maintenance or DDoS mitigation.
- **IGP routes** traffic to the closest resolver within Region 1 (e.g., PoP3 to PoP1 cost 10, PoP2 cost 20) and to CoreR1 (cost 10) for inter-region failover.
- **Failure Cases:**
 - Single PoP Failure: If PoP1 fails, PoP3 routes to PoP2 (OSPF cost 20, 65000:1102, LP 100).
 - Both PoPs Fail: PoP3 routes to CoreR1 (cost 10), then to Region 2, PoP3 via BGP (65000:2103, LP 90).
- **Load Balancing:** ECMP to distribute load across primary and secondary resolvers within each PoP (e.g., PoP1's resolvers).
- **BGP Optimizers:** ExaBGP or GoBGP with custom scripts to dynamically adjust BGP communities or AS-path prepending for load balancing and DDoS mitigation.
 - **E.g.** AS-path prepending with community 65000:2103 (LP 90) to steer traffic from overloaded PoP1 to Region 2, PoP3.



- Redundancy: Two resolvers per PoP, two PoPs per region
- Failover: OSPF prioritizes PoP1 (cost 10) over PoP2 (cost 20); BGP for inter-region
- Load Balancing: ECMP within PoPs.
- Overflow: BGP 65000:2101, LP 90

Region1-PoP3 IGP

To PoP1: cost 10

To PoP2: cost 20

Region1-PoP3 BGP

Accepts 65000:1101, 65000:1102 with LP 100

Accepts 65000:2103 with LP 90

Accepts 65000:2105 with LP 80 (distant regions)

Other Considerations

- Identifying Which Server is Giving an End-User Trouble - network issue (BGP, latency) or DNS software issue?
 - Latency & Error Tracking (per pop metrics, traceroutes, active monitoring - RIPE Atlas).
 - Logging & Request Tagging (use of TXT records)
 - Diagnostic endpoints (dig +short *whichpop.anycast.example.com*)
 - Use of lower TTLs to allow fast failover.
 - Monitoring query rates per POP(Response Rate Limiting / iptables)
 - Long-lived flows (e.g., DNS-over-TCP, DoH/DoT) can break if routes change mid-session ('*TCP RST / Connection: close headers*' to encourage re-resolution).
- Periodically see node performance & scaling (CPU, RAM, Network)
- The service will perform best if servers are widely distributed, with higher density in and surrounding high demand areas
 - Start small (2-3 nodes), test failover, tune performance, validate and expand.

Anycast in action (.bd)

Key	UDpv4 Queries ▾	UDpv4 Responses	TCPv4 Queries	TCPv4 Responses	UDpv6 Queries	UDpv6 Responses	TCPv6 Queries	TCPv6 Responses
Hamburg	3,744,082	2,191,412	1,330,811	1,198,701	2,950,644	1,531,484	1,385,060	1,242,106
Turin	1,375,343	1,097,994	336,689	336,243	1,112,165	709,537	339,495	339,053
Santiago	860,882	860,632	1,315	1,291	726,364	724,685	927	911
Ashburn	764,707	753,434	10,009	9,946	376,904	376,471	431	428
Hong Kong	720,560	695,968	45,321	45,253	569,686	503,579	44,956	44,955
Tokyo	669,916	669,707	930	921	299,508	299,451	98	86
Bucharest	612,708	611,106	731	617	227,962	227,703	354	298
Zurich	483,697	481,161	3,705	3,642	295,009	287,768	2,230	2,187
Seattle	471,279	464,992	4,297	4,244	326,510	326,294	74	74
Montreal	387,208	372,467	11,699	11,692	277,905	277,710	175	175
Singapore	348,341	313,435	33,176	32,869	197,547	195,797	646	634
Dallas	319,220	318,775	166	166	314,943	314,550	126	126
Frankfurt	281,679	281,302	216	199	220,130	219,916	72	71
San Jose	241,736	191,745	42,474	42,423	74,886	74,840	10	10
Johannesburg	218,384	216,005	481	469	190,742	188,134	213	206
Los Angeles	204,655	203,814	392	389	219,353	218,312	2,043	1,628
Seoul	194,092	193,447	330	321	55,177	54,499	316	316
San Francisco	181,687	179,959	1,269	1,252	138,718	137,105	1,058	1,056
Queretaro	171,870	171,848	41	41	107,179	107,163	27	27
Dhaka	166,769	159,014	502	500	28,606	21,628	94	94
Sydney	143,528	143,165	613	610	74,497	74,447	98	98









Anycast in action (.np)

Key	UDpv4 Queries ▾	UDpv4 Responses	TCPv4 Queries	TCPv4 Responses	UDpv6 Queries	UDpv6 Responses	TCPv6 Queries	TCPv6 Responses
Kathmandu	567,767	565,634	367	366	213,739	212,505	276	275
Ashburn	407,859	407,570	277	272	192,777	192,122	275	269
San Francisco	359,859	359,742	58	58	324,621	324,522	23	23
Jakarta	303,867	302,872	957	952	26,345	25,603	79	59
Tokyo	176,140	176,055	128	128	60,478	60,428	114	113
Dallas	168,487	168,333	8	8	135,236	135,078	1	1
Seattle	150,171	150,057	43	43	227,959	227,794	36	36
Frankfurt	147,357	147,150	323	258	110,450	110,346	39	39
Hong Kong	131,419	131,356	69	69	178,371	178,276	115	115
Singapore	130,800	130,144	34	34	119,310	118,705	12	12
Turin	128,485	128,041	635	630	122,559	120,507	608	605
Santiago	115,203	115,139	266	266	92,077	92,005	180	180
Hamburg	111,837	110,526	1,936	1,851	24,234	14,080	1,866	1,827
Los Angeles	93,062	92,845	98	98	108,545	108,158	231	186
Montreal	92,940	88,509	147	137	47,083	46,995	78	77
San Jose	77,499	69,374	5,283	5,283	50,080	50,020	6	6
Kualalumpur	71,359	70,541	193	183	45,860	45,410	339	338
Bucharest	70,982	70,809	192	192	24,307	24,294	30	30
Seoul	67,877	67,585	191	191	46,707	46,425	190	190
Warsaw	58,536	58,345	17	17	46,743	46,587	7	7
Johannesburg	54,501	53,868	177	175	39,385	39,017	207	206
New York	47,421	47,244	49	47	29,530	29,455	52	52
Zurich	46,584	46,193	271	271	32,080	31,762	167	166
Osaka	44,074	43,867	210	210	30,859	30,494	402	402

Measuring from Outside of the network

- Using RIPE atlas probes (<https://atlas.ripe.net>)
- Trace-route to “valid destination - anyns.pch.net”
 - up and running
 - the service is operating
 - we *always* peer with the route server

Routing in NP

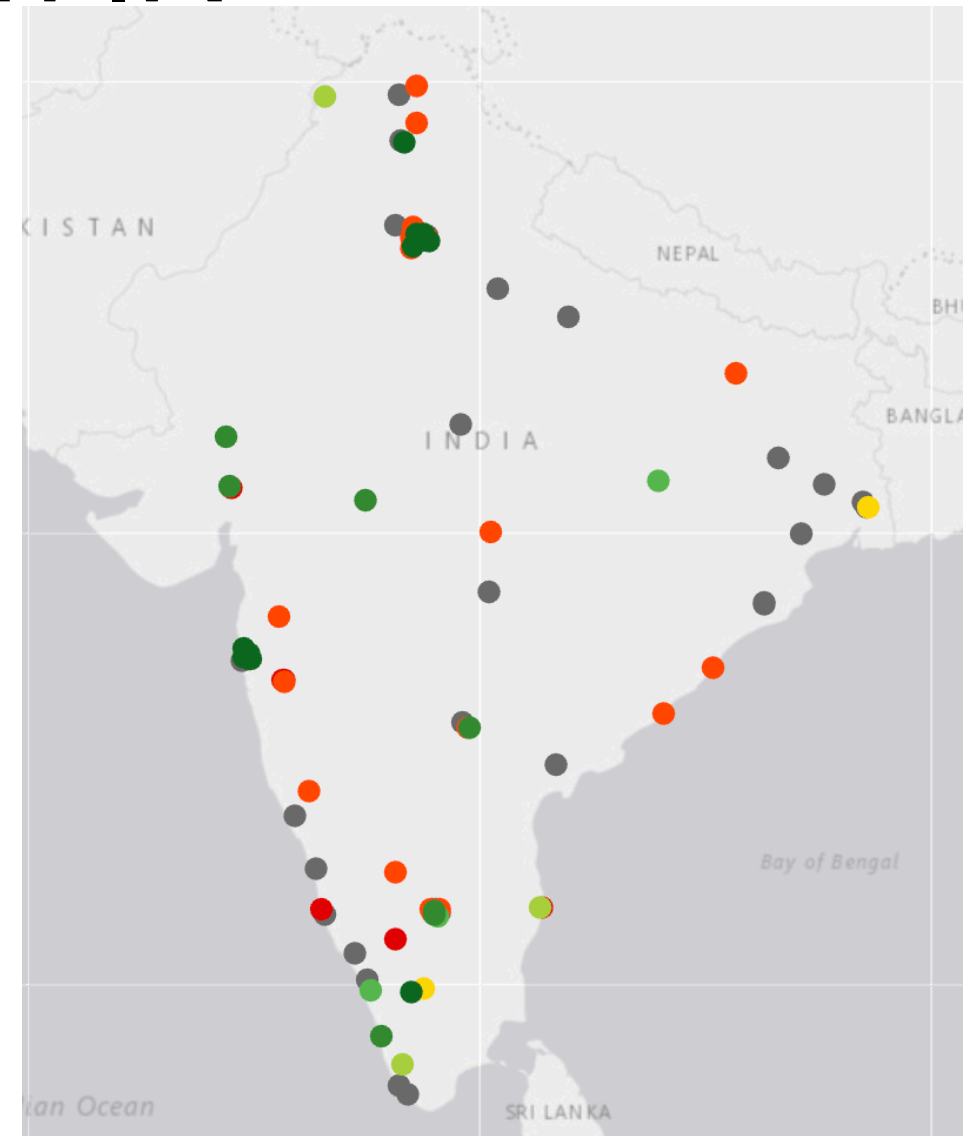
ASN	Connected to the IX	PCH
17501		
4007		
153566		
140050		



Routing in IN

ASN	Connected to local IX	PCH
151704	✓	✓
24560	✗	✗
134674	✓	✗
24186	✓	✗
133982	✓	✓

61.663 ms



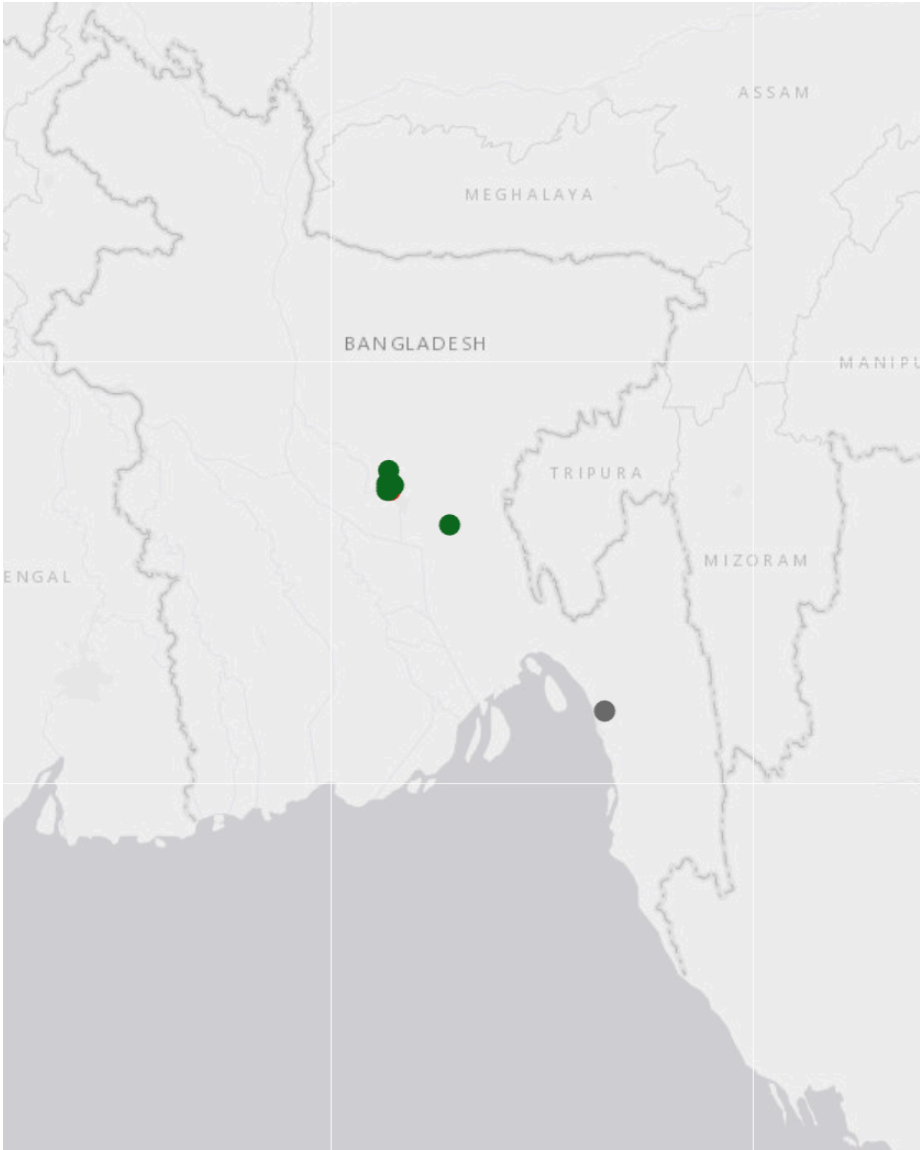
Traceroute for Probe 1001549

Hop	IP Address	Reverse DNS	ASN	RTT 1	RTT 2	RTT 3
1	192.168.1.1			0.332 ms	0.296 ms	0.285 ms
2	100.98.128.1			5.678 ms	5.365 ms	5.344 ms
3	*	*	*	*	*	*
4	100.100.107.232			11.819 ms	11.982 ms	11.936 ms
5	172.31.222.44			16.752 ms	16.868 ms	17.059 ms
6	172.31.222.45			16.689 ms	17.686 ms	16.315 ms
7	125.20.121.17		9498	16.227 ms	17.083 ms	16.973 ms
8	116.119.158.254			61.577 ms	79.258 ms	61.468 ms
9	27.111.228.1	42.sgw.equinix.com		76.883 ms	76.049 ms	76.777 ms
10	204.61.216.4	anydns.pch.net	42	62.109 ms	61.663 ms	62.218 ms

Routing in BD

ASN	Connected to the IX	PCH
63961		
147181		
141988		
58717		

226.485 ms



```
1.1.router.dac.woodynet.net# sh ip bgp 103.15.246.0
% Network not in table
```

Traceroute for Probe 6785

Hop	IP Address	Reverse DNS	ASN	RTT 1	RTT 2	RTT 3
1	*	*	*	*	*	*
2	<u>103.102.43.12</u>			0.949 ms	0.901 ms	0.915 ms
3	*	*	*	*	*	*
4	<u>204.61.210.134</u>	tun729.router.dac.woodynet.net	<u>715</u>	219.854 ms	219.714 ms	219.574 ms
5	<u>204.61.216.4</u>	anyns.pch.net	<u>42</u>	226.945 ms	226.485 ms	226.623 ms

Routing in BT

Host	Packets		Rings				
	Loss%	Snt	Last	Avg	Best	Wrst	StDev
1. 172.20.10.1	0.0%	10	5.4	16.5	3.9	64.9	21.2
2. 172.24.255.254	0.0%	10	40.4	43.9	14.3	100.4	25.7
3. 172.23.0.17	0.0%	10	99.2	46.2	10.4	148.5	44.3
4. 172.23.15.121	0.0%	10	31.9	30.1	12.0	78.0	21.9
5. 172.23.15.105	10.0%	10	83.2	33.0	11.9	83.2	23.5
6. leasedline-202-144-131-41.thimphu-core-1-bmobile.druknet.bt	10.0%	10	24.8	44.1	14.9	120.5	39.7
7. p2p-103.245.242.232.tr1.thimphu.druknet.bt	11.1%	9	13.8	39.6	13.8	79.7	21.6
8. p2p-103.245.242.228.tr1.pling.druknet.bt	0.0%	9	88.0	40.7	14.5	88.0	23.2
9. 103.245.243.157	0.0%	9	129.3	110.5	92.8	137.5	18.5
10. 42.sgw.equinix.com	0.0%	9	128.9	106.1	88.5	131.8	15.4
11. ns3.druknet.bt	11.1%	9	96.6	94.9	90.4	110.1	6.4



Why!

- Networks don't peer at the IXP
- Mis-configurations
- Wrong AS-SET and Prefix Filters

Remember!

- `peering_local_pref > transit_local_pref`
- Advertising more specifics to your transit hurts *you*
- Use modern tools to automate building filters from AS-SETs
- Connect more Atlas Probes and Anchors

“every time you send a packet to international destination that could be served locally, you are subsidizing the internet eco-system in that country ”



Thanks for your attention

Dibya Khatiwada
Packet Clearing House

dibya@pch.net