Peering issues of an Anycast DNS provider with a heterogeneous peering setup

Klaus Darilion, RcodeZero DNS

klaus.darilion@nic.at

Who we are: RcodeZero DNS

- Subsidiary of nic.at, the .at domain registry
- We needed an Anycast DNS service for ourselves (.at)
- We offered the service to other TLDs
- We offered secondary DNS service for registrars/ISPs/enterprises too

Our Anycast history

- 6 locations: US+EU
- Anycast Performance: horrible
- We were DNS experts, but we had no idea how routing works on the Internet
- Now: 50+ nodes, good routing but continuous tuning is required

How we build an Anycast location

- We need BGP to announce your prefixes
  - With transit providers
  - With local peers (e.g. to bypass some routing quirks between Tier1 providers)
- We need a server
  - Colocation + our own hardware
  - Rented Baremetal
  - Rented a VM

  -> We do all that possibilities, depending on **availability** and **pricing**

Pricing

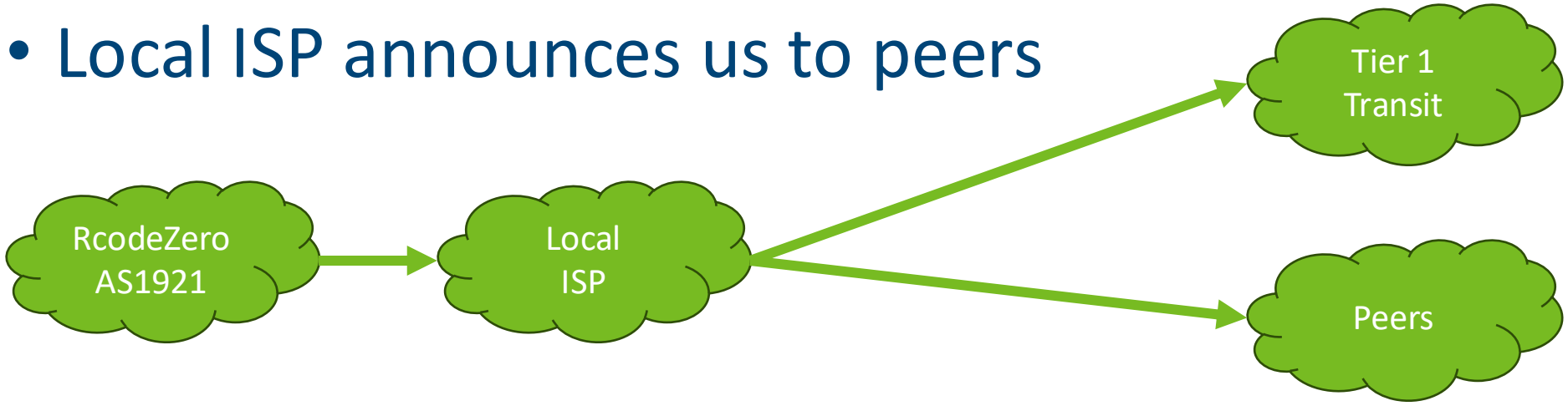
- Doing everything yourself
  - Rent Colocation + power in data center: min. 300 EUR/month
  - Transit Link: min 300 EUR/month
  - IX Link: ~ 500 EUR/month

  → min. 1000 EUR/month without a server
- Rent a server/VM from a local ISP
  - around 400 EUR/month incl. traffic
  - preferred

→ At most locations we do not peer directly, but our local ISP peers

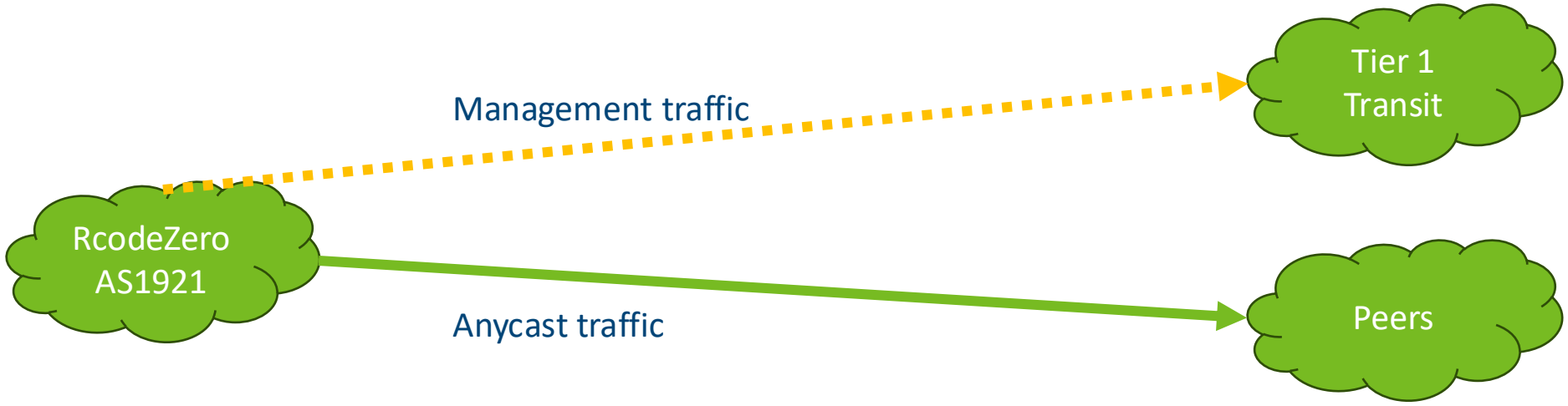
“Global Nodes”

- VM or Baremetal (rented or colocation)
- Our prefixes announced to local ISP
- Local ISP announces us to transit
- Local ISP announces us to peers



“Local Nodes”

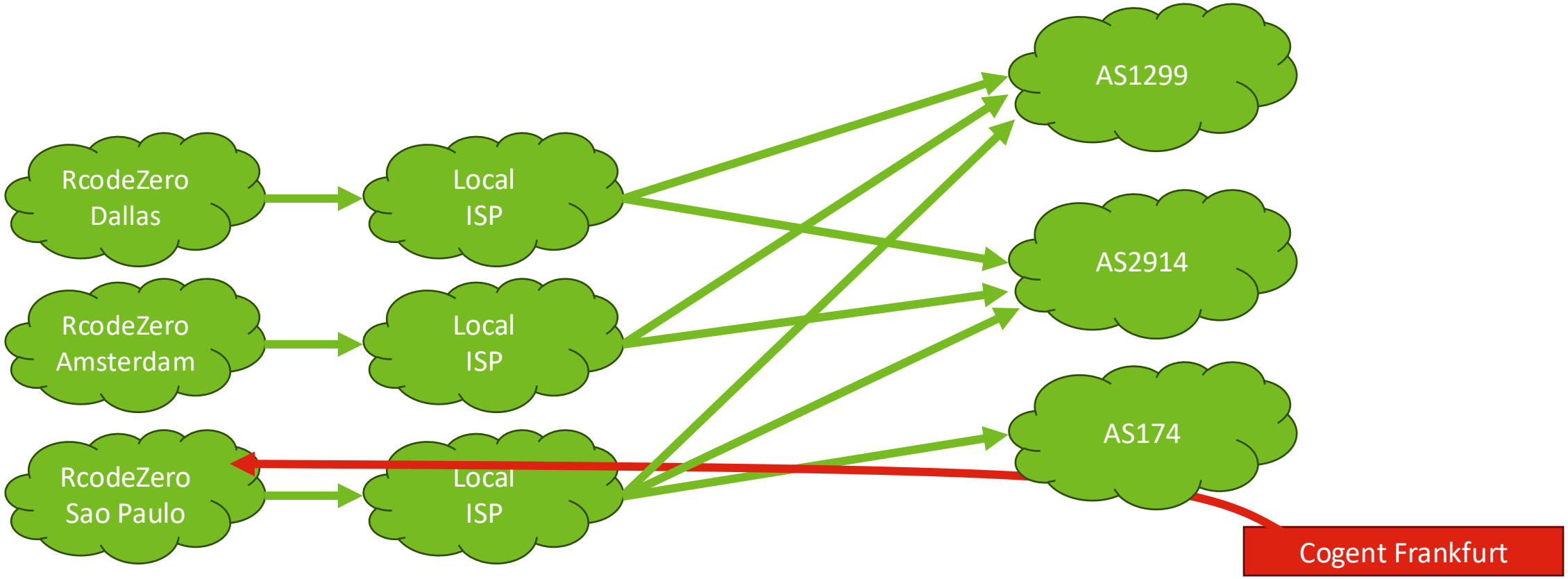
- Pro Bono Nodes
- Transit only for management traffic
- Anycast Prefixes only announced to local IX



Issue 1: Homogenous Transit

- We use AS1299 and AS2914 as transit
- Example: Possible new location with Local ISP
  - ISP has transit AS1299, AS2914 and AS174

Issue 1: Homogenous Transit



Cogent Transit only in Sao Paulo -> Cogent would route traffic from the whole world to Sao Paulo

Solution: Homogenous Transit

- Disable transits that we only have on a single location
- Option 1: ISP provides a „do not announce to ASxxxx community“
- Option 2: use transit‘s community „set local-pref below peer“
- What if local ISP and transit do not provide communities?
  - we can not use this local ISP

Which Transits does an ISP have?

- RIPE RIS Looking Glass + some scripting
  - Very efficient way to see if an ISP matches our routing policy

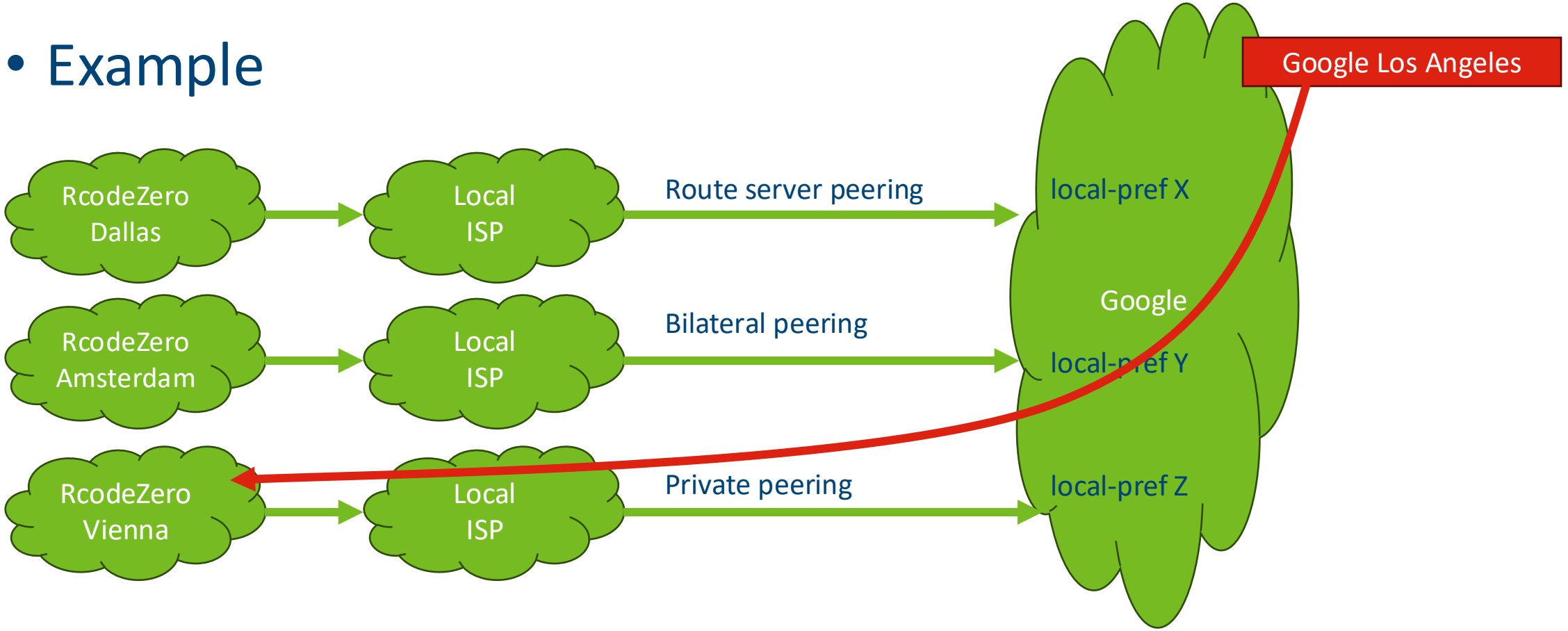
```
Upstreams for 103.6.86.98
36236 Netactuate/HostVirtual
   3491 PCCW
      1299 Telia
      7018 AT&T
      852
      3356 Level3
         13830
      3257 GTT/Tinet
         29680
      1828
      6830 UPC
      13030 Init7
      174 Cogent
         212934
   24482 sg.gs
      24482 sg.gs
      199524 G-Core Labs S.A.
         28917
         41095 IPTP HK
      6453 Tata (Tata Worldwide Tier1)
         9002 RETN
            212271
         1257 Tele2
      3257 GTT/Tinet
      1853 ACOnet
         6720
      1280
   2914 NTT
      15562
      3356 Level3
         3549 Level3/GBLX
      6461 Zayo/AboveNet
      3320 DTAG
      12956 Telxius=Telefonica Wholesale Network
   38008
      46997
   9002 RETN
   49544
```

Issue 2: Peering

- Peering is in the hand of our local ISP
  - Good: lots of peerings from day 1
  - Bad: ISP decides how and with whom to peer
- How?
  - Route server peering at IX
  - Bilateral peering at IX
  - Private peering
- Problem: Many peers give different local-pref depending on peering option

Issue 2: different local-pref

- Example



X < Y < Z → Private peering in Vienna is preferred, Google routes traffic from the whole World to Vienna

Issue 2: different local-pref

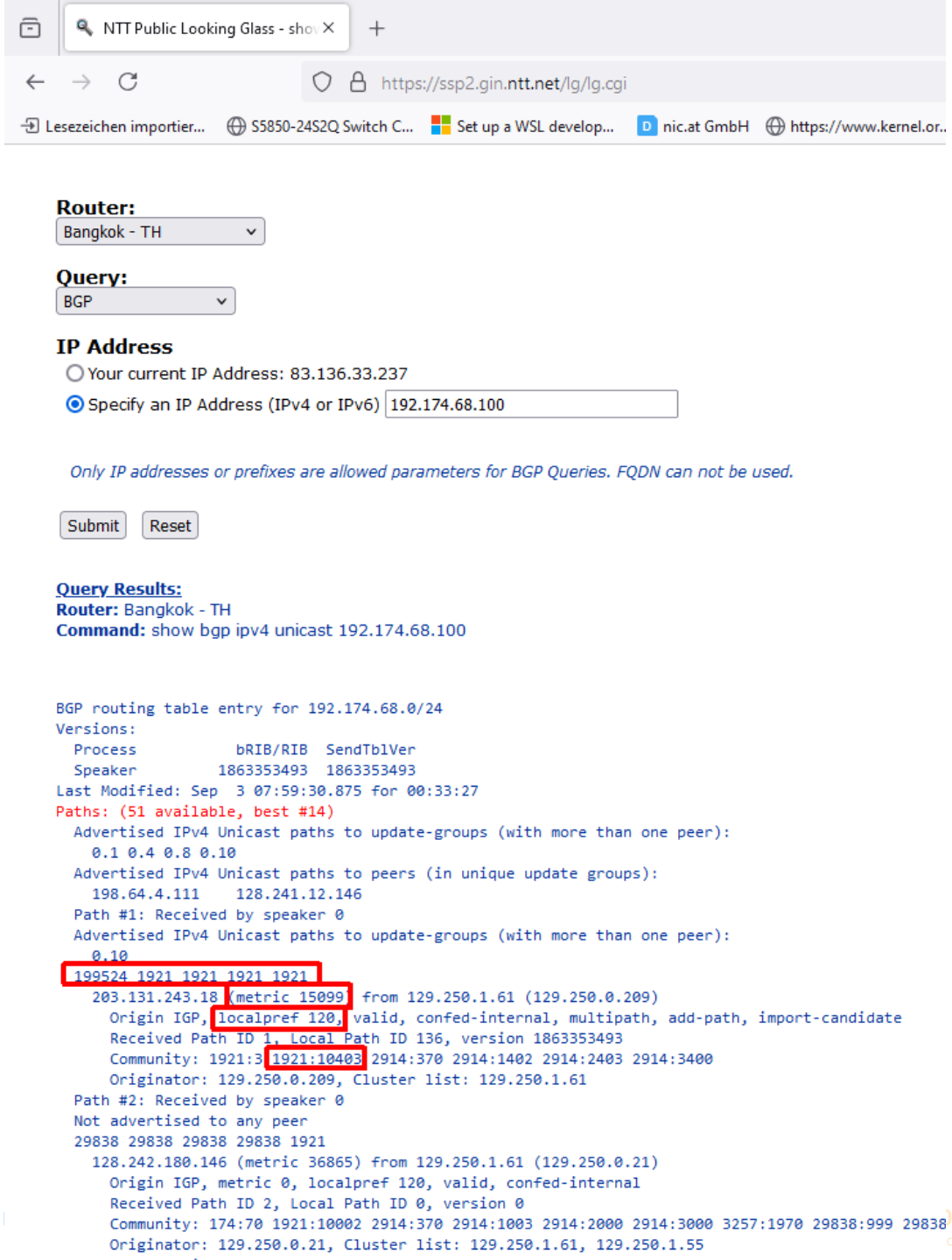
- Solution
  - no solution
  - local-pref changes via BGP community are not allowed on peering links (some providers allow it on transit links)
- Workaround
  - Find local ISPs with private peerings

Issue 3: Debugging

- Example: Traffic is not routed to nearest location. **Why?**
- We only know what we announce to our local ISP
  - We can not see how our ISP announces to transit and peers
- Traceroute shows the current path, but not why that path is used
- Some providers block traceroute (Google, Azure …)
- RFC1918 IP addresses and Bogons are missing in Trace
- To debug routing decisions of an AS we need a „BGP Looking Glass“, ideally with all routes and local-pref and BGP communities

Great Looking Glass

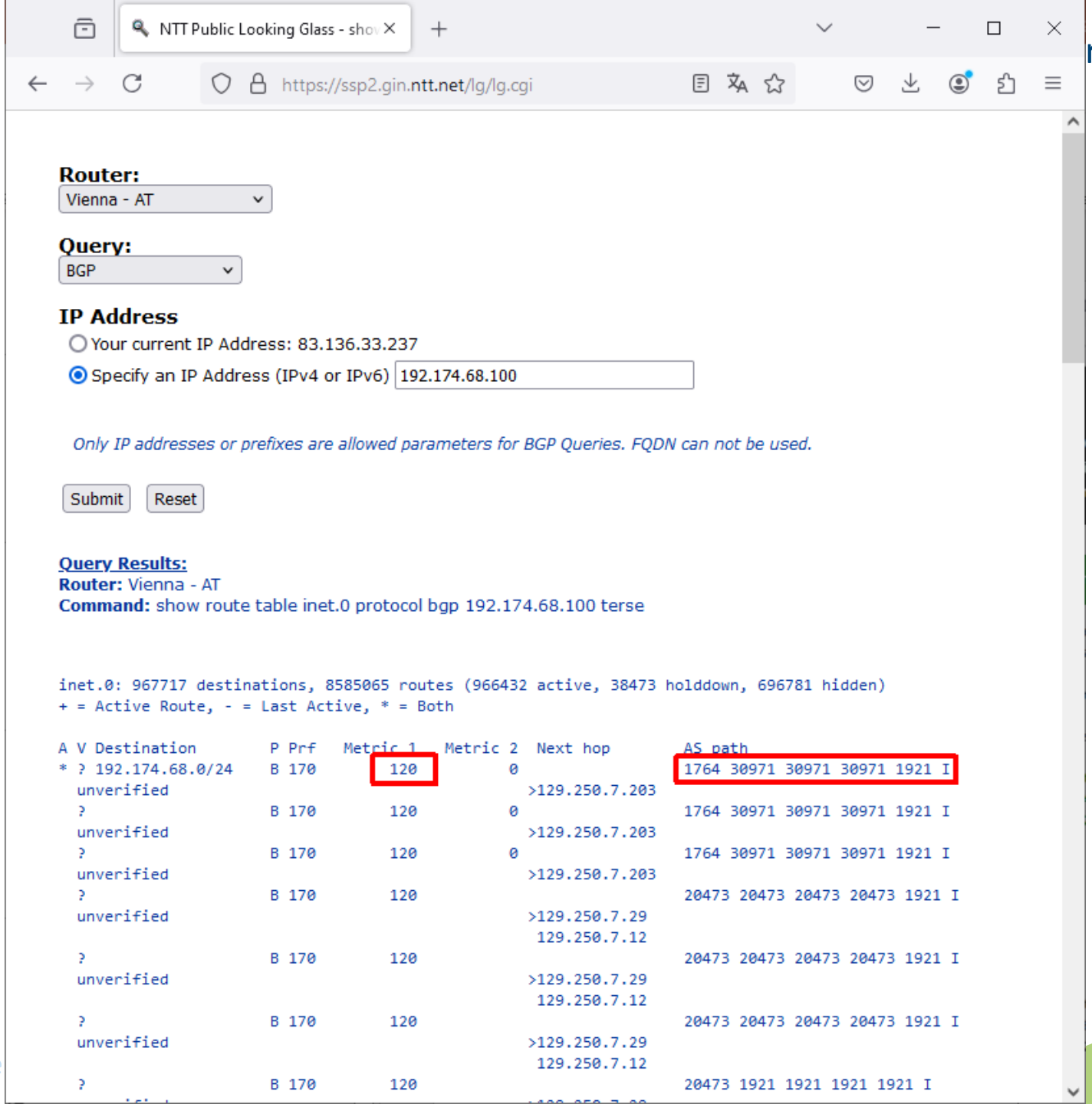
AS Path
local-pref
IGP metric
BGP communities



Good Looking Glass

Communities missing
Metric missing
→ difficult to map Path to location

NTT: please improve it ;-)



IX Debugging

- Looking Glass of Route Server
  - Not always public, does not show bilateral/private peerings
- Looking Glass of problematic peer
  - Hyperscalers do not offer Looking Glass
- Alternative: Looking Glass of well-peered ISPs
  - AS6939 HE
  - Packet Clearing House

HURRICANE ELECTRIC
INTERNET SERVICES

Looking Glass

Welcome to Hurricane Electric's Network Looking Glass. The information provided by and the support of this service are on a best effort basis. These are some of our routers at core locations within our network. We also operate a public route server accessible via telnet at route-server.he.net.

Show options

core3.fra1.he.net> show ip bgp routes detail 192.174.68.100

Matching Routes	16
Status Codes	**A** - Aggregate **B** - Best **b** - Not Install Best **C** - Confederation eBGP **D** - Damped **E** - eBGP **H** - History **I** - iBGP **L** - Local **M** - Multipath **m** - Not Installed Multipath **S** - Suppressed **F** - Filtered **s** - Stale **x** - Best-External

Status	Network	Next Hop	Learned	Metric	LocPrf	MED	Weight	Path	Origin	ROA
BME	192.174.68.0/24	80.81.194.3	80.81.192.157 (6695	0	100	0	0	44066, 1921x4	IGP	✓
ME	192.174.68.0/24	80.81.194.3	80.81.193.157 (6695)	0	100	0	0	44066, 1921x4	IGP	✓
ME	192.174.68.0/24	80.81.194.3	80.81.194.3 (44066)	0	100	0	0	44066, 1921x4	IGP	✓
I	192.174.68.0/24	80.249.212.38	216.218.252.67 (6939)	68	100	0	0	20473x4, 1921	IGP	✓
I	192.174.68.0/24	185.1.240.204	216.218.253.242 (6939)	73	100	0	0	20473x4, 1921	IGP	✓
I	192.174.68.0/24	37.49.237.90	216.218.252.66 (6939)	100	100	0	0	20473, 1921x4	IGP	✓
I	192.174.68.0/24	91.206.52.164	216.218.252.82 (6939)	104	100	0	0	1921x5	IGP	✓
I	192.174.68.0/24	185.0.21.31	216.218.253.45 (6939)	120	100	0	0	1764, 30971x3, 1921	IGP	✓
I	192.174.68.0/24	193.203.0.22	216.218.253.15 (6939)	125	100	0	0	1764, 30971x3, 1921	IGP	✓
I	192.174.68.0/24	195.66.227.115	216.218.253.11 (6939)	130	100	0	0	1921x5	IGP	✓
I	192.174.68.0/24	194.68.123.194	216.218.253.27 (6939)	180	100	0	0	42708x2, 1921x3	IGP	✓
I	192.174.68.0/24	185.6.36.47	216.218.253.44 (6939)	188	100	0	0	42310, 1921x4	IGP	✓
I	192.174.68.0/24	195.208.210.28	216.218.253.57 (6939)	230	100	0	0	42473, 1921x4	IGP	✓
I	192.174.68.0/24	195.149.232.196	216.218.252.224 (6939)	235	100	0	0	8308, 1921x4	IGP	✓
I	192.174.68.0/24	185.1.192.56	216.218.253.73 (6939)	310	100	0	0	20473, 1921x4	IGP	✓
E	192.174.68.0/24	195.219.220.130	195.219.220.130 (6453)	0	100	0	0	6453, 4755, 20473x3, 1921x4	IGP	✓

Last Update 23h27m53s ago

Entry cached for another 60 seconds.

2024-09-03 09:03:19 UTC

IXP, City, Country

DE-CIX Chicago , Chicago , United States
DE-CIX DUS, Düsseldorf, Germany
DE-CIX Dallas, Dallas, United States
DE-CIX Frankfurt, Frankfurt, Germany
DE-CIX Hamburg, Hamburg, Germany
DE-CIX Istanbul, Istanbul, Turkey
DE-CIX Kuala Lumpur, Kuala Lumpur, Malaysia
DE-CIX Lisbon, Lisbon, Portugal
DE-CIX Madrid, Madrid, Spain
DE-CIX Marseille, Marseille, France
DE-CIX Mumbai, Mumbai, India
DE-CIX Munich, Munich, Germany

Sort By IXP | Sort By City | Sort By Country

Query

- show ip bgp summary
- show ipv6 bgp summary
- (•) show ip bgp <prefix> [netmask|prefixlength]
- show ipv6 bgp <prefix>
- show ip bgp regex <regex>

Argument(s)

192.174.68.0 255.255.255.0

Examples: 204.61.216.0 255.255.255.0

or

204.61.216.0 24

Submit

Looking

Result

```
BGP routing table entry for 192.174.68.0/24, version 59177078
Paths: (3 available, best #2, table default)
  Not advertised to any peer
  44066 1921 1921 1921 1921
    80.81.194.3 from 80.81.193.157 (80.81.193.157)
      Origin IGP, metric 0, valid, external
      Community: 3856:54800
      Large Community: 6695:1911:90 6695:1912:0 6695:1913:276 6695:1914:150
      Last update: Fri Aug 23 10:10:24 2024
  44066 1921 1921 1921 1921
    80.81.194.3 from 80.81.192.157 (80.81.192.157)
      Origin IGP, metric 0, valid, external, best (Router ID)
      Community: 3856:54800
      Large Community: 6695:1911:90 6695:1912:0 6695:1913:276 6695:1914:150
      Last update: Fri Aug 23 10:10:24 2024
  44066 1921 1921 1921 1921
    80.81.194.3 from 80.81.194.3 (212.224.104.231)
      Origin IGP, metric 0, valid, external
      Community: 3856:54800
      Last update: Fri Aug 23 10:10:24 2024
```

Look up IP adresses in whois/peeringdb

Router Server Peering

Route Server Peering

Bilateral Peering

Issue 4: IX Peering ourselves

- Our “local nodes” are directly connected to IX
  - SWISS-IX, RIX, TREX, SIX, GigaPIX …
- Some big players do not peer with route servers
- Peering requests are very often not answered ☹
- Asymmetric IX Routing: we need transit for DNS responses where destination IP is not reachable via IX

Issue 5: DDoS Mitigation

- On demand DDoS mitigation (Voxility, Cloudflare …)
  - Activated by ourself
  - Activated by local ISP
- AS path and peering policy of our prefixes changes dramatically
- Careful planning to avoid routing problems during mitigation

Conclusion

- Routing optimizations would be much easier if we do the interconnects ourselves, but too expensive for our size
- Heavy use of BGP communities, AS path prepending and ISP pre-selection for close-to-optimal routing
- Unsolvable problems
  - different local-pref for route-server/bilateral/private peering
  - transit does not offer BGP communities (or they do not work)
  - please tell me if you have an idea

nic.at

RcodeZero DNS

klaus.darilion@nic.at