Anycast DNS Local Nodes and Routing Problems due to Asymmetric Routing on Internet Exchanges

Klaus Darilion · Head of Operations · klaus.darilion@nic.at

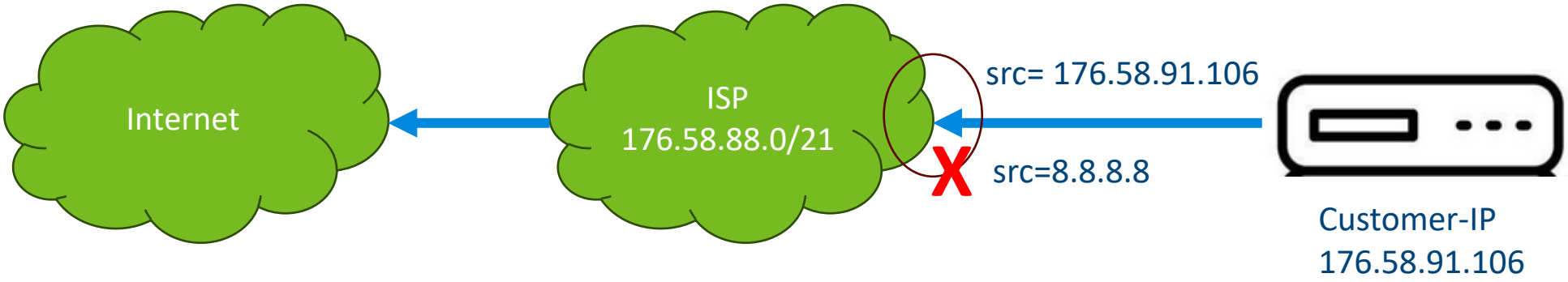
Executive Summary

1. BCP 38 / RFC 2827: Network Ingress Filtering
   – Prevent IP address spoofing
2. Anycast DNS Local Node
   – Anycast Prefix only announced to IX
3. Asymmetric Routing
   – Request received via IX
   – No route to send response on IX

➔ If that 3 meet together -> Problem

1. BCP 38

- ISP assigns one or more IP address to a customer
- ISP verifies src IP of packets sent by the customer
  - must be one of the assigned addresses
  - then ISP will route the packet



- customer uses a wrong source IP address
  - packets are dropped

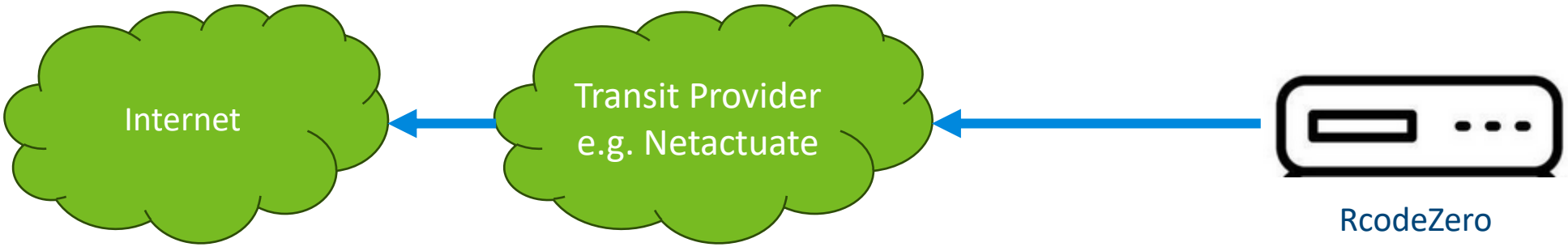
Many of our providers perform BCP38 filtering

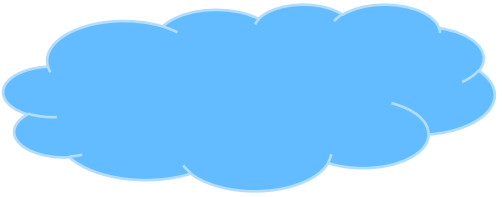
2. „Local Node“

- Let‘s describe terminology

"Transit"

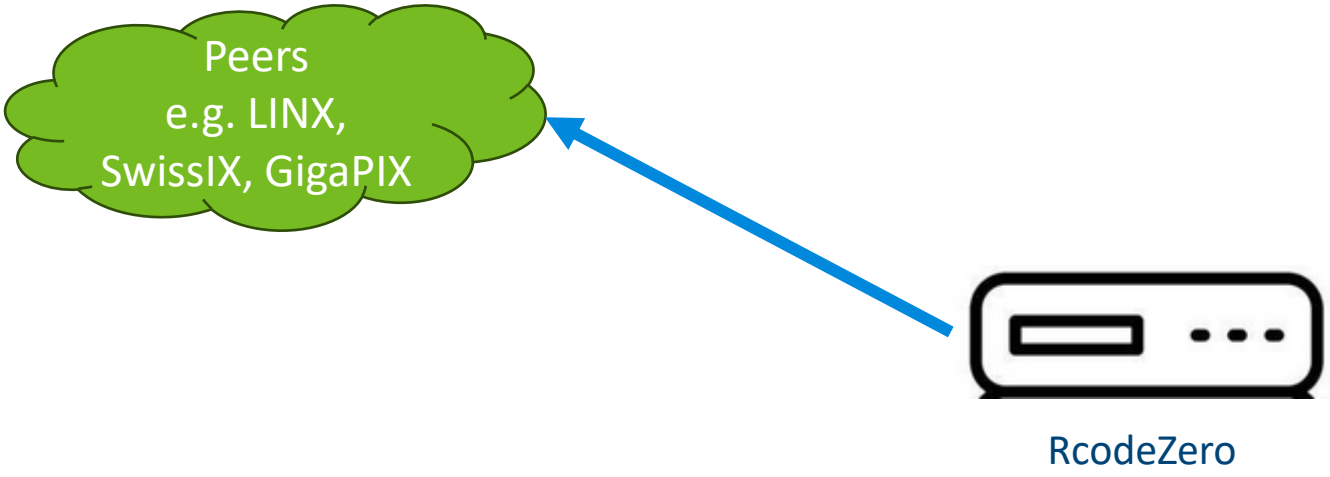
- To talk with the "whole" Internet, we need a provider that forwards (transit) our packets to the destintations
- This provider is called "transit provider" (or "upstream provider")
- We have to pay the transit provider





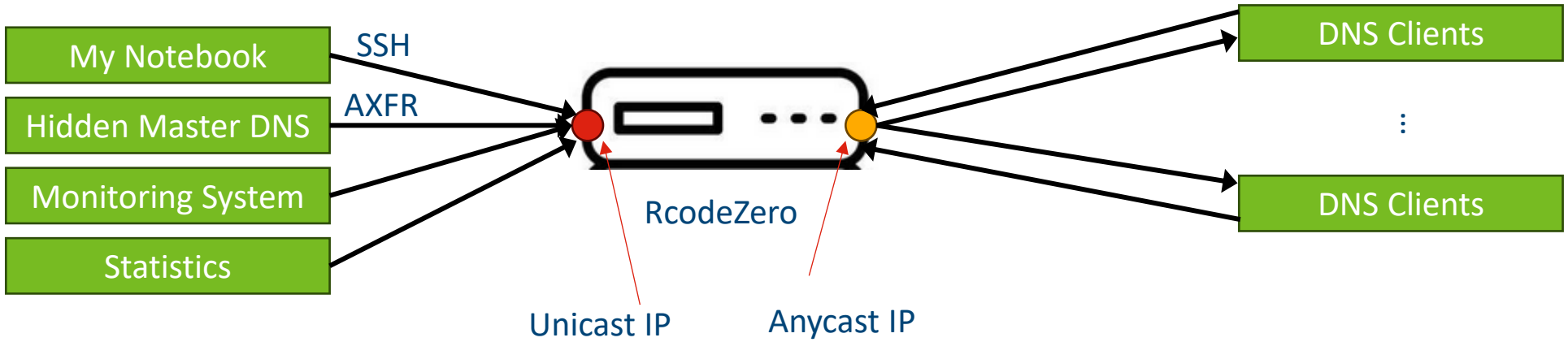
"Peering"

- Connectivity to some other networks (only that networks)
- No connectivity to the remaining Internet
- Usually free of costs (except switch port costs at the IX)



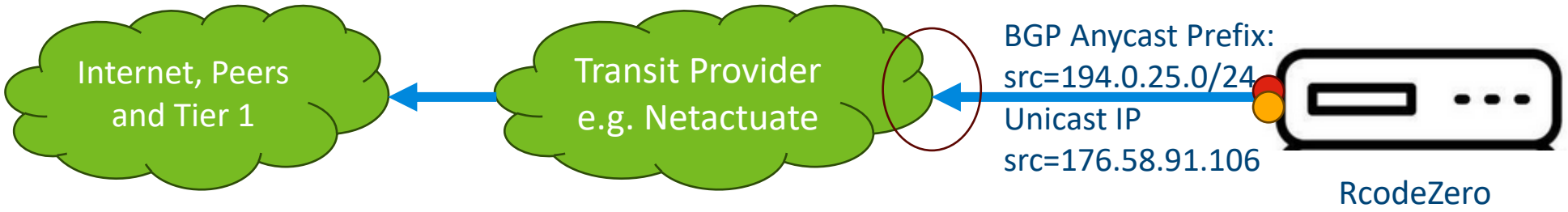
Anycast DNS Server Addressing

- Every server has 2 categories of IP addresses
  - A Unicast (globally unique) IP address for the server management
  - The Anycast IP addresses for the DNS service



RcodeZero „Global Node“

- Single network link for:
  - Management traffic
  - Anycast DNS traffic
- Transit for mgmt-traffic and anycast traffic
  - The provider knows our mgmt IP
  - The provider learns our anycast IP addresses (BGP)
  - Allows outgoing data with mgmt and anycast IP as source

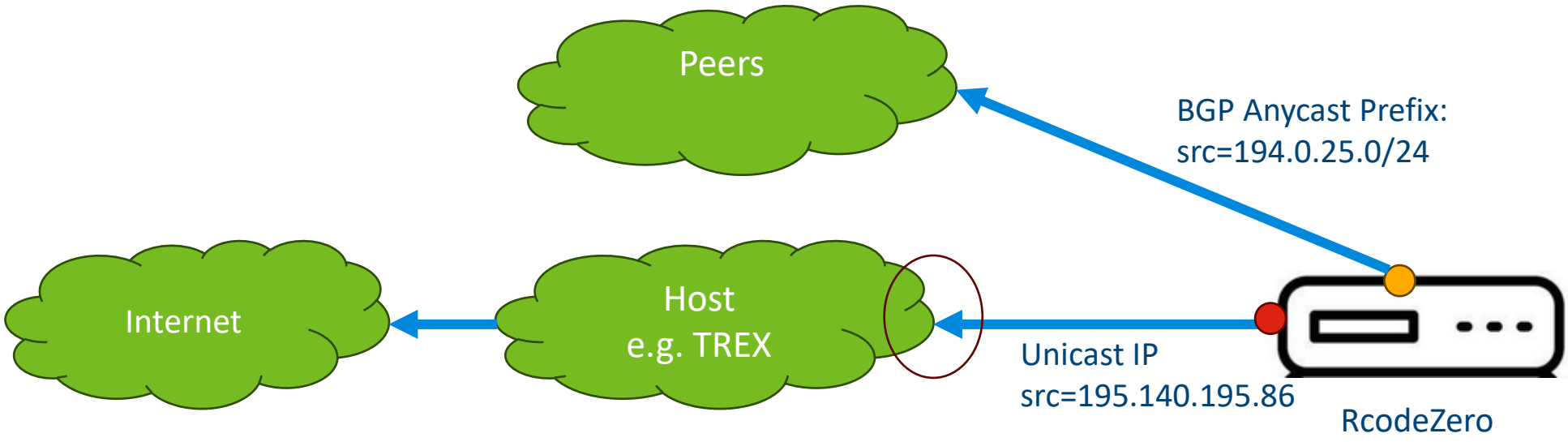


RcodeZero Local Node (IX nodes)

- The IX provides free colocation and IP connectivity for the server management
- Additionally the DNS server is directly connected to an Internet Exchange (peerings)
- The server has 2 network links
  - Management traffic with transit
  - Service traffic (Anycast DNS) with peerings only

Anycast Local Node

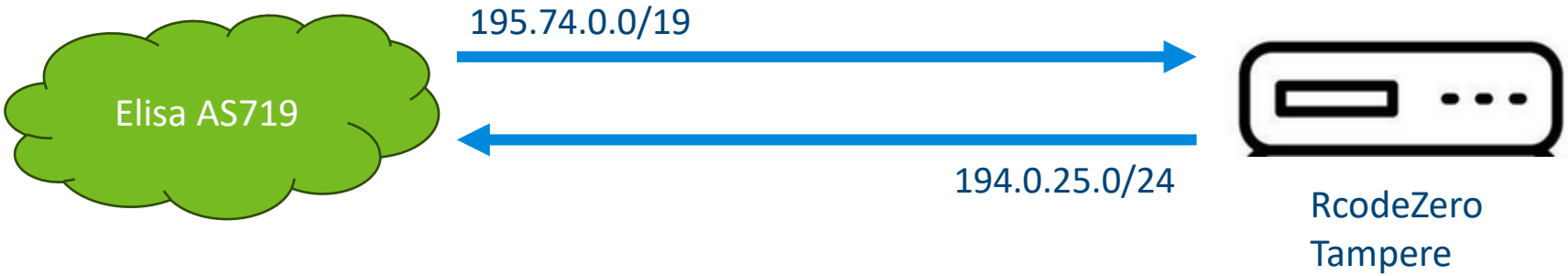
- The transit provider on the management link only knows the unicast IP address
  - BCP38? On the manamgent link only traffic originating from that mgmt IP is accepted



3. Symmetric vs Asymmetric Routing

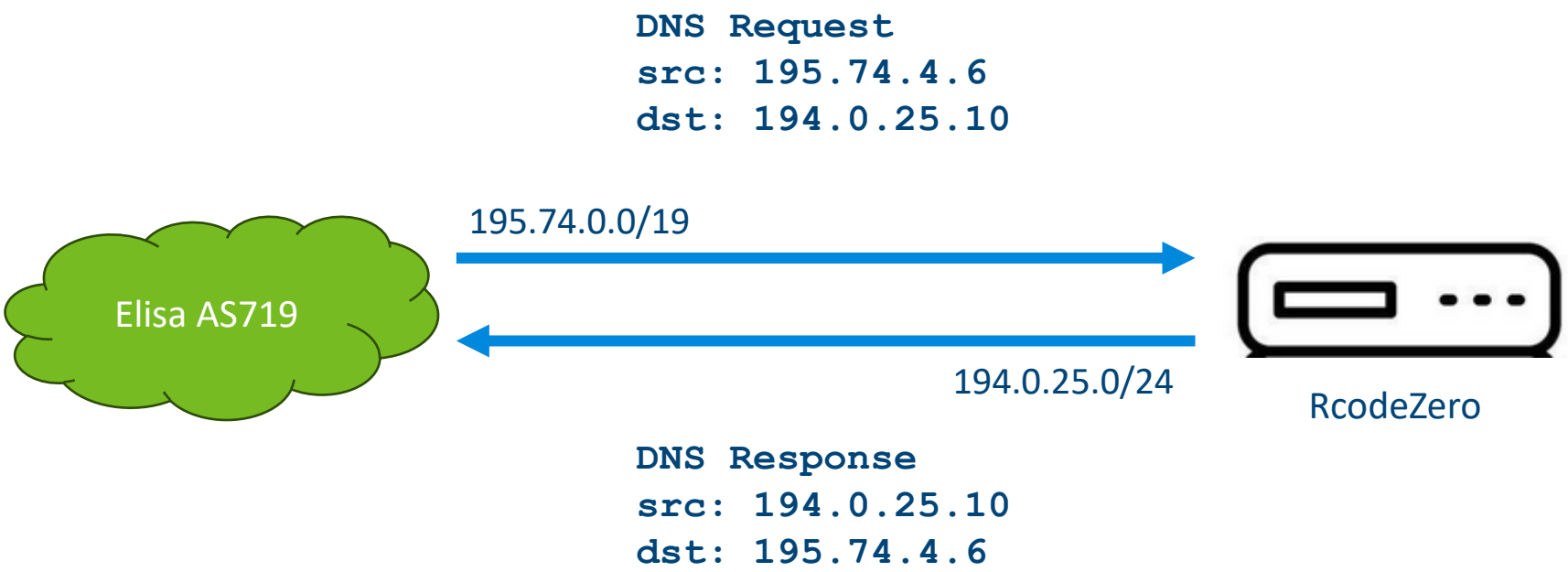
Peering

- Every AS announces its own (and customer) prefixes to peers



Symmetric Routing

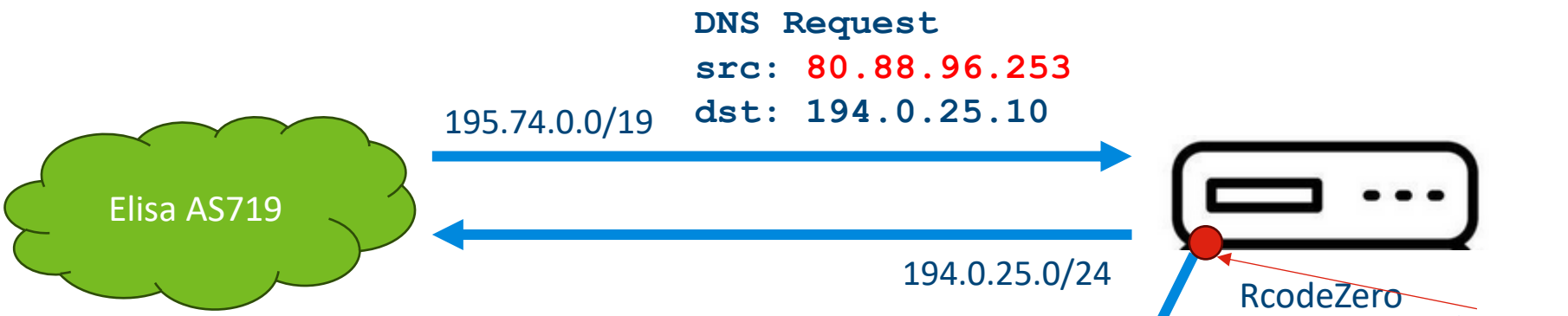
- Peer sends DNS request to us via IX
- We respond with DNS Response via IX



Clients will experience timeouts! (That's how we get aware of this issue.)

Asymmetric Routing

- Peer sends DNS request to us via IX
- No Route to send Response via Internet Exchange



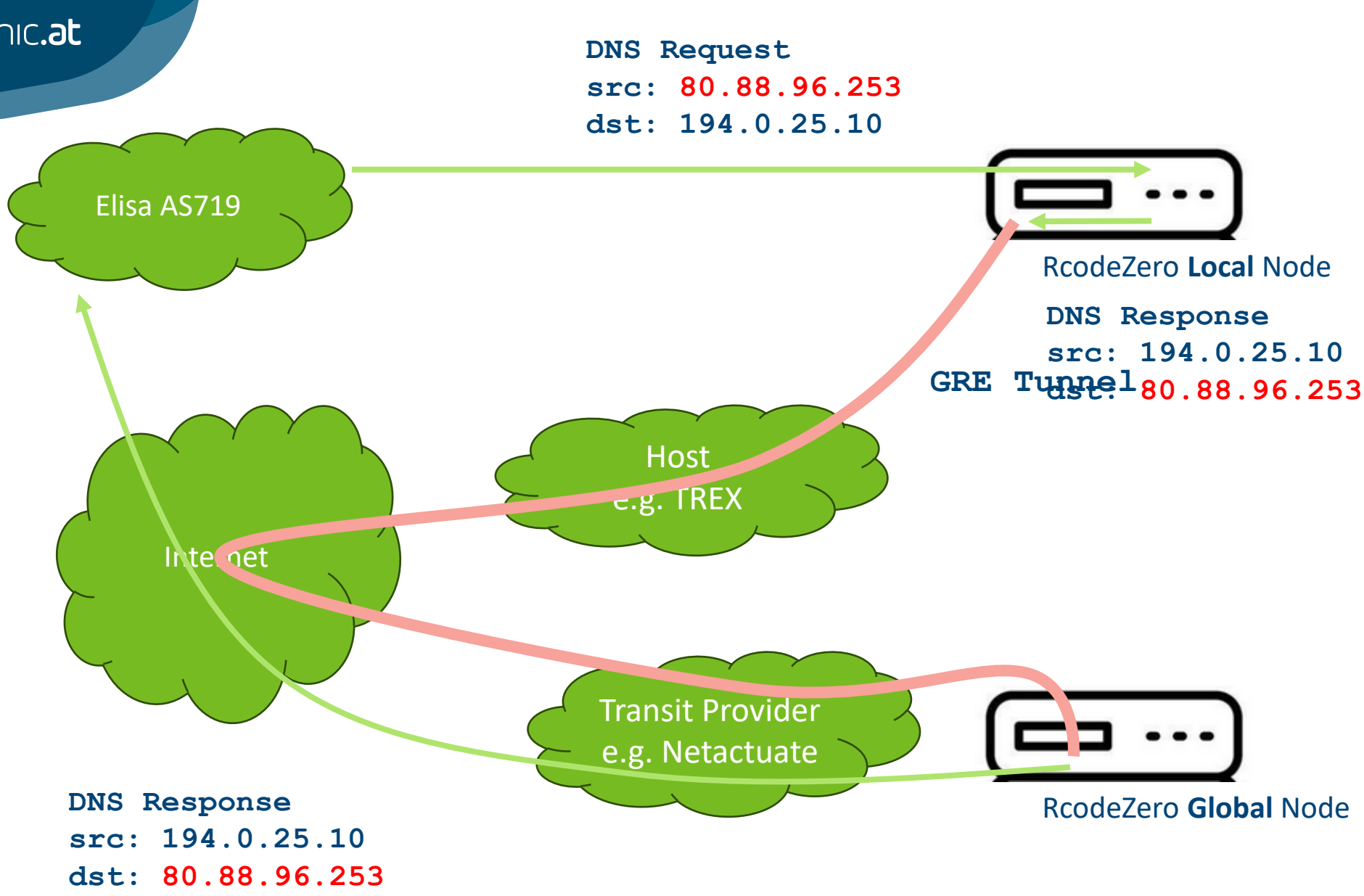
- How to route the response?

How to solve the issue?

- We must forward the response to someone that routes it

1. Mgmt traffic provider should tweak BCP38 filter
   - disable BCP38, static allow-list for our Anycast prefixes, ..
   - preferred
   - not always possible or simple not wanted
2. Beg for or buy transit from some ISP
   - potential extra costs for transit or cross connect patches
   - def. route via IX may break IX term
   - takes time
3. Tunnel traffic to a global node

- Or a combination of 2+3


DNS Request
src: 80.88.96.253
dst: 194.0.25.10
Elisa AS719
RcodeZero Local Node
DNS Response
src: 194.0.25.10
GRE Tunnel
dst: 80.88.96.253
Host
e.g. TREX
Internet
Transit Provider
e.g. Netactuate
RcodeZero Global Node
DNS Response
src: 194.0.25.10
dst: 80.88.96.253

Implementation Routing

- 2 separate routing tables (idea from Aleksi Suhonen)
  - default routing table for mgmt traffic
  - table "2" for anycast DNS traffic
- Feed IX routes into table 2

```
route-map from-ix permit 10
set table 2
```

- Activate table 2 for Anycast DNS traffic

```
routing-policy:
  - from: 194.0.25.0/24
    table: 2
```

Implementation GRE Tunnel

- Create GRE Tunnel

```
tunnels:
  ip6gre-toglobal:
      mode: ip6gre
      local: 2a00:11c0:4a:10::165
      remote: 2a02:850:ffe7::1
```

- Add default route into tunnel

```
! Make a static rule with higher distance
!(eBGP=20, iBGP=200) to act only as fallback
! May be overruled by a BGP default route
ip route 0.0.0.0/0 ip6gre-toglobal 250 table 2
ipv6 route ::/0 ip6gre-toglobal 250 table 2
```

Implementation GRE Tunnel

- Create GRE Tunnel

```
tunnels:
  ip6gre-toglobal:
      mode: ip6gre
      local: 2a00:11c0:4a:10::165
      remote: 2a02:850:ffe7::1
```

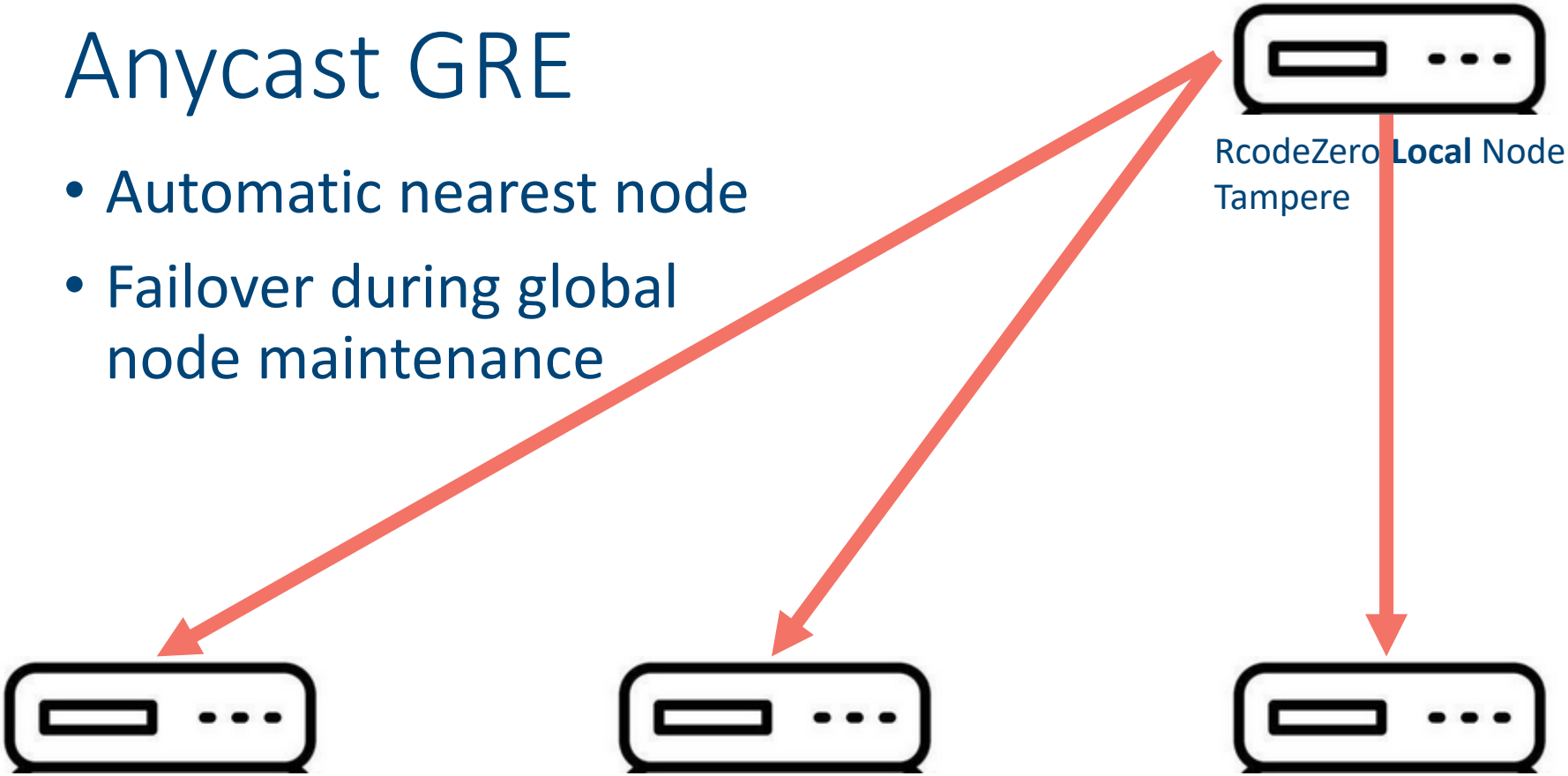
Which global node should be used?

- Add default route into tunnel

```
! Make a static rule with higher distance
!(eBGP=20, iBGP=200) to act only as fallback
! May be overruled by a BGP default route
ip route 0.0.0.0/0 ip6gre-toglobal 250 table 2
ipv6 route ::/0 ip6gre-toglobal 250 table 2
```

Anycast GRE

- Automatic nearest node
- Failover during global node maintenance



Summary

- Anycast GRE tunnel from local nodes to global nodes
- Local host does BCP38?
  - Default route for anycast DNS traffic into GRE tunnel
  - Optionally add BGP-based default route sponsored by a local ISP (can be added later)
- We still support IXs with local nodes

- How do others anycast providers with local nodes solve the problem?

Klaus Darilion · Head of Operations

klaus.darilion@nic.at