RcodeZero DNS

Operational insights

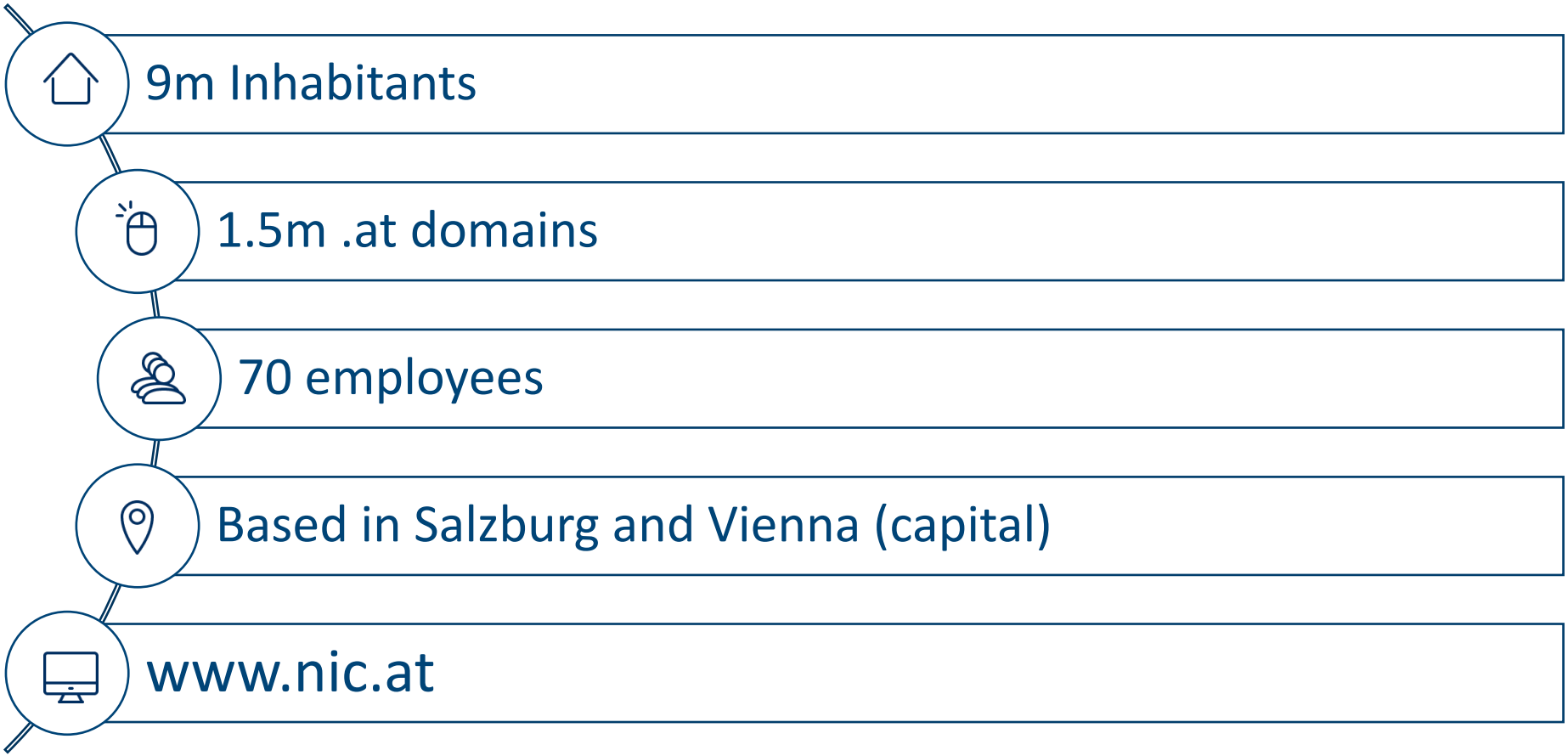
2024-05 · Christian Schöpp

Christian Schöpp

- PO RcodeZero DNS
- @nic.at


Tampere
Austria

nic.at – Registry for .at



Agenda

1. What we do
2. Why we do it
3. How we do it and
4. What are the challenges
5. Recommendations

What we do

- Authoritative DNS
- DNS Anycast Service
  - For us and other TLDs
  - For registrars, ISPs and corporations
- Anycast: Global fleet of nodes available under the same (set of) ip-adresse(s).

Why we do it

- Diversification
  - You (as a business) should not rely on just one product.
- Resiliency/redundancy
  - You (with responsibility to a service) should not rely on just one provider.
- Latency
  - You want similar experience around the globe.

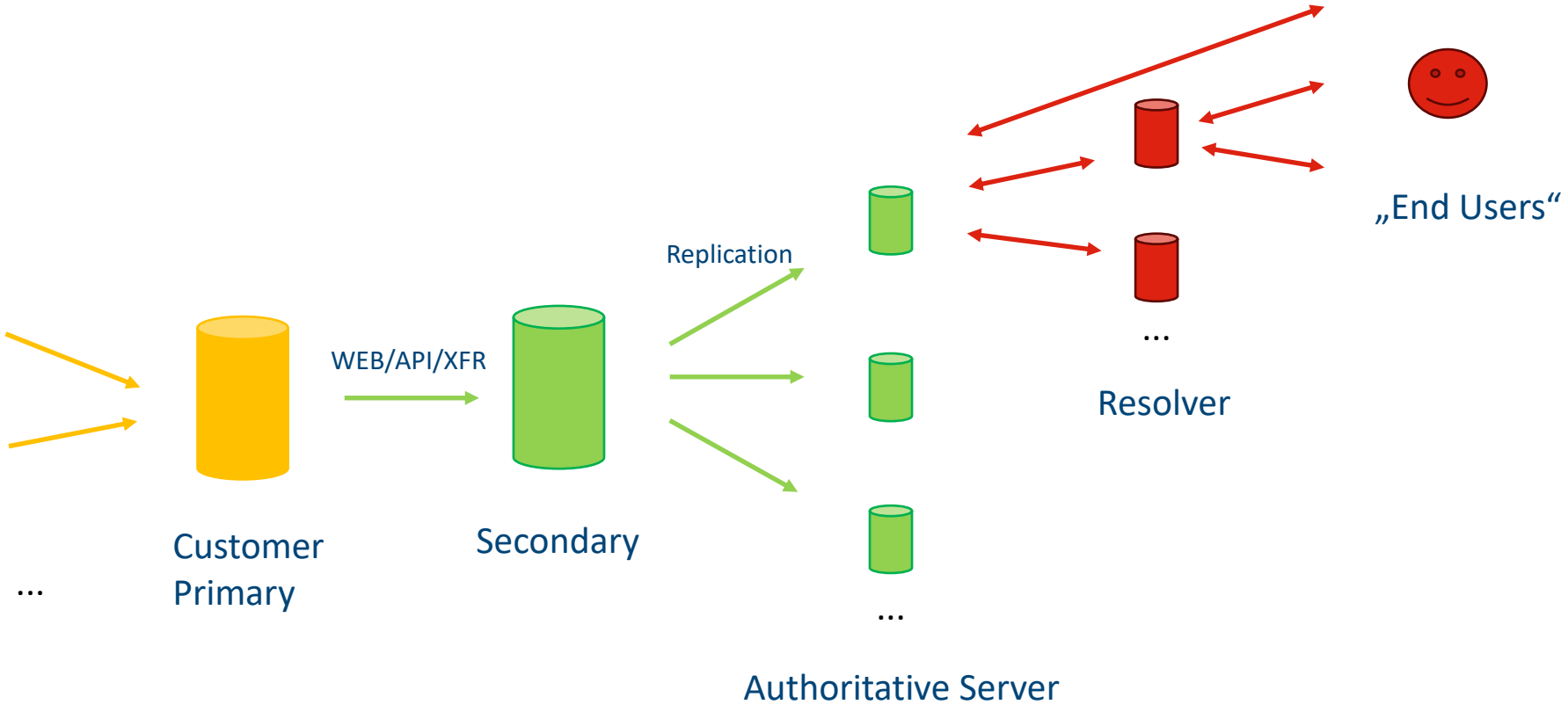
How we do it

- Basics
- Overview
- Nodes

Basics

1. We receive zones and content (data) from our customers
2. We distribute them around the globe
3. We (constantly) keep them in sync
4. We answer upon queries
5. We log queries (statistics)

Overview (simplyfied)



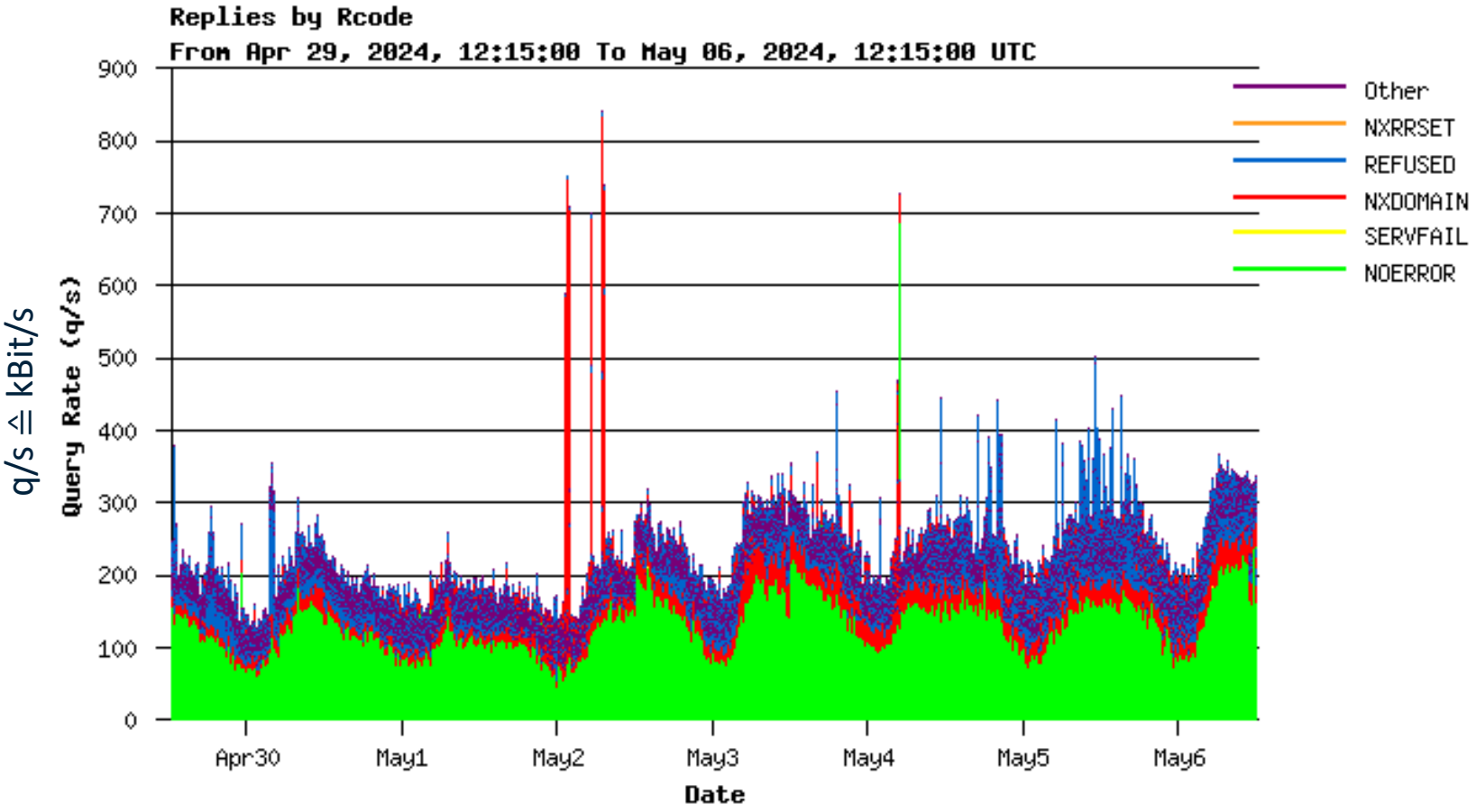
Architecture (TLD vs. Secondary)

	TLD (e.g. .fi)	Secondary (e.g. nog.fi)
Fleet/clouds	1	2
Nodes	35	52
Number of AS	1	2
Anycast IP´s	1	2
Client 2 RcodeZero	XFR	Webinterface, REST, XFR
Zones per Client	Few	Up to millions
Delegations per Zone	Up to millions	Few

Where to place a node?

- Nodes by population
- Nodes by customers
- Nodes by traffic
- Nodes by speed
- Nodes by costs
- Nodes by tactical reasons

Tampere



q/s ≙ kBit/s

Tampere

- Local node
- Little traffic
- Bilateral peering is necessary
  - Big players are often not on route server
- Mostly beneficial for the country it is placed at
  - Low latency
  - Less DDoS attacks

What are the challenges

- KPI's
- Measurement
- Routing
- Attacks

Key Performance Indicators

1. Performance (ms)
   - Routing issues

2. Uptime (%)
   - Attacks

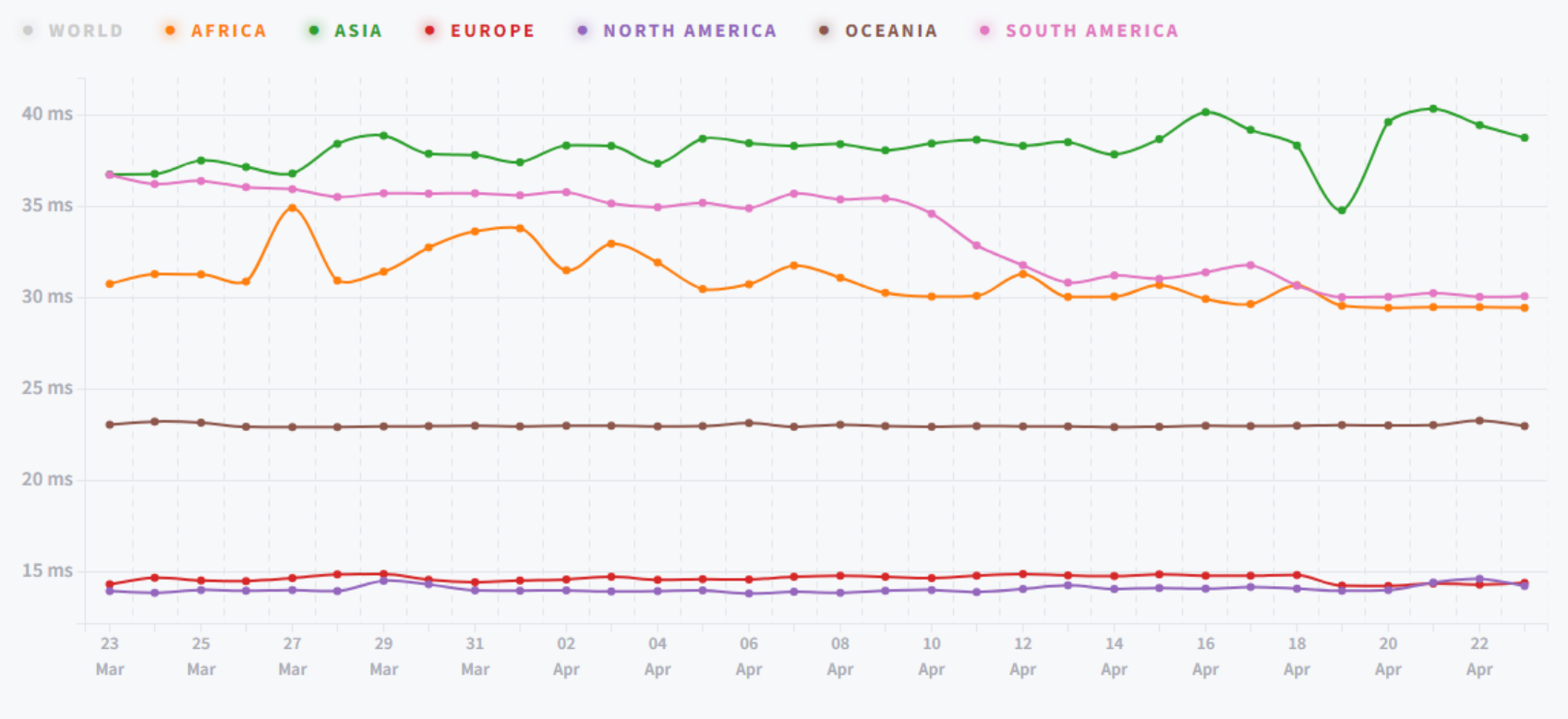
3. Propagation delay
   - Architecture
   - Global connectivity issues

Performance

Strategies

1. Nodes everywhere
2. Smart node placement and optimized routing
3. Don’t care

Performance



Uptime

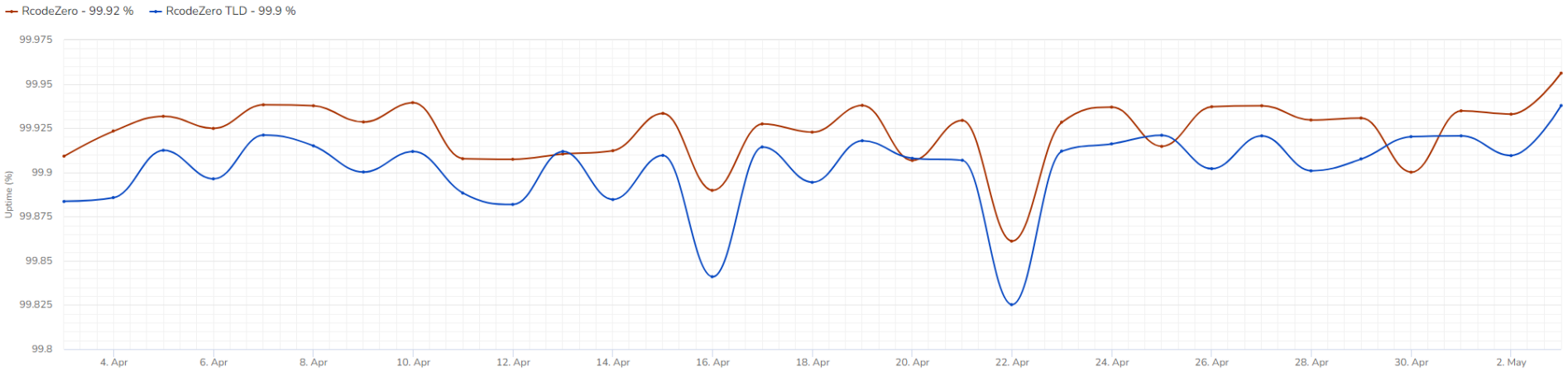
DNS has built in redundancy

- nog.fi lists 5 authoritative nameservers (dig nog.fi NS)
- Resolvers apply their own strategy
  - Test them regularly
  - Rank them
    - Speed
    - Reliability
  - Distribute DNS queries (somehow)

Uptime

- Routing issue at upstream provider
  - Black hole every three months
  - Queries or answers get lost
- Multi vendor strategy(TLD)
- 2 independent clouds (2nd level)
  - Different routing policy
  - Different transit providers
  - Different node

Uptime


RcodeZero - 99.92 %
RcodeZero TLD - 99.9 %
Uptime (%)
99.975
99.95
99.925
99.9
99.875
99.85
99.825
99.8
4. Apr
6. Apr
8. Apr
10. Apr
12. Apr
14. Apr
16. Apr
18. Apr
20. Apr
22. Apr
24. Apr
26. Apr
28. Apr
30. Apr
2. May

Measurement

- "Wer viel misst, misst viel Mist."
- If you measure a lot, you measure a lot of nonsense.
- Measurement is influenced by the point of view.

Measurement

- DNS = mostly UDP, sometimes TCP
- UDP might get lost

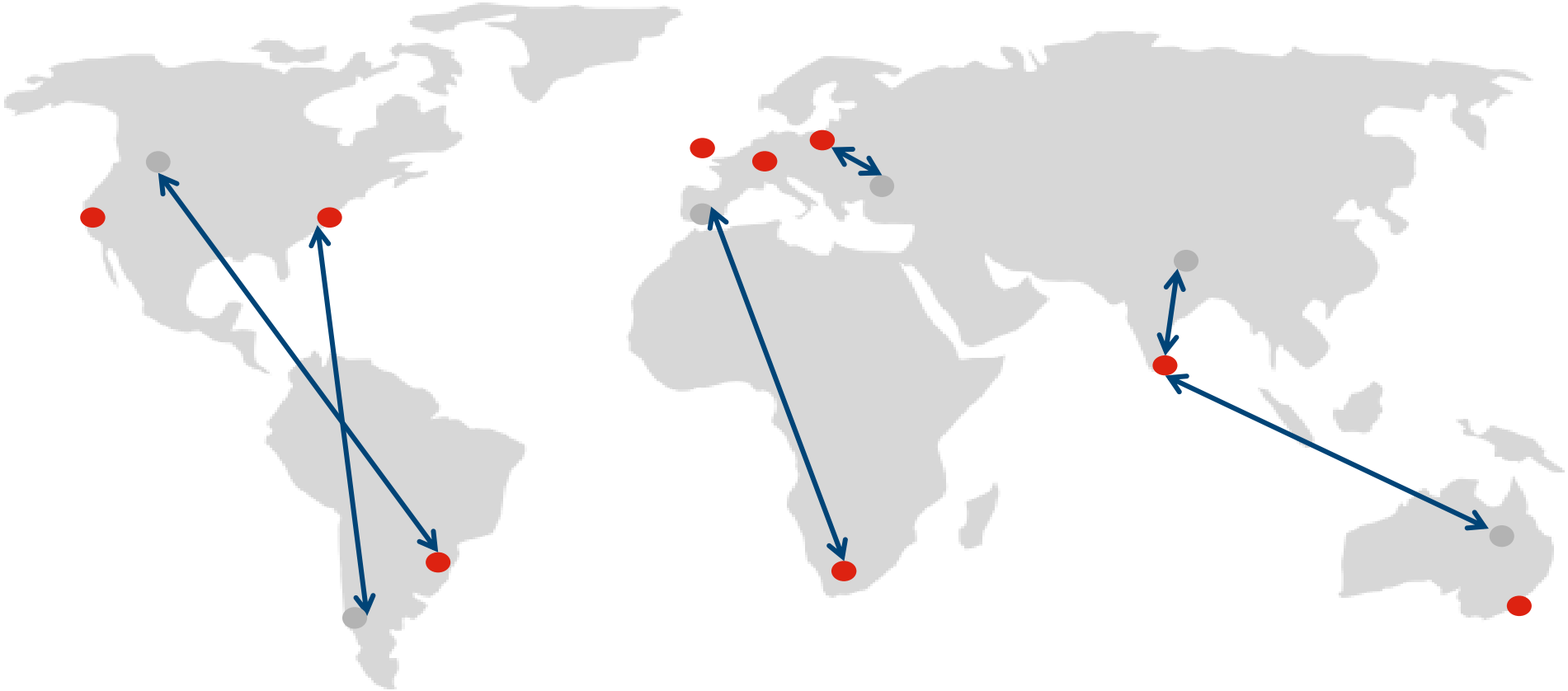
- UDP ≠ ICMP
- route A ≠ route B

- We have outsourced measurement

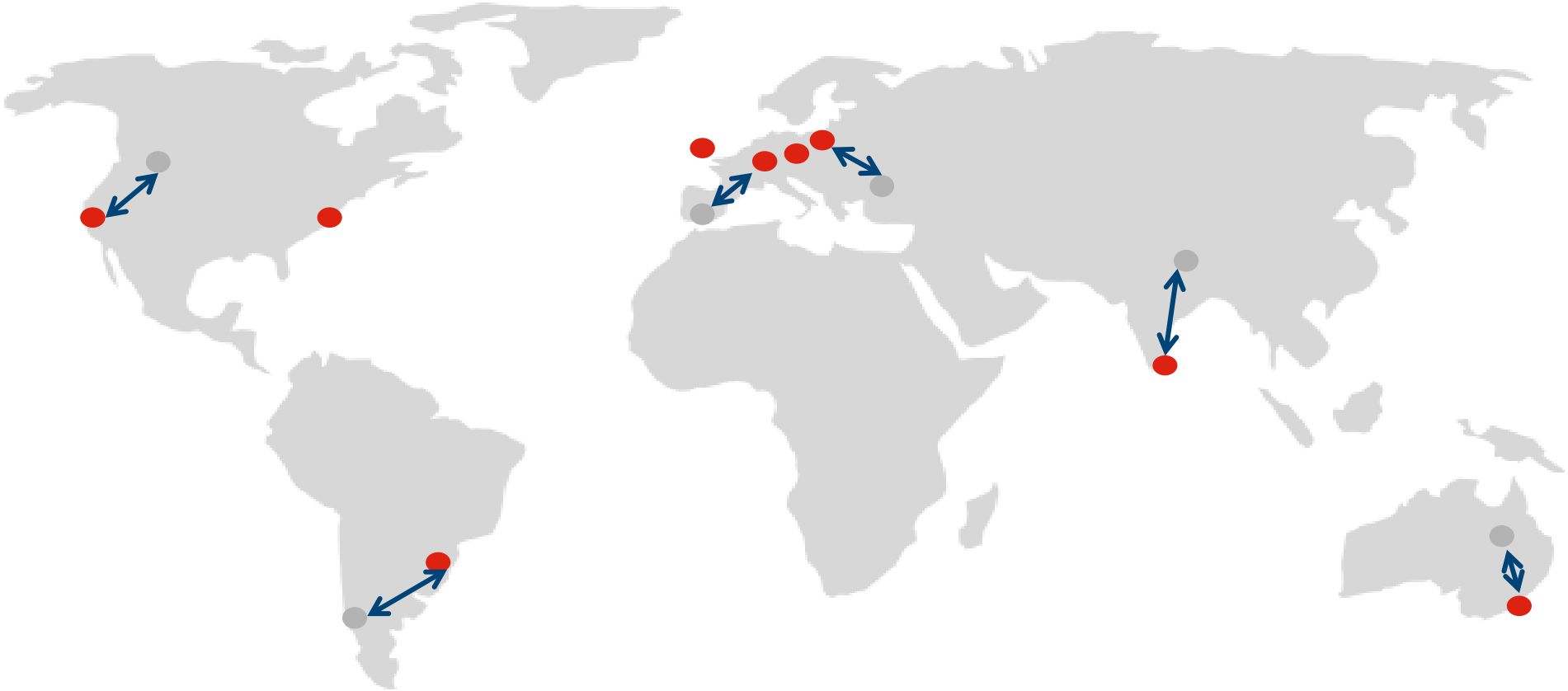
Routing challenges

- We want traffic as local as possible
  - For low latency
  - To allow load balancing
  - Prerequisite to scale

Bad optimization



Good optimization

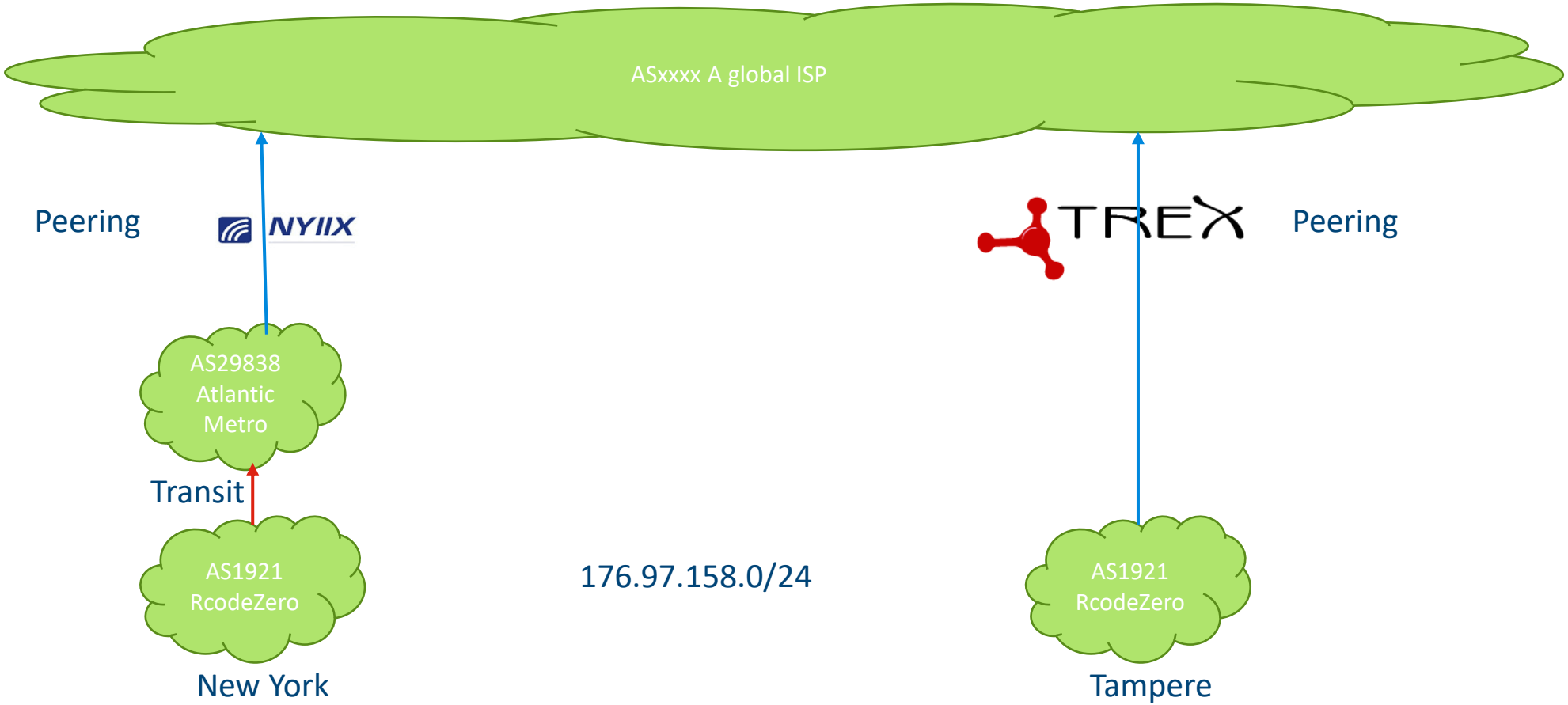


Routing challenges

- Heavy traffic engineering
  - Announce or not to announce to transits/peers?
  - Path prepending
  - Use of upstream BGP communities
  - Asymetric routing (local nodes/exchanges)
- Individual peerings do not scale
  - We love open peerings via route servers
- Never ending story (globally seen)

Why do you prepend?

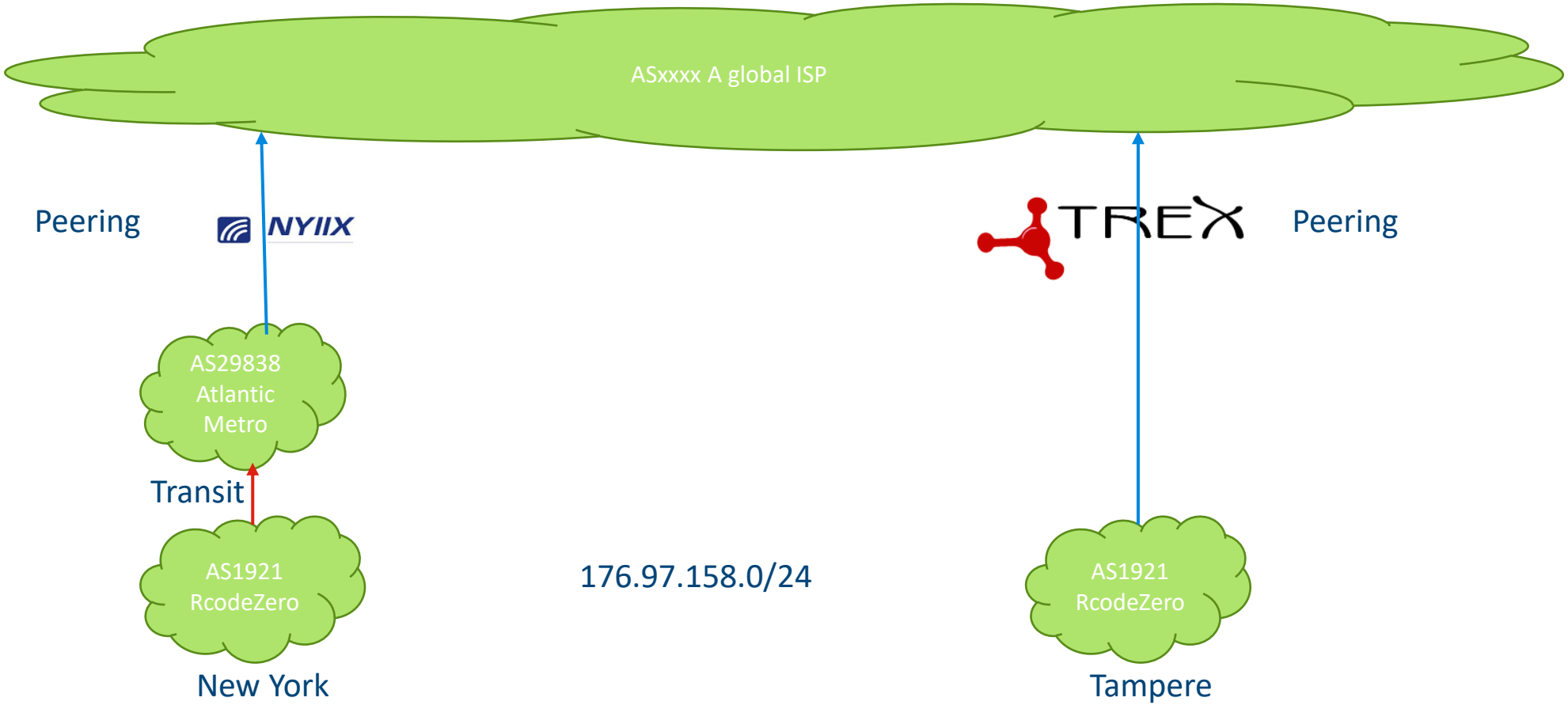
Example of unoptimized BGP



Example of unoptimized BGP

New York: AS-Path length=2: 30971 1921

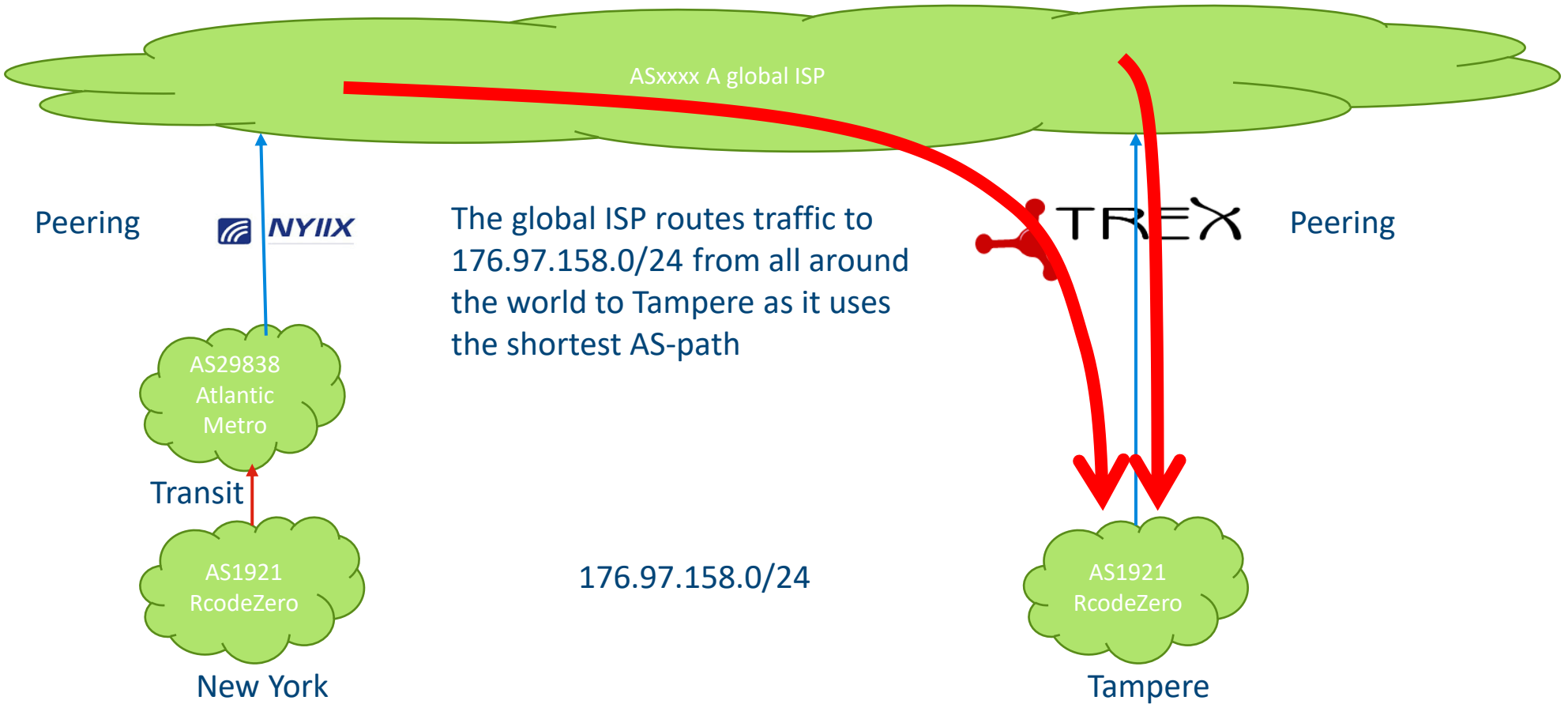
Tampere: AS-Path length=1: 1921



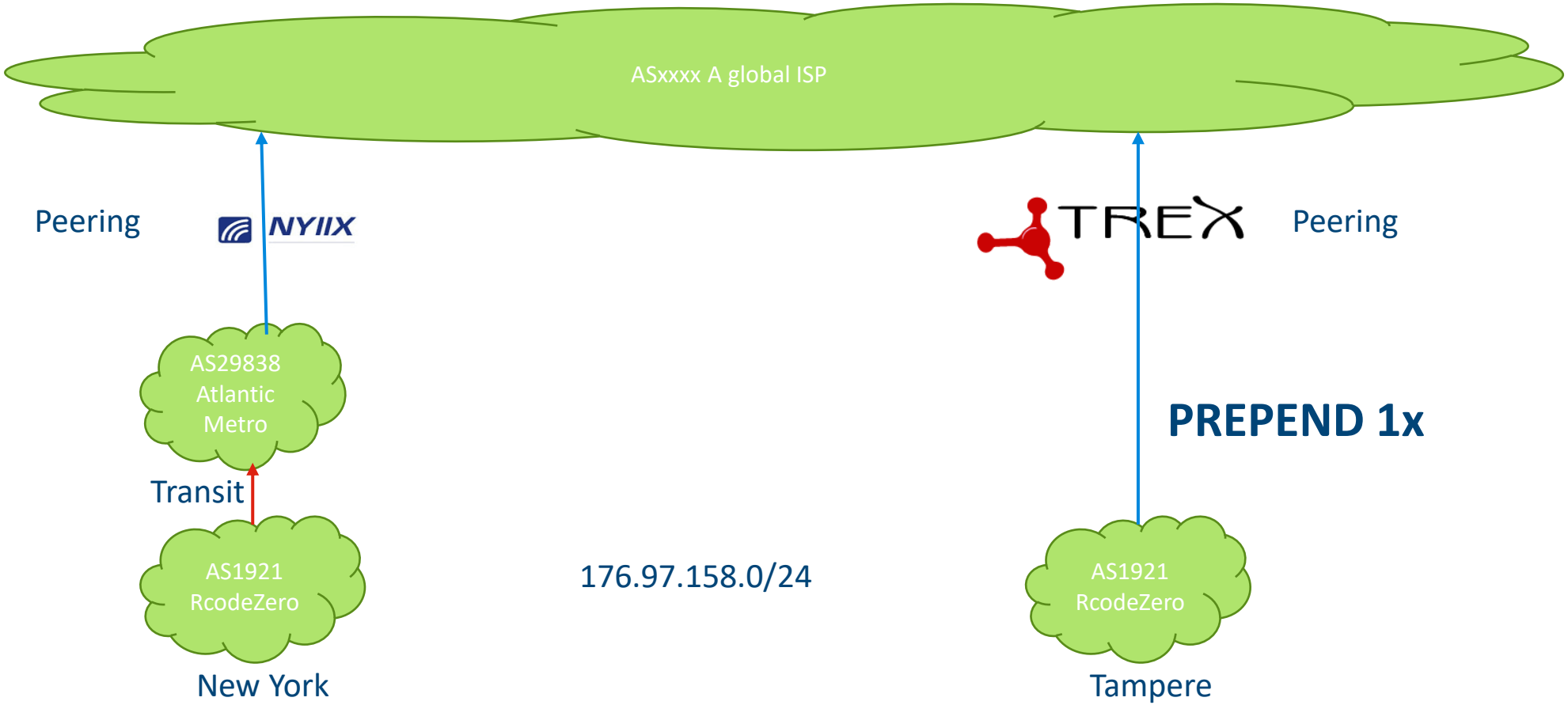
Example of unoptimized BGP

New York: AS-Path length=2: 30971 1921

Tampere: AS-Path length=1: 1921



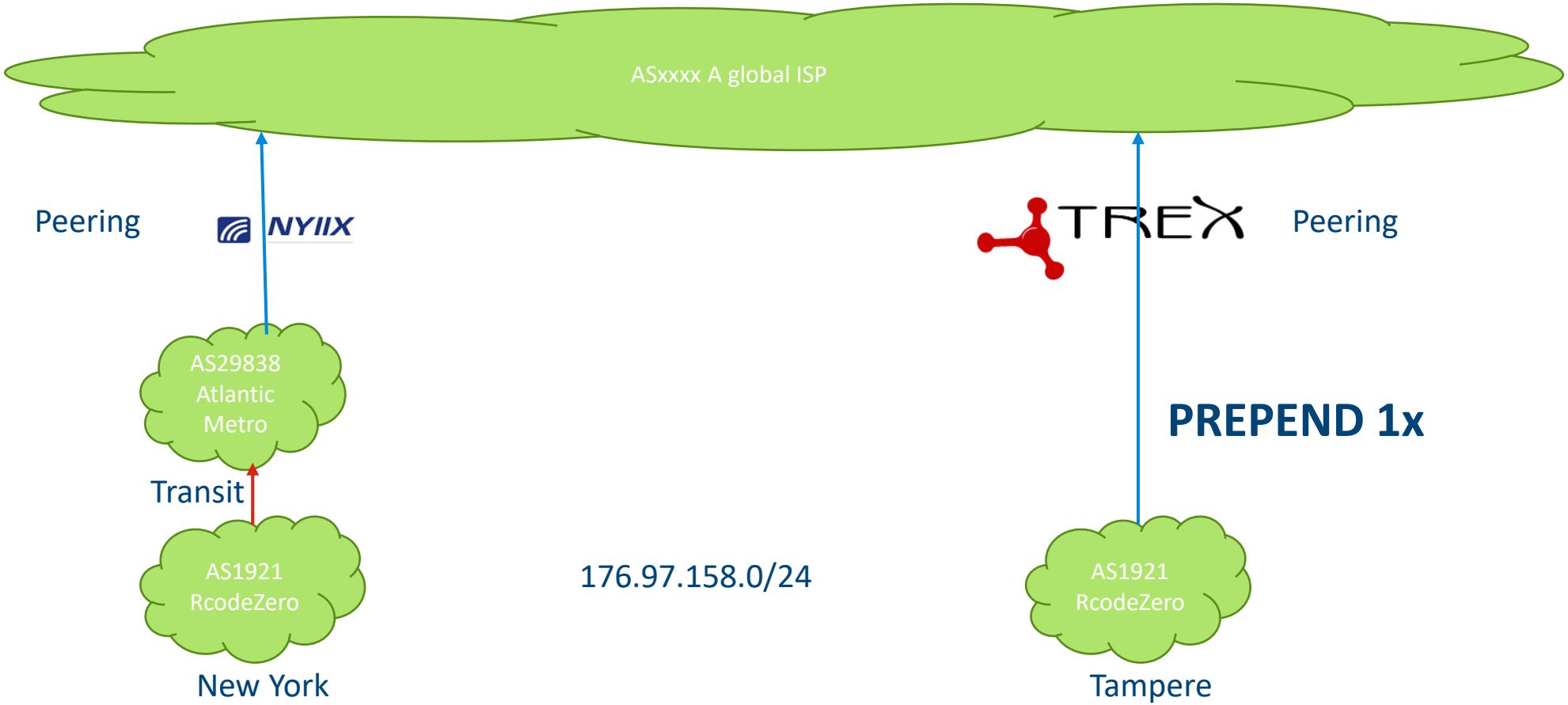
Example of OPTIMIZED BGP



Example of OPTIMIZED BGP

New York: AS-Path length=2: 30971 1921

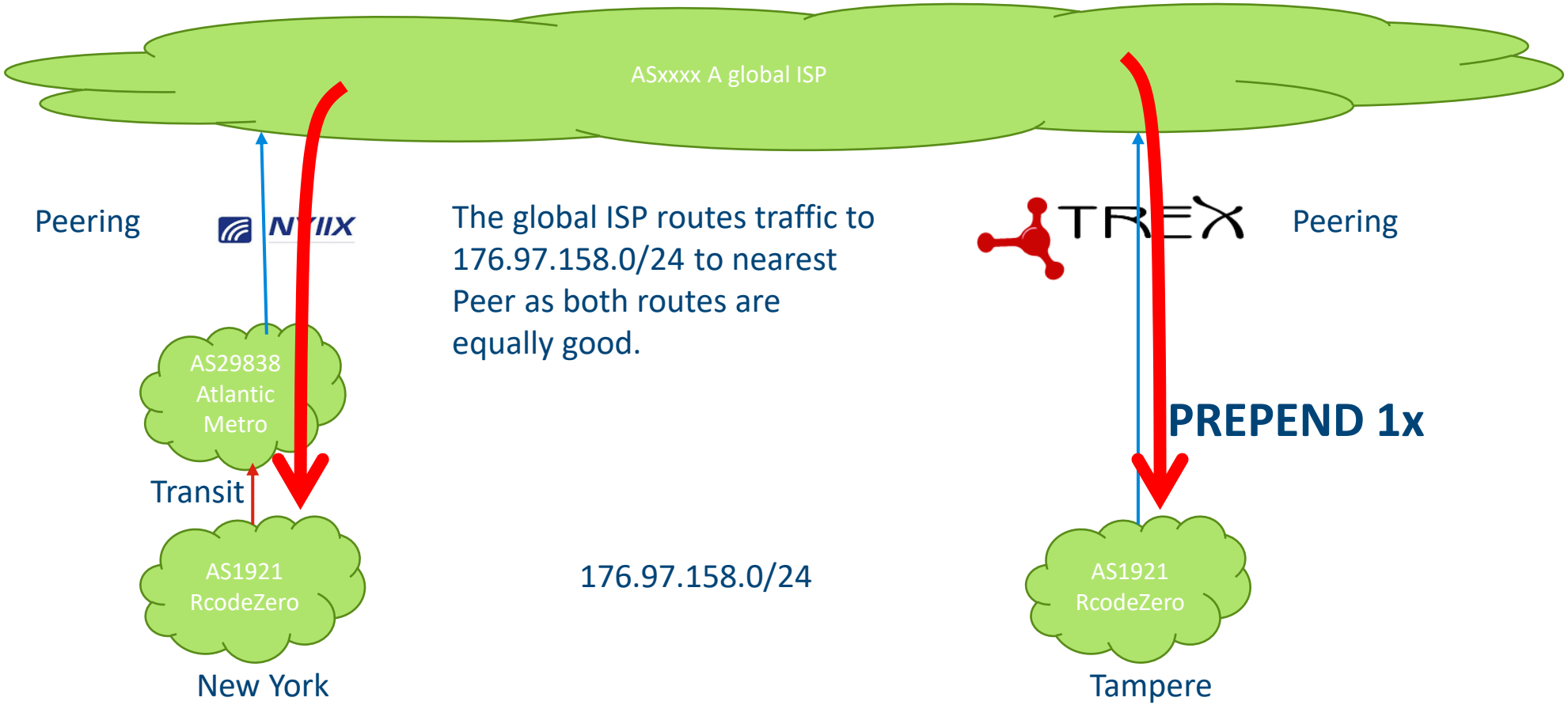
Tampere: AS-Path length=2: 1921 1921



Example of OPTIMIZED BGP

New York: AS-Path length=2: 30971 1921

Tampere: AS-Path length=2: 1921 1921



Prepending is needed

- To control traffic
  - by making the shortest AS path as long as the longest
    - on all our anycast locations
    - to peers
    - global transit providers
- To consider special routing situations
  - DDoS mitigation provider is activated
    - Traffic should be routed via DDoS mitigation provider, not directly to us
    - Extend AS path even one more time
- So we ended up having an AS path length of 5 towards IX/Peers and Tier1 transit providers.

Routing challenge (real life)

- An incumbent is the local key player. At home they peer with nobody - they want to sell transit.

- Somewhere else they are a small players and (needs to) peer with everybody.

-> Traffic is going round the world instead of going to a node close by.

Exchange or Provider

- Exchange
  - Full control over peerings / routing
  - Only if invited

- Provider
  - Colocation/Server/VM included
  - Peerings included
  - Transit included
  - Traffic shaping trough BGP communities
    - Therefore not every provider suitable

Common attacks (authoritative DNS)

- Volumetric
- Application layer

Volumetric Attack

- Garbage to fill up links or nodes
- Outsourced
  - Automatic detection per node
  - Our prefix announced by the provider
  - Scrubbed and anycasted back to “nearest” node
  - Very little impact on latency and load distribution

Application Layer Attacks

- DNS queries (try to) overload our service
  - Real attacks
  - Configuration mistakes
  - Research/Security/Penetration tests

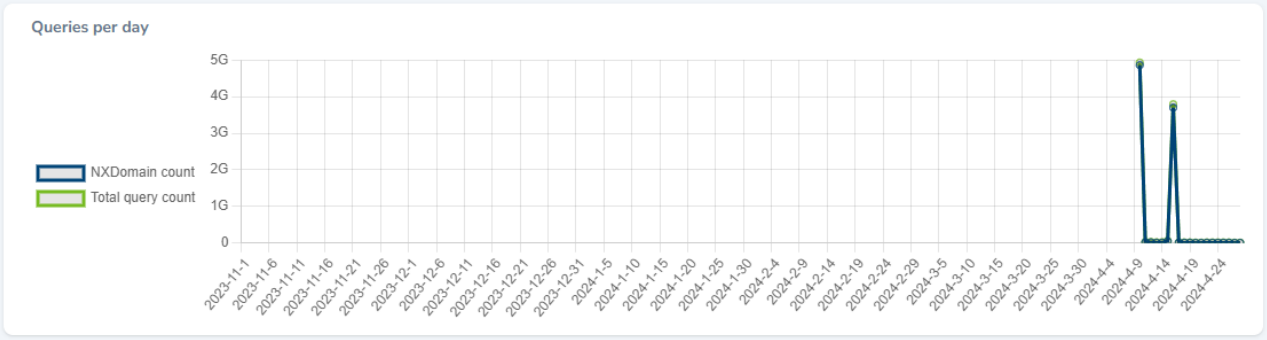
Application Layer Attacks

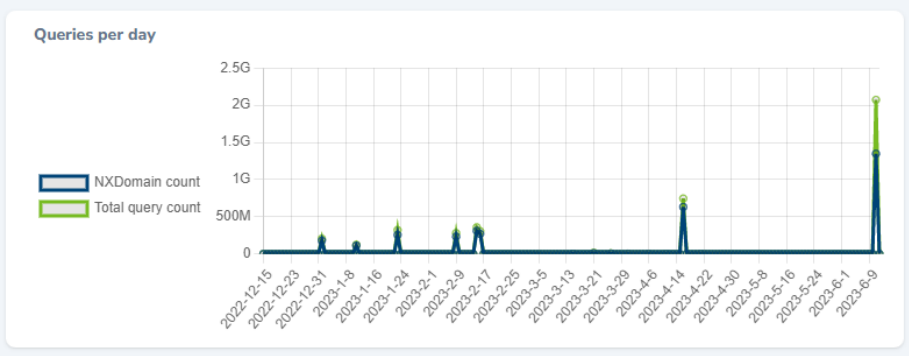
- Query (random) 2nd or 3rd level domain
  - nslookup 123xyz.nog.fi, abc789.nog.fi…
  - **NXDOMAIN**
  - **"I know that it does not exist"**
- Query name server (more or less) directly
  - nslookup abc.bca ns.nog.fi
  - **REFUSED**
  - **"You have come to the wrong place"**

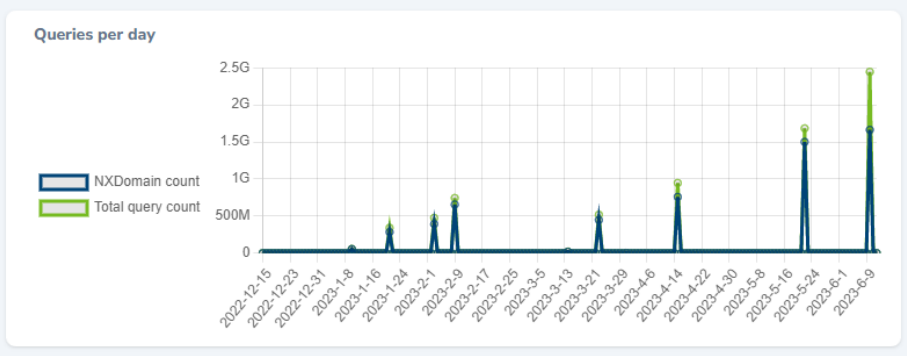
Application Layer Attacks

- Sunny day vs. rainy day
  - Factor 100 – 1000
- Size matters
  - More nodes are better
  - Stronger nodes are better
  - But not all nodes get equal amount of traffic

Attacks (per domain)

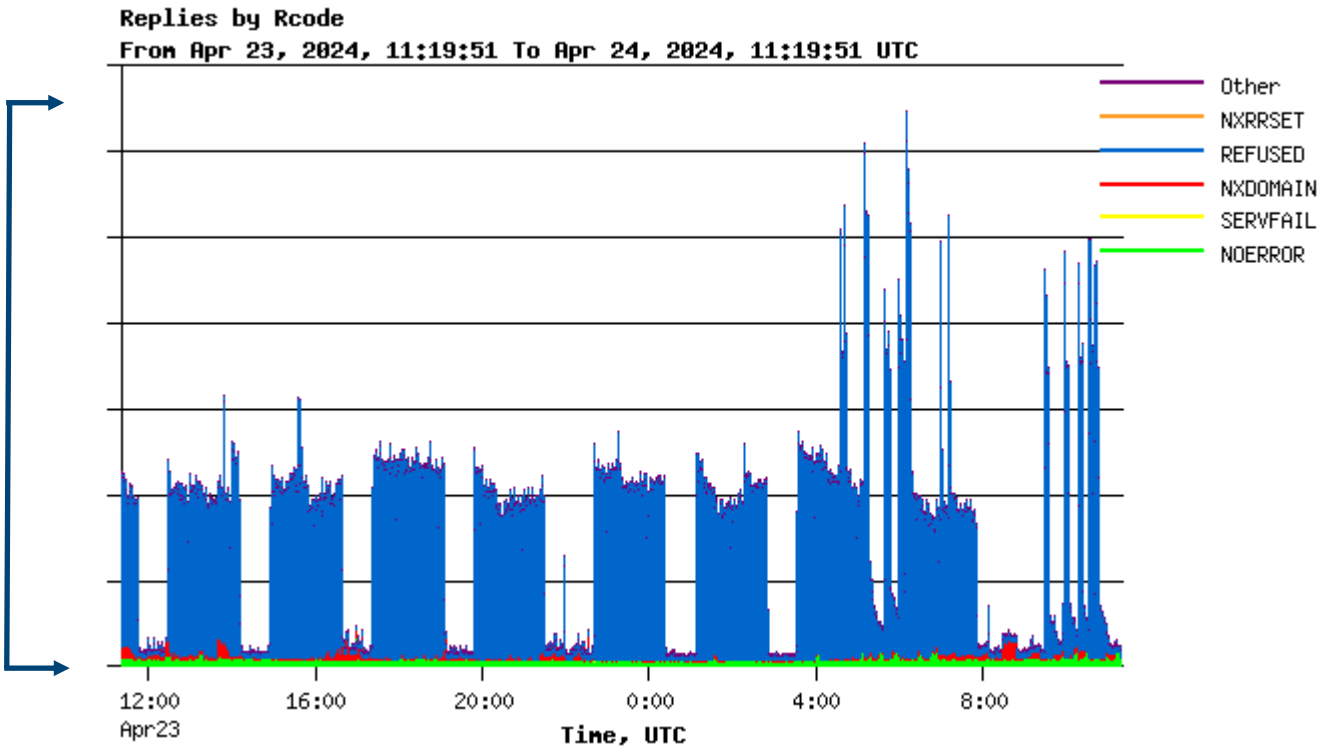






Attacks (per node)

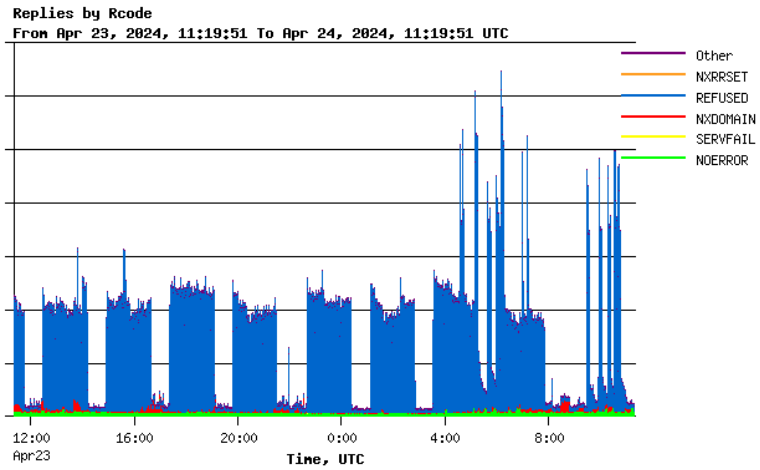
~Factor 100



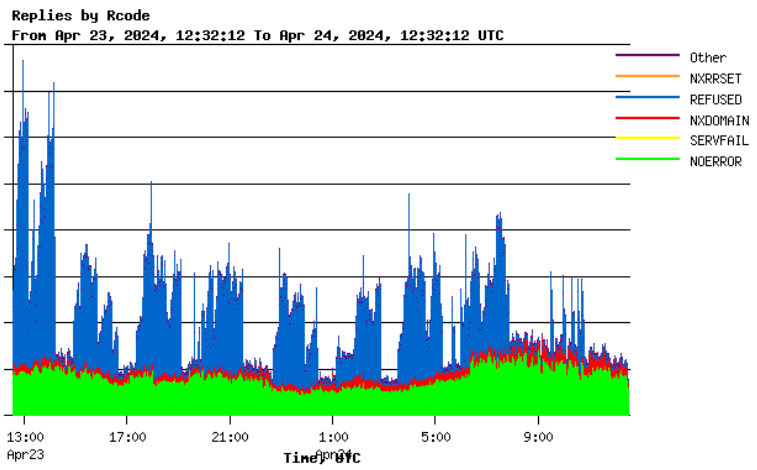
Example Node

Attacks

- Which one is closer to the source?

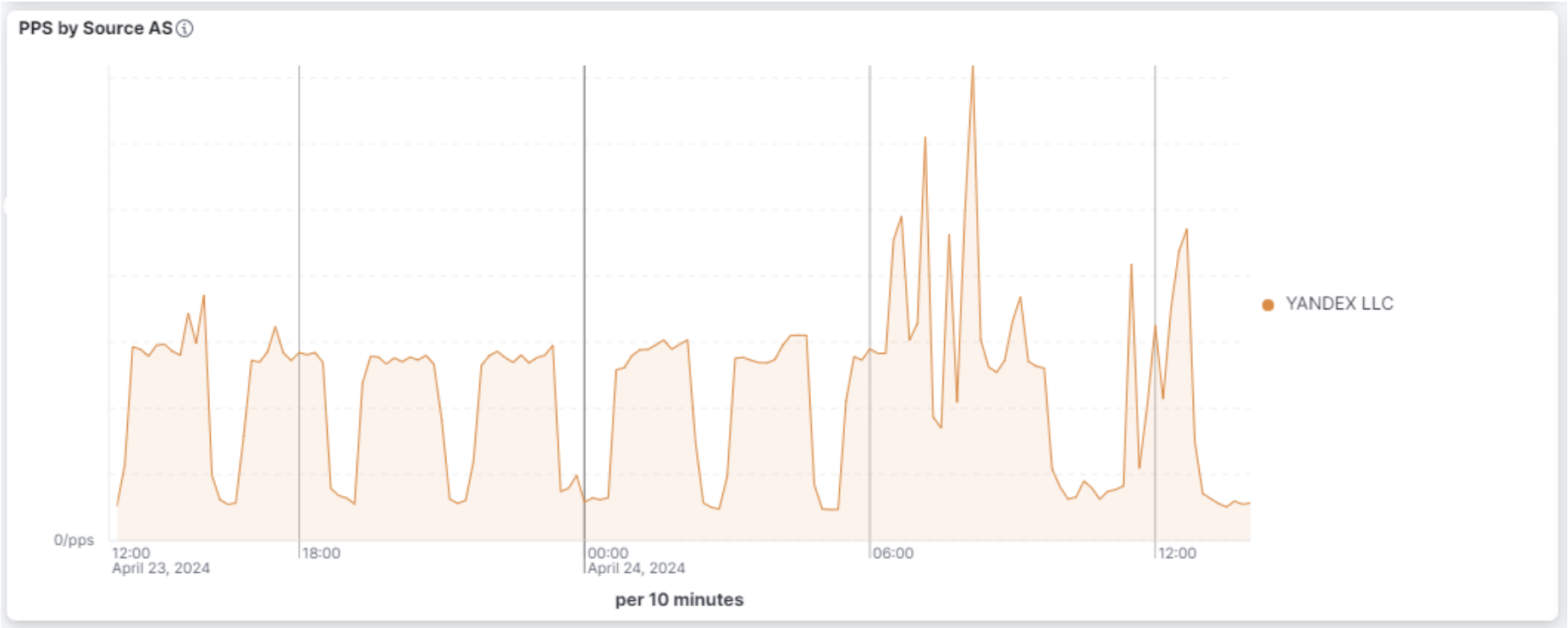


Moscow



London

Attacks


PPS by Source AS
YANDEX LLC
0/pps
12:00
April 23, 2024
18:00
00:00
April 24, 2024
06:00
12:00
per 10 minutes

Attacks

- Most attack traffic comes via public DNS providers
  - Hard/impossible to block
- Target is either you or your customer
- Source of the attack ≠ attacker
- REFUSED are usually configuration mistakes
- NXDOMAIN are usually attacks

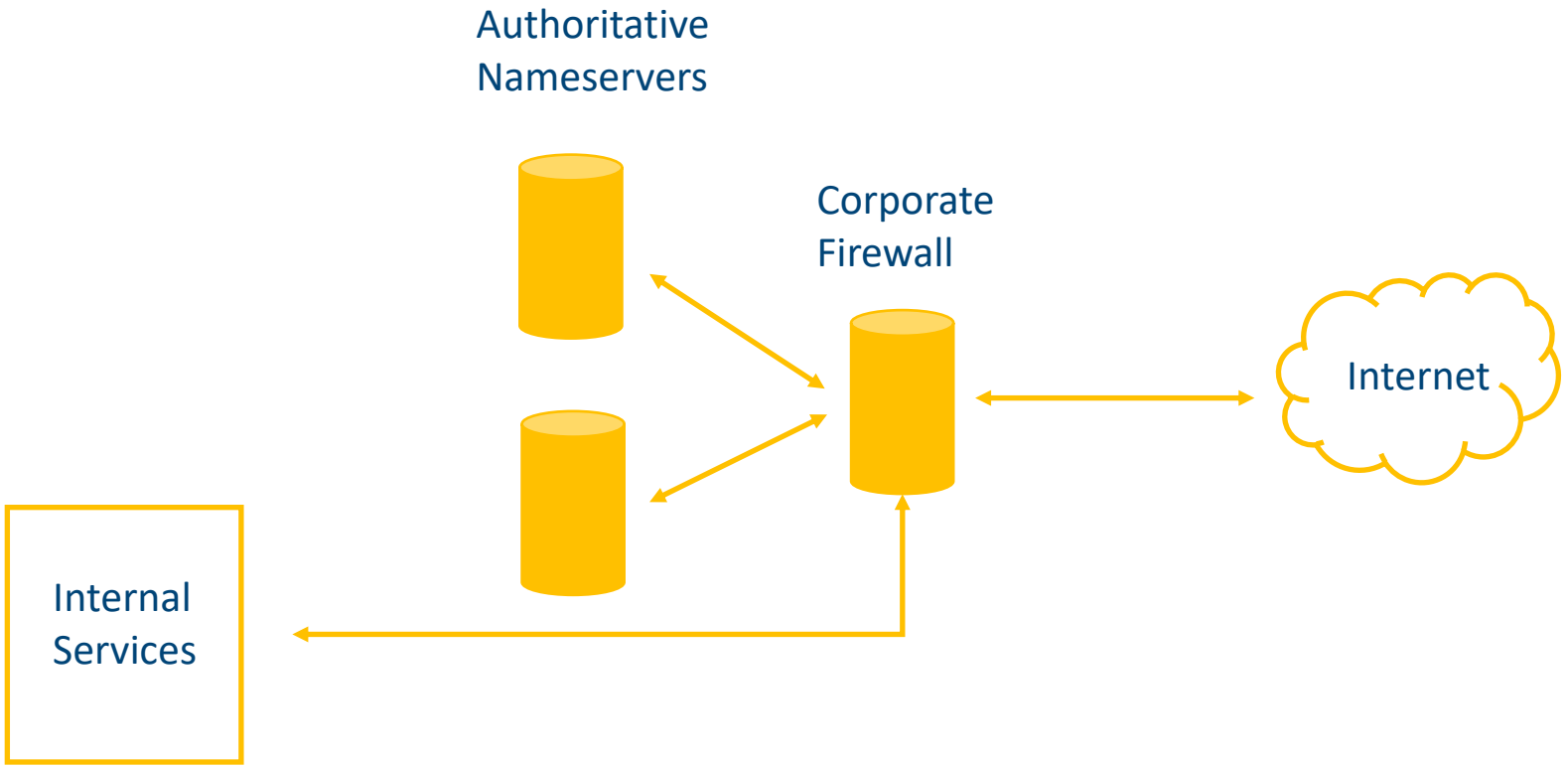
Fun facts

- Most zones are never queried.
- Most queries are for zones that do not exist.

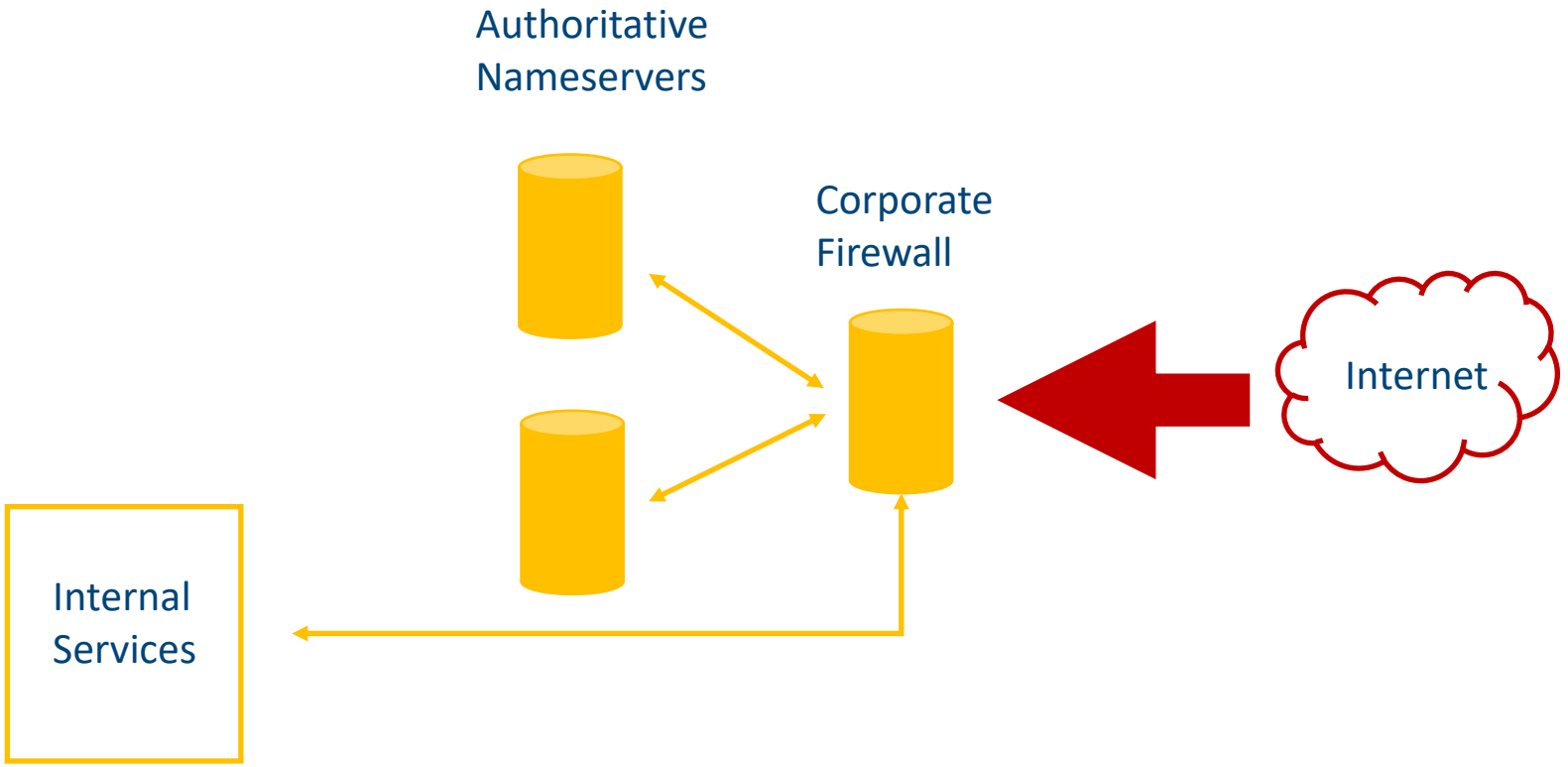
Recommendation

- Usually, customers come to us because of
  - bad architecture
  - being too small
  - or both

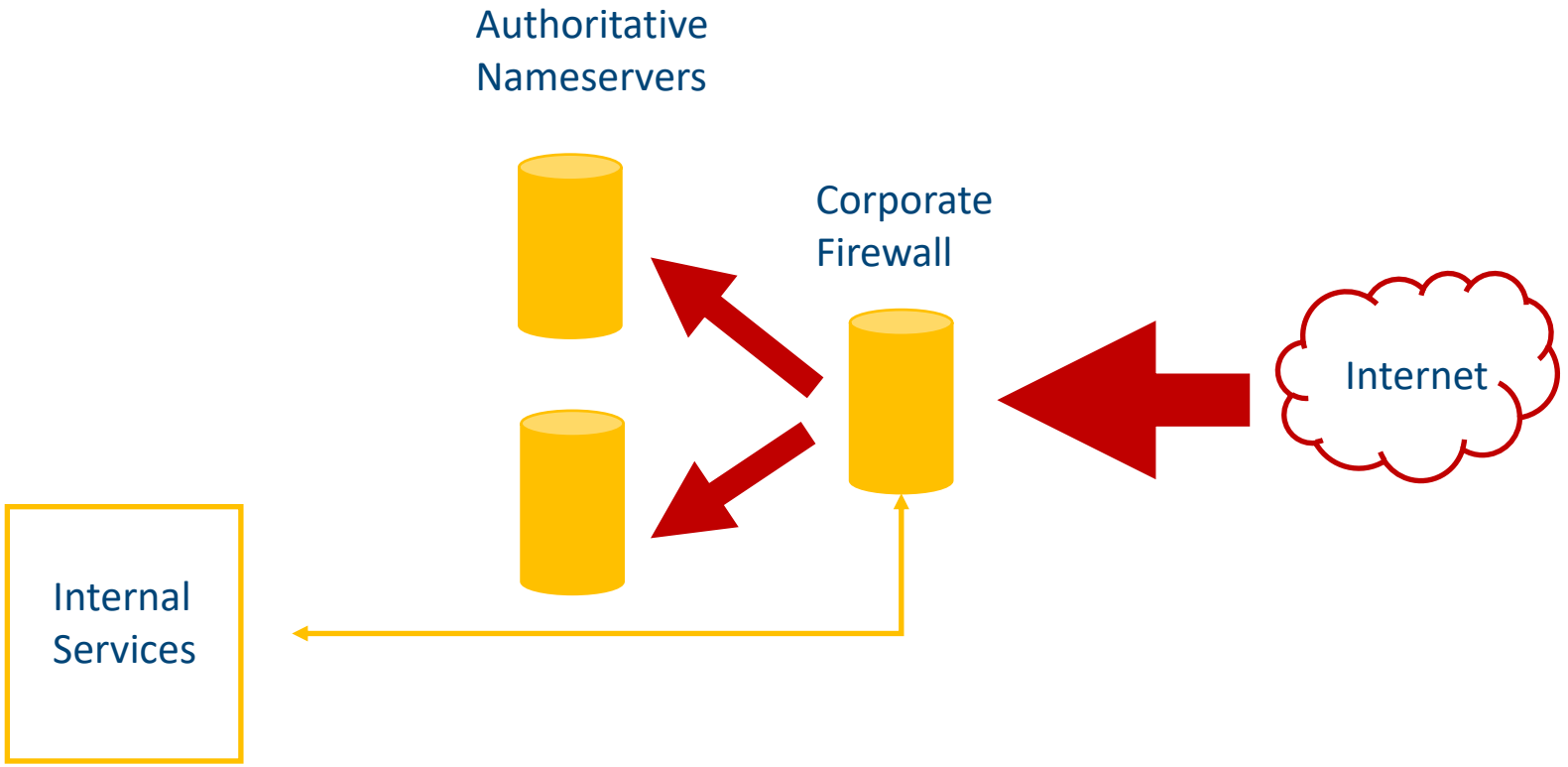
Architecture



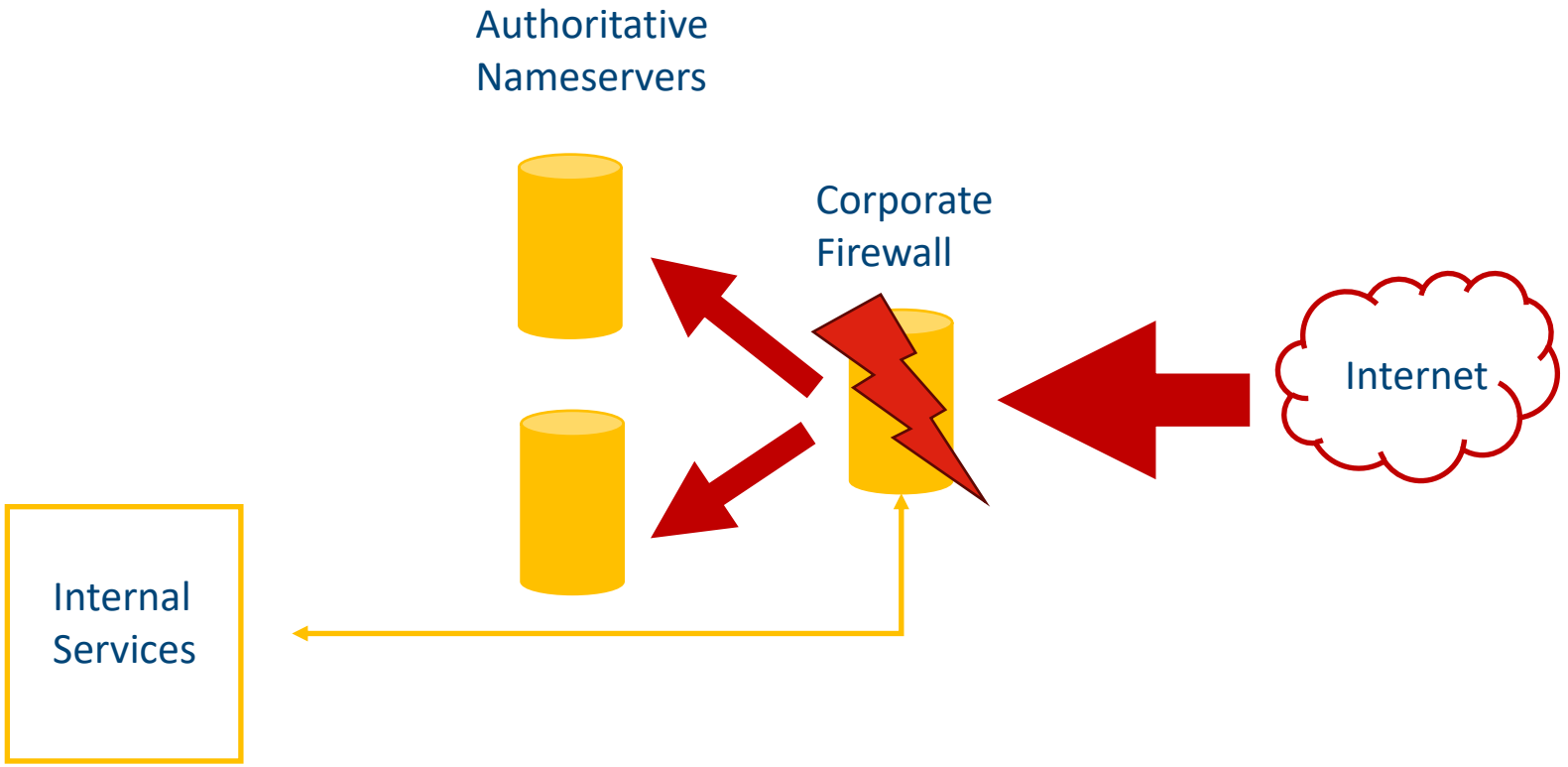
Architecture



Architecture



Architecture



Architecture

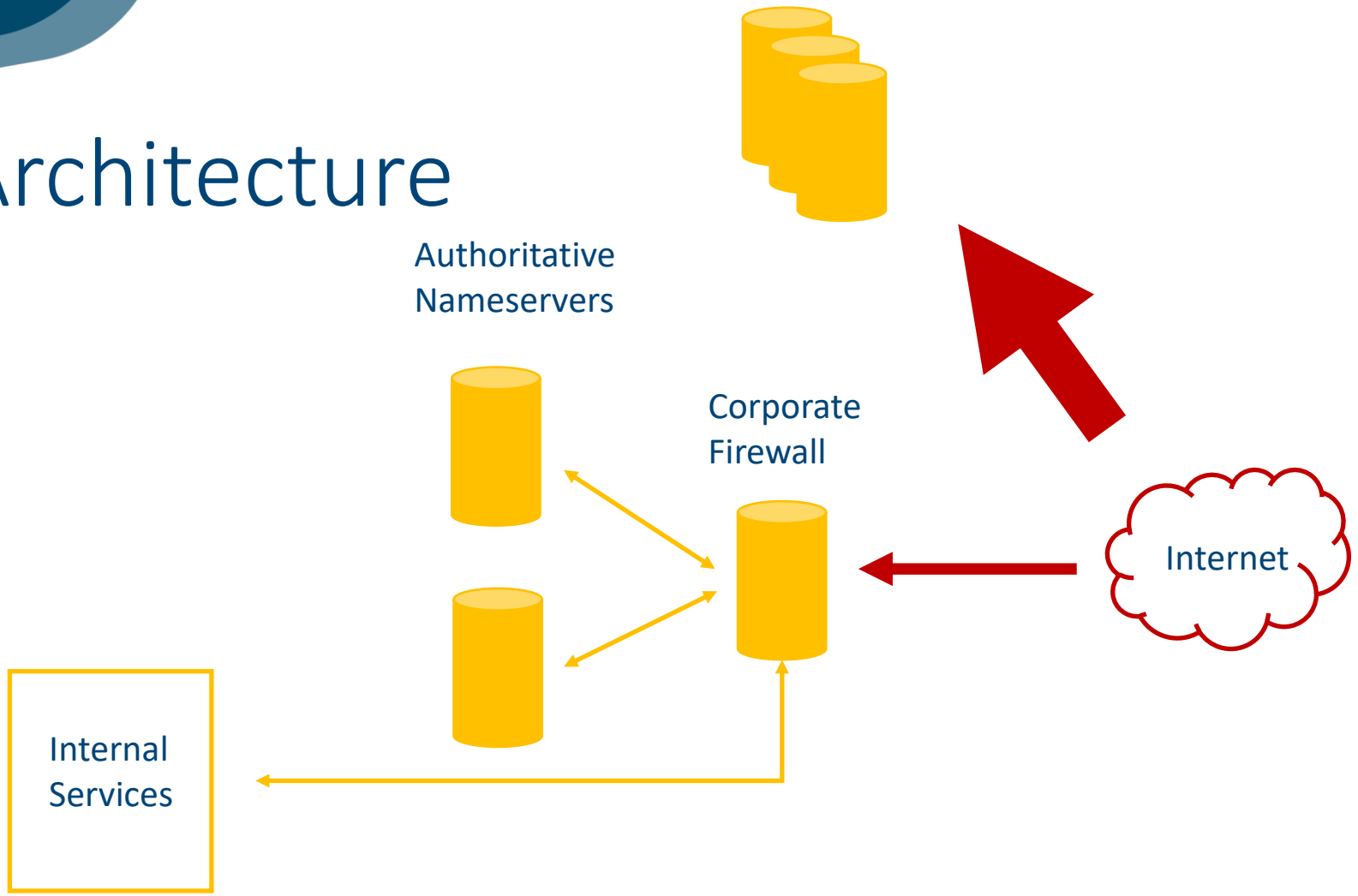
DO NOT

LOG OR INSPECT

DNS TRAFFIC

(PORT 53)

Architecture



Recommendation

- Stateless Firewall
- Zones should be always available from the inside
  - (Hidden Master, forward/slaving, splitDNS)
- Redundancy
  - with volumetric DDoS protection
  - flat rate for queries and traffic

RcodeZero DNS

christian.schoepp@nic.at