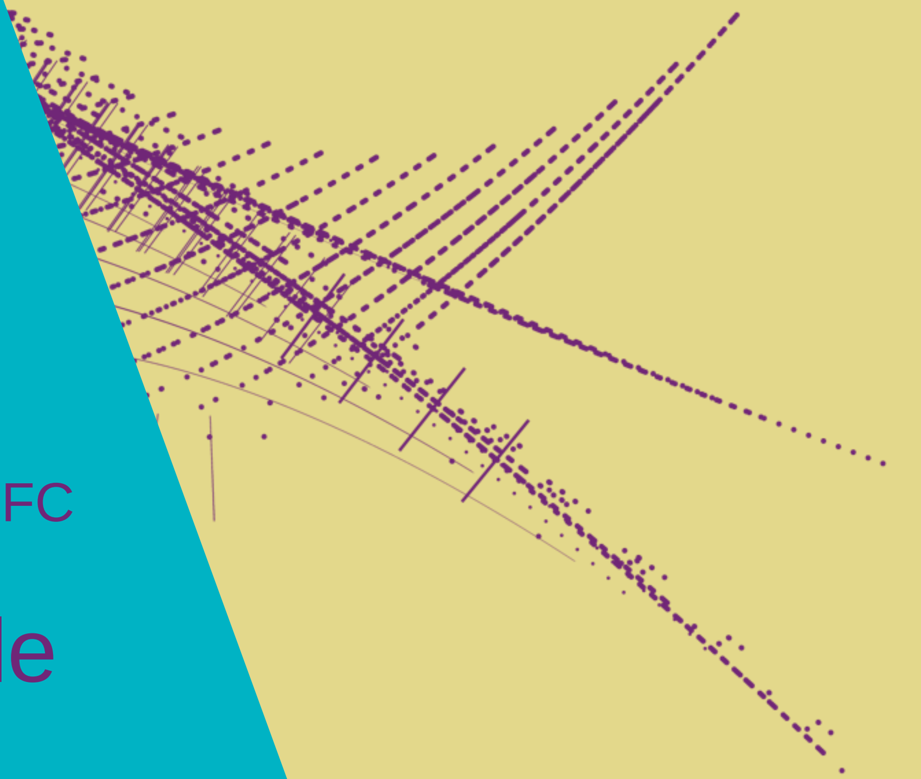
Nikhef

Maastricht University

RCauth.eu at Nikhef, GRNET, and STFC

# Building highly-available stateful services using IP anycast



David Groep

EUGridPMA56 meeting
October 2022

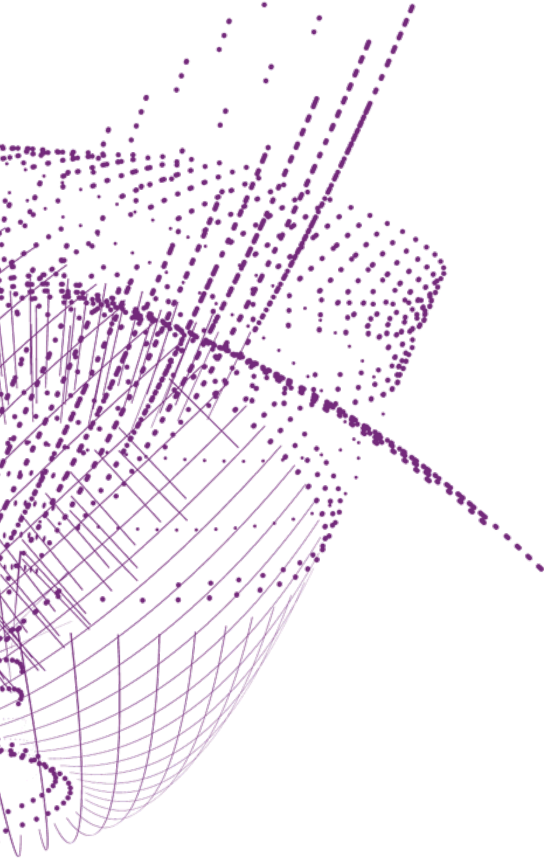
# Why go here?

Creating a PKIX credential translation service for the AARC BPA
with high-availability, scalability, and pan-European redundancy

**“provide a highly-available credential bridging services (RCauth.eu)”**

And at the same time

- demonstrate that also stateful services can be effectively anycasted
- find minimum viable anycast environment still having global properties
- provide a reference HA architecture for EOSC core services
- dispel arguments that building IP anycasted services is complex



# The RCauth.eu service

# AAI evolution of Research and e-Infrastructures

Most infrastructures move to community proxies

- Less credentials to manage, appearing ‘simpler’ to the user
- support both augmenting attributes as well as credential translation
- but non-web access remains challenging for ‘SAML’ federations

EEPKI + RFC3820 did solve both the CLI and delegation use case nicely!

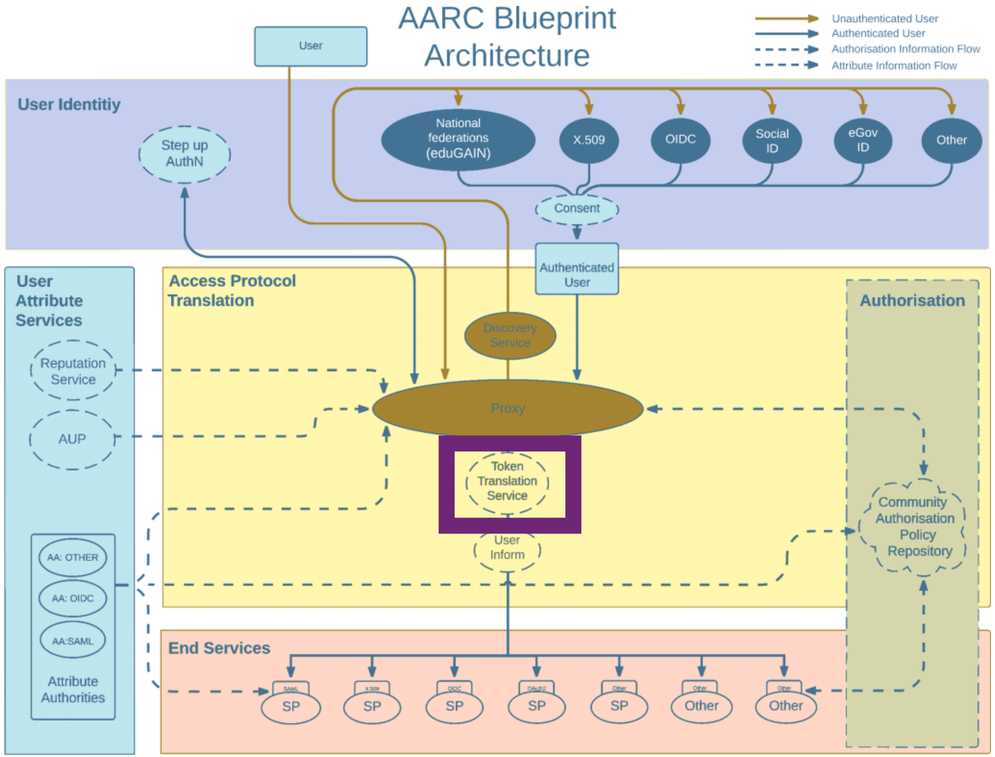
OIDC + OAuth2 can do the same, provided clients gets better, SAML ECP? no…

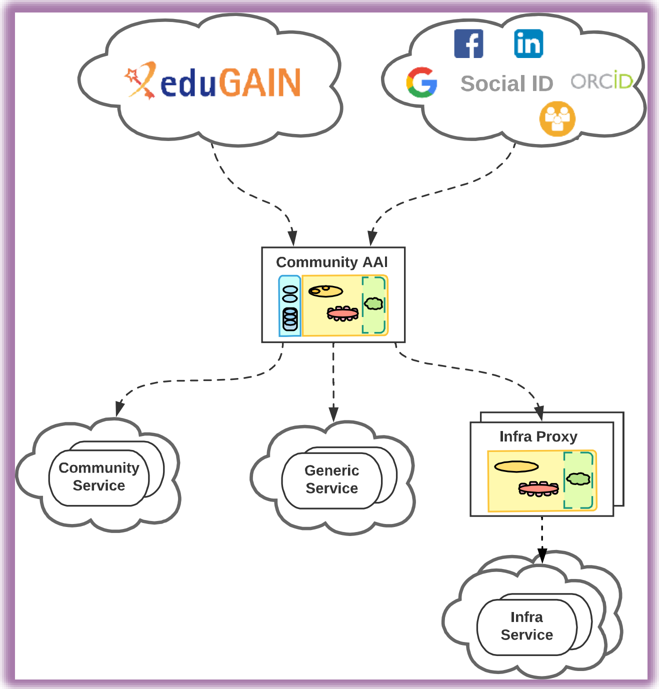
Bridging and translation

SAML->OIDC, SAML->X509, X509->OIDC, X509->SAML, OIDC->X509, …

- Does not require major technical changes in existing R&E federations
- Allows community-centric identifiers

# AARC Blueprint Architecture





AARC-G045 – https://aarc-community.org/guidelines/aarc-g045/

# A *-to-X509 Token Translations Service for Europe

Ability to serve a large pan-European user base without national restrictions

- without having to rely on specific national participation exclusively for this service
- serve needs of cross-national communities that have large but sparse user base

Use existing resources and e-Infrastructure services

- without the needs for security model changes at the resource centre or national level

Allow integration of this system in science gateways and portals with minimal effort

- only light-weight industry-standard protocols

Permit the use of the VOMS community membership service

- attributes for group and role management  in attribute certificates
- also for portals and science gateways access the e-Infrastructure

Concentrate service elements that require significant operational expertise

- not burden research communities with care for security-sensitive service components
- keep a secure credential management model
- coordinate compliance and accreditation

# RCauth.eu – a ubiquitous federated IOTA

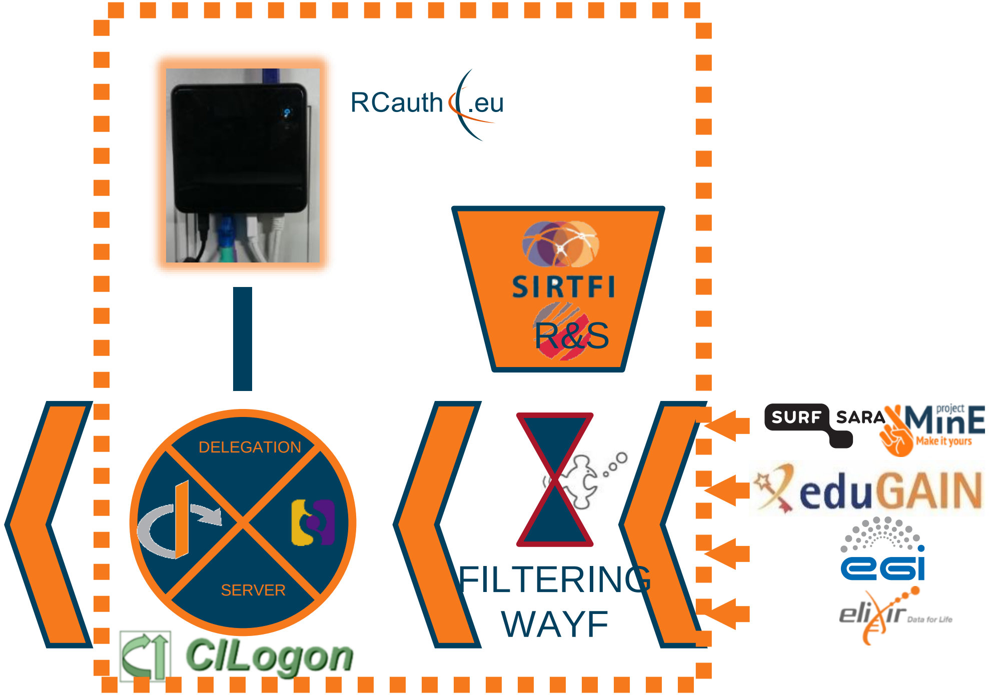
- RCauth is an IGTF accredited IOTA (DOGWOOD class) CA
  - Online credential conversion
  - Connected to eduGAIN (R&S+Sirtfi) plus direct, e.g. EGI Check-in and eduTEAMS

RCauth.eu

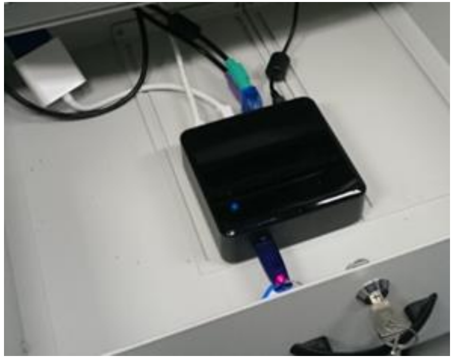
- Inspired by and leveraging the delegation service from CILogon

*For CILogon, Jim Basney et al, NCSA/UIUC for NFS – see https://cilogon.org/* CILogon Service

# Long ago, in a drawer, far far away

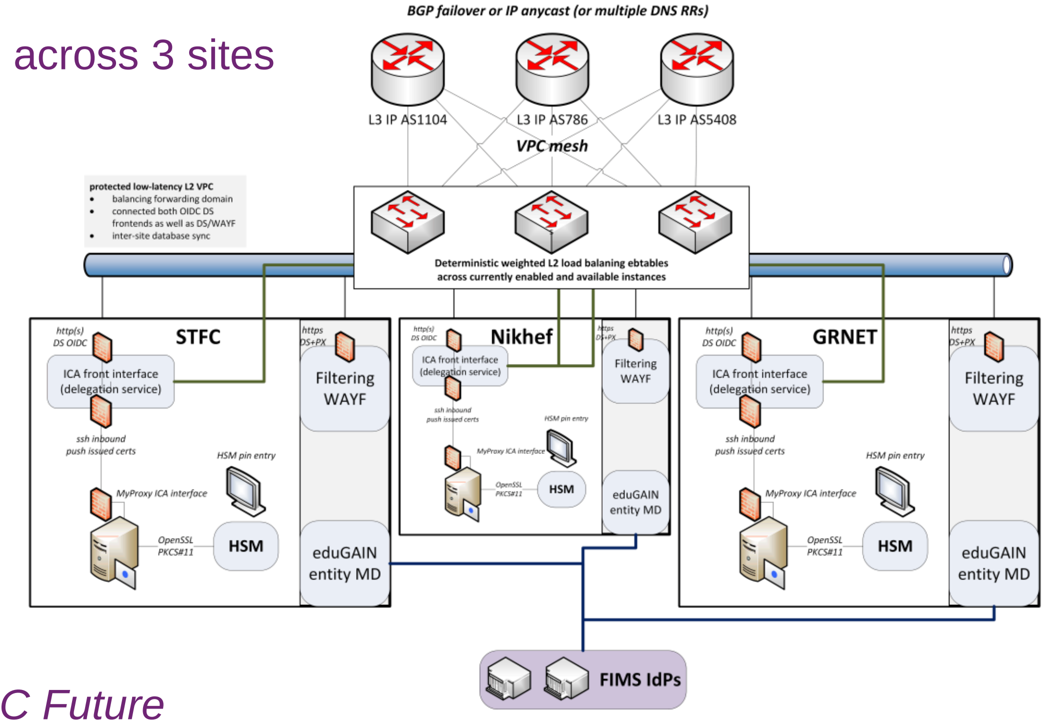


RCauth.eu



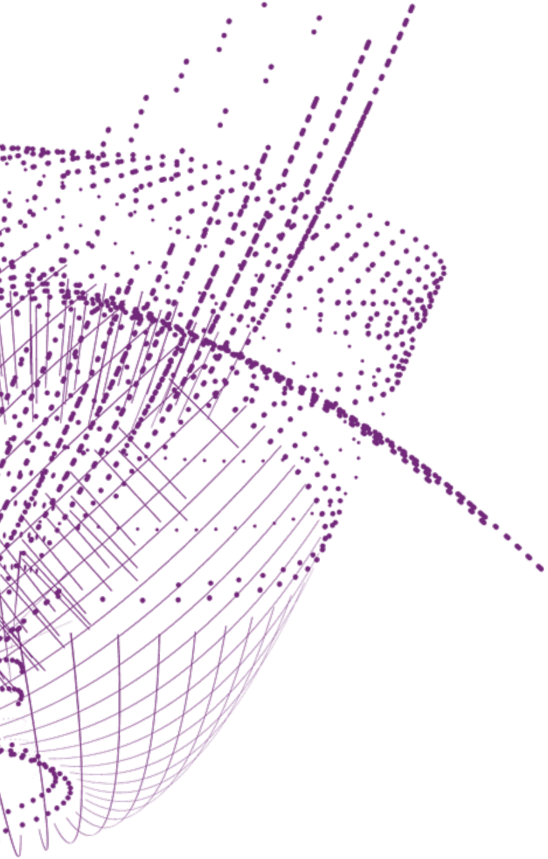
# Since we do not like SPOFs …

Implement a High Availability setup across 3 sites



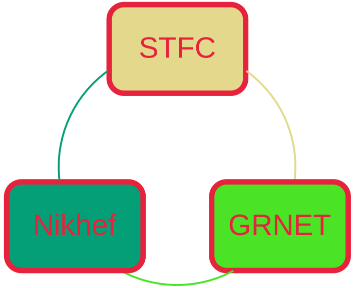
BGP failover or IP anycast (or multiple DNS RRs)
L3 IP AS1104
L3 IP AS786
L3 IP AS5408
VPC mesh
protected low-latency L2 VPC
balancing forwarding domain
connected both OIDC DS frontends as well as DS/WAYF
inter-site database sync
Deterministic weighted L2 load balaning ebtables across currently enabled and available instances
STFC
http(s) DS OIDC
ICA front interface (delegation service)
ssh inbound push issued certs
HSM pin entry
MyProxy ICA interface
OpenSSL PKCS#11
HSM
https DS+PX
Filtering WAYF
eduGAIN entity MD
Nikhef
http(s) DS OIDC
ICA front interface (delegation service)
ssh inbound push issued certs
HSM pin entry
MyProxy ICA interface
OpenSSL PKCS#11
HSM
https DS+PX
Filtering WAYF
eduGAIN entity MD
GRNET
http(s) DS OIDC
ICA front interface (delegation service)
ssh inbound push issued certs
HSM pin entry
MyProxy ICA interface
OpenSSL PKCS#11
HSM
https DS+PX
Filtering WAYF
eduGAIN entity MD
FIMS IdPs

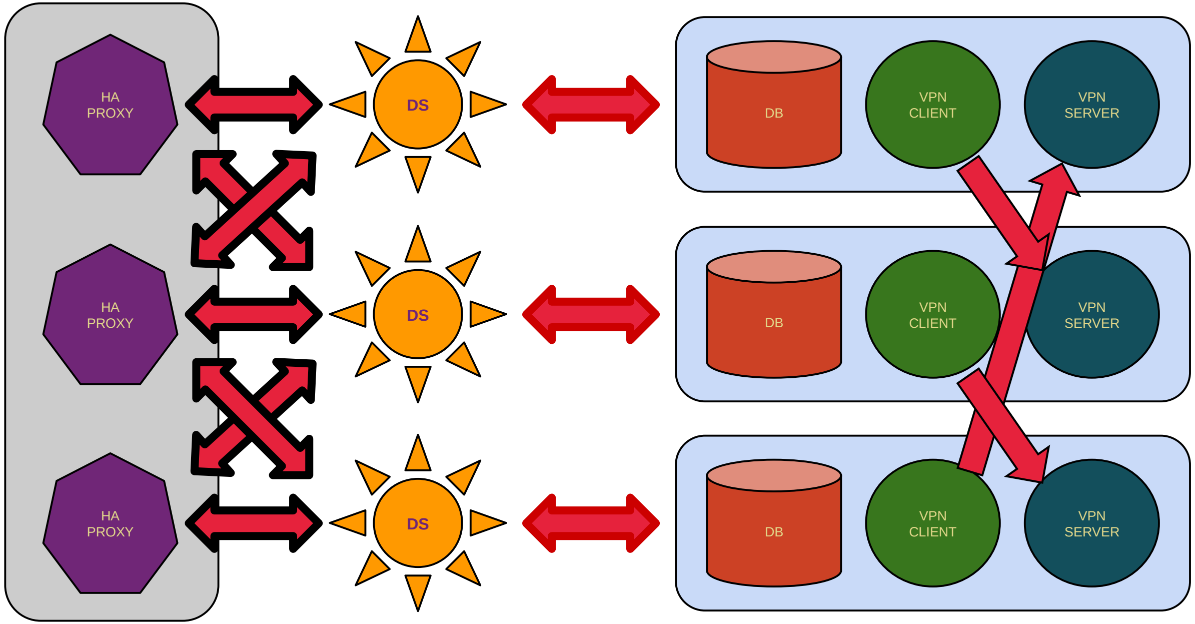
*Supported by EOSC Hub and EOSC Future*



# towards a pan-European distributed service

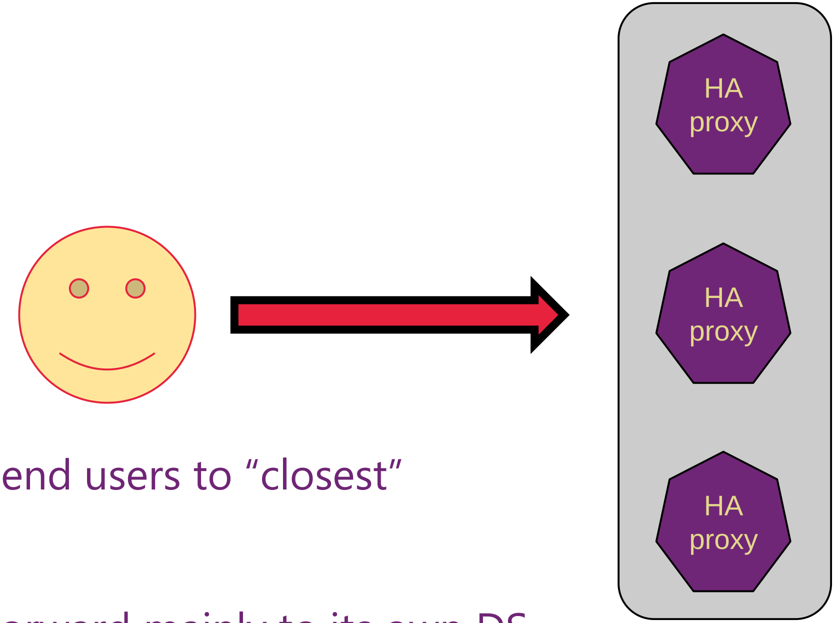
# Distributed RCauth service





selected imagery: Mischa Sallé, Jens Jensen, Nicolas Liampotis

# A transparent multi-site setup?



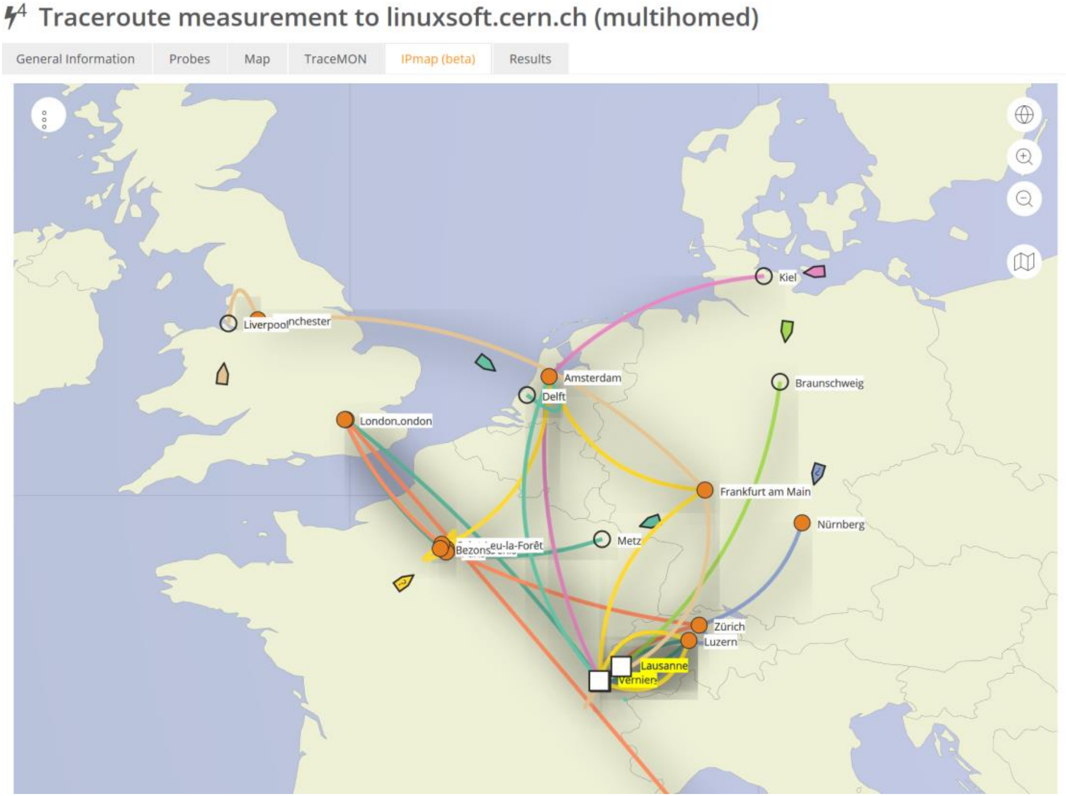
Need a way to send users to “closest” working service

Each HA proxy forward mainly to its own DS

If a HA loses its backend DS, it can still route to the other DS’es

selected imagery: Mischa Sallé, Jens Jensen, Nicolas Liampotis

# Intermezzo – BGP routing principles



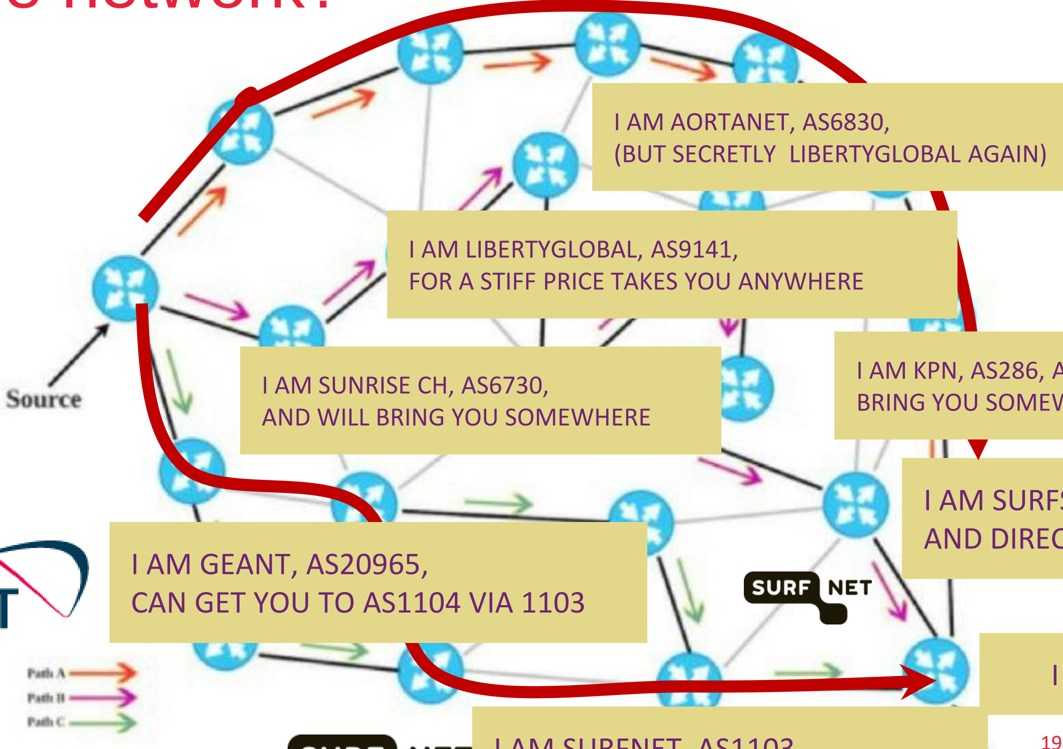
Traceroute measurement to linuxsoft.cern.ch (multihomed)

*Data: TraceMON IPmap from RIPE NCC Atlas atlas.ripe.net measurement 9249079*

# A labyrinthine network?

I AM CERN, AS513 …
AND WANT TO TALK TO
*E.G.* 194.171.96.128/25

188.184.38.9



I AM AORTANET, AS6830,
(BUT SECRETLY LIBERTYGLOBAL AGAIN)

I AM LIBERTYGLOBAL, AS9141,
FOR A STIFF PRICE TAKES YOU ANYWHERE

I AM SUNRISE CH, AS6730,
AND WILL BRING YOU SOMEWHERE

I AM KPN, AS286, AND WILL
BRING YOU SOMEWHERE NEAR

I AM SURFSARA, AS1162,
AND DIRECTLY TALK TO AS1104!

I AM GEANT, AS20965,
CAN GET YOU TO AS1104 VIA 1103

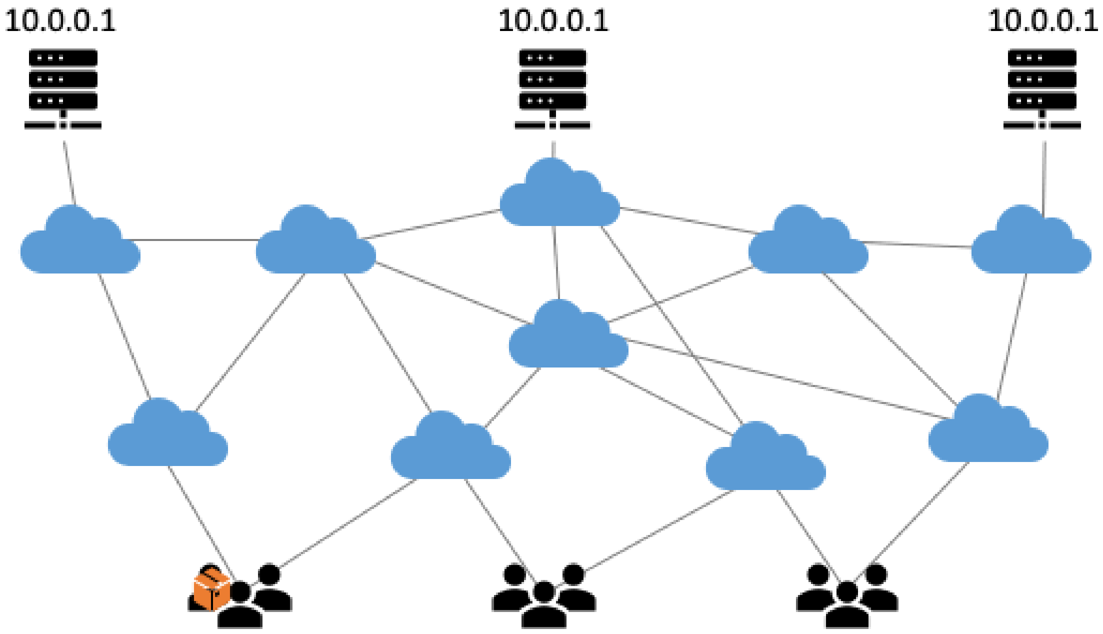
I AM NIKHEF, AS1104!

194.171.96.128/25

I AM SURFNET, AS1103,
AND CAN BRING YOU TO AS1104

“SEGMENT ROUTING”
IMAGE BY DAVID PENALOZA, CISCO

# Anycast: when the same place exists many times



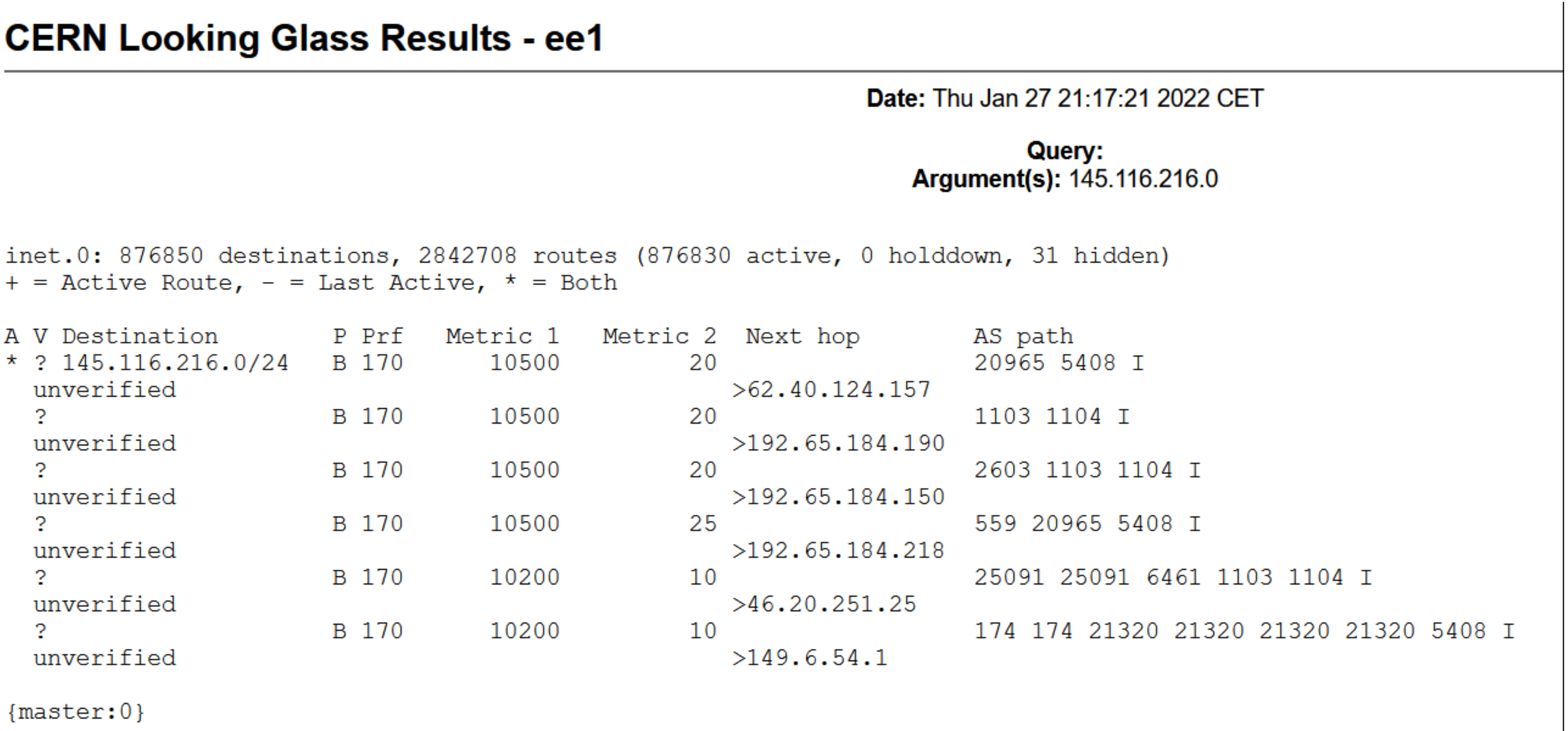
## So we used

- 3 (now: 2) sites
- one VM at each site exposing 145.116.216.1
- smallest v4 subnet (/24)
- bird + a service probe
- each site’s own ASN
- some IRR DB editing
- v6 is similar, with a /48

*and some monitoring*

routing image: SIDNlabs - https://www.sidnlabs.nl/en/news-and-blogs/the-bgp-tuner-intuitive-management-applied-to-dns-anycast-infrastructure

# Same address, two paths



## CERN Looking Glass Results - ee1

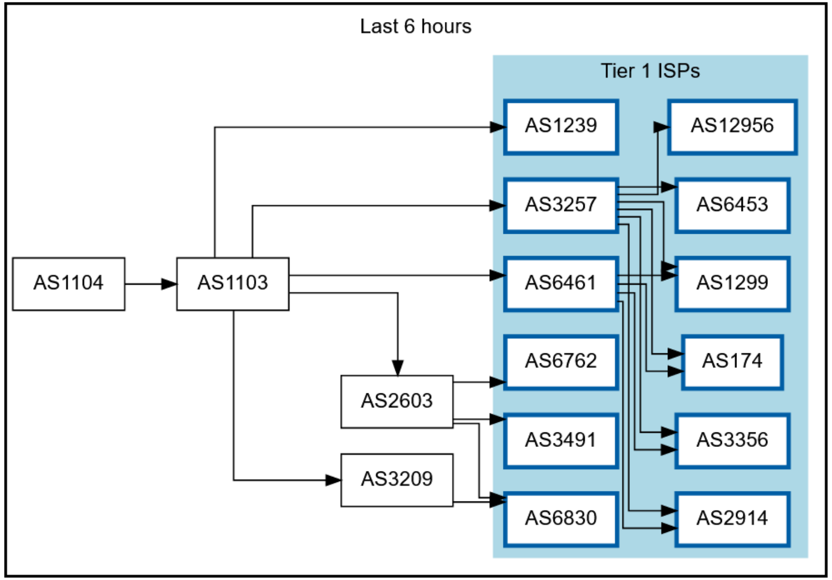
**Date:** Thu Jan 27 21:17:21 2022 CET

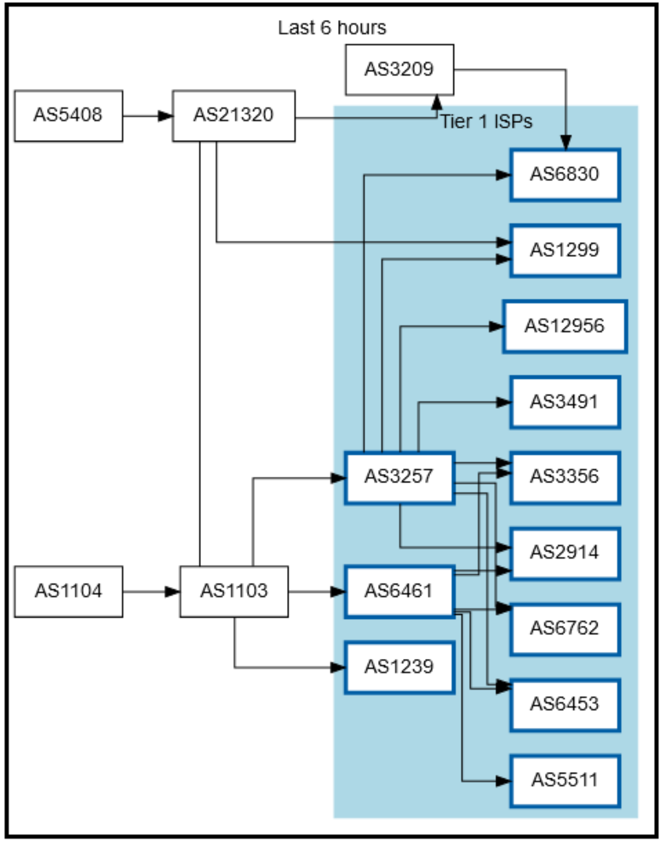
**Query:**
**Argument(s):** 145.116.216.0

```
inet.0: 876850 destinations, 2842708 routes (876830 active, 0 holddown, 31 hidden)
+ = Active Route, - = Last Active, * = Both

A V Destination        P Prf   Metric 1   Metric 2  Next hop        AS path
* ? 145.116.216.0/24   B 170      10500         20                  20965 5408 I
  unverified                                       >62.40.124.157
  ?                    B 170      10500         20                  1103 1104 I
  unverified                                       >192.65.184.190
  ?                    B 170      10500         20                  2603 1103 1104 I
  unverified                                       >192.65.184.150
  ?                    B 170      10500         25                  559 20965 5408 I
  unverified                                       >192.65.184.218
  ?                    B 170      10200         10                  25091 25091 6461 1103 1104 I
  unverified                                       >46.20.251.25
  ?                    B 170      10200         10                  174 174 21320 21320 21320 21320 5408 I
  unverified                                       >149.6.54.1

{master:0}
```

assigned 2a07:8504:1a0::/48 and 145.116.216.0/24 to RCauth anycast

# Getting 2a07:8504:1a0::/48 and 145.116.216.0/24 out there

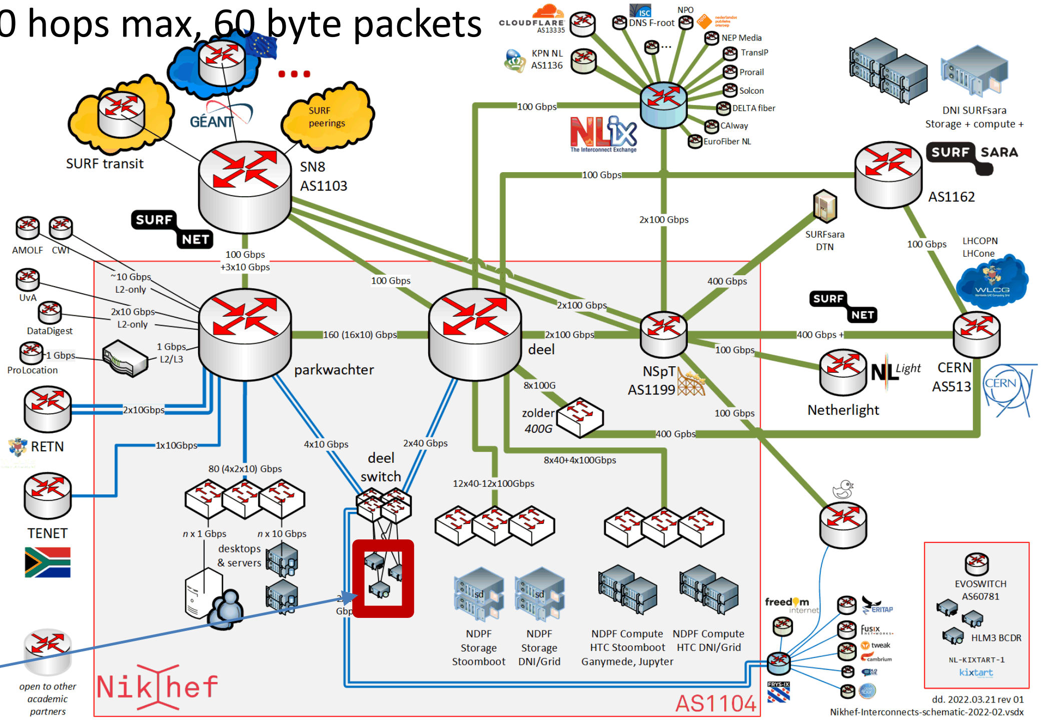




route maps: bgp.tools for 145.116.216.0/24 – IPv6 would be similar

# Shortest path, also when mixing with the default-free zone

```
[root@kwark ~]# traceroute -IA 145.116.216.1
traceroute to 145.116.216.1 (145.116.216.1), 30 hops max, 60 byte packets
 1  cmbr.connected.by.freedominter.net
    (185.93.175.234) [AS206238]
 2  connected.by.freedom.nl
    (185.93.175.240) [AS206238]
 3  et-0-0-0-1002.core1.fi001.nl.freedomnet.nl
    (185.93.175.208) [AS206238]
 4  as1104.frys-ix.net (185.1.203.66) [*]
 5  parkwachter.nikhef.nl
    (192.16.186.141) [AS1104]
 6  gw-anyc-01.rcauth.eu
    (145.116.216.1) [AS786/AS5408/AS1104]
```



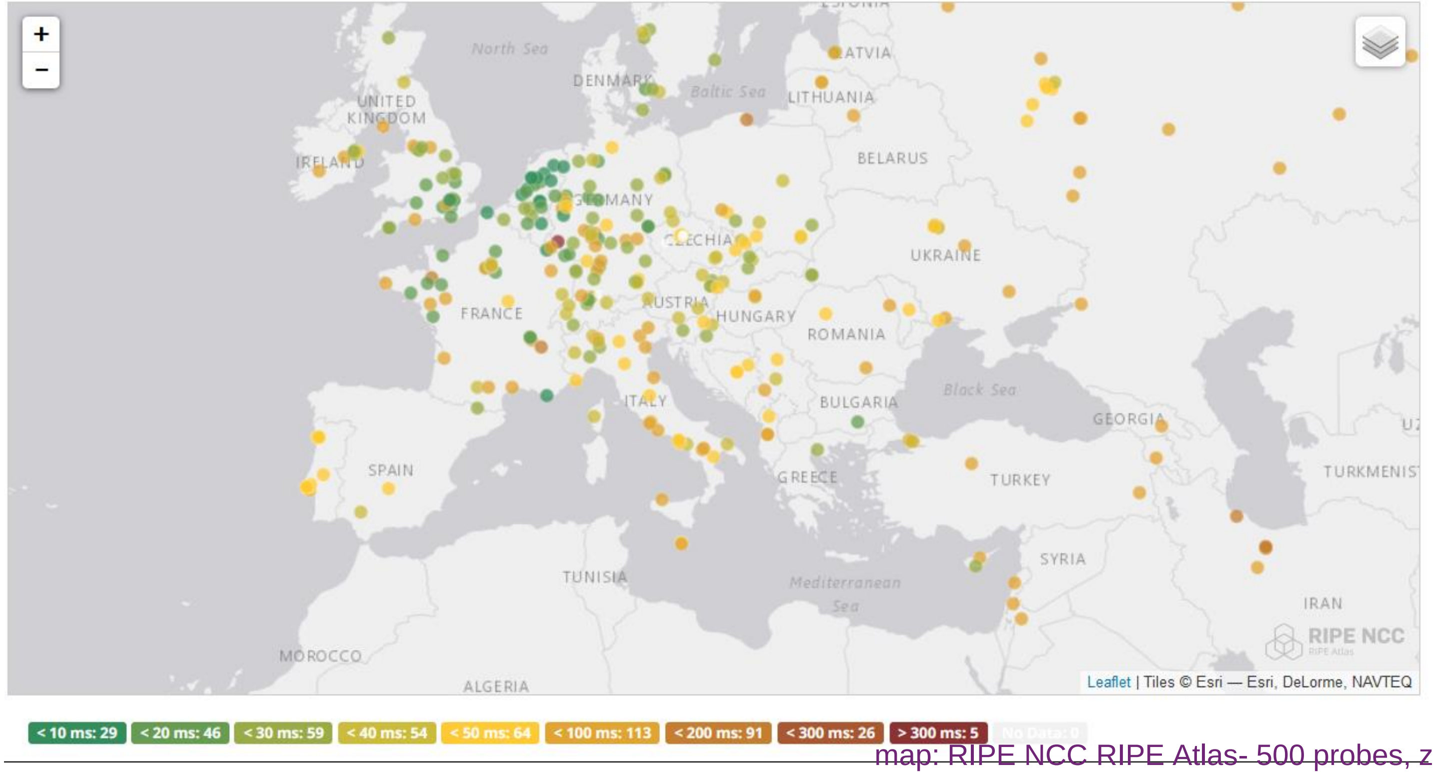
*rcauth.eu HA proxy*

# Prerequisites are relatively simple

- an IPv4 /24 netblock (and, or) an IPv6 /48
- your own, or a friendly AS
- a set of IRR route objects, and either none, or a correct RPKI VRP
  (easily done in your local RIR registry: APNIC, RIPE, ARIN, AfriNIC, LACNIC)
- bird, or quagga, with a monitoring plugin (to flap the route in case of downtime)

- But you don’t per-se need:
- a unique AS just for this anycast activity (it works equally well without it)
- a balanced AS path length (unless you want load balancing as well as redundancy)
- your own AS (if you have a friendly AS willing to re-announce your specific route)

# And you get reasonable load balancing



map: RIPE NCC RIPE Atlas- 500 probes, zoomed in on Europe

# Other HA options

- Local HA with an HA proxy and pacemaker/CRM failover works on the local network – and can be meshed with two signing systems
this is the local Nikhef RCauth instance setup

- DNS-based fast-failover – the method used for InAcademia
automatic updating of DNS a distributed set of servers, auto-updating each other
But does require that the DNS domain level operator remains available, since you need *very* short TTLs (and of course your ccTLD/gTLD as well)

- Add a dedicated HA link for the back-end databases
e.g. multiple redundant circuits over an MPLS cloud

# The hard challenge: when *is* a service actually ‘up’?

STFC has a delegation service and issuance system, but no filtering wayf so its traffic is sent through Nikhef. But it *is* part of the gallera cluster

Nikhef has an internally-redundant DS+issuance system (4 boxes), and if either of these is down, the ‘service’ at Nikhef is still ‘up’

With a gallera cluster with 3 nodes, when the links are severed, on reconnection it cannot form a majority quota unless all come up at the same time. A tie-breaker would be needed, but where?

And: now operational monitoring and SLA monitoring are different …

This work has also been co-supported by projects that have received funding from the European Union's Horizon research and innovation programmes under Grant Agreement No. 856726 (GN4-3), 101017536 (EOSC Future), 777536 (EOSC-hub), 730941 (AARC2).

# Still here? Thanks!

*In collaboration with Mischa Sallé and Tristan Suerink (Nikhef), Nicolas Liampotis and Kyriakos Gkinis (GRNET), and Jens Jensen (STFC RAL)*

David Groep
davidg@nikhef.nl
https://www.nikhef.nl/~davidg/presentations/
https://orcid.org/0000-0003-1026-6606

cc BY