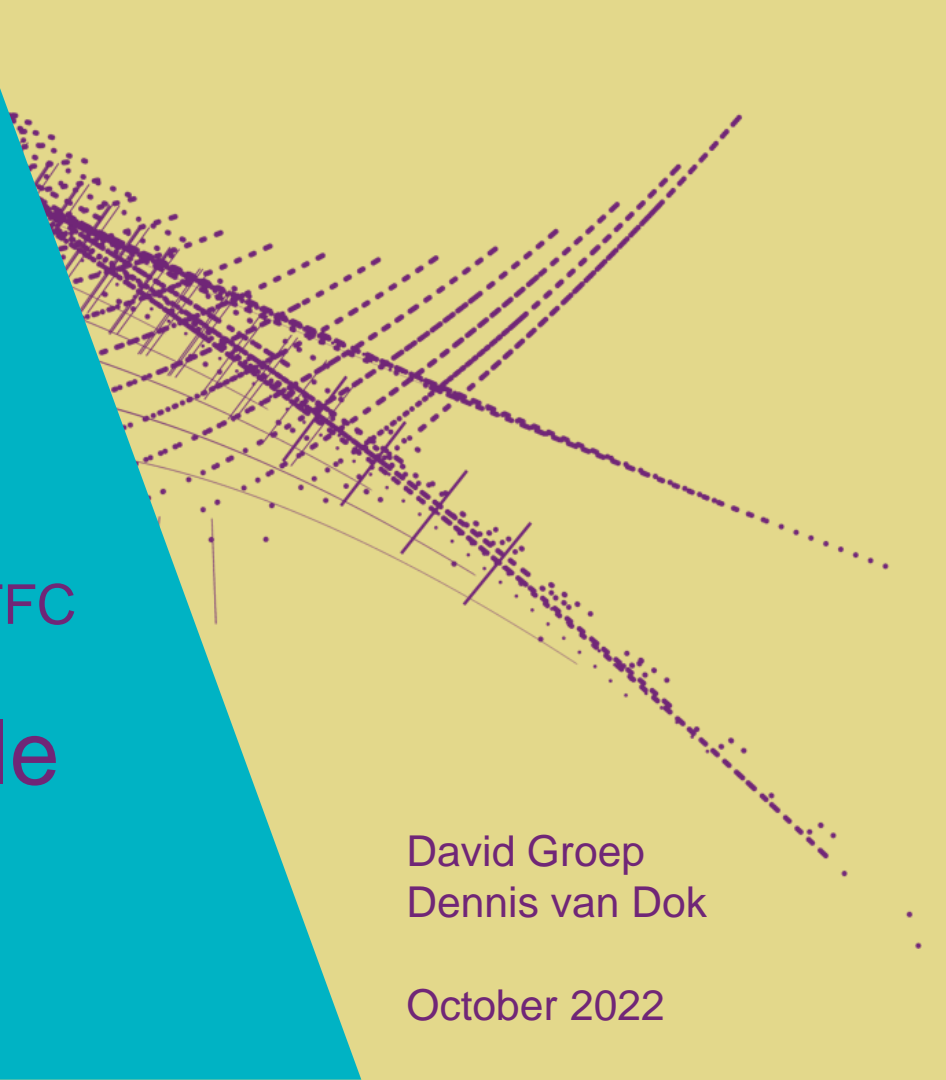# Nikhef

Maastricht University

RCauth.eu at Nikhef, GRNET, and STFC

Building highly-available stateful services using IP anycast

David Groep
Dennis van Dok

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
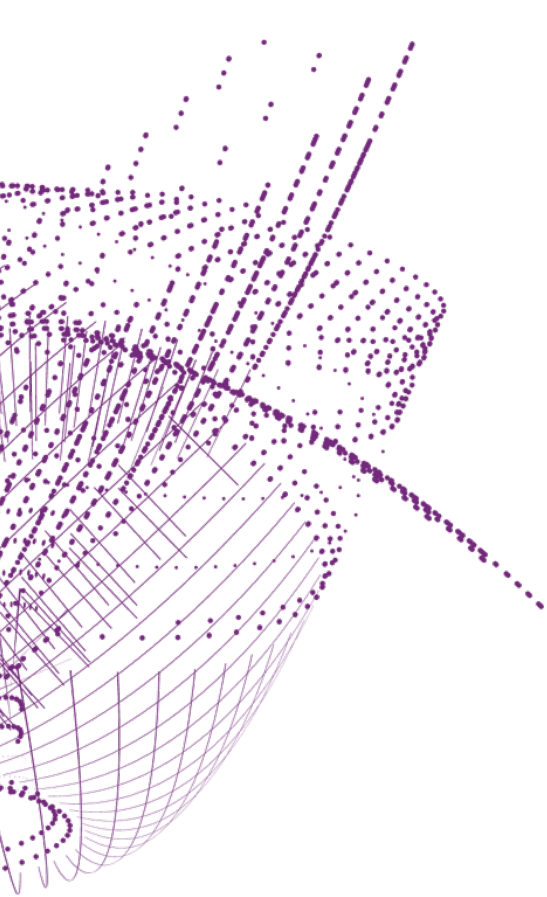
The RCauth.eu service

# AAI evolution of Research and e-Infrastructures

Most infrastructures move to community proxies, with *credential translation*

- less credentials to manage, appearing 'simpler' to the user
- support both augmenting attributes as well as credential translation

but since non-web access remains challenging for 'SAML' federations

- 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/

Buiding stateful HA services using IP anycast for RCauth.eu
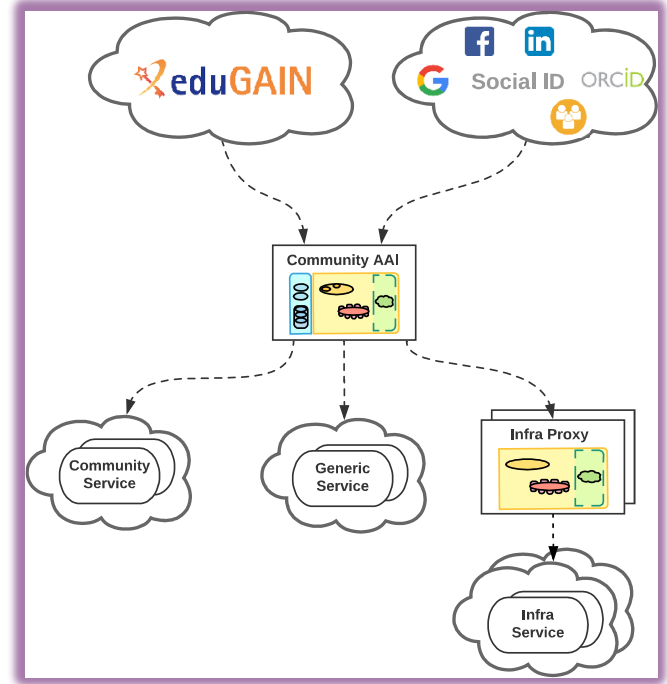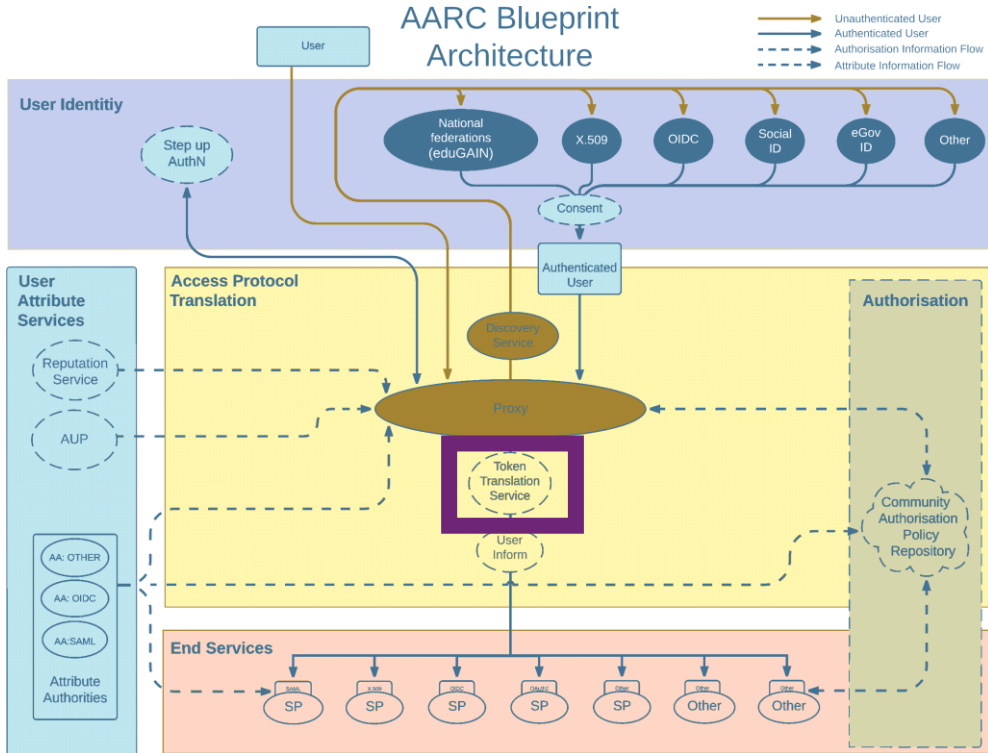
One element: RCauth.eu – a ubiquitous federated IOTA CA

- Credential translation from (SAML) federation to X.509

- IGTF accredited IOTA (DOGWOOD class) CA
  - online credential conversion based on basic assurance
  - connected to eduGAIN (R&S+Sirtfi) plus direct,
    e.g. EGI Check-in and eduTEAMS
  - support credential management systems for portals (MasterPortal)

- 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**

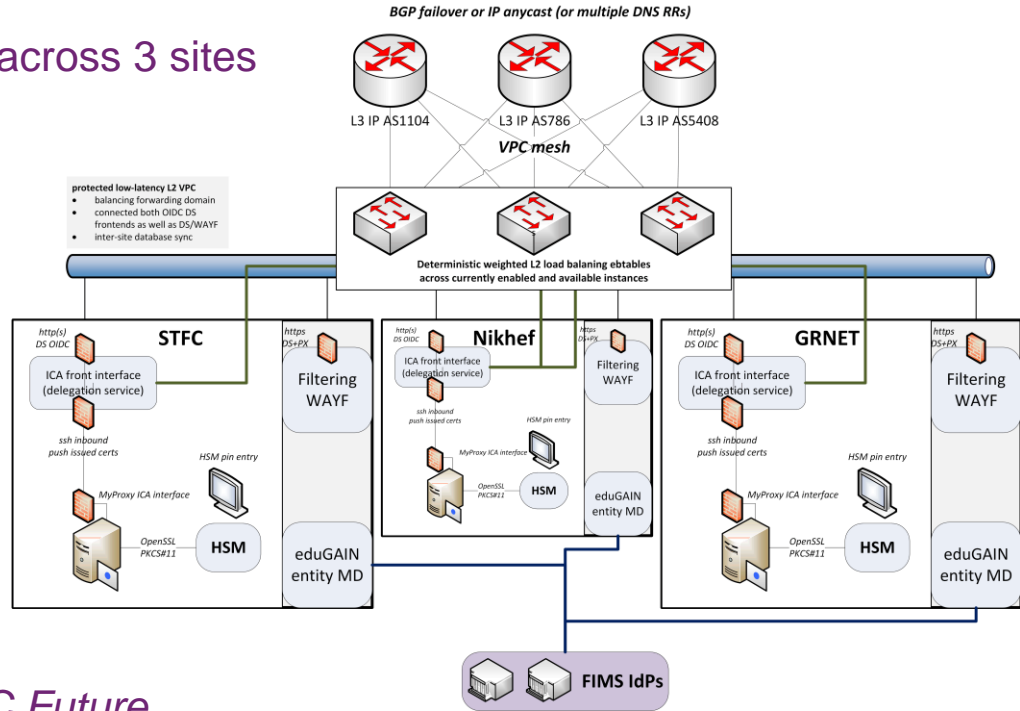**RCauth** **.eu**

Nik|hef

Building the first RCauth – at a single site
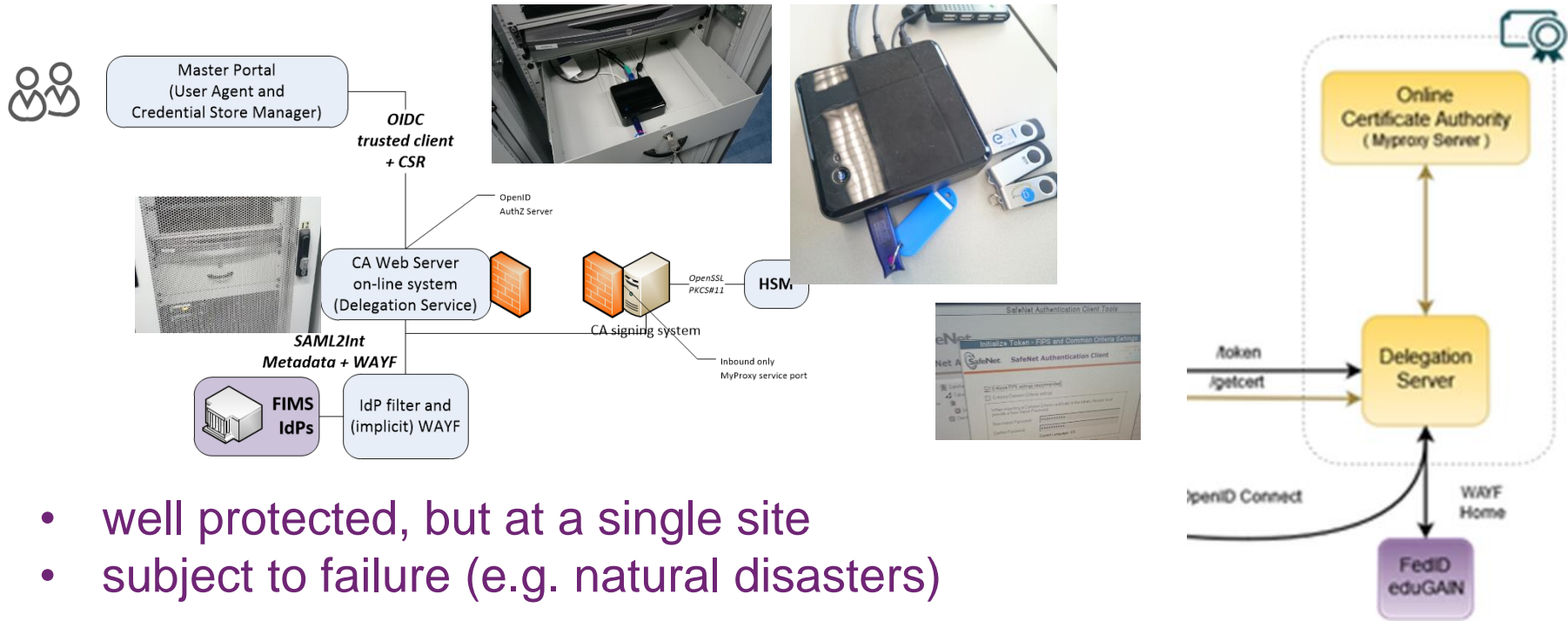
- well protected, but at a single site
- subject to failure (e.g. natural disasters)

Since we do not like SPOFs …

Implement a High Availability setup across 3 sites

Supported by EOSC Hub and EOSC Future

Buiding stateful HA services using IP anycast for RCauth.eu
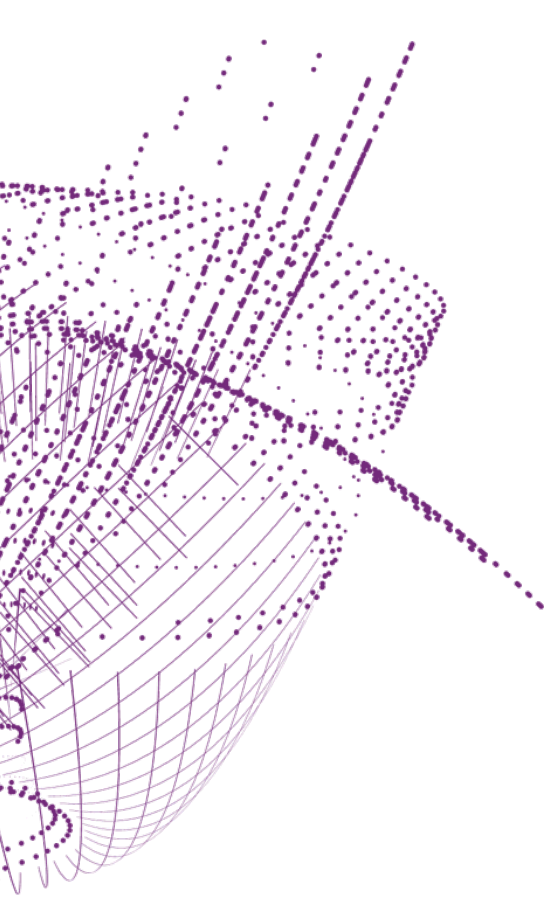
towards a pan-European distributed service

Distributed RCauth service

selected imagery: Mischa Sallé, Jens Jensen, Nicolas Liampotis

Buiding stateful HA services using IP anycast for RCauth.eu
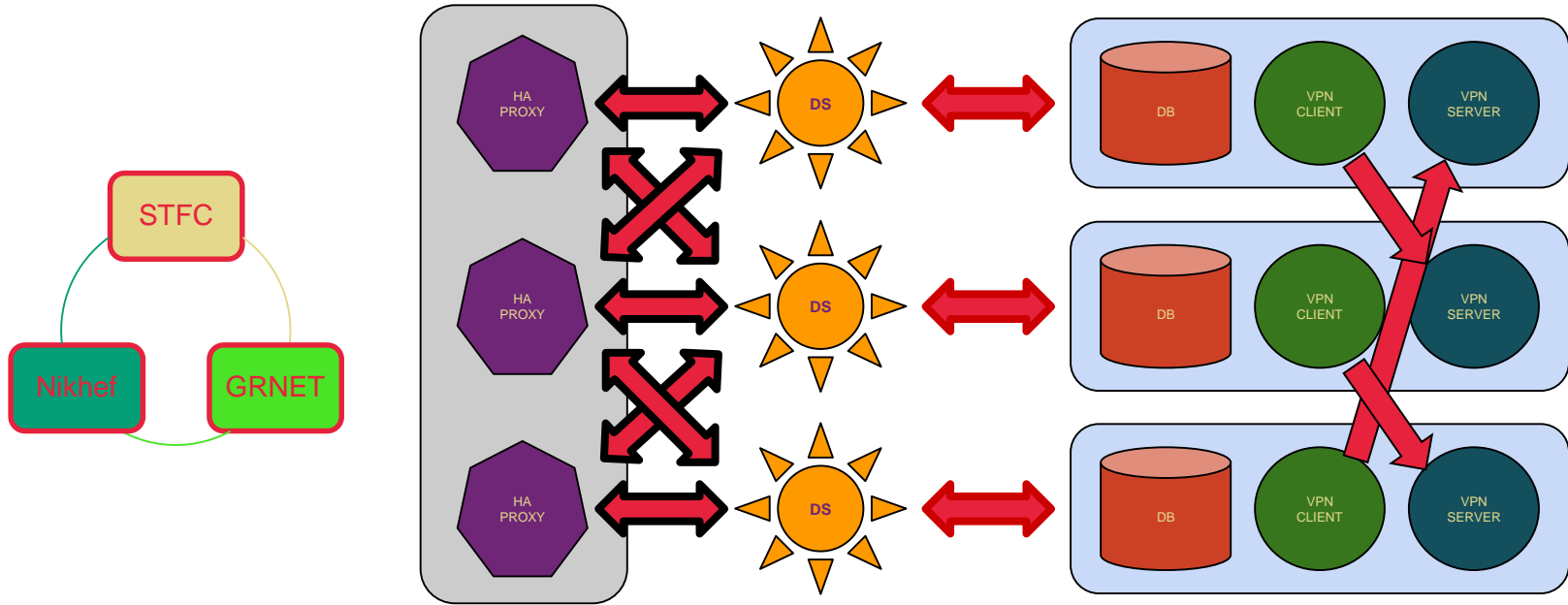
A transparent multi-site setup?
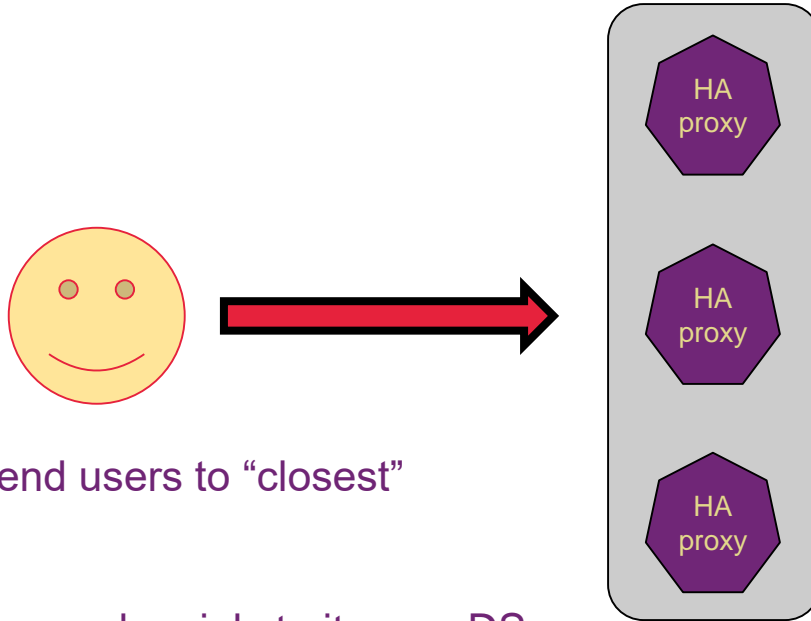
Need a way to send users to "closest" working service

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

Buiding stateful HA services using IP anycast for RCauth.eu

Intermezzo – BGP routing principles

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

Buiding stateful HA services using IP anycast for RCauth.eu

A labyrinthine network?

I AM CERN, AS513 … and want to talk to e.g. 145.116.216.1

`188.184.38.9` Source

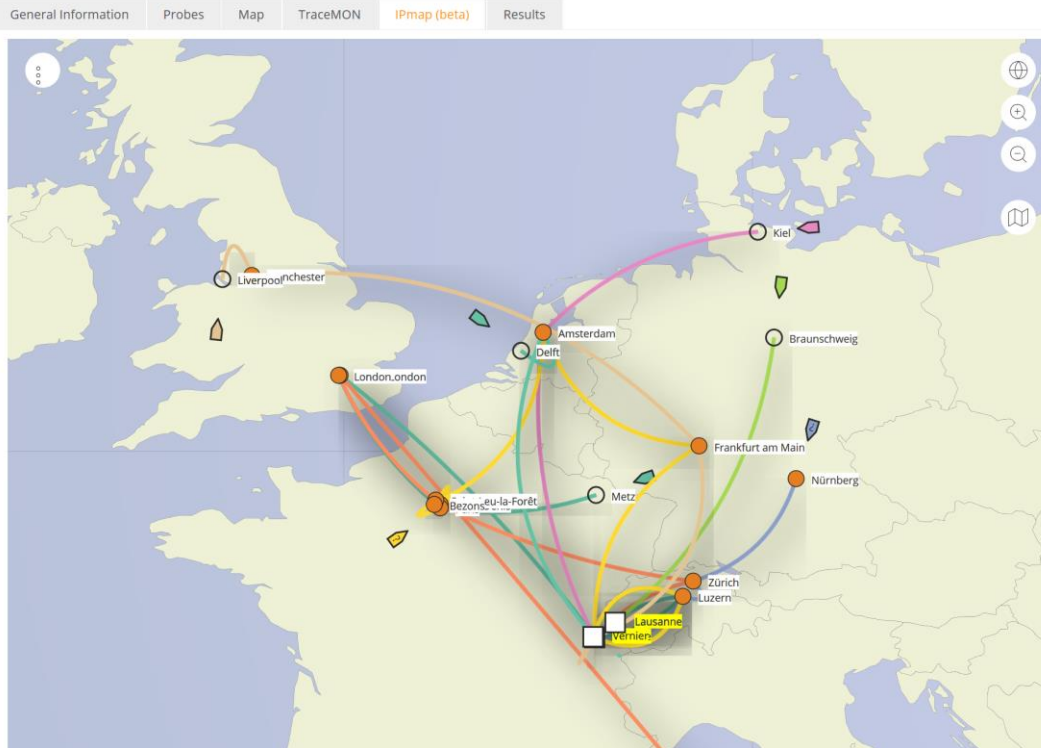*I am IP Max, AS51530, and like to take Swiss things*

I am LibertyGlobal, AS9141, and for a stiff price will take you anywhere

I AM Sunrise CH, AS6730, and will bring you somewhere – I hope

I am KPN, AS286, and will bring your somewhere near

I am SURF Amsterdam, AS1162, and I can talk directly AS1104!

I'm GEANT, AS20965, and I can get you to AS1104, but via AS1103

I AM Nikhef, AS1104!

I am SURFnet, AS1103, and can bring you to AS1104

Destination
`145.116.216.0/24`

Path A →
Path B →
Path C →

"segment routing" background image by David Penaloza, Cisco; content: only artificial routing example shown

Anycast: when the same place exists many times

10.0.0.1   10.0.0.1   10.0.0.1
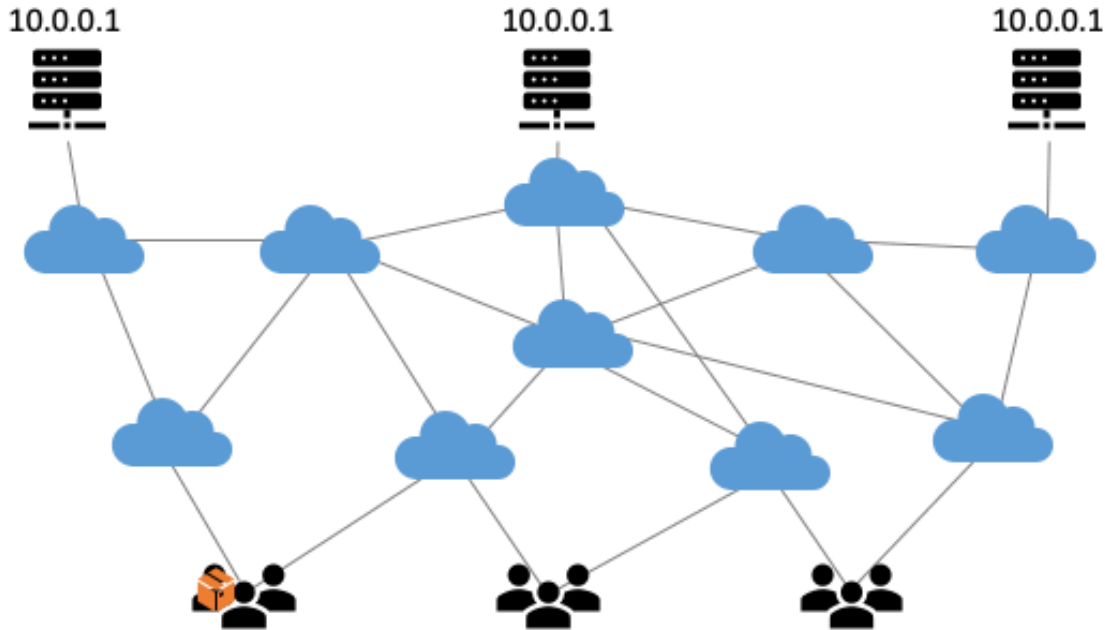
So we used

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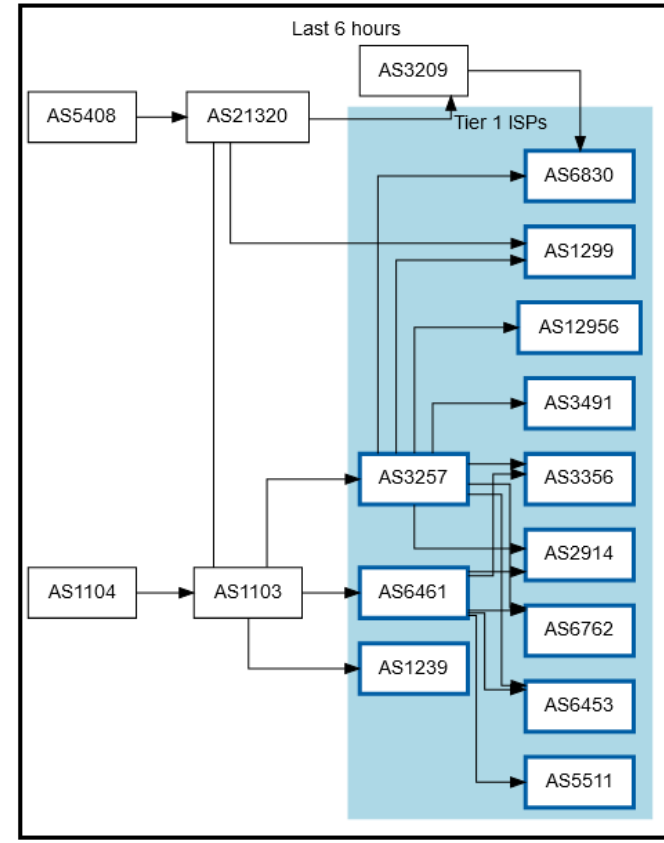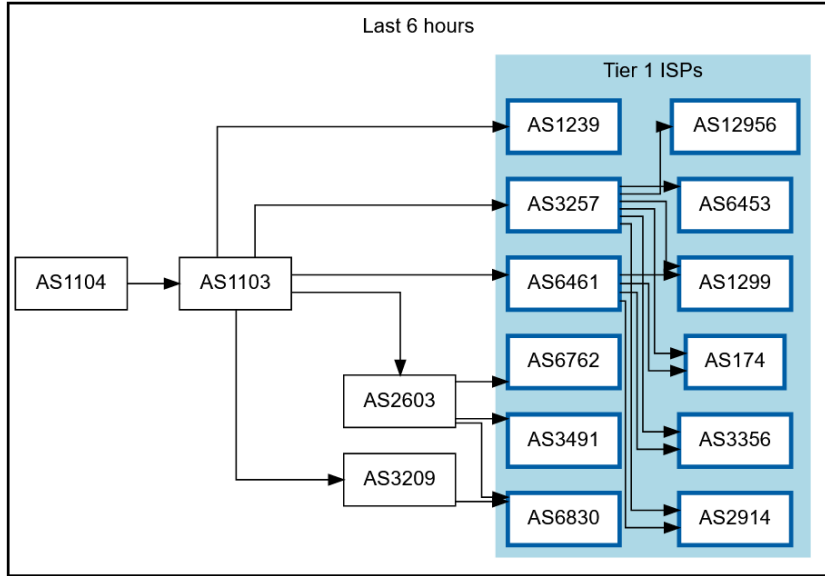
assigned 2a07:8504:1a0::/48 and 145.116.216.0/24 to RCauth anycast

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

Buiding stateful HA services using IP anycast for RCauth.eu

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

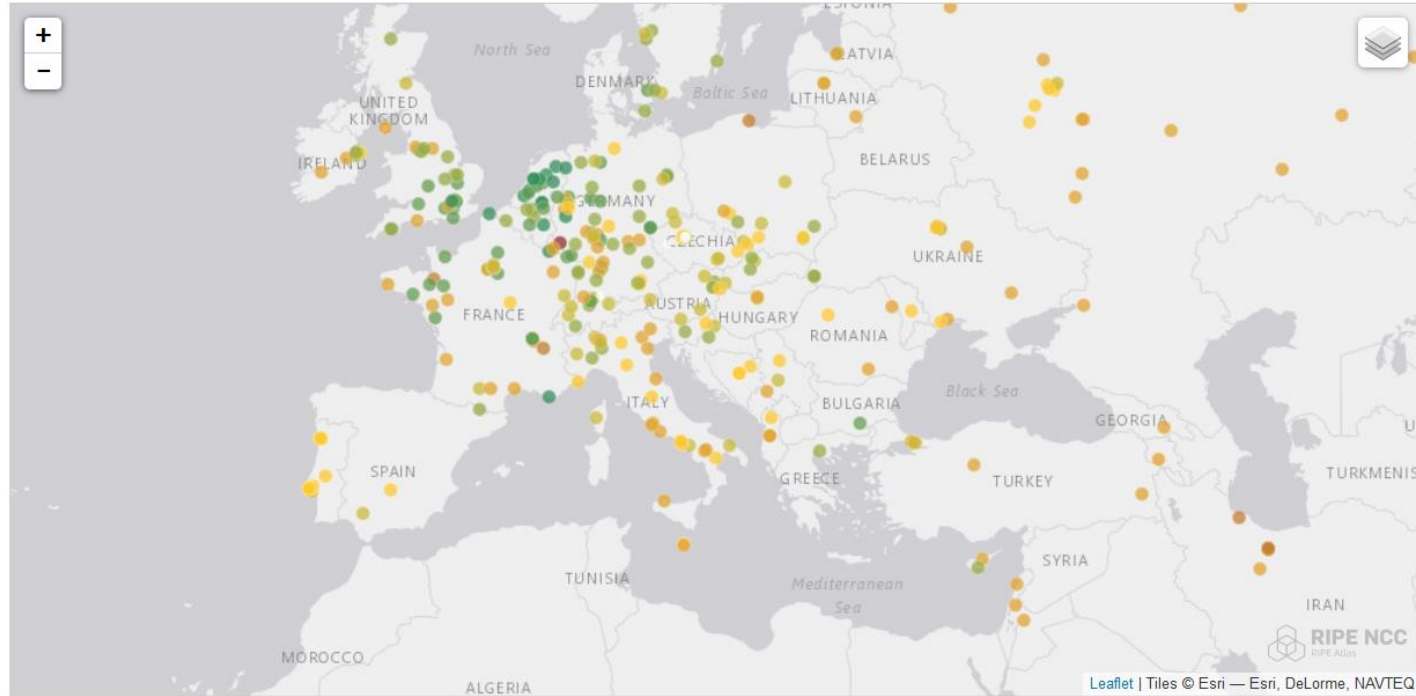Buiding stateful HA services using IP anycast for RCauth.eu
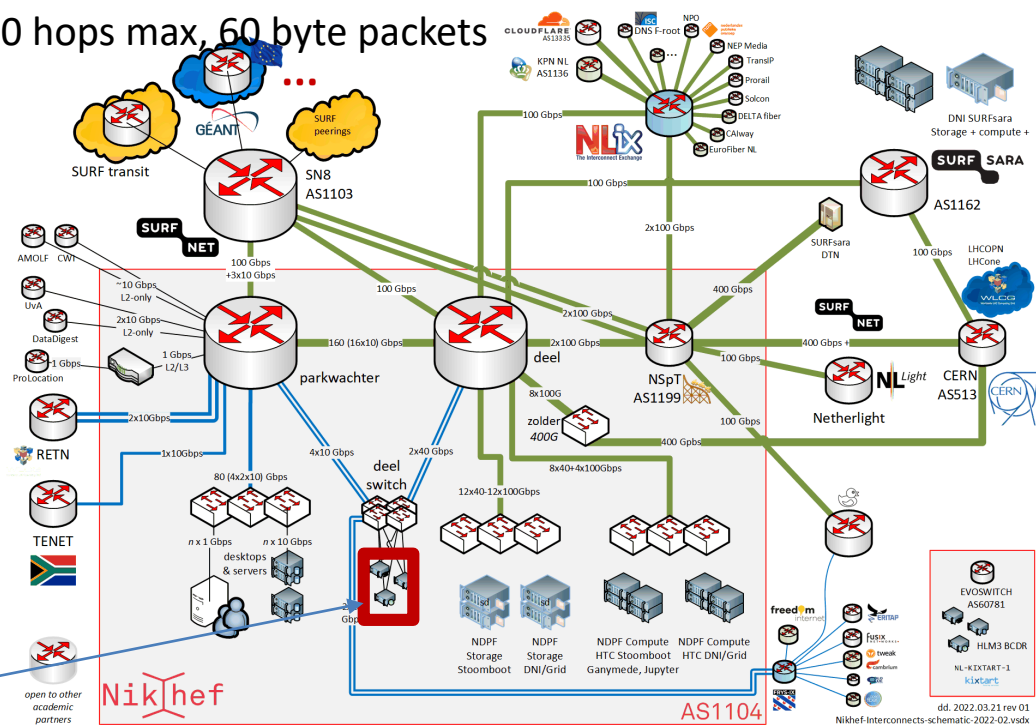
Prerequisites are relatively simple

- an IPv6 /48 block (and a legacy IPv4 /24 netblock)
- your own, or a friendly, Autonomous System Number (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)

Buiding stateful HA services using IP anycast for RCauth.eu

And you get reasonable load balancing

<10 ms: 29   <20 ms: 46   <30 ms: 59   <40 ms: 54   <50 ms: 64   <100 ms: 113   <200 ms: 91   <300 ms: 26   >300 ms: 5   No Data: 0

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

Buiding stateful HA services using IP anycast for RCauth.eu

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

Buiding stateful HA services using IP anycast for RCauth.eu

Nikhef

The hard challenge: when *is* a service actually 'up'?

STFC has a delegation service and issuance system,
but for now 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 galera 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 …

Buiding stateful HA services using IP anycast for RCauth.eu

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

Maastricht University

Nik|hef