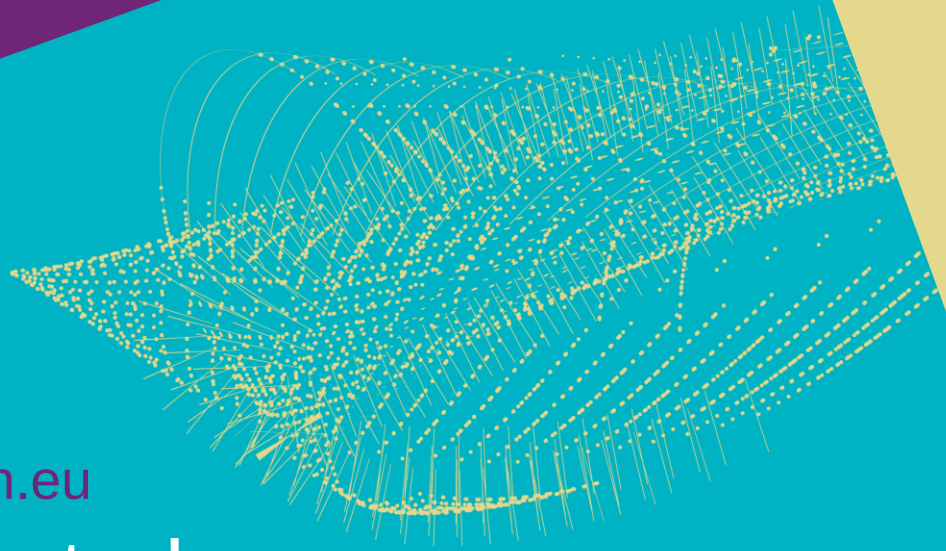
Nikhef

Maastricht University



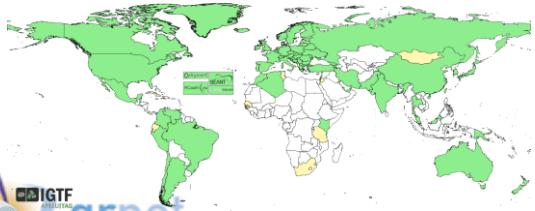
a multi-domain anycasted
high availability solution
for stateful services in RCauth.eu

Building the anycasted RCauth Federated authentication proxy

David Groep
Nikhef PDP programme
UM Dept. of Advanced Computing Sciences

ISGC Taipei, March 2023

We live in a federated world!



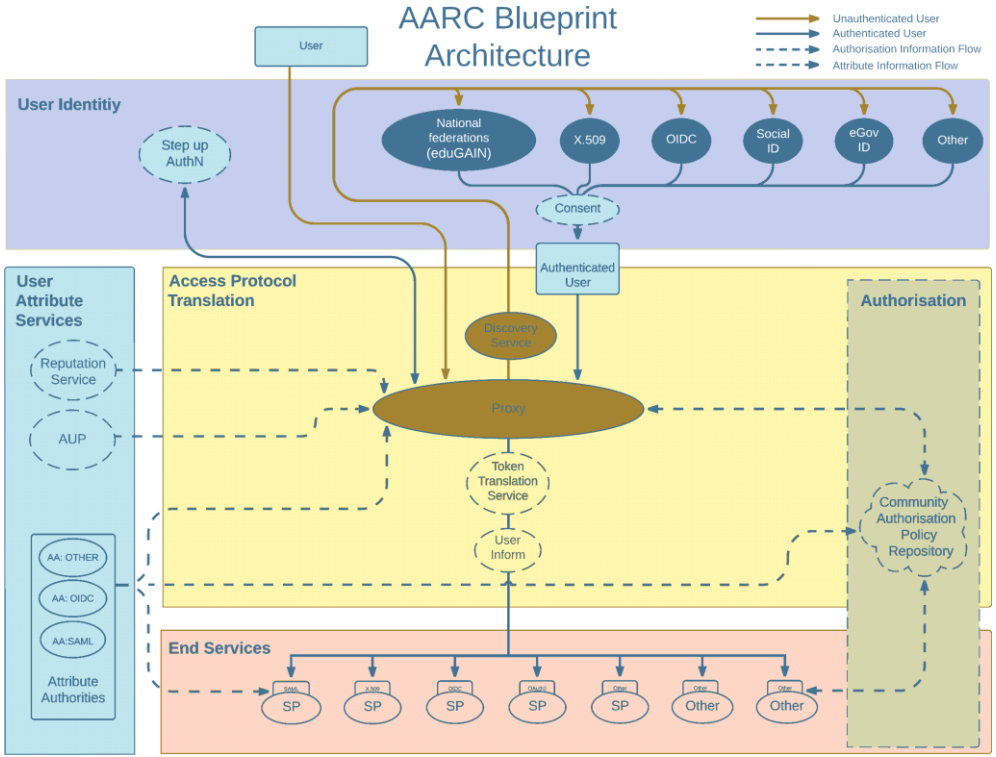
slide inspiration: Licia Florio, NORDUNET

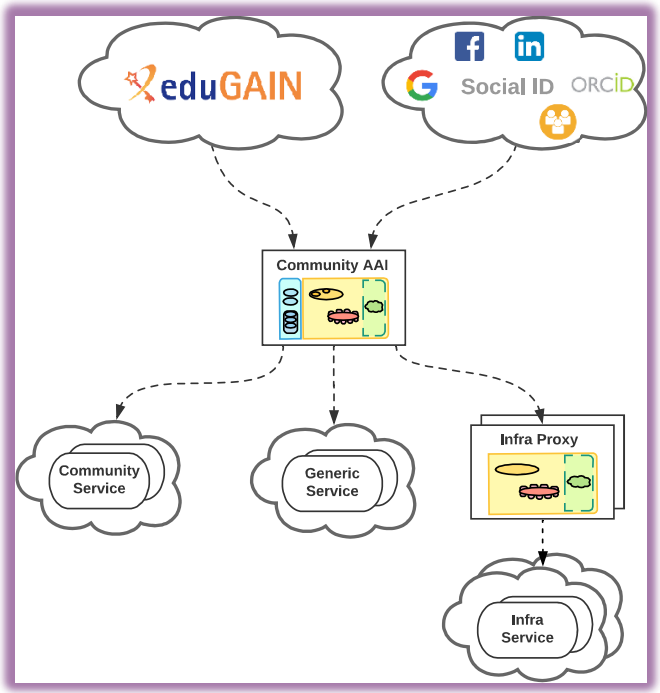
But just identity federation with your home organisation is not enough

- Access services using **identities from their Home Organizations.**
- **Access** services **based on role(s)** users have in the collaboration. This info is not known to IdPs/eduGAIN.
- Secure integration of **guest identity solutions** and **support for stronger authentication** mechanisms.
- Requirement for **one persistent identity** across all the community's services when needed and **account linking**.
- **Web** and **non-web** resources
- **Hide complexity** of multiple IdPs/feds/At Auth/ technologies.

slide design: Licia Florio, NORDUNET

Most trust flows from the (research) community

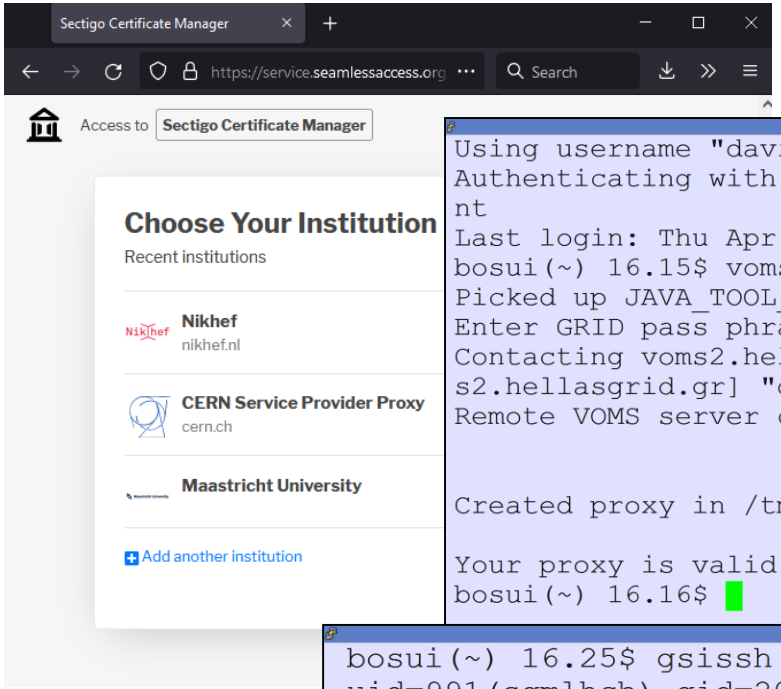




AARC Blueprint Architecture (2019) AARC-G045 https://aarc-community.org/guidelines/aarc-g045/; stacked proxies: EOSC AAI Architecture
EOSC Authentication and Authorization Infrastructure (AAI), ISBN 978-92-76-28113-9, http://doi.org/10.2777/8702

Seamless (eduGAIN) Access to (non-Web) Resources using PKIX?

Traditional workflow – using a client-held credential



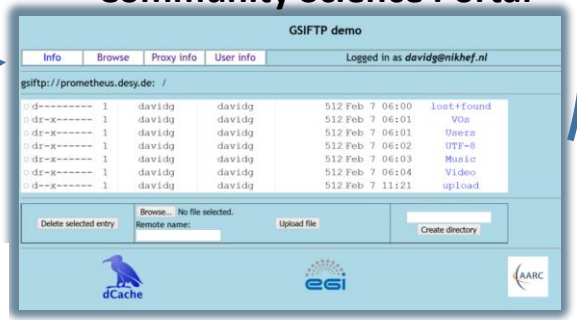

```
Using username "davidg".
Authenticating with public ke
nt
Last login: Thu Apr 13 17:43:46 2017 from 2a07:8500:120:e05b:
bosui(~) 16.15$ voms-proxy-init -voms dteam
Picked up JAVA_TOOL_OPTIONS: -Xmx512M
Enter GRID pass phrase for this identity:
Contacting voms2.hellasgrid.gr:15004 [/C=GR/O=HellasGrid/OU=h
s2.hellasgrid.gr] "dteam"...
Remote VOMS server contacted succesfully.


Created proxy in /tmp/x509up_u5917.

Your proxy is valid until Wed Apr 19 04:16:05 CEST 2017
bosui(~) 16.16$
```

Works great *provided* the user understand the technology – and we may have found all users that know how to manage this ☹

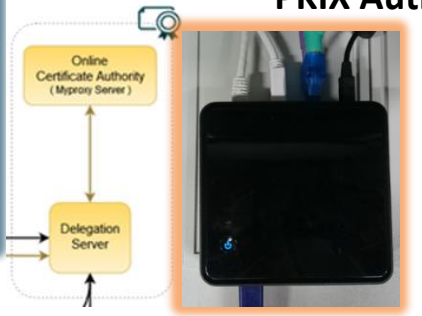
```
bosui(~) 16.25$ gsissh sgmlhcb@kot.nikhef.nl -p 1975 'id -a && hostname -f'
uid=991(sgmlhcb) gid=2015(lhcbsgm) groups=2015(lhcbsgm)
kot.nikhef.nl
bosui(~) 16.25$
```

In-line token translation services SAML-to-PKIX?

Community Science Portal

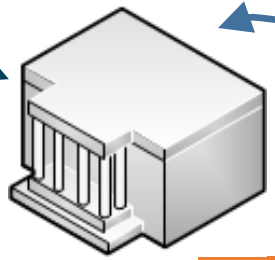


Infrastructure Master Portal Credential Store

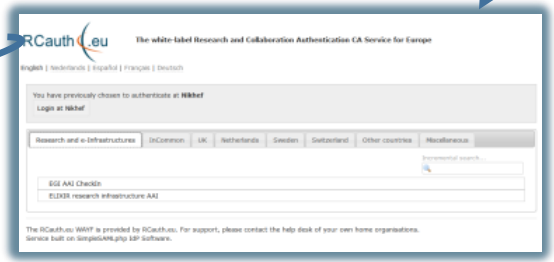


Accredited PKIX Authority

User Home Org
or Infrastructure IdP



REFEDS R&S
Sirtfi Trust

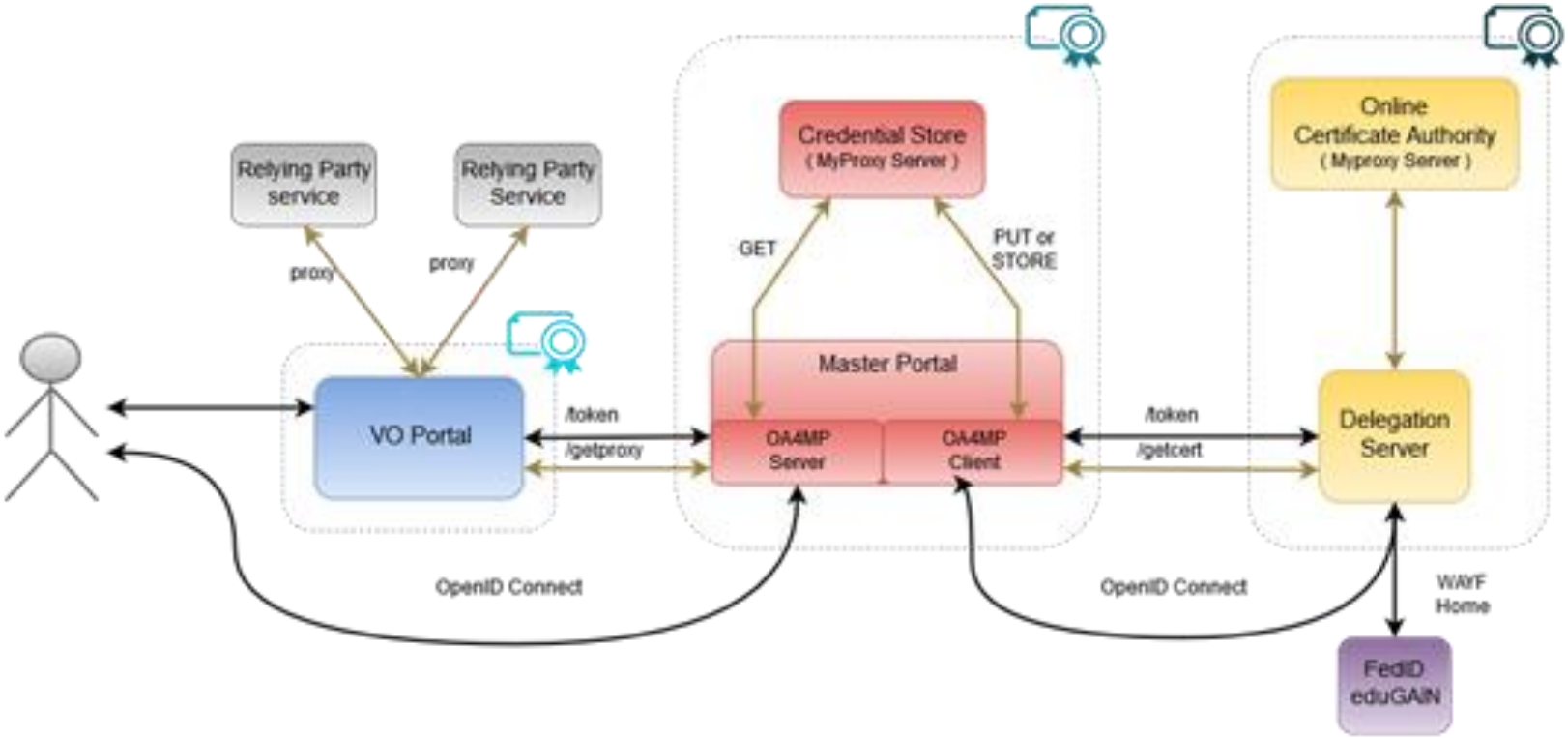


Policy Filtering WAYF / eduGAIN

Built on CILogon and MyProxy
www.cilogon.org

see also https://rcdemo.nikhef.nl/

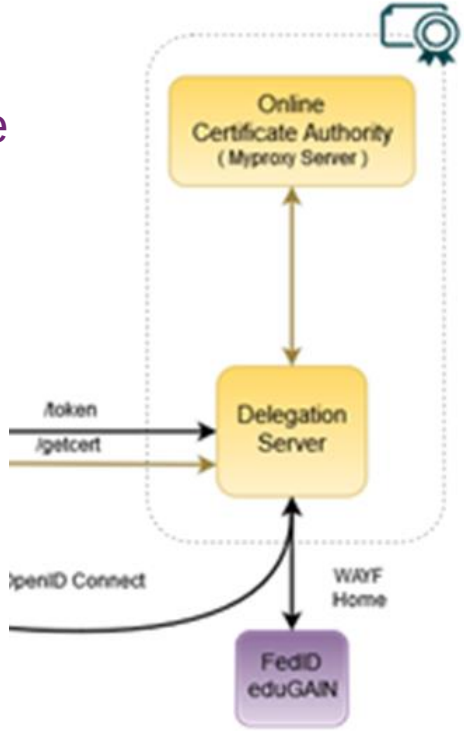
RCauth.eu and the MasterPortal OIDC credential manager



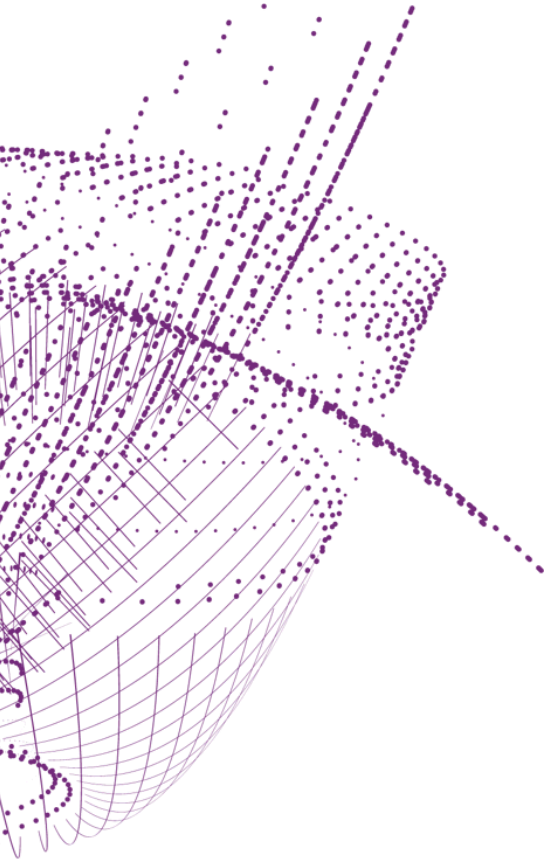
RCauth.eu – a white-label IOTA CA in Europe

- Cover as much as R&E Federated (Europe++) as possible
- Scoped to research and collaborative use cases
- In a scalable and sustainable deployment model

https://rcauth.eu/ https://rcdemo.nikhef.nl/



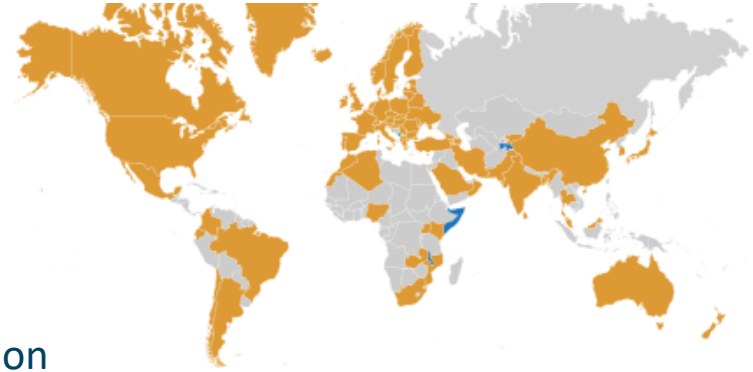
Service inspired by and using components (such as the DS) from CILogon Service
Jim Basney's CILogon, see https://www.cilogon.org/docs/20141030-basney-cilogon.pdf



global IdPs in eduGAIN
the quest for a reasonable, non-reassigned name

The joys of global interfederation

Our Registration Authorities: the Federated IdPs



- RAs are the eligible IdPs connected through a Federated Identity Management System (FIMS)
- primarily: ensemble of IdPs in eduGAIN that meet the policy requirements of this CA
- Eligible applicants are all affiliated to an RA

Three eligibility models

1. Direct relationship CA-IdP, with agreement declaration
2. Rest of eduGAIN: – "Sirtfi" security incident response and OpSec capabilities plus
   – REFEDS "R&S section 6" non-reassigned identifiers and applicant name

   are required, and tested via statement in 'meta-data' and by releasing the proper attributes
3. within the Netherlands, SURFconext Annex IX* already ensures compliance for all IdPs

"IdPs within eduGAIN [#3] are deemed to have entered materially into an agreement with the CA"

Unique certificated from FIM via eduPerson and REFEDS R&S

Sources of naming and uniqueness, that work *today*

- **eduPersonPrincipalName** – scoped point-in-time unique identifier, which could be, but usually is not, privacy preserving: “davidg@nikhef.nl”, “P70081609@maastrichtuniversity.nl”
- **eduPersonTargetedID** – scoped transient non-reassigned identifier, like urn:geant:nikhef.nl:nikidm:idp:sso!*27c8d63ed42c84af2875e2984*
- **subject-id** - a scoped persistent non-reassigned identifier, which should be privacy-preserving: 44f7751265a6e8b228f9@nikhef.nl

Plus the (domain-name based) schacHomeOrganisation and a ‘**representation of the real name**’

/DC=eu/DC=rcauth/DC=rcauth-clients/O=*orgdisplayname*/CN=*commonName +uniqeness*

uniqueness will added to commonName via hashing of *ePPN, ePTID, subject-id*, so that an enquiry via the issuer allows unique identification of the vetted entity”

commonName – should be readable element in printable 7-bit chars

‘REFEDS R&S’ gives a subset of attributes that should be released:

1. the *displayName* attribute from the IdP
2. the *givenName* attribute, followed by a space, followed by the *sn* attribute from the IdP
3. the *commonName* (cn) attribute from the IdP

but we need to make it printable in ASCII

We tried using *java.text.Normalizer.Form.NFD* and map the remainder to “X”, which gives:

If IdP sends us this UTF-8	Representation in CN RDN
Jőzsi Bácsi	Jozsi Bacsi
Guðrún Ósvífursdóttir	GuXrun Osvifursdottir
Χρηστος Κανελλοπουλος	XXXXXXX XXXXXXXXXXXXX
簡禎儀	XXX

Oops!

but also Νικόλας Λιαμπότης may not quite like that … and I understand …

- *java.text.Normalizer.Form.NFD* and 'X-ing' the rest particularly bad for Greeks, Bulgarians, Chinese, Georgians, Thai, Armenians, Serbians, …

ICU - International Components for Unicode (icu-project.org) appears to be better, but:

- there are many options for transliteration
- some code points shared between different languages, that prefer different transliterations
- some code points are absent even in UTF-8 causing ambiguity

So we moved to the ICO, but even then the mapping is not trivial:

ICU ICU regex

UTF-8 → Latin-1 → ASCII → IA5String (we need PrintableString + "@" and minus [:/=])

thanks go to Mischa Sallé for the transliteration studies (and much more)

But straightforward translation is not always good

Just Any-Latin fails for Slavonic unique “sh” sounds. E.g. for ‘Миша’

- with *Any-Latin* becomes ‘Miša’ which then translates into ‘Misa’ after the Latin-Ascii

but quite some people called ‘Миша’ want to see ‘Mischa’, but not all, so you need

- first *Russian-Latin/BGN*, making it ‘Misha’, which is slightly better, then do *Any-Latin* (1-to-1)
- but “*Russian-Latin/BGN+Serbian-Latin/BGN”* is different from the reverse …

First Any-Latin/BGN, then Any-Latin, to fix mapping to → š and the → s

- Բարեւ աշխարհ → Barev a**sh**kharh (with the /BGN, to ensure the “sh”)
- ישראל → ysr'l (taken care of without the /BGN, otherwise the ש never makes it)

And Unicode does not distinguish the *diaeresis* and the *umlaut*

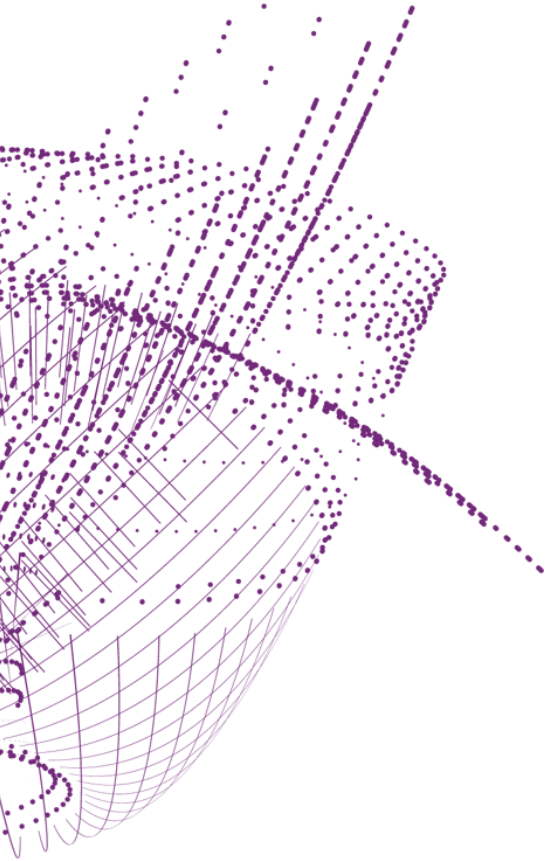
- Günter Strauß ⟶ Gunter Strauss *should* have been ‘Guenter Strauss’
- Daniëlle ⟶ Danielle is good, you definitely don’t want ‘Danieelle’

As the so for stability, we keep Any-Latin here and treat all as a diaeresis

What will we get?

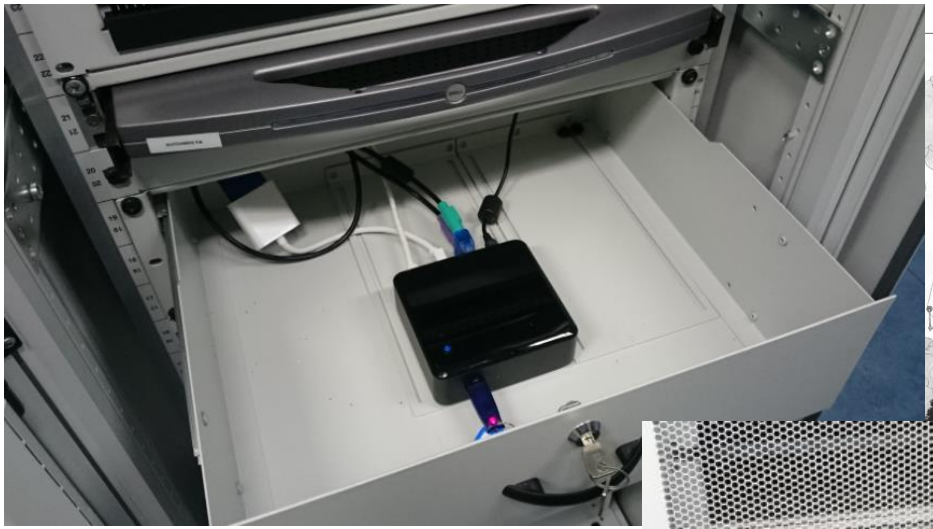
```
$ java -cp icu4j-59_1.jar:. transliterate2 [...]
 "Jőzsi Bácsi" "Guðrún Ósvífursdóttir" \
 "Χρηστος Κανελλοπουλος" "簡禎儀"
```

```
Input:   Jőzsi Bácsi
Output:  Jozsi Bacsi
Input:   Guðrún Ósvífursdóttir
Output:  Gudrun Osvifursdottir
Input:   Χρηστος Κανελλοπουλος
Output:  Christos Kanellopoulos
Input:   簡禎儀
Output:  jian zhen yi
```



Building the initial RCauth.eu

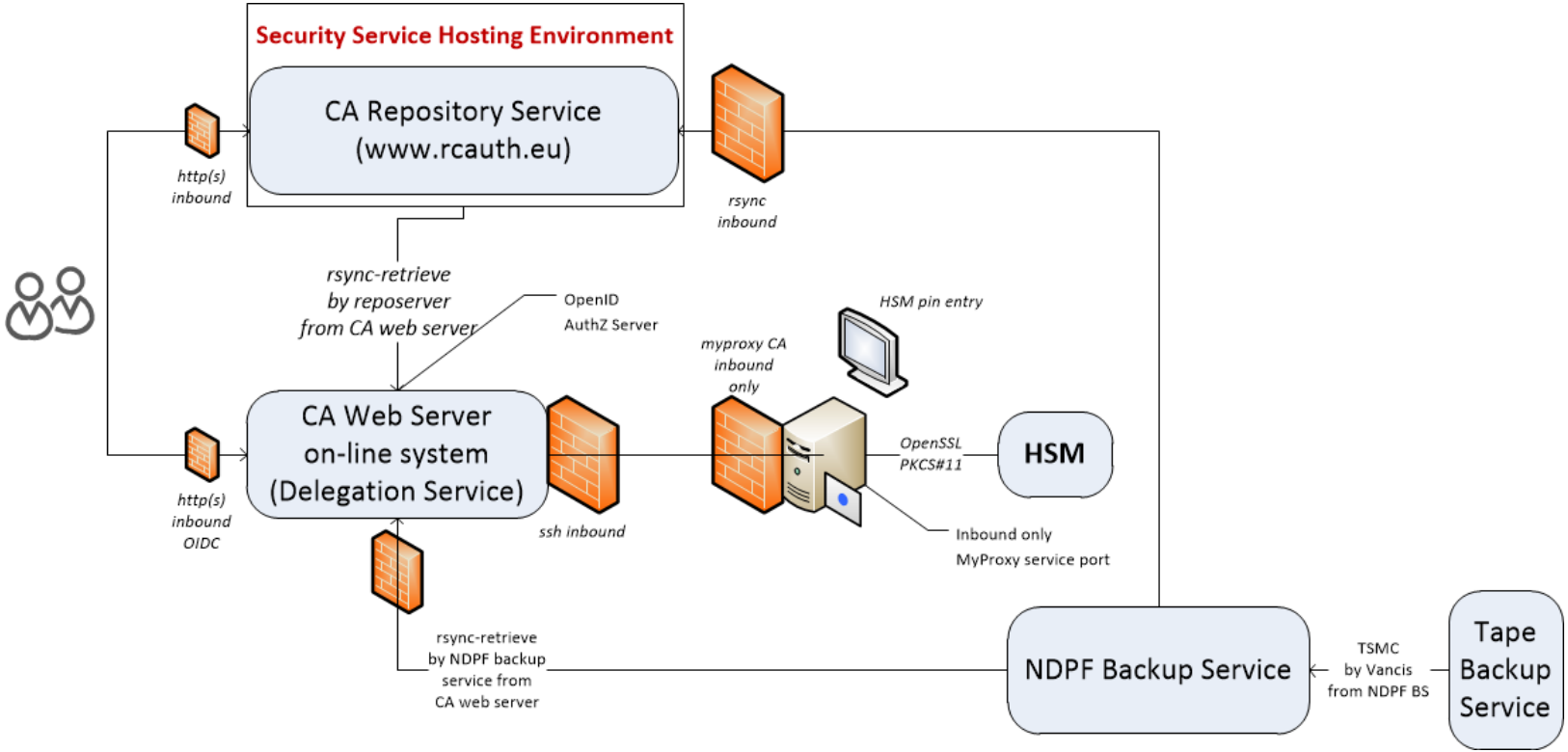
A fully compliant ‘Heath Robinson’ CA



Physical controls

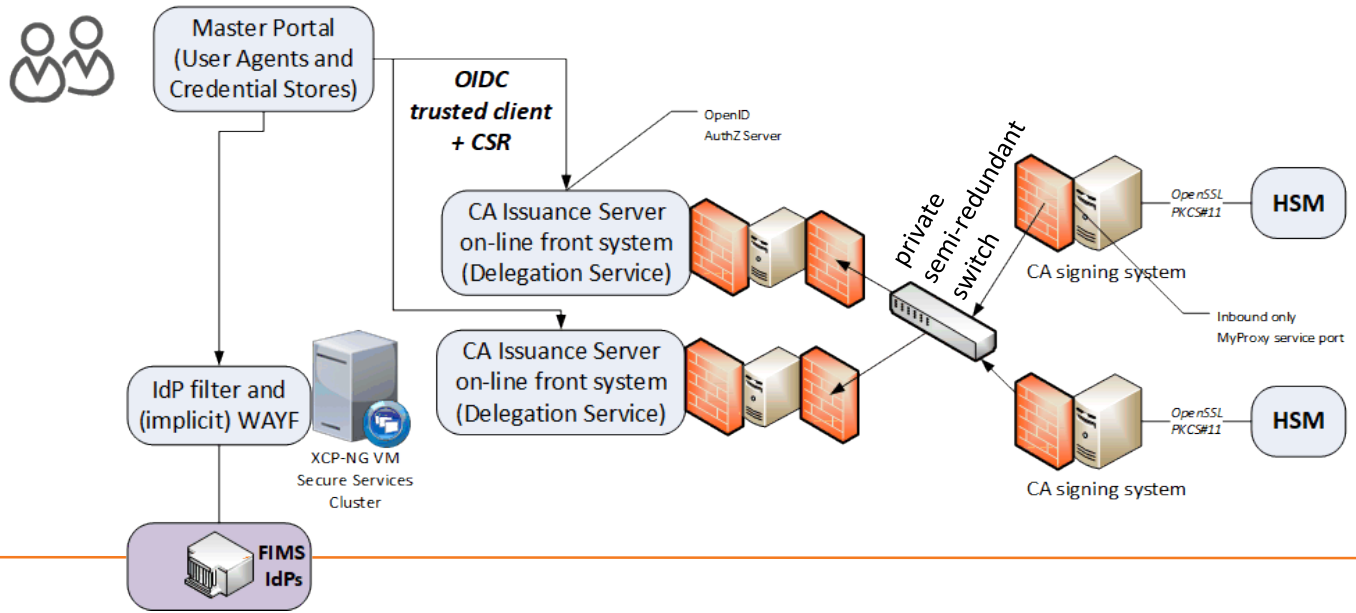
- Located at Nikhef, Amsterdam, NL
- Scientific Data Centre part of the NikhefHousing Facilities
- ID based access control, 24hr guard on-site
- CA and security systems in locked dedicated cabinet on 2nd floor
  On-line CA signing system in locked drawer

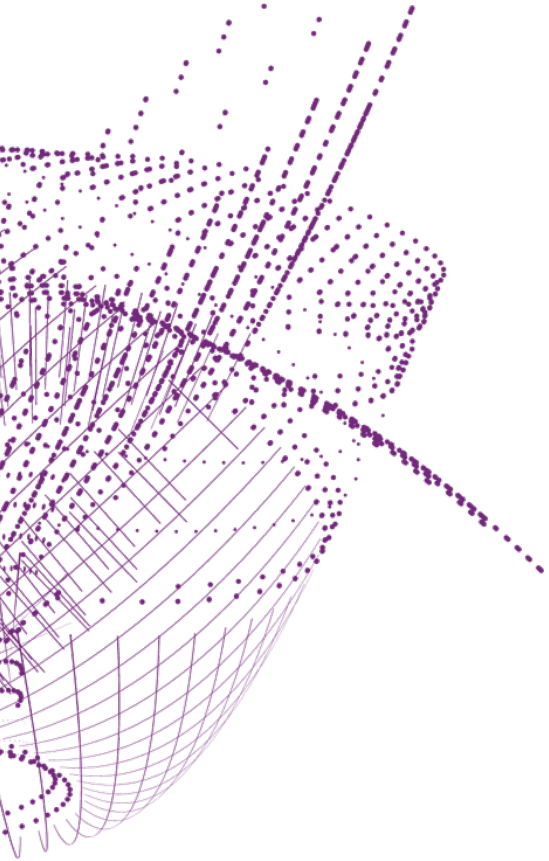
Logical set-up



A local highly-available setup at Nikhef Amsterdam

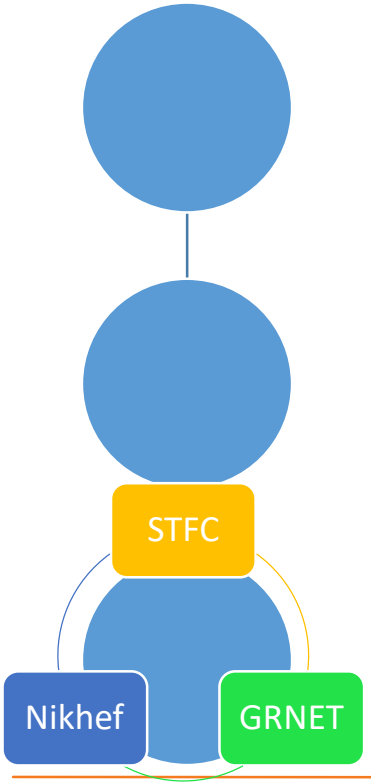
- Most ‘fault-prone’ components are
  - Intel NUC (single power supply)
  - HSM (can lock itself down, and the USB connection is prone to oxidation)
  - DS front-end servers (they are physical hardware, albeit with redundant disks and powersupplies)
- Eliminated first using ‘local HA’





towards a pan-European distributed service

RCauth.eu Governance



Governance Board

Representatives (and one alternate) from each Materially Contributing Stakeholder
EGI.eu, EUDAT (ETFC), GÉANT, Nikhef (SURF)

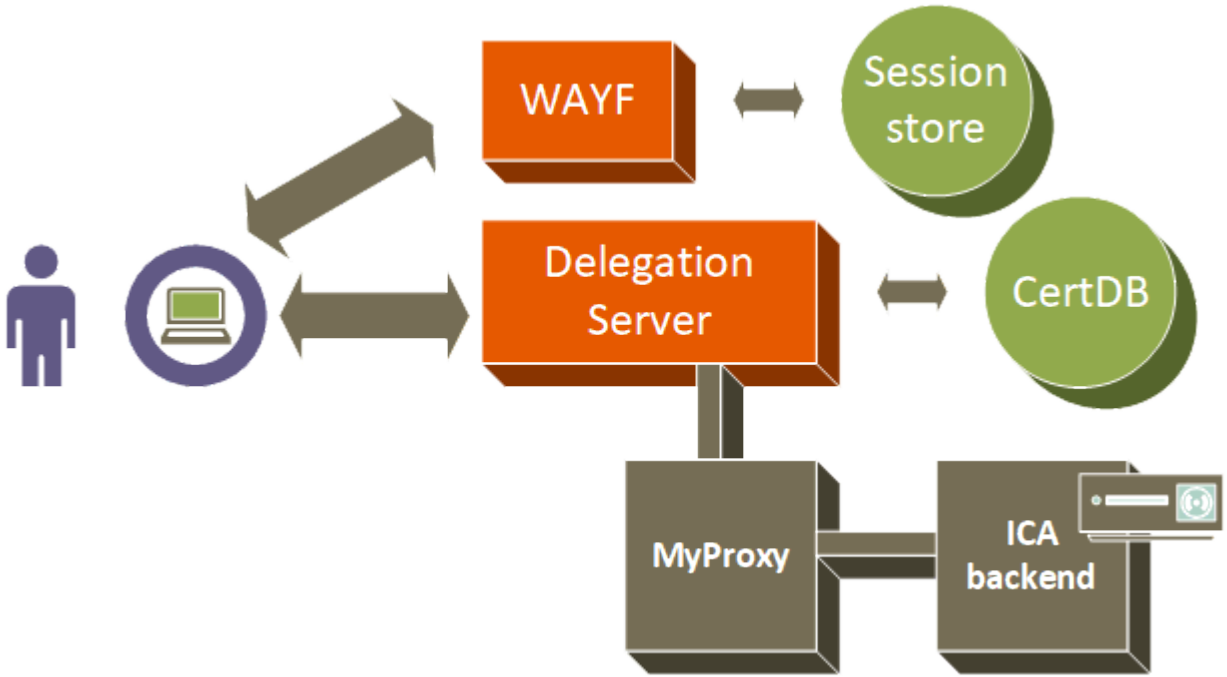
RCauth PMA

Individuals drawn from the wide community […] experts in the field of identity management for research and collaboration, PKI technology and identity bridging

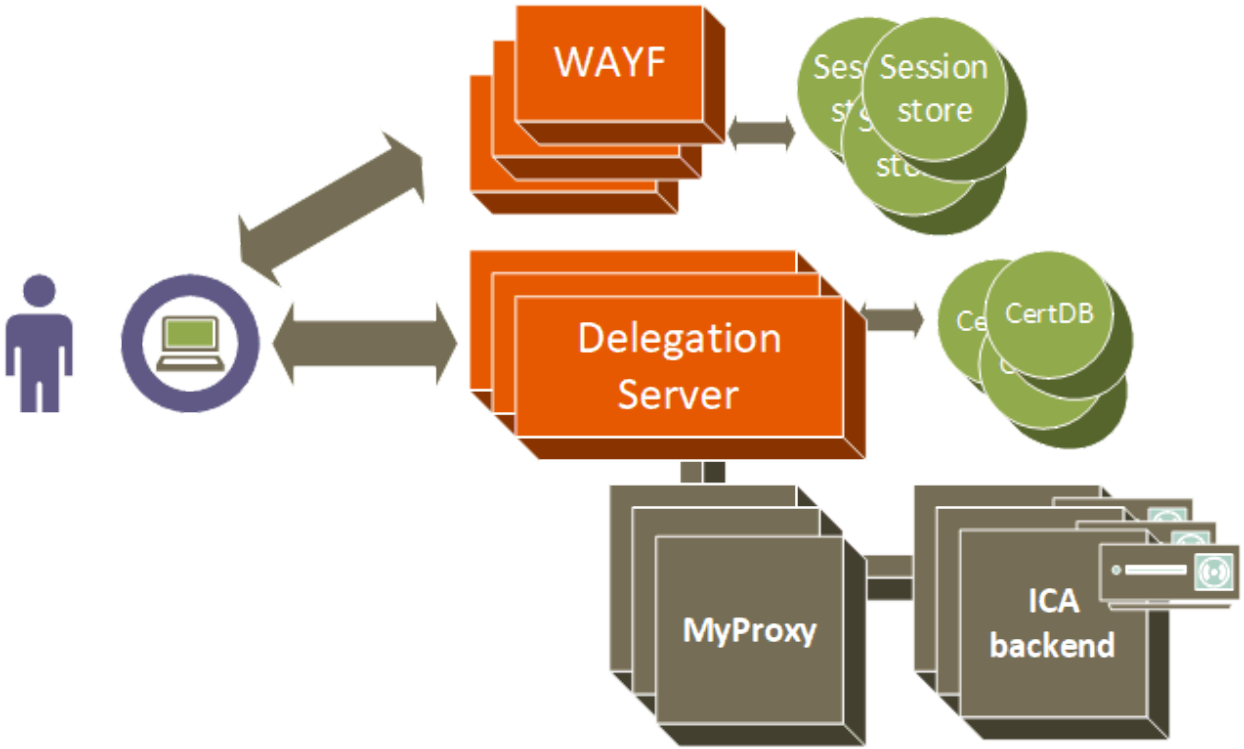
Ops Coordination Team

Operations people from each of the **hosting partners** with a (copy of) the RCauth.eu signing key, and those partners otherwise involved in OPS

From a single instance …



… to a 3-fold continuously-consistent setup



HA solutions

Local high availability, three distinct providers?

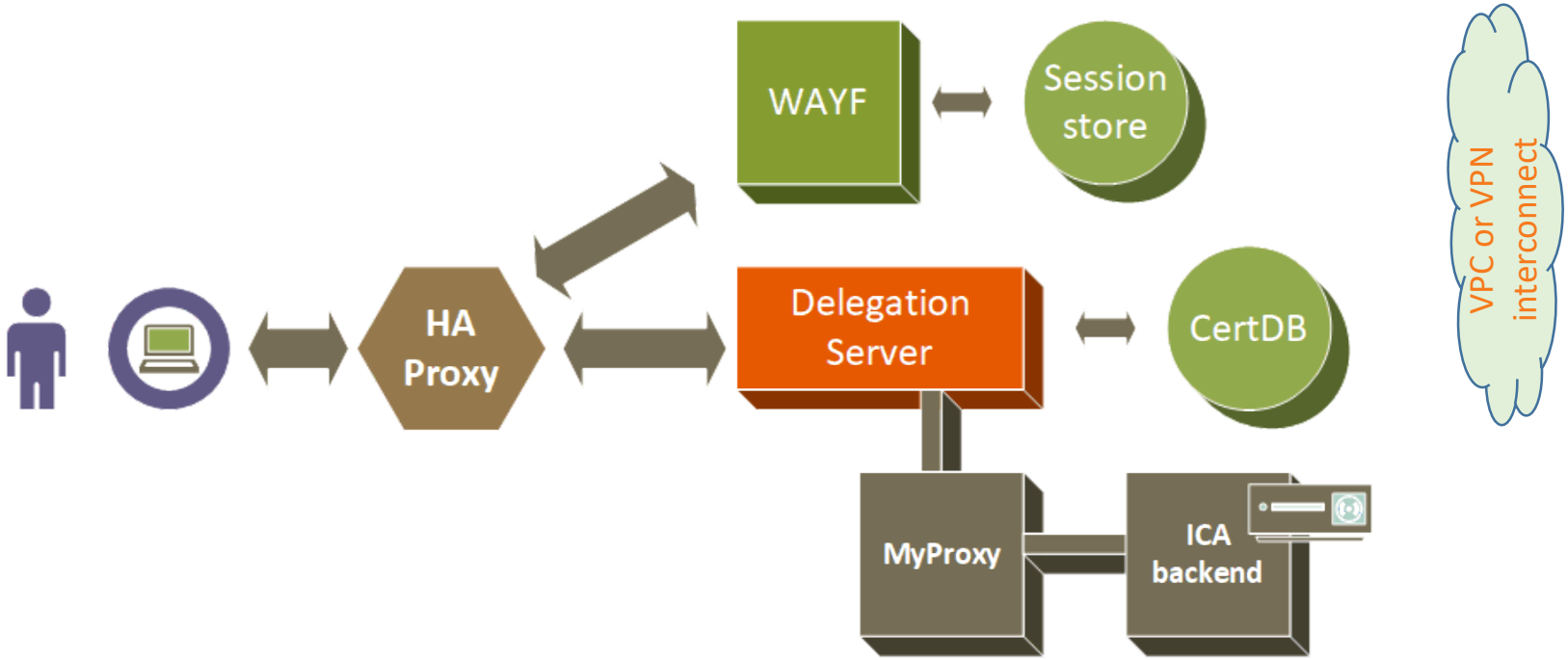
- pushes account linking burden to the relying parties/service providers
- users may have 3 credentials, which is confusing
- a single identifier would require ‘ensured’ database synchronization – no true independence

DNS-based fail-over?

- the ‘trivial’ model relies on the client not to cache answers for long, ***and*** not to round-robin the DNS answers - since the WAYF and DS go together
- short TTL is quite bad for reliance, since both service and domain name provider must be up
- ‘advanced’ DNS-based solutions (like for InAcademia) – with near-realtime updates of a distributed DNS may appear better, but still: need a overly-low TTL, and move the HA problem to the DNS provider (or ccTLD), rather than solve it

So we looked at network-layer resilience, the ‘go-to’ solution for large CDN providers

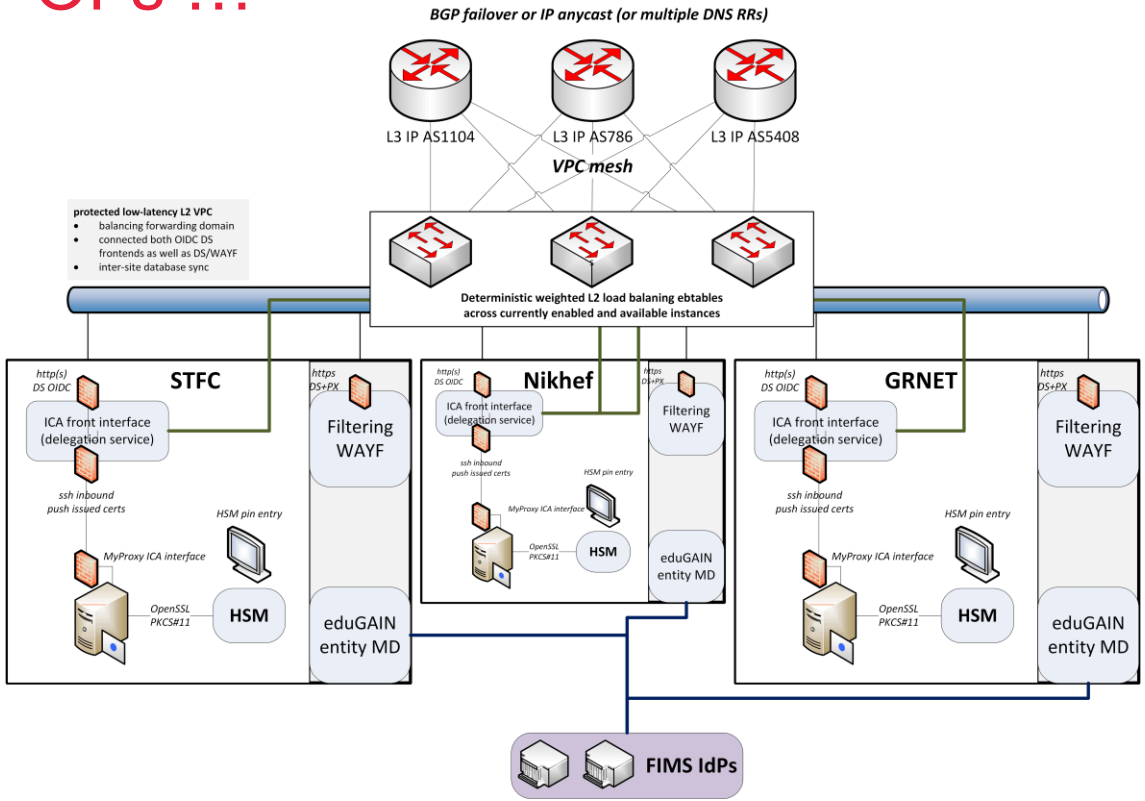
Services at a site go up and down together - adding an HAproxy



Since we do not like SPOFs …

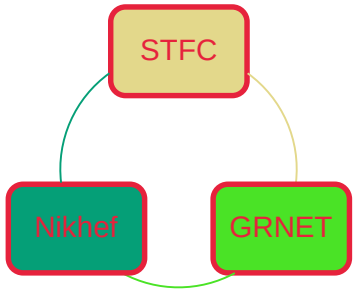
Implement a High Availability setup

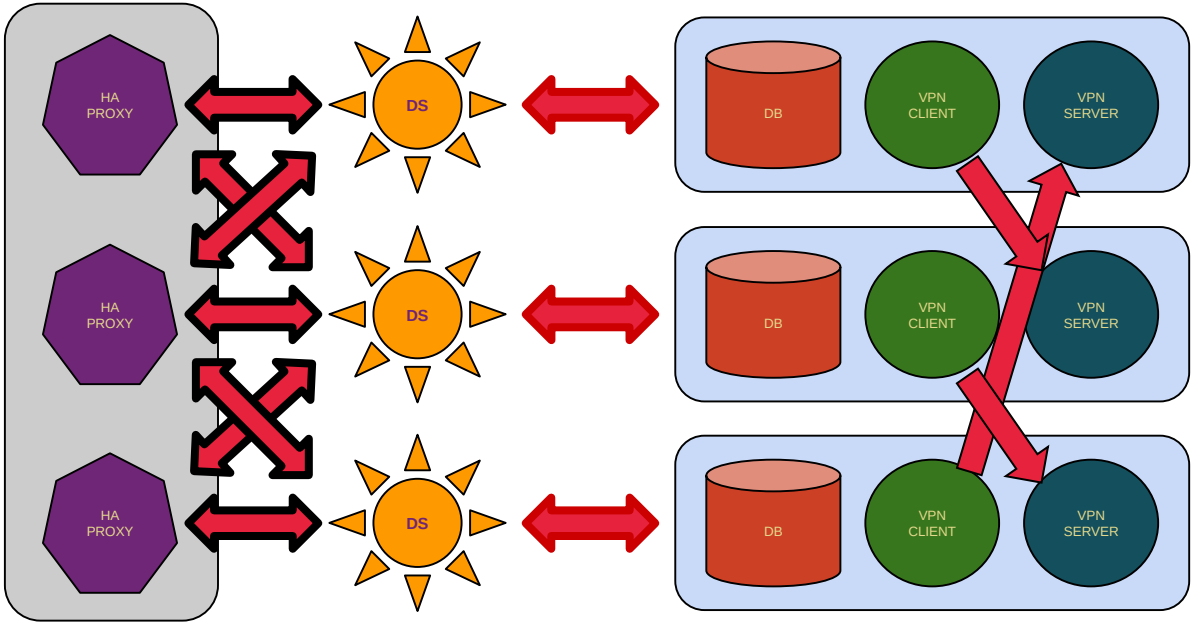
- across the 3 sites
- using IP anycast
- L3 VPN or L2 VPC
- with minimal effort

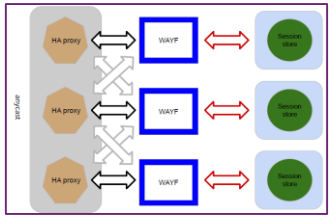


work supported by EOSC Hub and EOSC Future

Distributed RCauth service





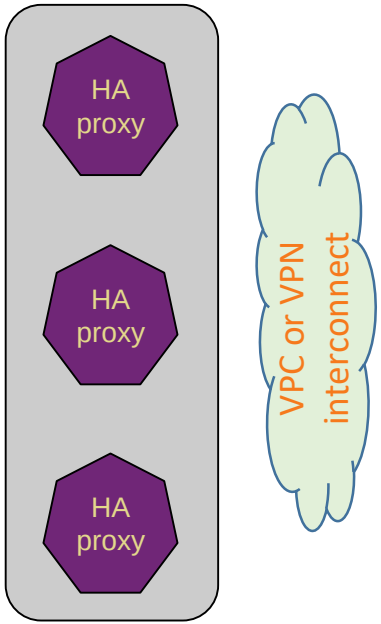


similar for the filtering WAYF

selected imagery: Mischa Sallé, Jens Jensen, Nicolas Liampotis

A transparent multi-site setup

User
- connects to HA proxy at {wayf,ica}.rcauth.eu
- HA proxy sends users to “closest” working service
- forward mainly to its own DS when available

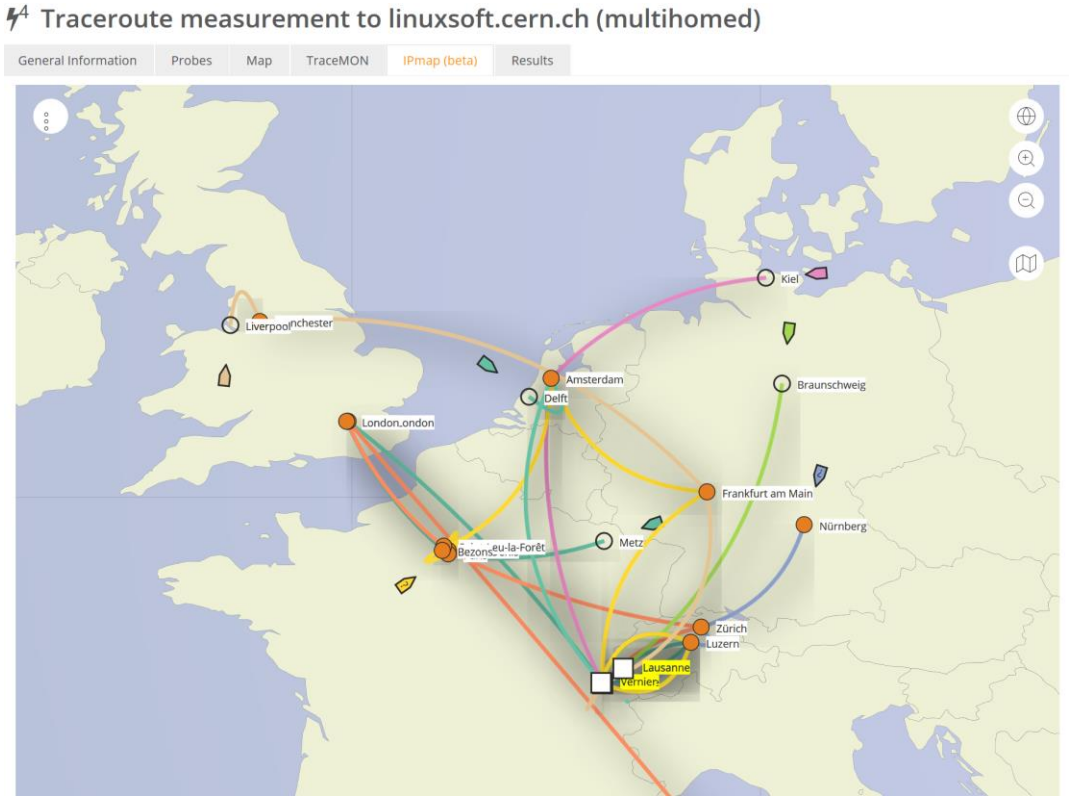


If a HA loses its backend DS, can still route to another DS over VPC/VPN backend

and wherever the user is, the service is at
- **2a07:8504:01a0::1**
- and **145.116.216.1** (for legacy IP users)

selected imagery: Mischa Sallé, Jens Jensen, Nicolas Liampotis

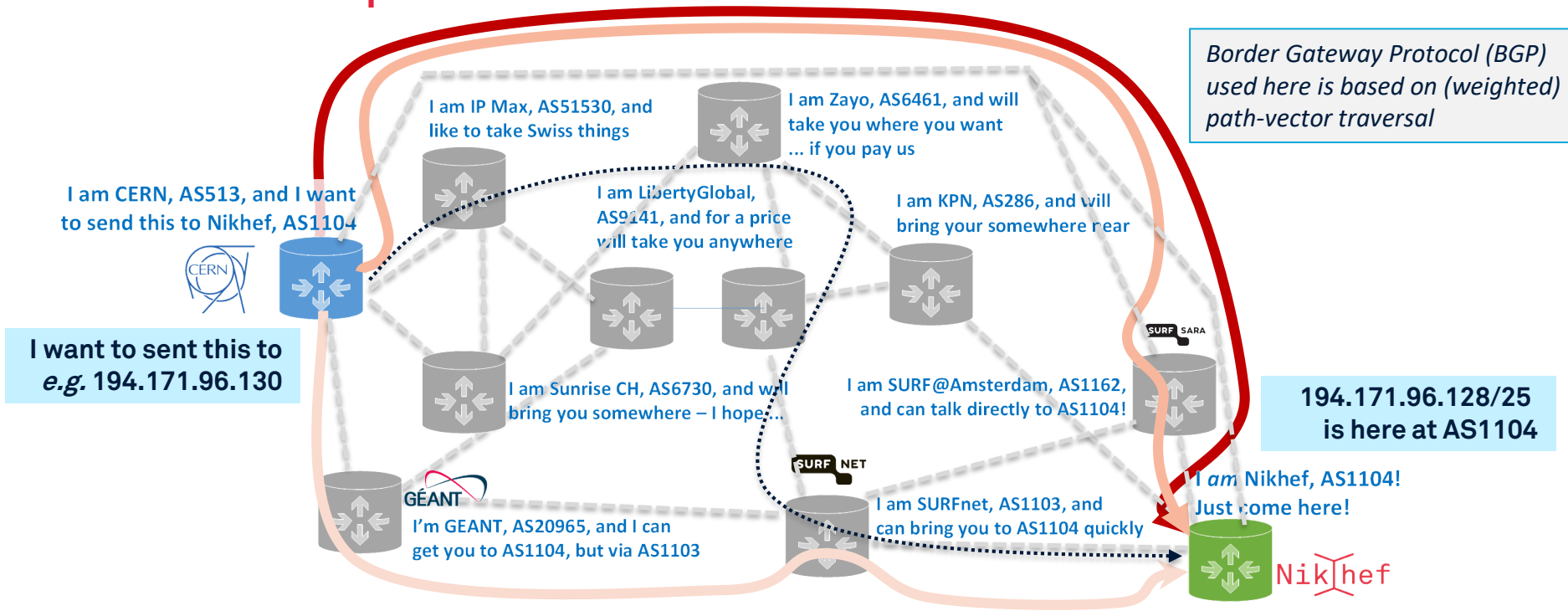
Intermezzo – BGP routing principles



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

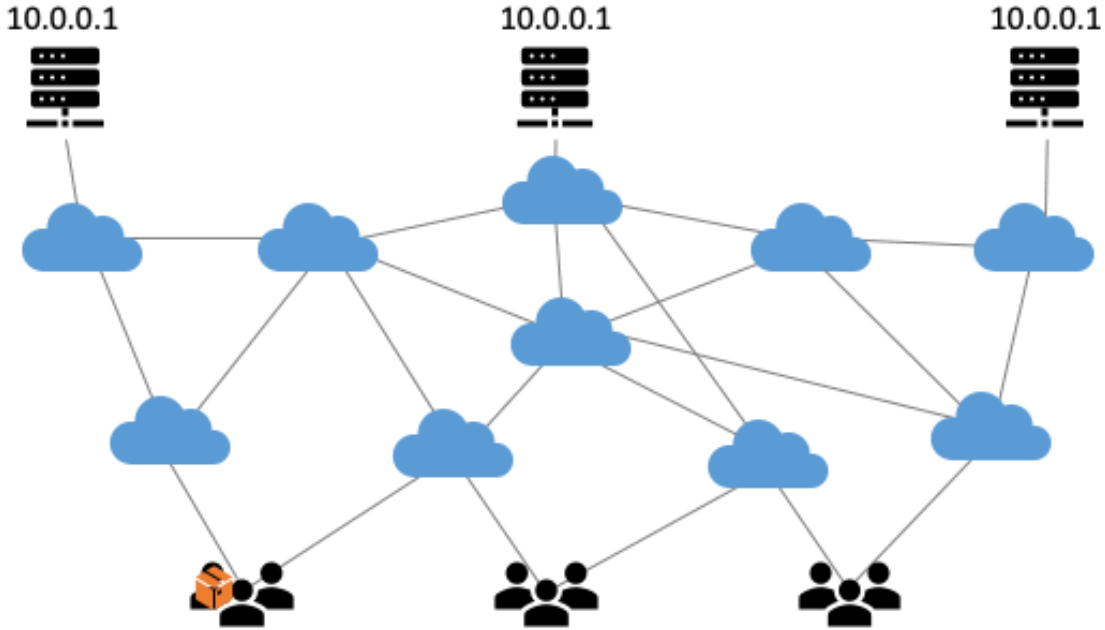
RIPE NCC
RIPE NETWORK COORDINATION CENTRE

How does a packet flow?



grey-dash lines for illustration only: may not correspond to actual peerings or transit agreements; red lines: the three existing LHCOPN and R&E fall-back routes; yellow: public internet fall-back (least preferred option)

Anycast: when the same place exists many times



So we used

- 3 (now: 2) sites
- one VM at each site exposing 2a07:8504:01a0::1
- smallest v6 subnet (/48)
- bird + a service probe
- each site's own ASN
- some IRR DB editing
- v4 is similar, with a /24

and some monitoring

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

BIRD config and probes

you need

- a health checker to drive the local BGP daemon
- a BGP talker, such as bird
- a *very* simple config

```
# Generated 2023-02-05 14:49:36.063331
# by anycast-healthchecker (pid=1299)
# 2001:db8::1/128 is a dummy IP Prefix.
# It should NOT be used and REMOVED
# from the constant.
define ACAST6_PS_ADVERTISE =
    [
        2001:db8::1/128,
        2a07:8504:1a0::1/128
    ];
```

```
include "/etc/bird.d/*.conf";

router id 194.171.98.77;

define ASN_OWN          = 65530;
define ASN_NEIGHBOUR    = 1104;
define ADDR_NEIGHBOUR4 = 194.171.98.94;
define ADDR_NEIGHBOUR6 = 2a07:8500:120:e011::1;

protocol device { scan time 10; }

protocol direct direct1 {
    interface "lo";
    ipv4 { import all; export none; };
    ipv6 { import all; export none; };
}

template bgp bgp_peers4 {
    local as ASN_OWN;
    ipv4 {
        import none;
        export filter match_route_filter;
    };
}
template bgp bgp_peers6 {
    local as ASN_OWN;
    ipv6 {
        import none;
        export filter match_route6_filter;
    };
}
protocol bgp BGP4 from bgp_peers4 { disabled no; neighbor ADDR_NEIGHBOUR4 as ASN_NEIGHBOUR; }
protocol bgp BGP6 from bgp_peers6 { disabled no; neighbor ADDR_NEIGHBOUR6 as ASN_NEIGHBOUR; }
```

But what is ‘healthy’?

Service status verification tool needed to ‘drive’ bird actions

- anycast_healthchecker by Pavlos Parissis

- with HAproxy
  on the front-end host
  on each site

```
Packager     : Mischa Sallé <msalle@nikhef.nl>
Vendor       : Pavlos Parissis <pavlos.parissis@gmail.com>
URL          : https://github.com/unixsurfer/anycast_healthchecker
Summary      : A healthchecker for Anycasted Services
Description :
Anycast-healthchecker monitors a service by doing periodic health
checks and based on the result instructs Bird daemon to either
advertise or withdraw the route to reach the monitored service. As
a result Bird will only advertise routes for healthy services.
```

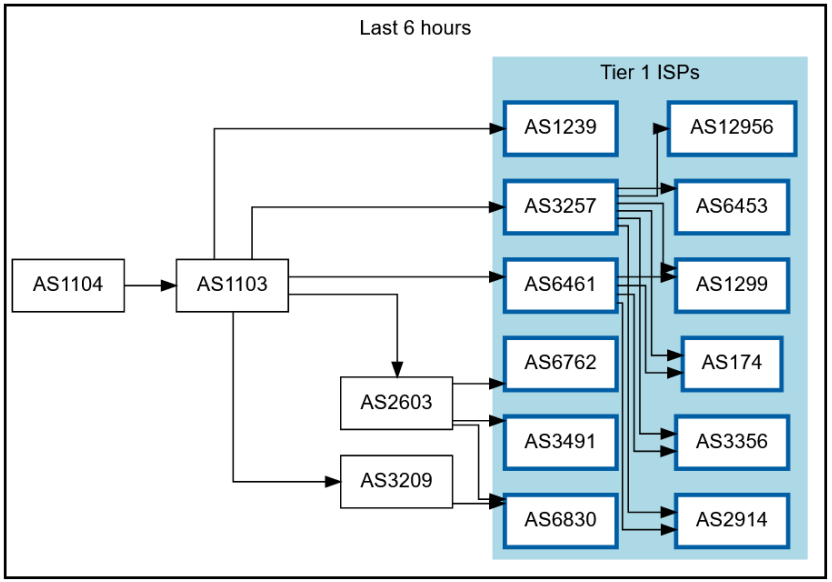
```
[haproxy]
check_cmd       = /usr/local/sbin/check_haproxy.sh
on_disabled     = withdraw
ip_prefix       = 145.116.216.1/32
[haproxy6]
check_cmd       = /usr/local/sbin/check_haproxy.sh
on_disabled     = withdraw
ip_prefix       = 2a07:8504:1a0::1/128
```

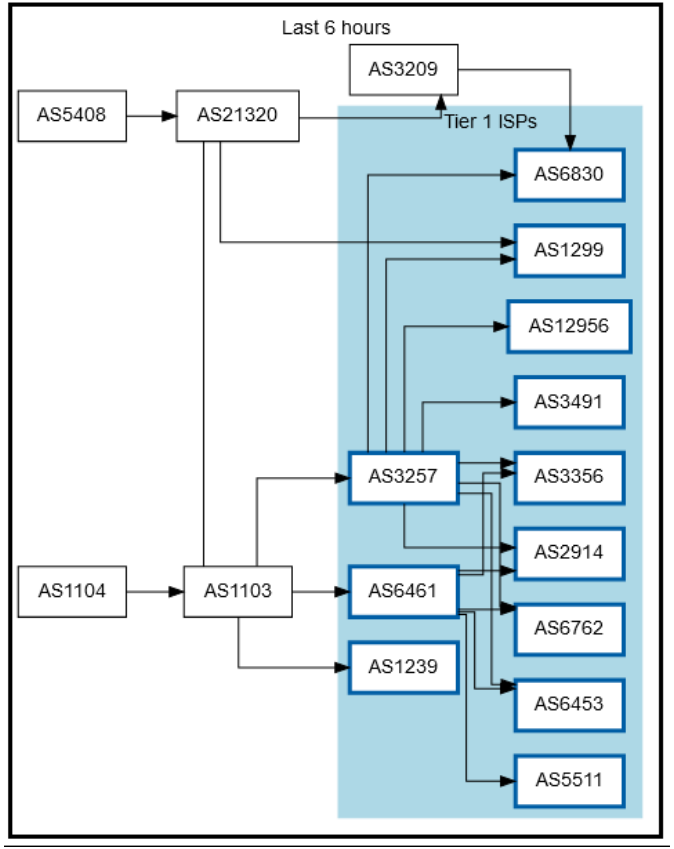
Both Delegation Service and filtering WAYF should be up

But since Nikhef also has local HA with two back-ends, either is OK!

```
# Checks WAYF backends, at least one should be up or starting
# i.e. in state 2 or 3 (see Section 9.3 Unix Socket commands in
# management.txt).
check_wayf()    {
    echo $state_cmd |\
        socat unix-connect:${haproxy_socket} stdio |\
        grep $wayf_pattern |\
        cut -d' ' -f${site_col},${state_col} |\
        while read wayf_site wayf_state
    do
        if [ "$wayf_state" -ge 1 -a "$wayf_state" -le 2 ];then
            # Found at least one up DS
            info "WAYF $wayf_site has state $wayf_state"
            return 1
        else
            warn "WAYF $wayf_site has state $wayf_state" >&2
        fi
    done
    return $((1-$?))
}
```

Getting 2a07:8504:01a0::/48 out there





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

CERN Looking Glass Results - ee1

```
inet6.0: 155476 destinations, 303862 routes (155437 active, 0 holddow
+ = Active Route, - = Last Active, * = Both

2a07:8504:1a0::/48 *[BGP/170] 01:08:50, MED 20, localpref 10500
                     AS path: 20965 5408 I, validation-state: unveri
                   >  to 2001:798:99:1::39 via irb.200
                   [BGP/170] 4d 23:13:16, MED 20, localpref 10500, f
                     AS path: 1103 1104 I, validation-state: unverif
                   >  to fe80::1a2a:d300:140f:bdb0 via irb.20
                   [BGP/170] 6d 23:17:01, MED 20, localpref 10500
                     AS path: 2603 1103 1104 I, validation-state: un
                   >  to 2001:1458:0:9::2 via irb.2903
                   [BGP/170] 01:08:26, MED 25, localpref 10500
                     AS path: 559 20965 5408 I, validation-state: un
                   >  to 2001:1458:0:2c::2 via irb.2902
                   [BGP/170] 01:08:49, MED 10, localpref 10200
                     AS path: 174 174 21320 5408 I, validation-state
                   >  to 2001:978:2:2::2a:1 via irb.3811
```

2a07:8504:1a0::/48

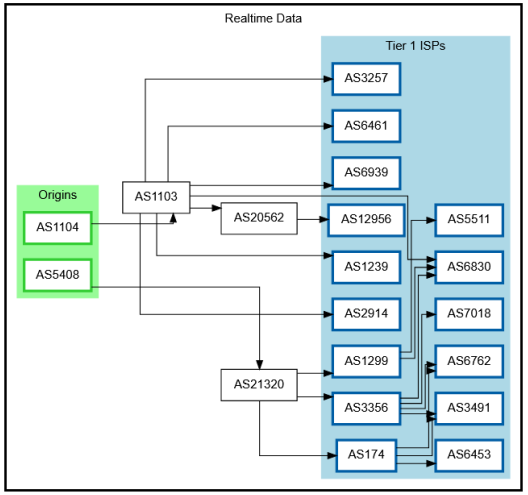
Announced by **AS1104, and 1 other**

Overview | Connectivity | Whois | DNS | Validation

Originators

	ASN	Description
	AS1104	Nikhef - Dutch National Institute for Sub-atomic Physics
	AS5408	National Infrastructures for Research and Technology S.A.

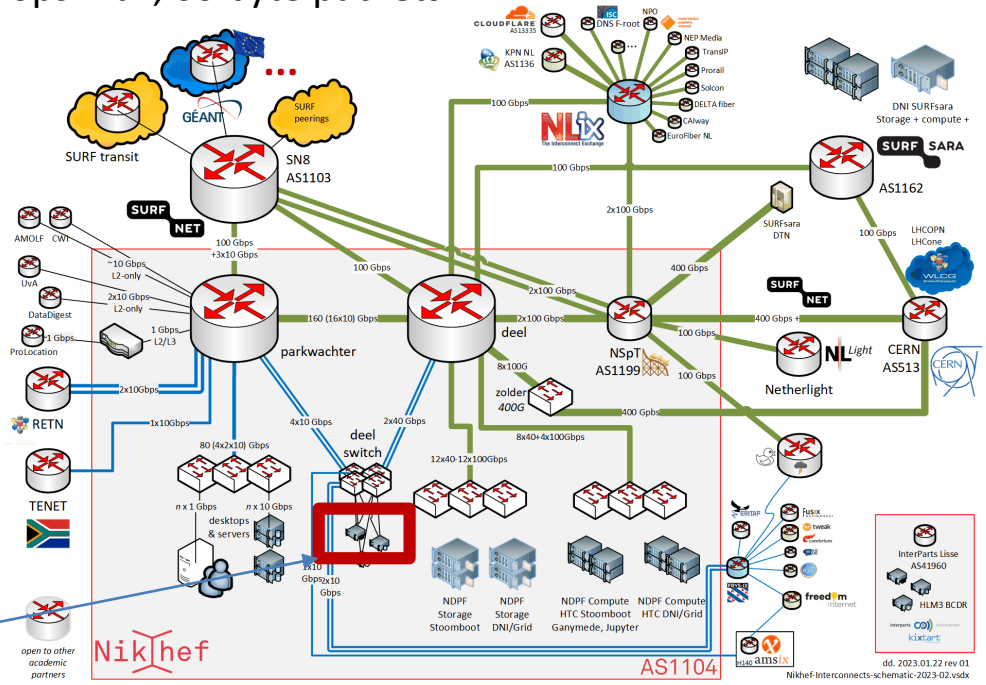
How can a prefix have multiple ASNs?



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

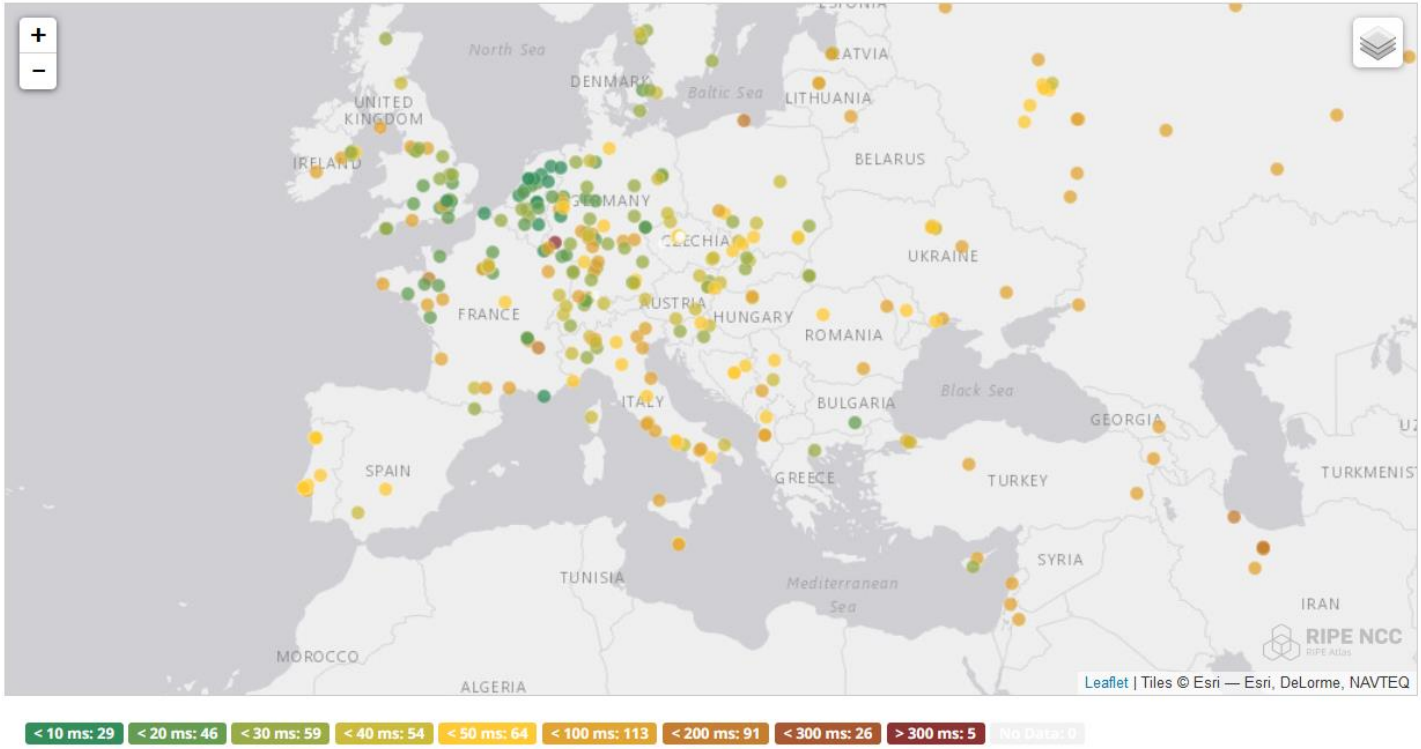
Prerequisites are relatively simple

- IPv4 ***/24*** netblock and IPv6 ***/48***
- your own, or a friendly, **ASN**
- a set of corresponding **IRR route objects**, and either none, or a correct RPKI
  (easily done in your local RIR registry: APNIC, RIPE, ARIN, AfriNIC, LACNIC)
- **front-end service (HAproxy)** for the Delegation Service and filtering WAYF
- **bird** (or quagga) with a service **health checker**

But you do not 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, distributed across Europe (https://atlas.ripe.net/measurements/50949024/)

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 used extensively (also active/passive) for other services at Nikhef

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

- use dedicated HA links for the back-end database connection or ip-forwarding
  e.g. multiple redundant circuits over an MPLS cloud emerging at each site

Current status

- All sites can sign production certificates
- DS databases cross-site replication using Galera over VPN
- HA CRL cross site synchronisation and issuance
- WAYF servers (GRNET and Nikhef)

Reuse the RCauth experience

All sources, Ansible playbooks, and materials are on GitHub
https://github.com/rcauth-eu

HA database and back-end VPN
- 3-node peer-peer redundant VPN with automatic failover
- extensible to >3, but then topology is less clear

Web services
- HAproxy stability and flexibility and coordinated ‘up-down’ status per site

HAHAP | BGP Anycast
- ‘bog-standard’ if service admins, cloud admins, and network people can collaborate and investigate incidents together

secure credential sharing and moving shared secrets is still cumbersome in practice

‘the difference between theory and practice is that, in theory, there is no difference’

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!

RCauth.eu distributed setup 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

Nikhef