Computing for Research & the Worldwide LHC Computing Grid

Building a global large-scale
ICT infrastructure
for research data processing

Nikhef

David Groep
DACS & Nikhef

5 November 2024

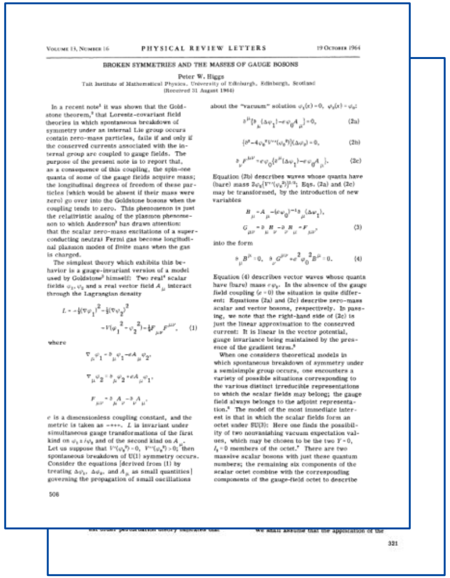
KEN3239 rev 1.2

Maastricht University | Department of Advanced Computing Sciences

Peter Higgs and Francois Englert at the 2013 Nobel prize press conference, Stockholm. Photo: Bengt Nyman, https://www.flickr.com/photos/97469566@N00

Exploding data? the Large Hadron Collider at CERN

1964



VOLUME 13, NUMBER 16 PHYSICAL REVIEW LETTERS 19 OCTOBER 1964

BROKEN SYMMETRIES AND THE MASSES OF GAUGE BOSONS

Peter W. Higgs
Tait Institute of Mathematical Physics, University of Edinburgh, Edinburgh, Scotland
(Received 31 August 1964)

In a recent note[1] it was shown that the Goldstone theorem,[2] that Lorentz-covariant field theories in which spontaneous breakdown of symmetry under an internal Lie group occurs contain zero-mass particles, fails if and only if the conserved currents associated with the internal group are coupled to gauge fields. The purpose of the present note is to report that, as a consequence of this coupling, the spin-one quanta of some of the gauge fields acquire mass; the longitudinal degrees of freedom of these particles (which would be absent if their mass were zero) go over into the Goldstone bosons when the coupling tends to zero. This phenomenon is just the relativistic analog of the plasmon phenomenon to which Anderson[3] has drawn attention: that the scalar zero-mass excitations of a superconducting neutral Fermi gas become longitudinal plasmon modes of finite mass when the gas is charged.

The simplest theory which exhibits this behavior is a gauge-invariant version of a model used by Goldstone[2] himself: Two real[4] scalar fields φ_1, φ_2 and a real vector field A_μ interact through the Lagrangian density

$$L = -\tfrac{1}{2}(\nabla\varphi_1)^2 - \tfrac{1}{2}(\nabla\varphi_2)^2 - V(\varphi_1^2+\varphi_2^2) - \tfrac{1}{4}F_{\mu\nu}F^{\mu\nu}, \qquad (1)$$

where

$$\nabla_\mu\varphi_1 = \partial_\mu\varphi_1 - eA_\mu\varphi_2,$$

$$\nabla_\mu\varphi_2 = \partial_\mu\varphi_2 + eA_\mu\varphi_1,$$

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu,$$

e is a dimensionless coupling constant, and the metric is taken as $-+++$. L is invariant under simultaneous gauge transformations of the first kind on $\varphi_1 \pm i\varphi_2$ and of the second kind on A_μ. Let us suppose that $V'(\varphi_0^2)=0$, $V''(\varphi_0^2)>0$; then spontaneous breakdown of U(1) symmetry occurs. Consider the equations [derived from (1) by treating $\Delta\varphi_1$, $\Delta\varphi_2$, and A_μ as small quantities] governing the propagation of small oscillations about the "vacuum" solution $\varphi_1(x)=0$, $\varphi_2(x)=\varphi_0$:

$$\partial^\mu\{\partial_\mu(\Delta\varphi_1) - e\varphi_0 A_\mu\} = 0, \qquad (2a)$$

$$\{\partial^2 - 4\varphi_0^2 V''(\varphi_0^2)\}(\Delta\varphi_2) = 0, \qquad (2b)$$

$$\partial_\nu F^{\mu\nu} = e\varphi_0\{\partial^\mu(\Delta\varphi_1) - e\varphi_0 A^\mu\}. \qquad (2c)$$

Equation (2b) describes waves whose quanta have (bare) mass $2\varphi_0\{V''(\varphi_0^2)\}^{1/2}$; Eqs. (2a) and (2c) may be transformed, by the introduction of new variables

$$B_\mu = A_\mu - (e\varphi_0)^{-1}\partial_\mu(\Delta\varphi_1),$$

$$G_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu = F_{\mu\nu}, \qquad (3)$$

into the form

$$\partial_\mu B^\mu = 0, \quad \partial_\nu G^{\mu\nu} + e^2\varphi_0^2 B^\mu = 0. \qquad (4)$$

Equation (4) describes vector waves whose quanta have (bare) mass $e\varphi_0$. In the absence of the gauge field coupling ($e=0$) the situation is quite different: Equations (2a) and (2c) describe zero-mass scalar and vector bosons, respectively. In passing, we note that the right-hand side of (2c) is just the linear approximation to the conserved current: It is linear in the vector potential, gauge invariance being maintained by the presence of the gradient term.[5]

When one considers theoretical models in which spontaneous breakdown of symmetry under a semisimple group occurs, one encounters a variety of possible situations corresponding to the various distinct irreducible representations to which the scalar fields may belong; the gauge field always belongs to the adjoint representation.[6] The model of the most immediate interest is that in which the scalar fields form an octet under SU(3): Here one finds the possibility of two nonvanishing vacuum expectation values, which may be chosen to be the two $Y=0$, $I_3=0$ members of the octet.[7] There are two massive scalar bosons with just these quantum numbers; the remaining six components of the scalar octet combine with the corresponding components of the gauge-field octet to describe

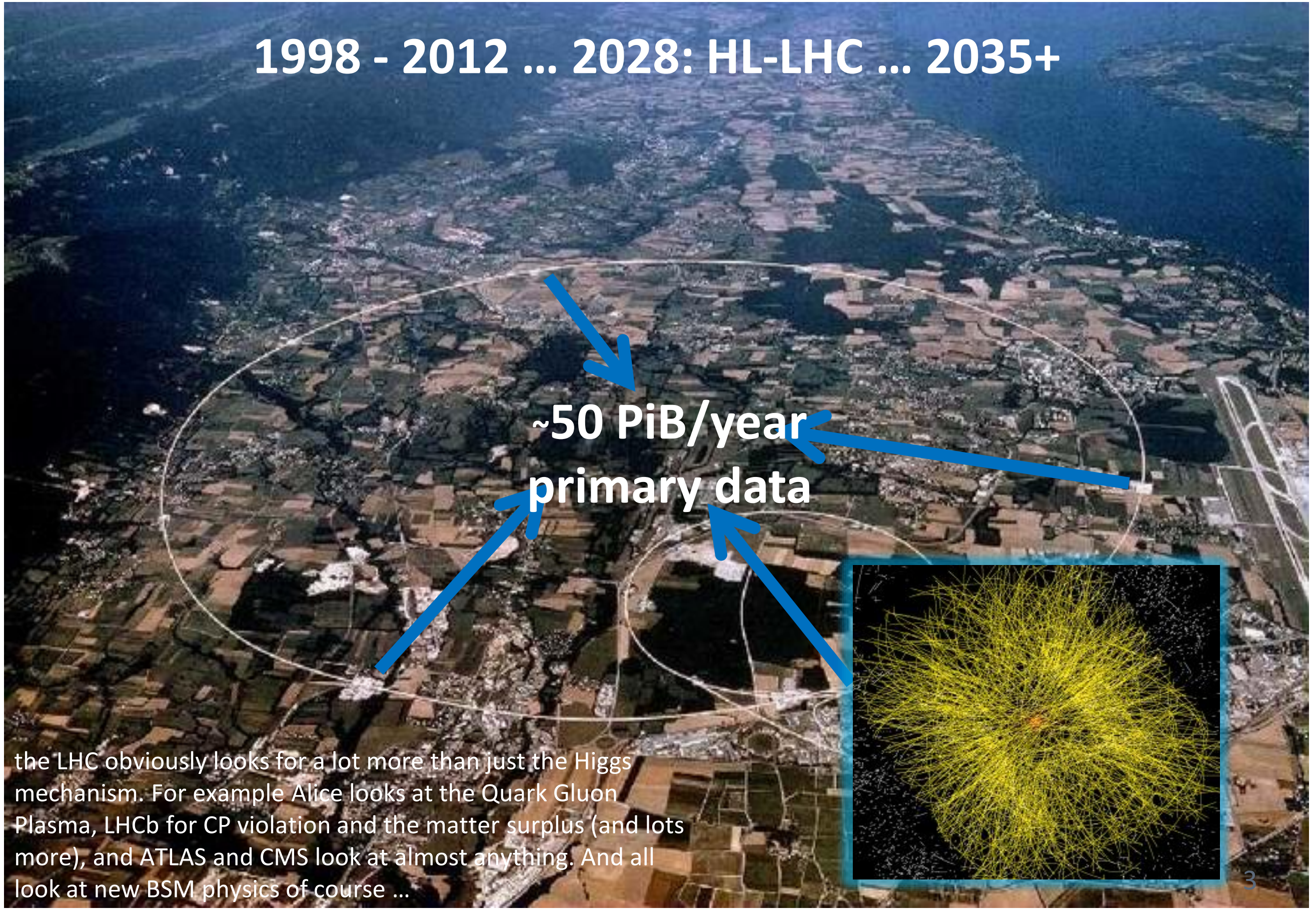
508

321

P. Higgs, Phys. Rev. Lett. 13, 508:

16823 characters, 165 kByte PDF

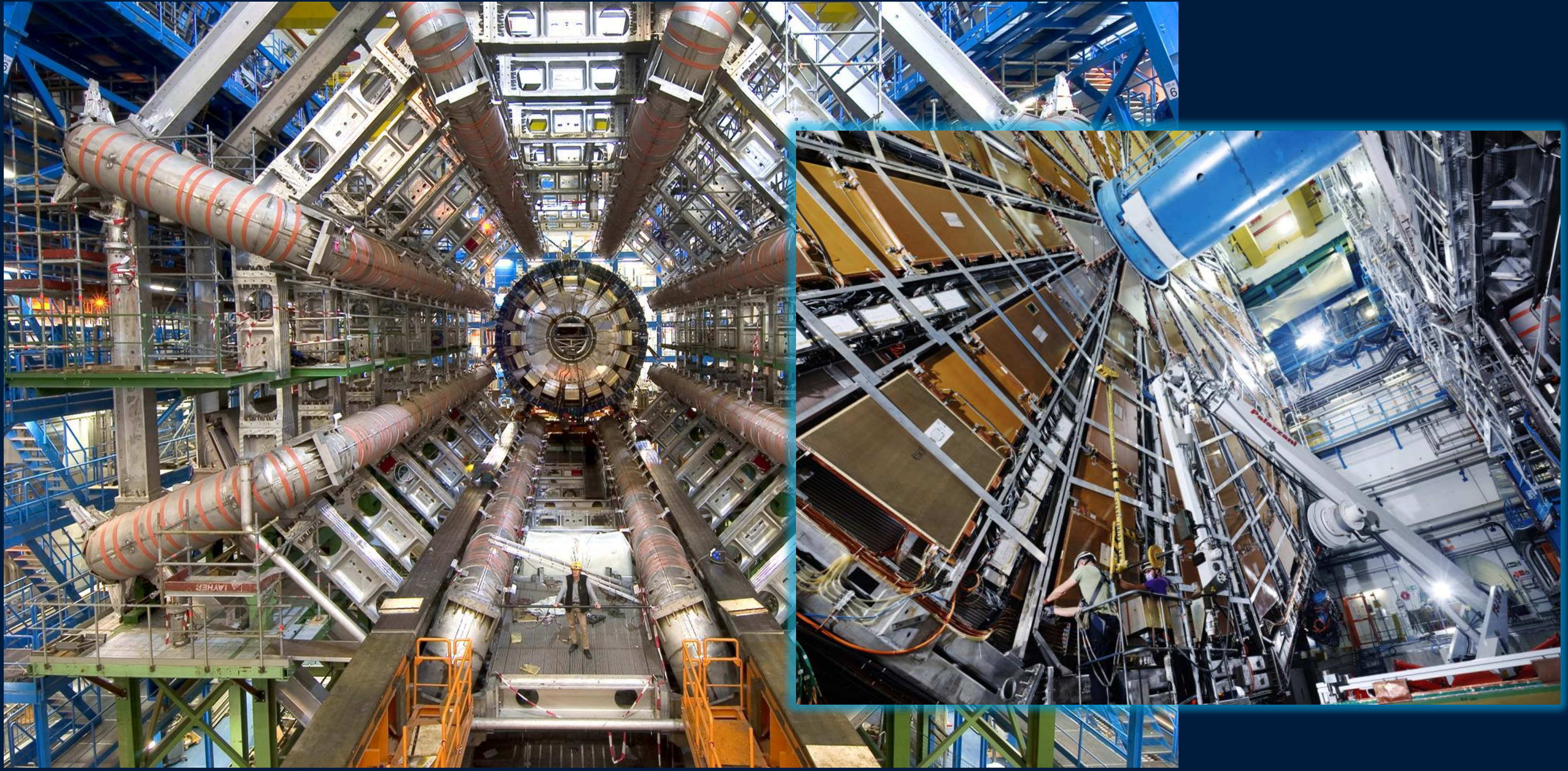
Maastricht University | DACS



1998 - 2012 … 2028: HL-LHC … 2035+

~50 PiB/year primary data

the LHC obviously looks for a lot more than just the Higgs mechanism. For example Alice looks at the Quark Gluon Plasma, LHCb for CP violation and the matter surplus (and lots more), and ATLAS and CMS look at almost anything. And all look at new BSM physics of course …



Images: ATLAS detector in the cavern at CERN. Source: CERN

Computing on lots of data – 40 million times/sec

ATLAS RAW single event
ROD File
1.60 MB

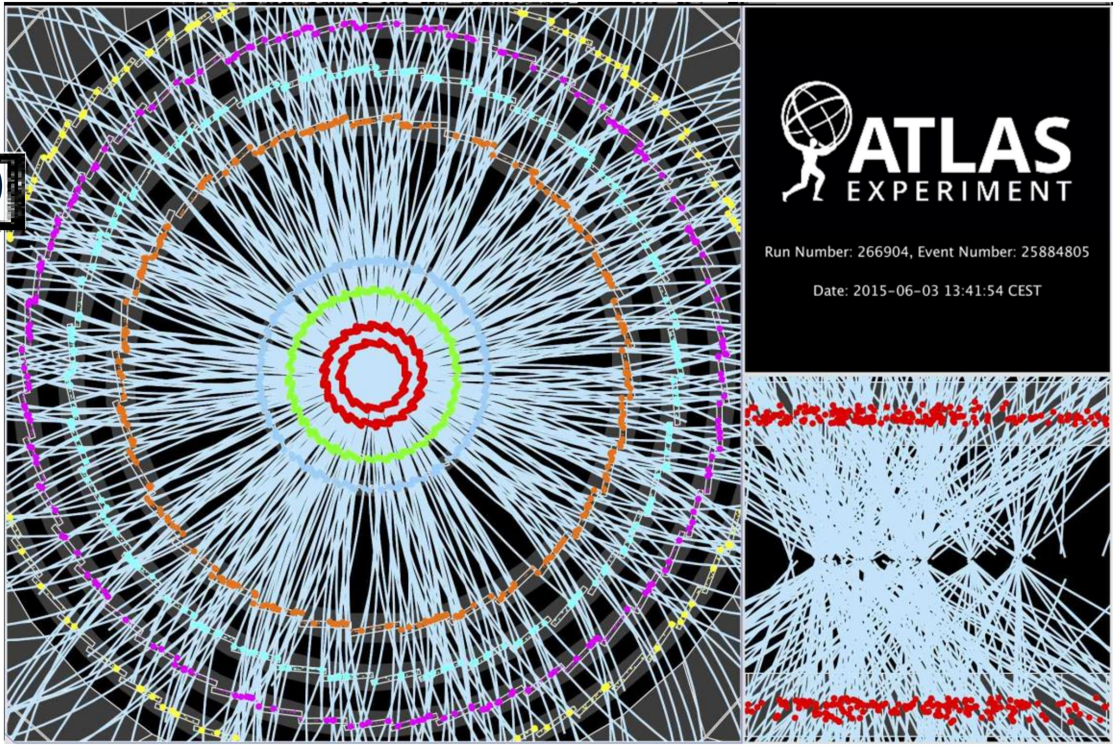
~60 TByte/s (compressed)

Trigger system selects 600 Hz ~ 1 GB/s data

~ 10 seconds compute for a single event at ATLAS with ‘jets’ containing ~30 collisions

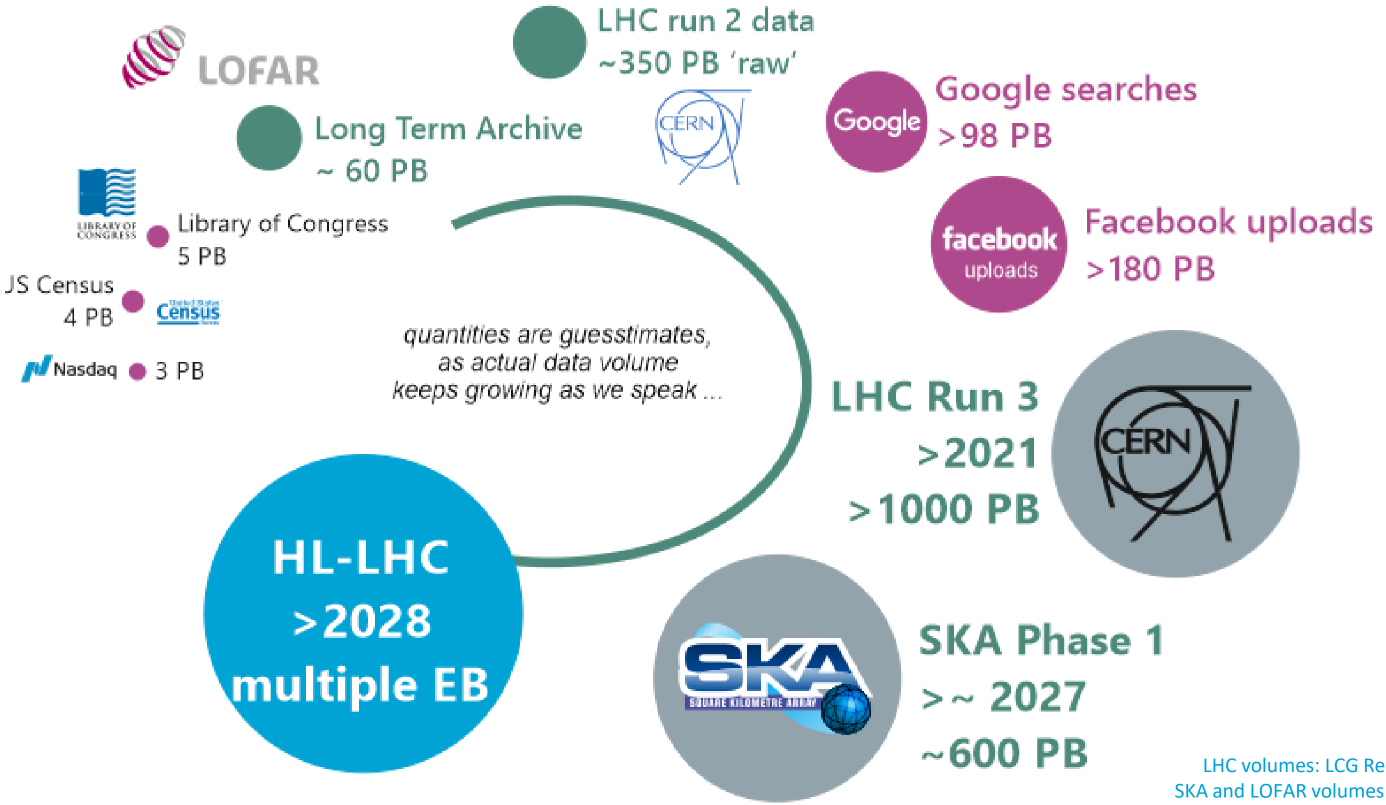
~10k researchers

CERN and ~170 institutes


ATLAS
EXPERIMENT
Run Number: 266904, Event Number: 25884805
Date: 2015-06-03 13:41:54 CEST

Display of a proton-proton collision event recorded by ATLAS on 3 June 2015, with the first LHC stable beams at a collision energy of 13 TeV;
Event processing time: v19.0.1.1 as per Jovan Mitrevski and 2015 J. Phys.: Conf. Ser. 664 072034 (CHEP2015)

Processing at scale for data intensive science



LOFAR

LHC run 2 data
~350 PB 'raw'

CERN

Google searches
>98 PB

Long Term Archive
~ 60 PB

Library of Congress
5 PB

JS Census
4 PB

Nasdaq 3 PB

facebook uploads

Facebook uploads
>180 PB

quantities are guesstimates,
as actual data volume
keeps growing as we speak ...

LHC Run 3
>2021
>1000 PB

HL-LHC
>2028
multiple EB

SKA Phase 1
>~ 2027
~600 PB

Data from various sources, for
public entities: data ca. 2018,
indicative, within ~ factor 2
LHC volumes: LCG Resource Scrutiny Group & CERN; 2020
SKA and LOFAR volumes: ASTRON/Michiel van Haarlem, 2020

So ‘big science’ needs some computing …

CERN Computing Centre B513, image: CERN, https://cds.cern.ch/record/2127440; tape library image CC-IN2P3 with LHC and LSST data; cabinets: Nikhef H234b

Scaling computing infra: volume is not the only thing that matters

COMPLETED PROJECTS

MODELLING INSTITUTIONAL DYNAMICS IN HISTORICAL COMMONS (MIDI)

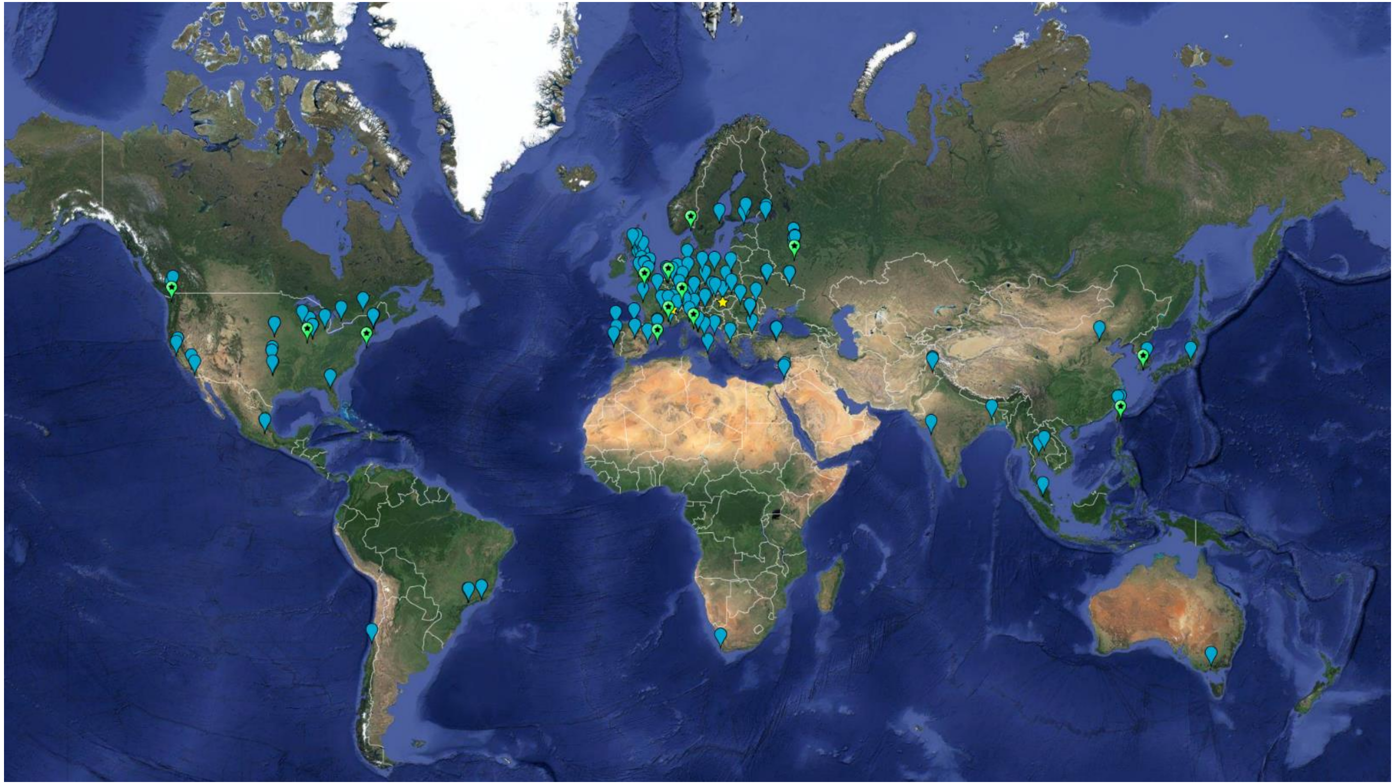
The Modelling institutional dynamics in historical commons (MIDI) project adopted an interdisciplinary perspective to contribute empirically-grounded and systematic knowledge of the mechanisms driving the process of institutional change. Its starting point is a dataset of European commons-management institutions composed by the Research Team Institutions for Collective Action in the Common Rules-project.

The goal of the project was to exploit our interdisciplinary competences to recode the dataset on the basis of a systematic institutional analysis framework, to fully exploit its potential through the use of data mining and evolutionary analysis techniques, and to embed the resulting knowledge in an agent-based model capturing the essential dynamics of institutional change.

The project involves researchers from various research disciplines from Linnaeus University (Sweden), Delft University of Technology (Netherlands), and Tine De Moor and René van Weeren from the *Institutions for Collective Action Research Team* (then at Utrecht University , The Netherlands). The project is funded by the Riksbankens Jubileumsfond and ran from 1 January 2018 until the end of 2019.

Sources: CERN https://wlcg.web.cern.ch/; HADDOCK, WeNMR, @Bonvinlab https://wenmr.science.uu.nl/; Virgo, Pisa, IT; SKAO: the SKA-Low observatory, Australia https://www.skatelescope.org/ - OpenMOLE simulation on EGI - https://cdn.egi.eu/app/uploads/2022/04/EGI_Use_Cases.pdf; agent-based modelling of ICAs: https://collective-action.info/research-on-icas/ Molood Dehkordi (TUDelft), Tine de Moor (EUR RSM)

Not in one place: the worldwide LHC Computing Grid



Earth background: Google Earth; Data and compute animation: STFC RAL for WLCG and EGI.eu; Data: https://home.cern/science/computing/grid

~ 1.4 million CPU cores
~ 1500 Petabyte
disk + archival

170+ institutes
40+ countries
13 'Tier-1 sites'
NL-T1:
SURF & Nikhef

largely based on
generic e-Infrastructures
EGI
EuroHPC
PRACE-RI
OpenScienceGrid
ACCESS-CI

For the LHC Computing Grid: wlcg.web.cern.ch, for EGI: www.egi.eu; ACCESS (XSEDE): https://access-ci.org/, for the NL-T1 and FuSE: fuse-infra.nl, https://www.surf.nl/en/research-it

Our journey today …

let's build some 'scalable' infrastructure for LHC computing, storage, networking, and a global AAI … *if we make it*

Using science use cases *from CERN's Large Hadron Collider, the SKA radio telescope, Gravitational Wave detection, structural biochemistry (WeNMR), and more …*

From the bottom up … of green fields, ships' diesels, and chilly corridors

Data intensive workflows that drive infrastructure development

- **why large-scale IT is distributed**: end of faster CPUs, thermal barrier, rise of parallelism

More than one …

- **High Performance & High Throughput**: distributed computing, storage and data placement
- **As a service:** herding systems, cloud platforms, containers, and service management

---

Networking the systems: linking 'more than one' globally

- **network design**: elephants vs. mice in shipping large quantities of data … and on cat videos
- *Optical Private Networks* and the *Open Networking Environment LHCone*

Networking the people

- **authentication and authorization** technologies
- **multilateral federation:** identity, community management & global trust

Putting it all together again *(and maybe an example of a federated anycasted authentication service)*

Start with …
a green field approach?

from field to facility

From field to facility

Trekkersveld IV, Zeewolde. From Zeewolde Actueel, https://www.zeewolde-actueel.nl/nieuws/gemeente/254432/bestemmingsplan-trekkersveld-4-ligt-ter-inzage; Microsoft DC Middenmeer, from https://nos.nl/l/2512478,

Feel the Power

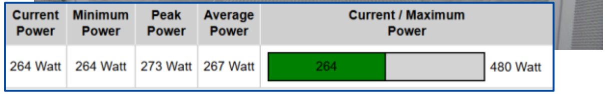
Images: Anton Mors, David Groep, Nikhef

Converting electricity into … chilled air & heat

Left-side image: frame from a movie by Anton Mors, people replaced by … Adobe Firefly (“without people”?, oh well, this was its best result ☹)

NikhefHousing: a cold aisle

Where to put large-scale IT: brief look at data centres



Current Power	Minimum Power	Peak Power	Average Power	Current / Maximum Power	
264 Watt	264 Watt	273 Watt	267 Watt	264	480 Watt

- ‘tier-1’ … ‘tier-4’ datacenters - increasingly redundant
- all systems are ‘lights out’, since the DC may be miles away
  - remotely controlled, incl. power-on, remote KVM
- small and large in terms of power and cooling capacity
  - smallish: Nikhef Housing Amsterdam is ~2.5 MW,
  - Meta Zeewolde (now cancelled) would have been 160 MW
- data centre efficiency metric: $PUE = \frac{E_{total}}{E_{IT\_equipment}}$

Reducing cost and impact by improving “Power Unit Efficiency” of the data centre:

- airflow engineering and efficient CRACs
- (free) cooling by changing inflow temperature
- Aquifer Thermal Energy Storage (ATES) to buffer heat (and re-use later for homes)

Typical PUEs vary from 1.03 (in Iceland) to 1.2 for ‘good’ datacenters in NL

Data centre tiering: Uptime Institute (Tunner, W.P.; Seader, J.H.; Brill, K.G. Tier Classifications Define Site Infrastructure Performance; White Paper)

Remote systems management: IPMI, Redfish and various vendor proprietary solutions – usually dedicated ‘out-of-band’ network connection, incl. remote KVM

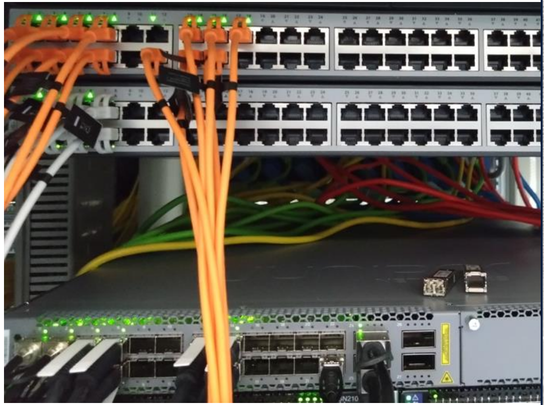
Every rack should have one

- A bare rack just lacks that nice and warm feeling, so you typically add

- some remotely monitored PDUs
- temperature sensor(s)

- out-of-band management switches
- systems installation net (managed)

- data, storage, and overlay networks dual 10/25/100GigE per system
  + optionally a low-latency fabric for HPC, like InfiniBand, RoCE, UltraEthernet …



Shown: H234b C06 ‘SOC’ cabinet, Nikhef, front and switches (at back)

Virtual and cloud services rely on this physical ‘stuff’

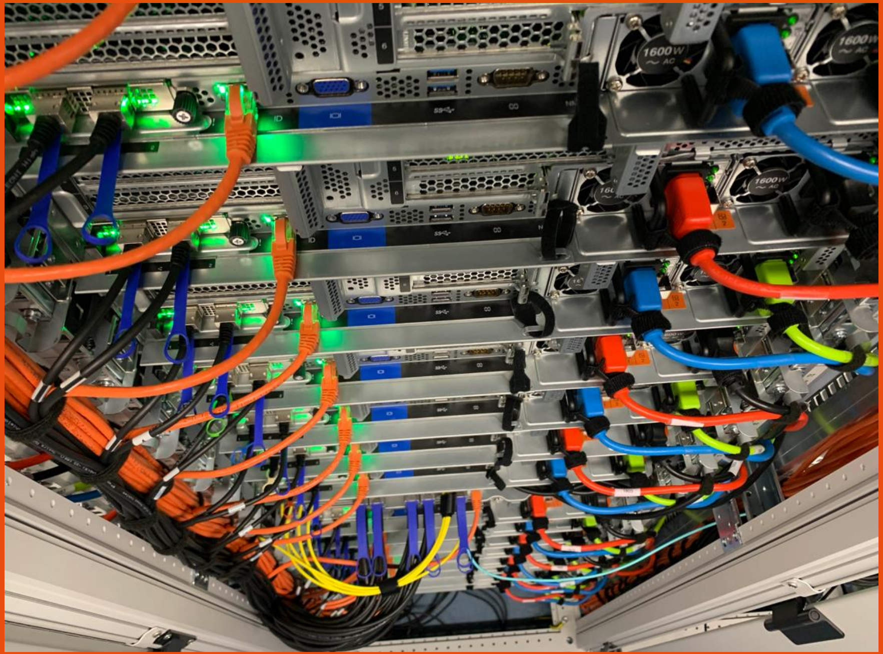
- HPC systems like the Dutch Snellius, a SuperMUC, LUMI, JUPITER, or Jules Verne,
- data-intensive computing like WLCG, radio astronomy, and so on
- your favourite (or not) typical hyperscalers like AWS, Azure, Google, OVH, Hetzner, …

and all those new AI systems and AI ‘factories’ that boost Nvidia stock nowadays …

DNI and NL-T1 capacity from 2023 DNI NWO, LOFAR, and WLCG; see https://www.surf.nl/onderzoek-ict/toegang-tot-rekendiensten-aanvragen ; fuse-infra.nl
SURF tape total: ~80 PByte by end 2022; image library at Schiphol Rijk from Sara Ramezani; NikhefHousing: https://www.nikhef.nl/housing/datacenter/floorplan/

Filling the Data Center

The challenges come
when you have ‘more than one’



Different types of large scale compute resources

- HPC and (computational) cluster computing:
  - modelling for weather/climate, fluid dynamics, but also e.g. QC-simulation

- HTC and data-intensive processing – horizontal scaling:
  - lots of data, as in High Energy Physics (HEP), *omics and protein docking, …
  - conveniently parallel,
    but (intensive) local I/O requirements on memory and scratch storage

- portals and many web applications:
  ‘horizontal’ scaling, often backed by cloud and virtualized resources
  - Cloud-native scaling and containers for ‘more of the same, different each time’
  - If it’s data at scale: object stores and ‘CDN’ web-scale caching

HPC: High Performance Computing; HTC: High Throughput Computing; K8S: Kubernetes; CDN: Content Delivery Network

Single CPU scaling stopped around 2004

- limitation is power, not circuit size
  - and clock frequency is most ‘power-hungry’
  - still some packages now @ TDP of 400W

- multiple cores on the same die helps:
  - AMD EPYC Genoa (Zen 4) has 96 cores/die
  - Intel Granite Rapids, Nvidia GraceHopper, …
  - but e.g. Intel Cascade Lake AP was less useful

- CPU design-level performance gains left
  - predictive and out-of-order execution
  - on-die parallelism (multi-core)
  - pre-fetching and multi-tier caching
  - execution unit sharing (‘SMT’)

  but at increased risk for security/integrity

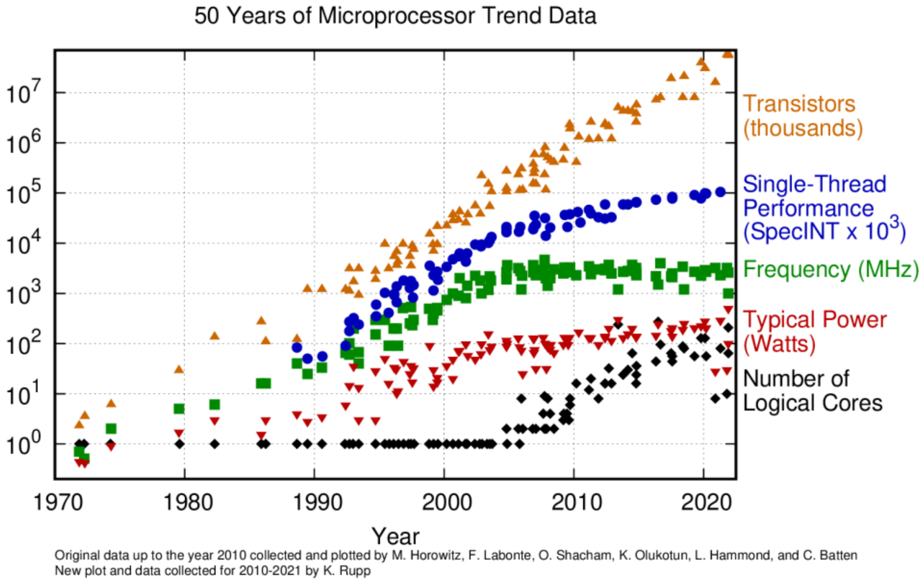
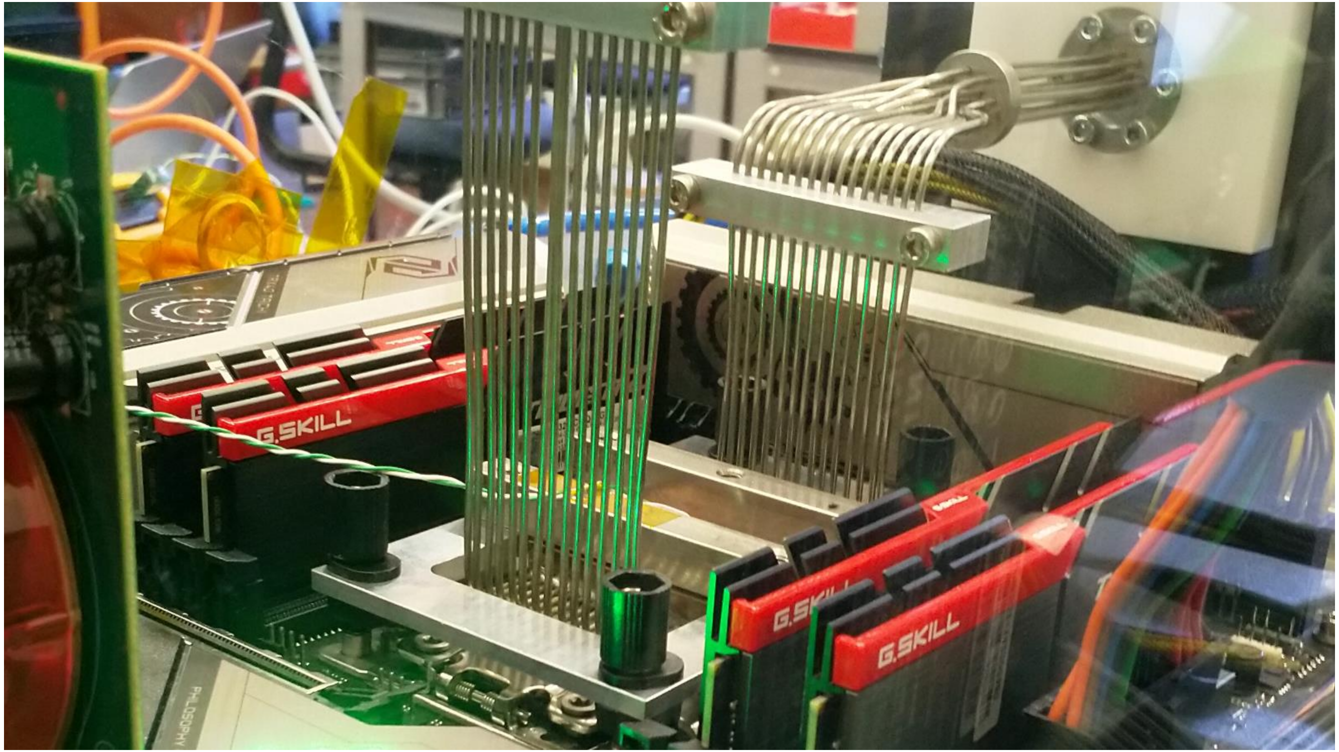


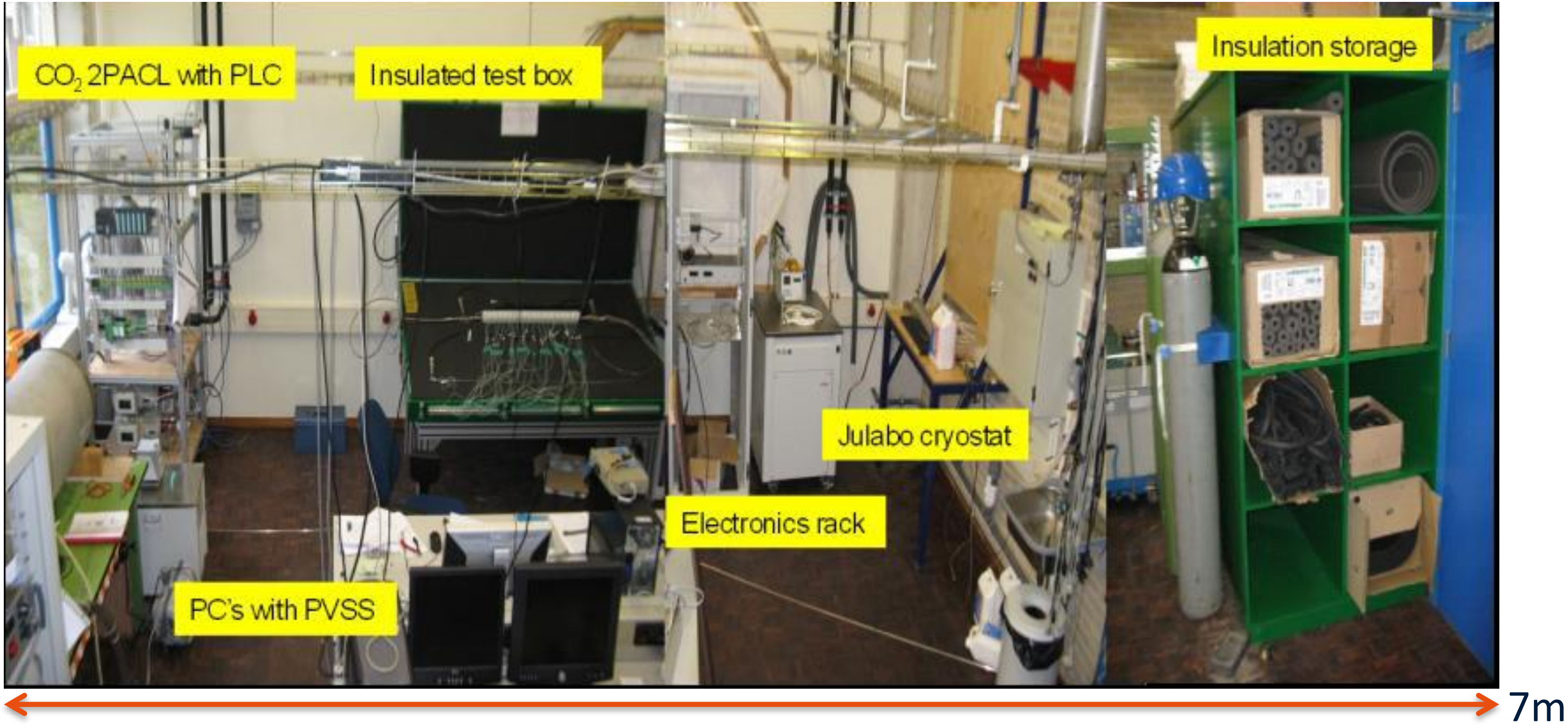
Image: K Rupp, https://github.com/karlrupp/microprocessor-trend-data

Fix the thing that didn’t scale well, CPU frequency??



LCO2 cooling of an AMD Ryzen Threadripper 3970X [56.38 °C] at 4600.1MHz processor (~1.25x nominal speed) sustained over all cores simultaneously, using the Nikhef LCO2 test bench system (https://hwbot.org/submission/4539341) - (Krista de Roo en Tristan Suerink)

… since you then need this around it …



Nikhef 2PA LCO2 cooling setup. Image from Bart Verlaat, Auke-Pieter Colijn *CO2 Cooling Developments for HEP Detectors* https://doi.org/10.22323/1.095.0031

proceed to clusters ▶

Step one: scale *inside* one system

- ‘trivial’ step-up is to do multiple sockets in one system
  2-socket, sometimes 4 socket on a motherboard

- to make it appear as a single shared memory system,
  cache coherency is required between the CPUs

- useful for tightly coupled parallel applications
  (weather forecasting, fluid dynamics, climate), but
  not needed for ‘trivially parallel’ high throughput needs

- depending on architecture cache coherency
  kills single-thread performance (although AMD did lot better here than the Intel *lakes)

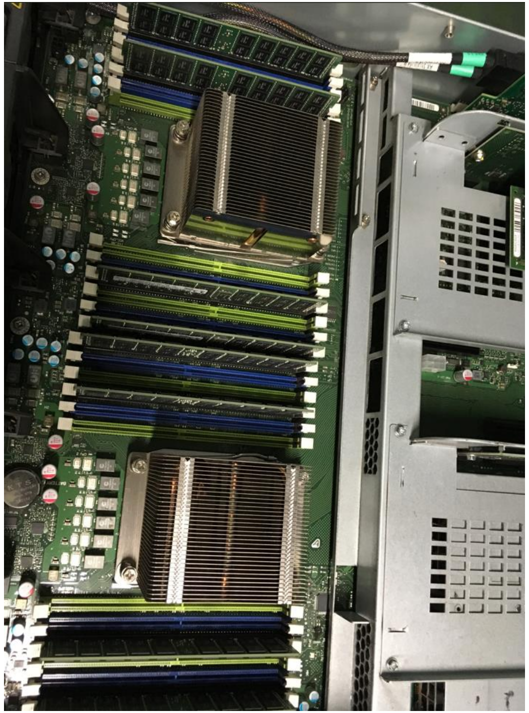


Image: dual-socket Fujitsu system at the Xenon experiment site, 2019. source: Tristan Suerink, Nikhef

CPU design changes may fit application, or not

AMD EPYC effective for applications like WLCG:

- Naples → Rome added shared memory die
- links all cores directly to memory

Rome-Milan improvement?

- shared L3 cache benefits tightly coupled HPC, but not ‘off-die memory’ limited HTC

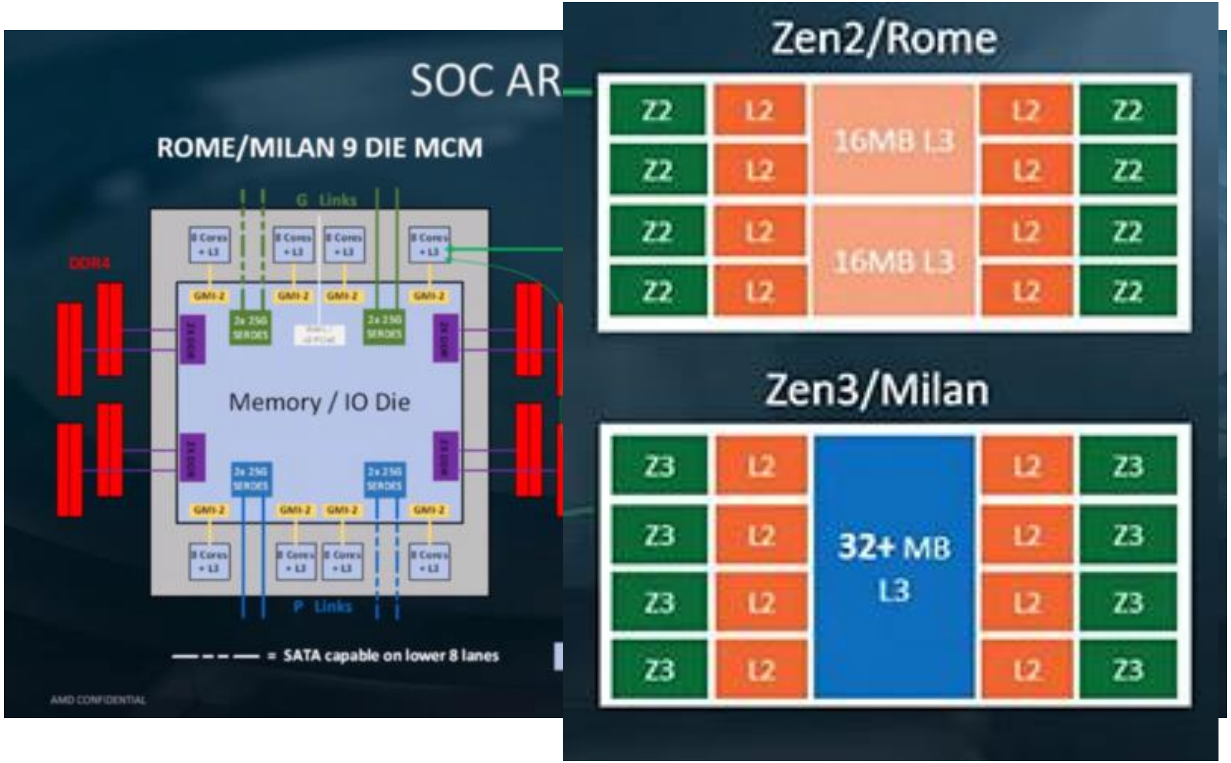
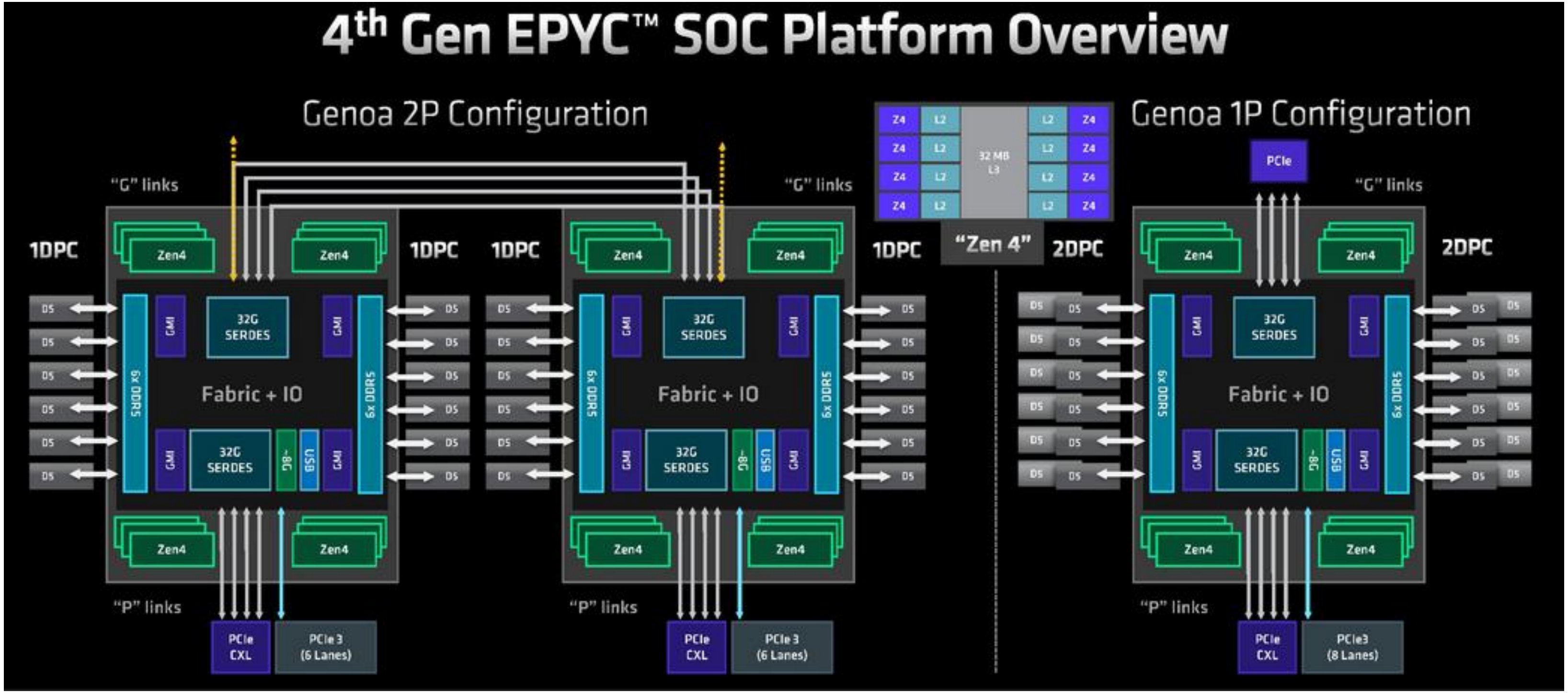


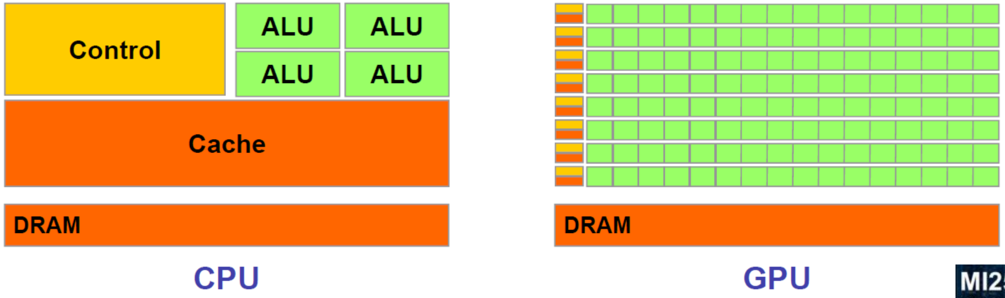
Image source: AMD, retrieved from https://m.hexus.net/tech/news/cpu/135479-amd-shares-details-zen-3-zen-4-architectures/

Scaling up, more examples



AMD EPYC Genoa platform, image from https://www.semianalysis.com/p/amd-genoa-detailed-architecture-makes

Accelerators – general purpose GPUs



leaving FPGAs out for a moment – but those are particularly useful in guaranteed-latency scenarios!

- but co-processing comes at a cost of moving data to and from the GPU
- often faster to keep computing and do selection & conditionals later
- computation speed heavily depends on precision (even 4-bit precision is used)
- quite power hungry!

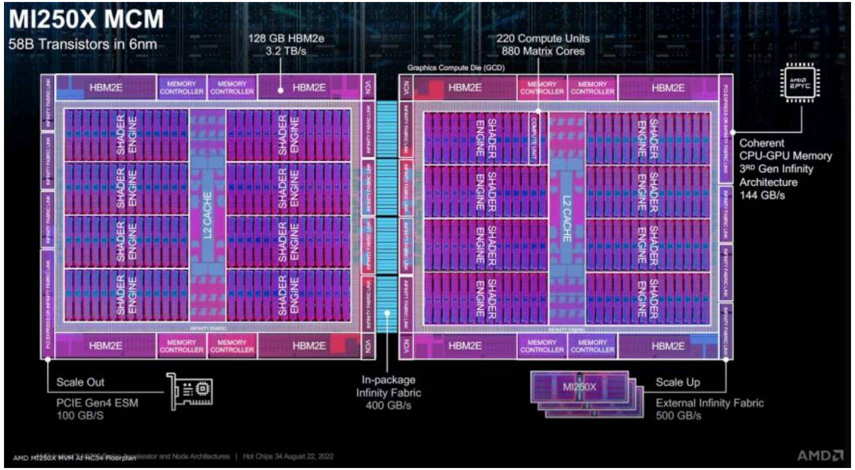
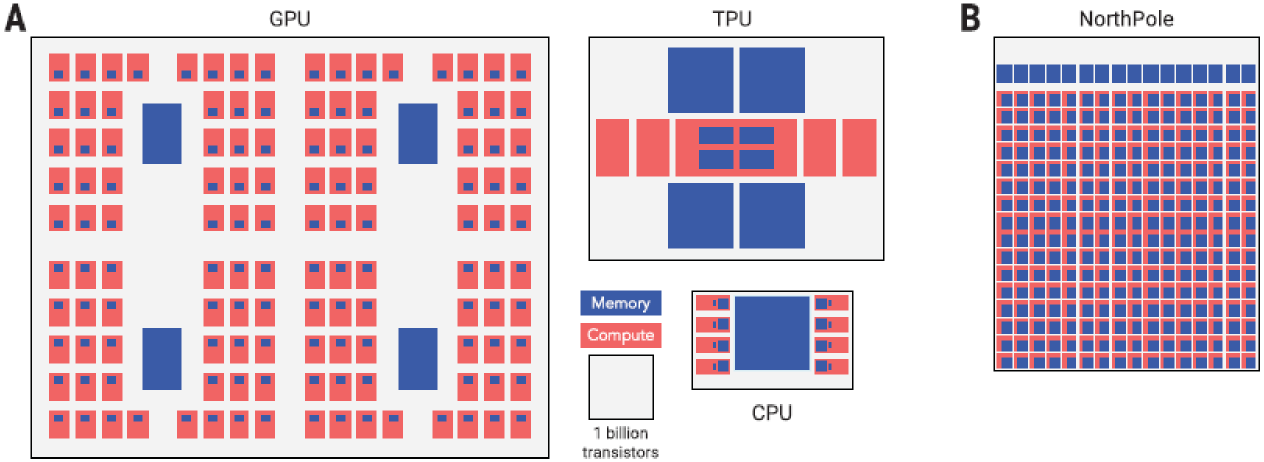


Image: ‘Massively Parallel Computing with CUDA’, Antonino Tumeo Politecnico di Milano, https://www.ogf.org/OGF25/materials/1605/CUDA_Programming.pdf
Floorplan image of die: AMD MI250 GPU, slide source: AMD

Aiming to remove the data access bottleneck

Separating memory from processing introduces the memory misses that slow down CPU processing as well GPUs due to need for (RDMA) main memory access

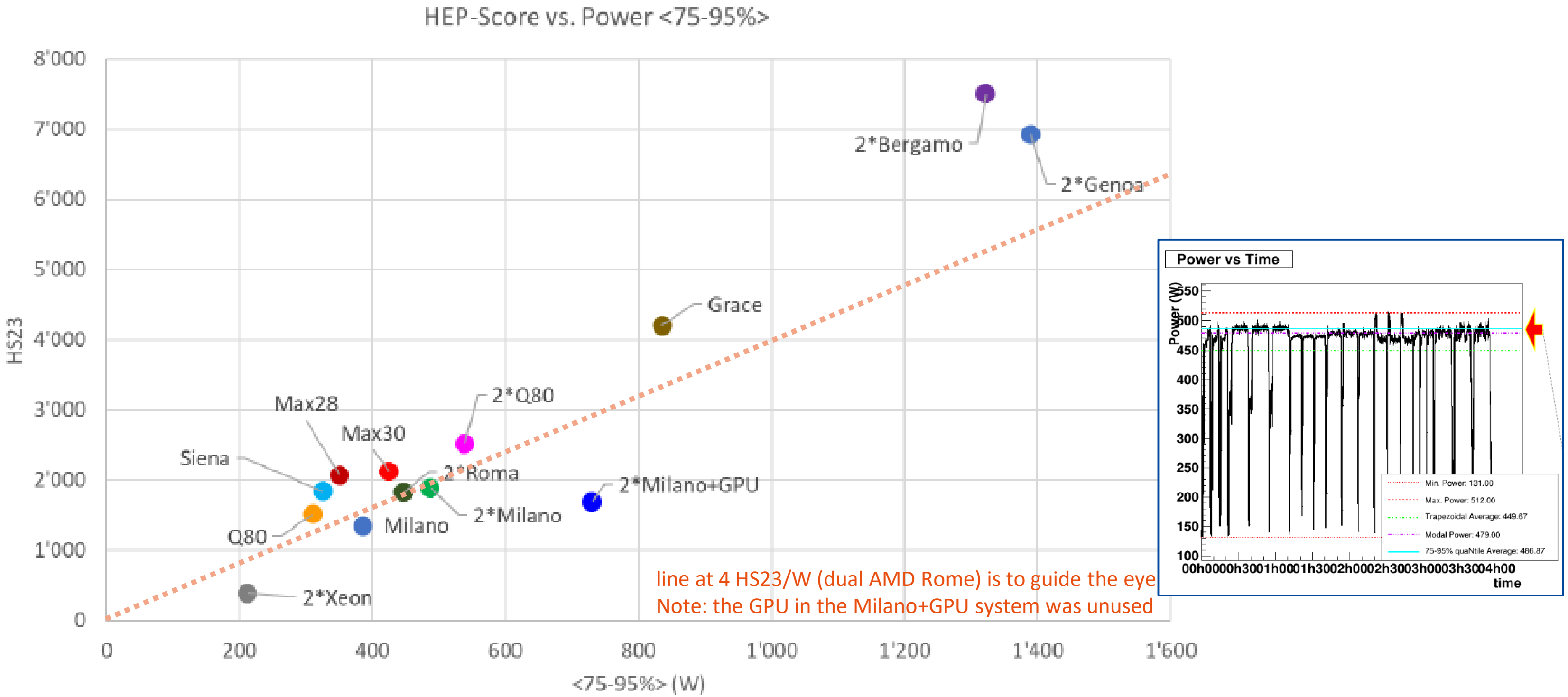
Some very recent designs aim to eliminate this by temporal co-location of program and memory (IBM NorthPole AI, Oct ‘23) with data-flow driven compute



Physical organization of on-chip memory (blue) and compute (red) are diagrammed for representative processors, scaled to constant transistors per unit area. From Modha’s paper Modha et al., *Science* **382**, 329–335 (2023)

Modha *et al.* https://doi.org/10.1126/science.adh1174 or read https://research.ibm.com/blog/northpole-ibm-ai-chip
PCIe card photo from https://www.ibm.com/blogs/solutions/jp-ja/northpole-ibm-ai-chip/

The energy bottleneck: architecture figure of merit



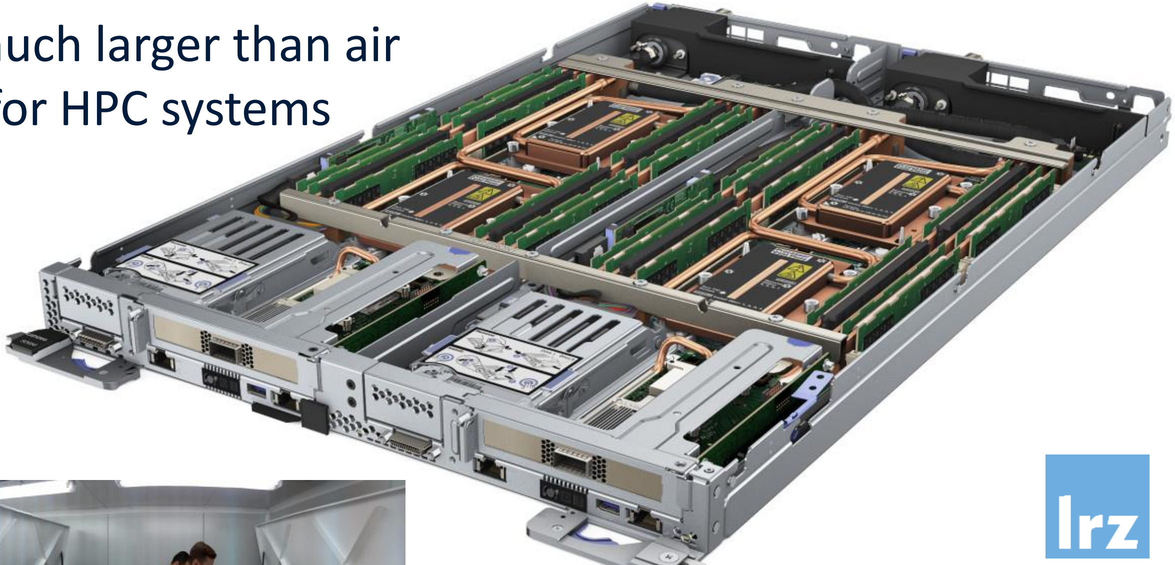
line at 4 HS23/W (dual AMD Rome) is to guide the eye
Note: the GPU in the Milano+GPU system was unused

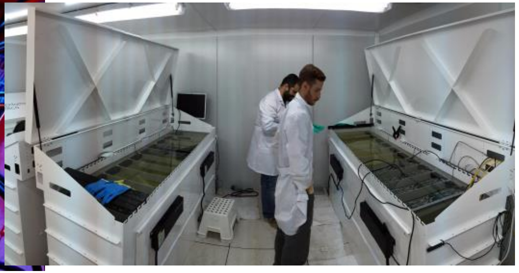
Data and graphs: Emanuele Simili, Glasgow University, at CHEP2024 (https://indico.cern.ch/event/1338689/contributions/6011562/)
HEPSPEC23 benchmark: https://gitlab.cern.ch/hep-benchmarks/hep-benchmark-suite ('memory-intensive' high throughput processing application benchmark)

How to get this heat out … in liquid form, maybe?

- Heat capacity of liquid is much larger than air
- by now (almost) standard for HPC systems

- immersive systems look cool, but are ‘a bit hard’ on maintenance





Strongly depends on systems engineering: when water inlet temperature can be >40 degC, you have almost always free cooling

Image source dual-board system: Lenovo, ThinkSystem SD650

immersive cooling image https://hypertec.com/blog/sustainable-emerging-tech-liquid-immersion-cooling/, PIC T1 centre, Barcelona, ES

And if large-scale IT does not quite fit … ahum …

Image source: https://lambdalabs.com/products/blade

SuperMicro (branded as ‘Lambda Blade’)
4U chassis, supporting 10 consumer-grade GPUs …
… with a bump

but there *is* a serious issue here!

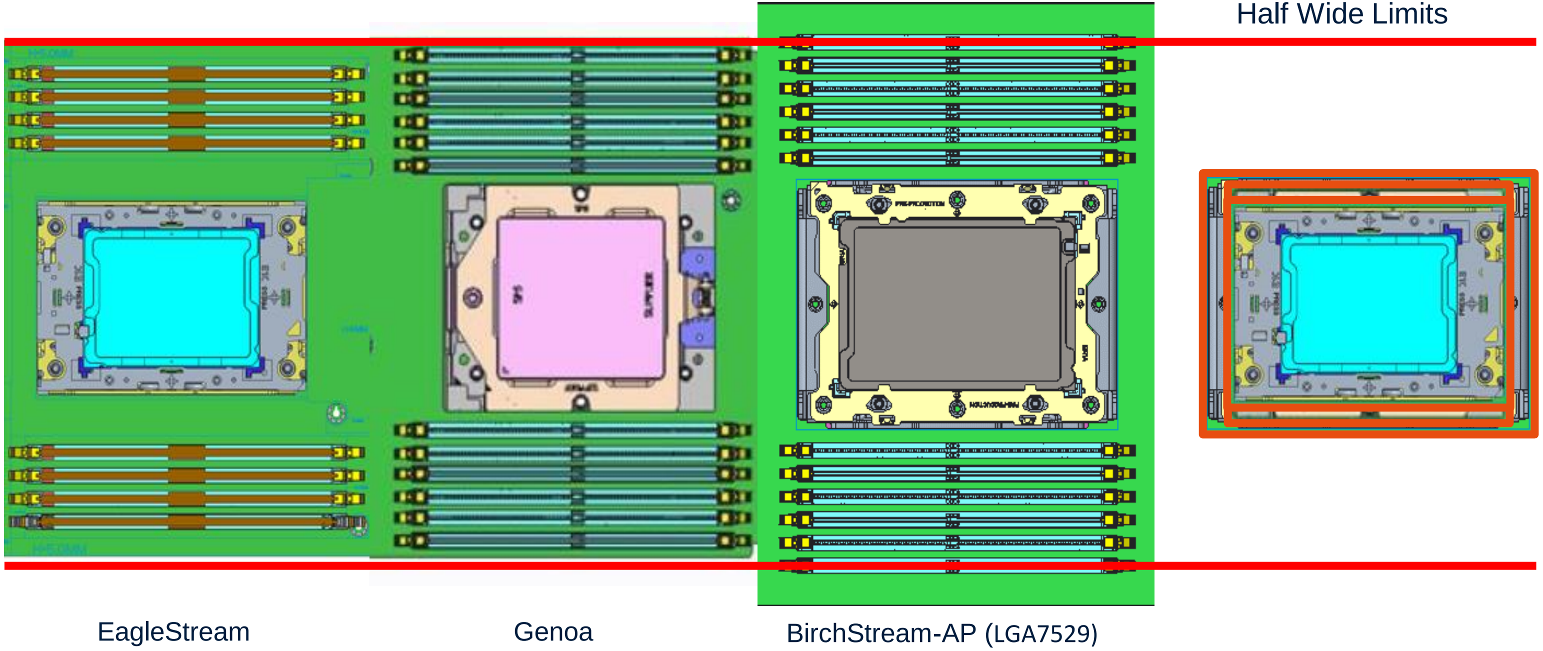


Image thanks go to Rick Koopman – Lenovo at the HTCondor Workshop 2024 https://indico.cern.ch/event/1386170/

Scaling up – beyond one lone motherboard

‘compute farming’: milking computers, in a balanced way

Data-driven workloads (like WLCG, SKA, WeNMR) need more than just compute

- **balanced features** for node throughput CPU, storage, memory bandwidth & latency, NIC & network speed

- **single-socket** multicore systems are fine, typically 64-128 cores per system
- **network**: 2x25Gbps (match #cores)
- **memory**: say ~ 8 GiB/core
- **local disk**: 8-16 TB NVME (~100GB/core)

+ space (physical + power) to add some **GPU**

Image: Cluster ‘Lotenfeest’ at the Nikhef NDPF, acquired March 2020. Lenovo SR655 with AMD EPYC 7702P 64-Core single-socket

But … fancy an interactive console install?

Images: Nikhef Housing H234b NDPF science processing data centre

Managing multiple system (physical or virtual)

Fabric (Configuration) Management

- do you know what is out there?
- update quickly & consistently when vulnerabilities are found?
- versioned repository for rollback?

note that not all tooling scales in itself

- **push**: ansible (using ssh logins), or home-brew scripting
- **pull**: each node runs its own actions, e.g. Saltstack, Ansible-agent, Quattor, Chef, …

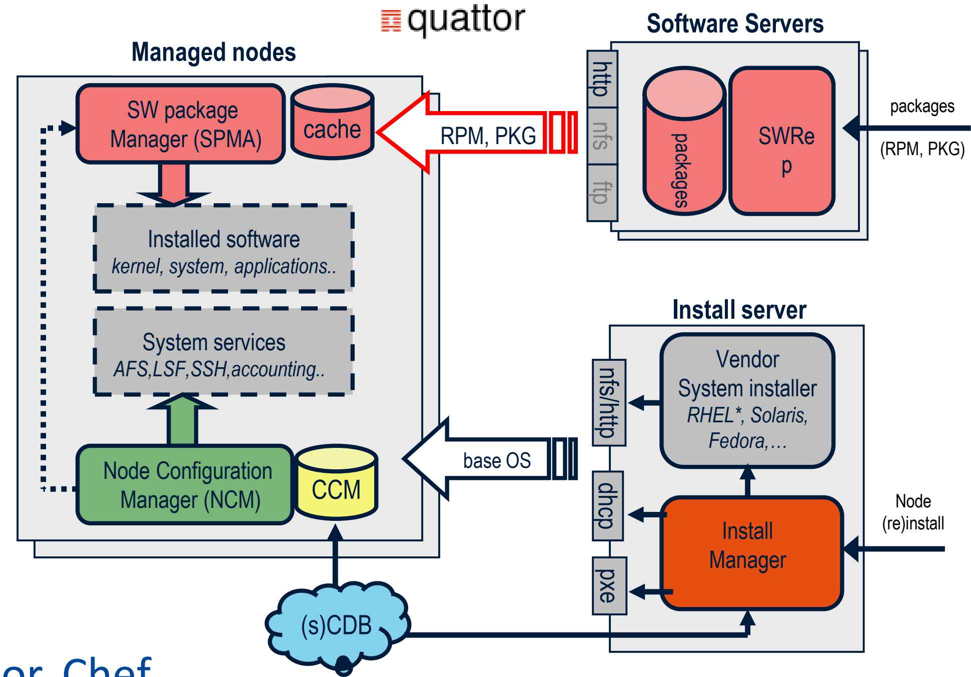
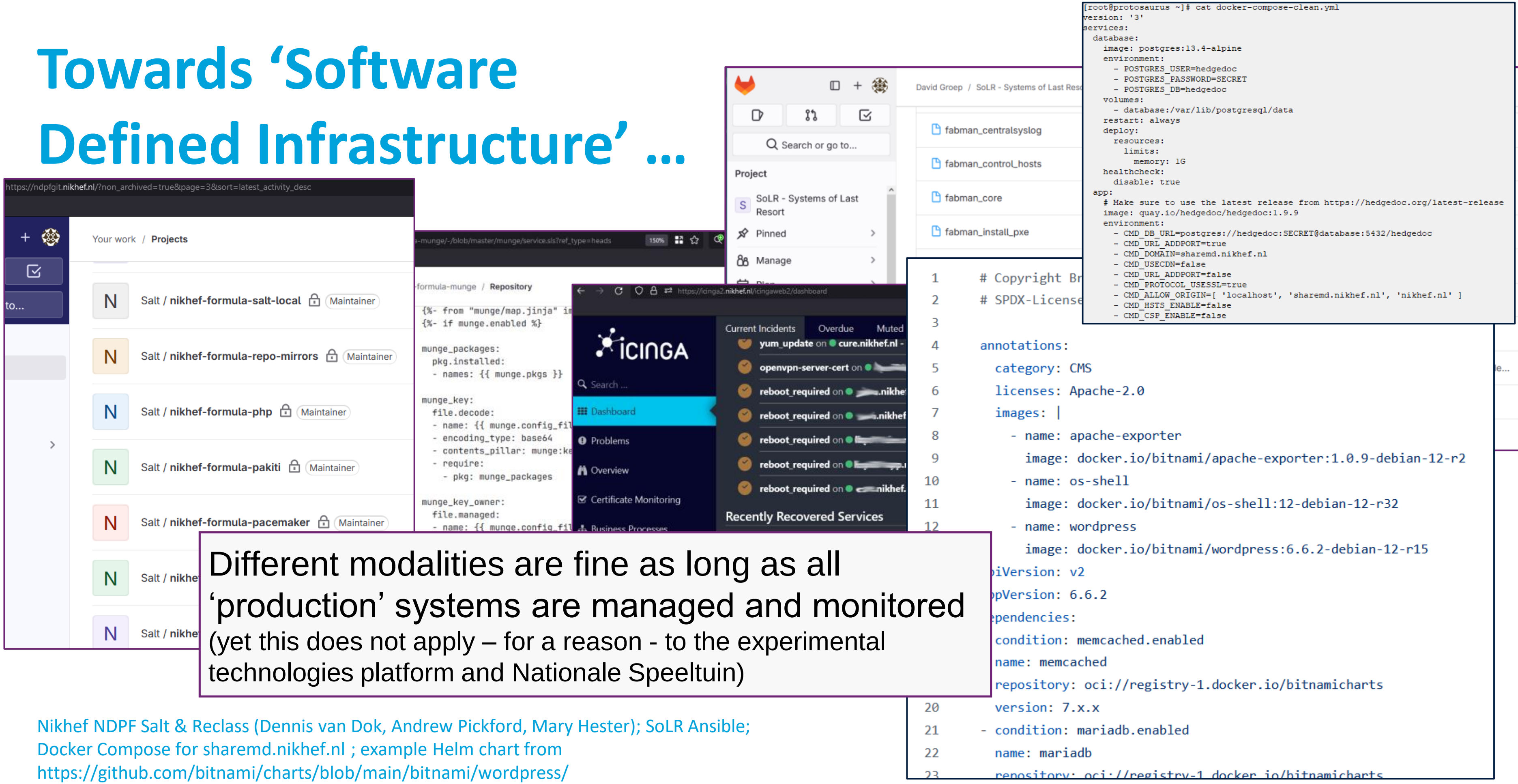


Illustration: German Cancio, CERN, quattor.org, used here as example; see also: ansible.com, saltproject.io, theforeman.org, cfengine.com, puppet.com, …

Towards 'Software Defined Infrastructure' …



https://ndpfgit.nikhef.nl/?non_archived=true&page=3&sort=latest_activity_desc

Your work / Projects

- N Salt / nikhef-formula-salt-local 🔒 Maintainer
- N Salt / nikhef-formula-repo-mirrors 🔒 Maintainer
- N Salt / nikhef-formula-php 🔒 Maintainer
- N Salt / nikhef-formula-pakiti 🔒 Maintainer
- N Salt / nikhef-formula-pacemaker 🔒 Maintainer
- N Salt / nikhe
- N Salt / nikhe

-munge/-/blob/master/munge/service.sls?ref_type=heads

-formula-munge / Repository

```
{%- from "munge/map.jinja" im
{%- if munge.enabled %}

munge_packages:
  pkg.installed:
  - names: {{ munge.pkgs }}

munge_key:
  file.decode:
  - name: {{ munge.config_fil
  - encoding_type: base64
  - contents_pillar: munge:ke
  - require:
    - pkg: munge_packages

munge_key_owner:
  file.managed:
  - name: {{ munge.config_fil
```

https://icinga2.nikhef.nl/icingaweb2/dashboard

ICINGA

Current Incidents | Overdue | Muted

- yum_update on ● cure.nikhef.nl -
- openvpn-server-cert on ● [redacted]
- reboot_required on ● [redacted].nikhef
- reboot_required on ● [redacted].nikhef
- reboot_required on ● [redacted]
- reboot_required on ● [redacted]
- reboot_required on ● [redacted]nikhef.

Recently Recovered Services

David Groep / SoLR - Systems of Last Reso

- fabman_centralsyslog
- fabman_control_hosts
- fabman_core
- fabman_install_pxe

```
[root@protosaurus ~]# cat docker-compose-clean.yml
version: '3'
services:
  database:
    image: postgres:13.4-alpine
    environment:
      - POSTGRES_USER=hedgedoc
      - POSTGRES_PASSWORD=SECRET
      - POSTGRES_DB=hedgedoc
    volumes:
      - database:/var/lib/postgresql/data
    restart: always
    deploy:
      resources:
        limits:
          memory: 1G
    healthcheck:
      disable: true
  app:
    # Make sure to use the latest release from https://hedgedoc.org/latest-release
    image: quay.io/hedgedoc/hedgedoc:1.9.9
    environment:
      - CMD_DB_URL=postgres://hedgedoc:SECRET@database:5432/hedgedoc
      - CMD_URL_ADDPORT=true
      - CMD_DOMAIN=sharemd.nikhef.nl
      - CMD_USECDN=false
      - CMD_URL_ADDPORT=false
      - CMD_PROTOCOL_USESSL=true
      - CMD_ALLOW_ORIGIN=[ 'localhost', 'sharemd.nikhef.nl', 'nikhef.nl' ]
      - CMD_HSTS_ENABLE=false
      - CMD_CSP_ENABLE=false
```

```
# Copyright Br
# SPDX-License

annotations:
  category: CMS
  licenses: Apache-2.0
  images: |
    - name: apache-exporter
      image: docker.io/bitnami/apache-exporter:1.0.9-debian-12-r2
    - name: os-shell
      image: docker.io/bitnami/os-shell:12-debian-12-r32
    - name: wordpress
      image: docker.io/bitnami/wordpress:6.6.2-debian-12-r15
piVersion: v2
pVersion: 6.6.2
ependencies:
  condition: memcached.enabled
  name: memcached
  repository: oci://registry-1.docker.io/bitnamicharts
  version: 7.x.x
- condition: mariadb.enabled
  name: mariadb
  repository: oci://registry-1.docker.io/bitnamicharts
```

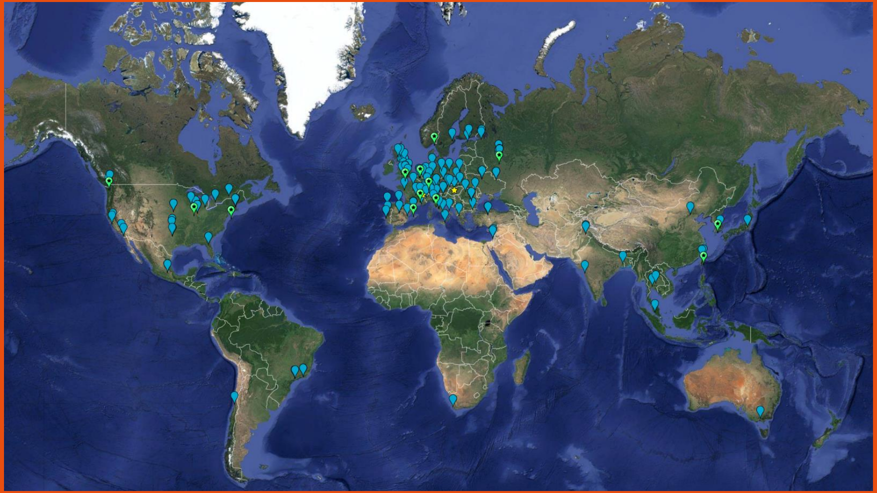
Different modalities are fine as long as all 'production' systems are managed and monitored

(yet this does not apply – for a reason - to the experimental technologies platform and Nationale Speeltuin)

Nikhef NDPF Salt & Reclass (Dennis van Dok, Andrew Pickford, Mary Hester); SoLR Ansible; Docker Compose for sharemd.nikhef.nl ; example Helm chart from https://github.com/bitnami/charts/blob/main/bitnami/wordpress/

More of *more than one …*

The physical layer … and managing software-defined infrastructure

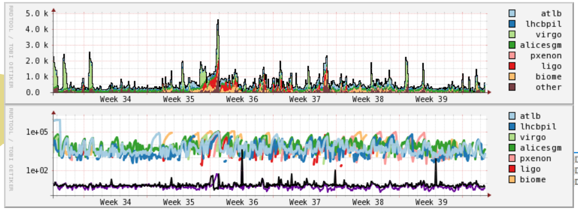


Large-scale IT: worldwide LHC Computing and beyond (2024 ed)

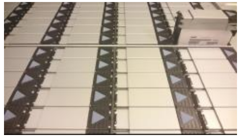
proceed to cloud management ▶

Cluster computing and ‘conveniently parallel’ HTC

?



```
DEF=lhcalice
DEF=lhcalice
DEF=lhcalice
GROUPCFG[auger]      FSTARGET=3    PRIORITY=200    MAXPROC=500     QDEF=augerbig
GROUPCFG[augsgm]     FSTARGET=1    PRIORITY=300    MAXPROC=2       QDEF=augerbig
QOSCFG[augerbig]     FSTARGET=3

# if these are queued, they will generally be of highest priority.
# limit their MAXIJOBs ... we really want two non-ATLAS VOs to be
# of rank higher than ATLAS before we drain the multicore pool.

GROUPCFG[virgo]      FSTARGET=25    PRIORITY=200    MAXPROC=2700   MAXIJOB=10 QDEF
=biggrid
GROUPCFG[ligo]       FSTARGET=23    PRIORITY=200    MAXPROC=2700   MAXIJOB=10 QDEF
=biggrid

# local groups

GROUPCFG[atlas]      FSTARGET=10   PRIORITY=200    MAXPROC=2200   QDEF=niklocal
```

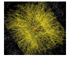


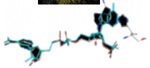
- ‘like milking cows’ (if you feed them lots of power first)
- parallel access to data comes at a cost of high IOPS

Batch system platform

Many things in life are *conveniently parallel*

- HEP events & simulation
- ligand matching & drug discovery
- structural biochemistry
- …



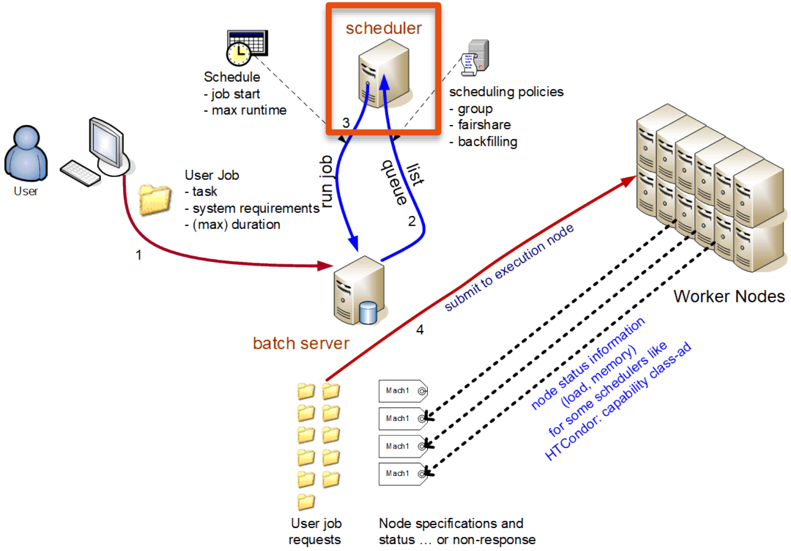


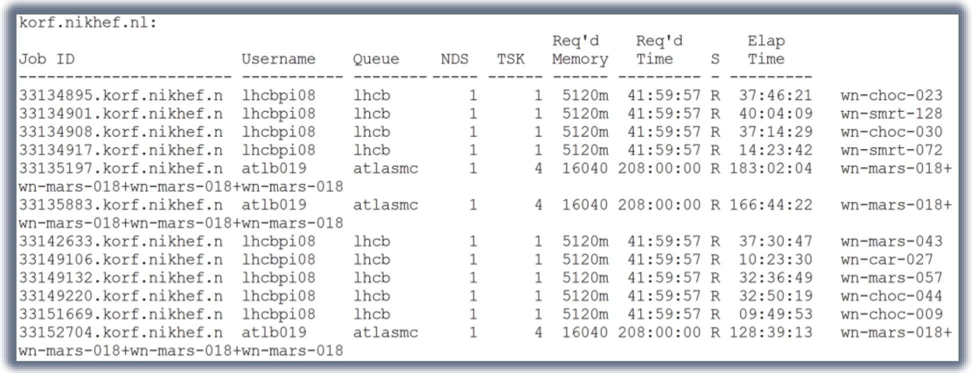
challenge not in parallelism itself

- we have had HPC systems for ages

but

- large numbers of (single-core) jobs
- heterogeneous workloads sharing the same set of worker nodes
- computing with concurrent data access





```
korf.nikhef.nl:
                                                          Req'd    Req'd       Elap
Job ID                  Username    Queue     NDS   TSK   Memory   Time    S   Time
----------------------- ----------- -------- ----- ------ ------ --------- - ---------
33134895.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  37:46:21   wn-choc-023
33134901.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  40:04:09   wn-smrt-128
33134908.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  37:14:29   wn-choc-030
33134917.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  14:23:42   wn-smrt-072
33135197.korf.nikhef.n  atlb019     atlasmc      1      4  16040 208:00:00 R 183:02:04   wn-mars-018+
wn-mars-018+wn-mars-018+wn-mars-018
33135883.korf.nikhef.n  atlb019     atlasmc      1      4  16040 208:00:00 R 166:44:22   wn-mars-018+
wn-mars-018+wn-mars-018+wn-mars-018
33142633.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  37:30:47   wn-mars-043
33149106.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  10:23:30   wn-car-027
33149132.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  32:36:49   wn-mars-057
33149220.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  32:50:19   wn-choc-044
33151669.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  09:49:53   wn-choc-009
33152704.korf.nikhef.n  atlb019     atlasmc      1      4  16040 208:00:00 R 128:39:13   wn-mars-018+
wn-mars-018+wn-mars-018+wn-mars-018
```

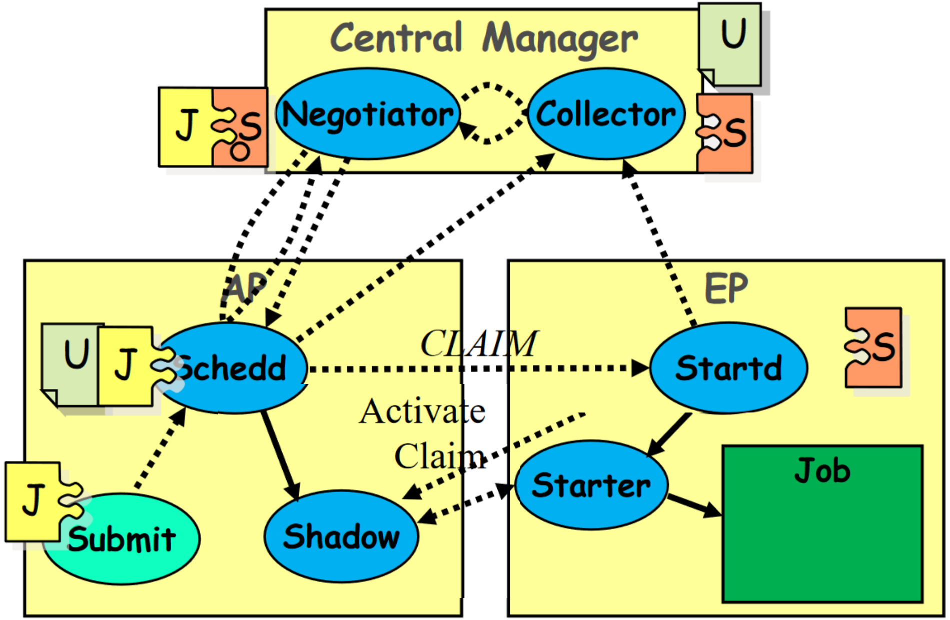
Scalable submission: HTCondor

Matchmaking based on ‘ClassAds’

- both jobs and machines advertise their requirements and capabilities in ‘classified advertisements’
- Matchmaking done by the negotiator

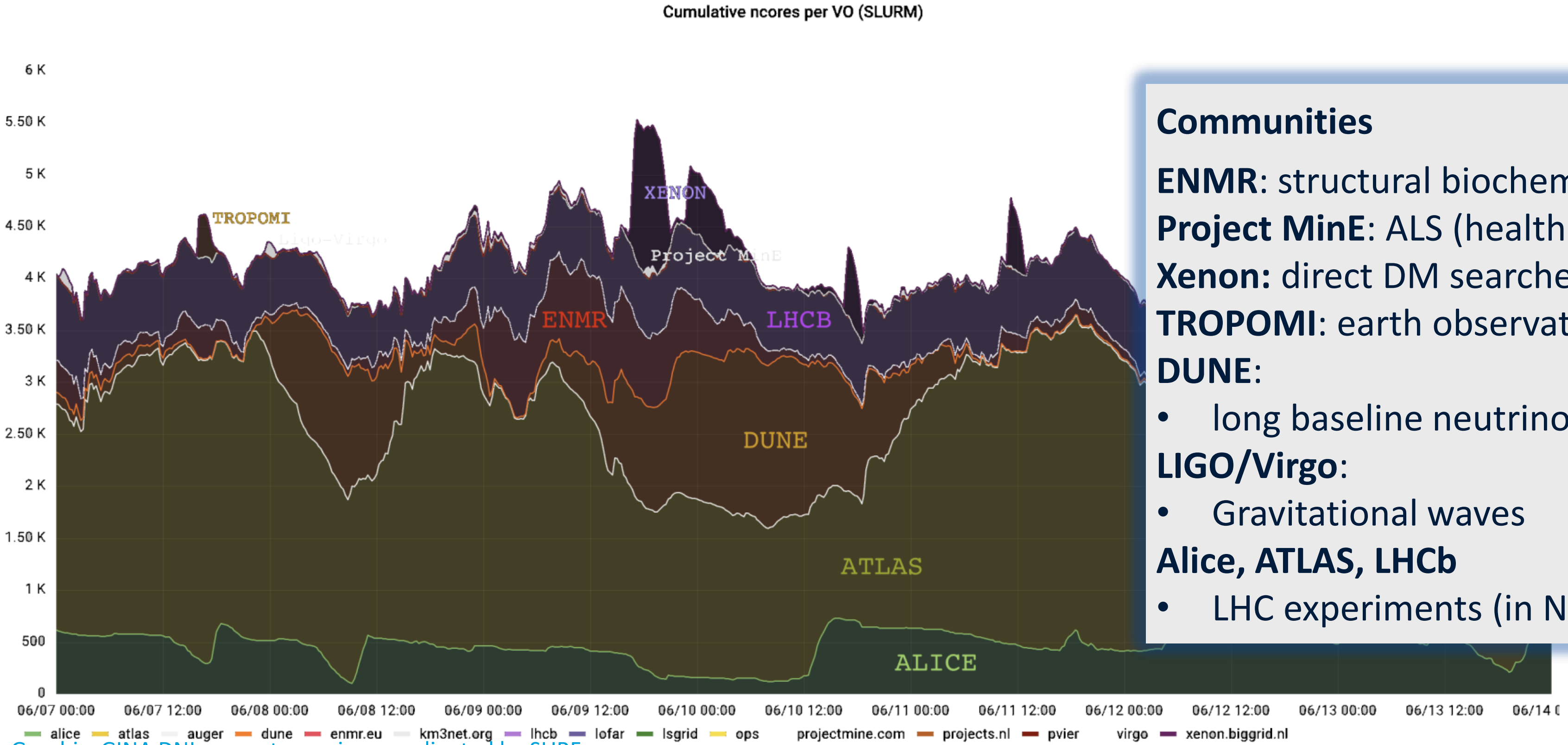
  execution nodes mostly autonomous

helps for scalability and resilience



HTCondor, Miron Livny *et al.*; Compiled from Todd Tannenbaum (2024 HTCondor Workshop) https://indico.cern.ch/event/1386170/contributions/6127903/

Dutch National e-Infrastructure: High Throughput GINA



Communities

ENMR: structural biochemistry
Project MinE: ALS (health)
Xenon: direct DM searches
TROPOMI: earth observation
DUNE:

- long baseline neutrinos

LIGO/Virgo:

- Gravitational waves

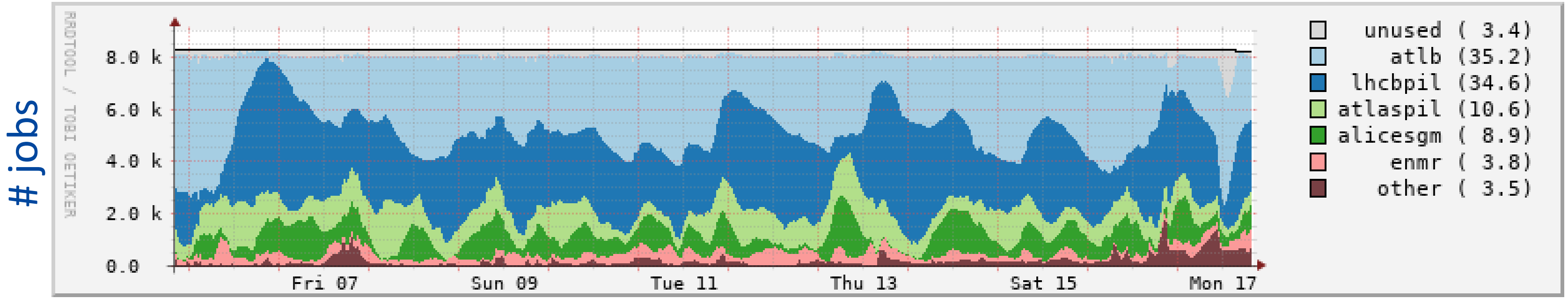
Alice, ATLAS, LHCb

- LHC experiments (in NL)

Graphic: GINA DNI compute service coordinated by SURF

Estimated Response Time (and predicting it)

- ‘Fair share’ – distributing load over time in a ‘continuous job supply’ system



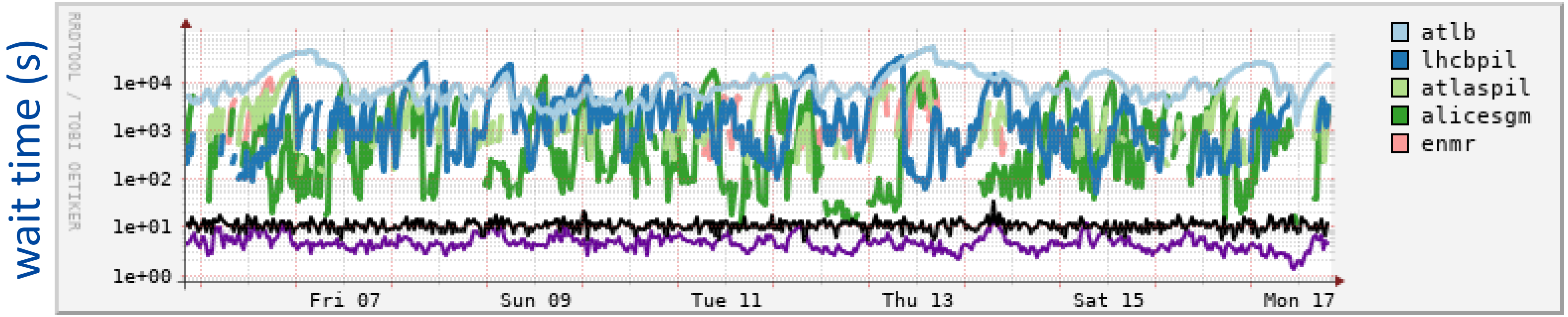


Image: Nikhef NDPF DNI “Grid” cluster. Period: October 6-17, 2022; top-5 communities; GRISview images: Jeff Templon

For work on run time prediction in high-occupancy clusters , see Hui Li *Workload characterization, modeling, and prediction …* https://hdl.handle.net/1887/12574

For occupancy, intended target audience makes a difference

For organized ‘production’ computing (planned months in advance in WLCG)

- ***predictable* scheduling** is more important (steady flow of results)
- **maximizing efficiency**: resource cost is the limiting factor in (physics) results
- co-scheduling with data (pre-placement) is required
- community-authorization based access to data sources only

For ‘local’ users, e.g. students whose progress tomorrow depends on results *today*

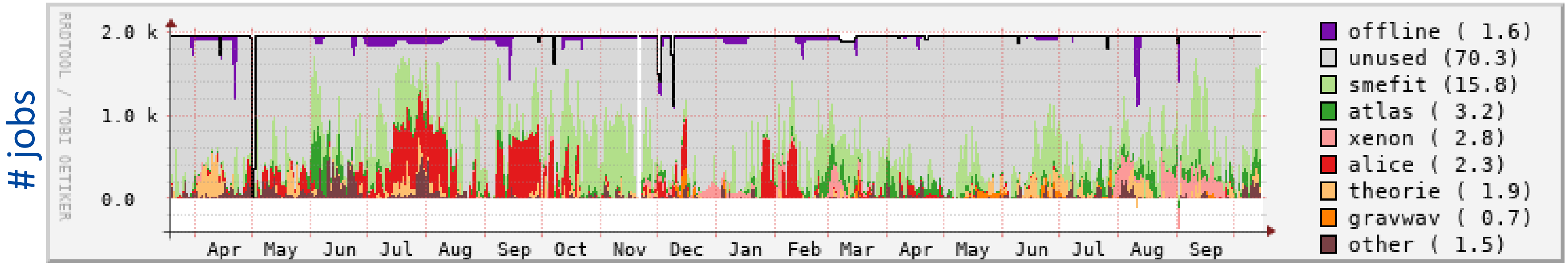
- ***response time*** is more important than efficiency
- fast turn-around/short waiting times by heterogeneous (‘competing’) user base
- data access must be parallelism-ready, but is ‘always’ local on-site
- local storage credentials and sharing with desktop and Jupyter environments

so offering two distinct classes of services is (in this case) intentional

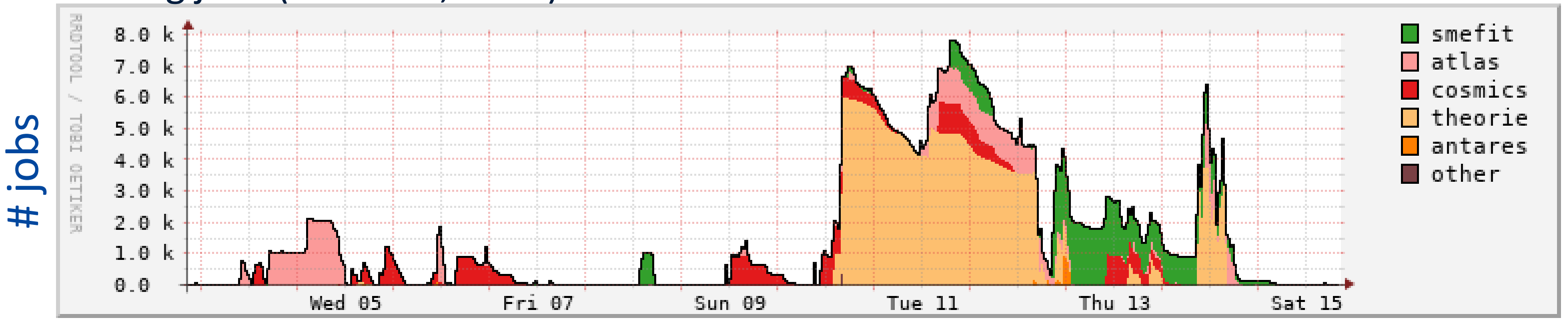
NDPF local analysis cluster ‘Stoomboot’

period: March 2021 .. October 2022

Running jobs:

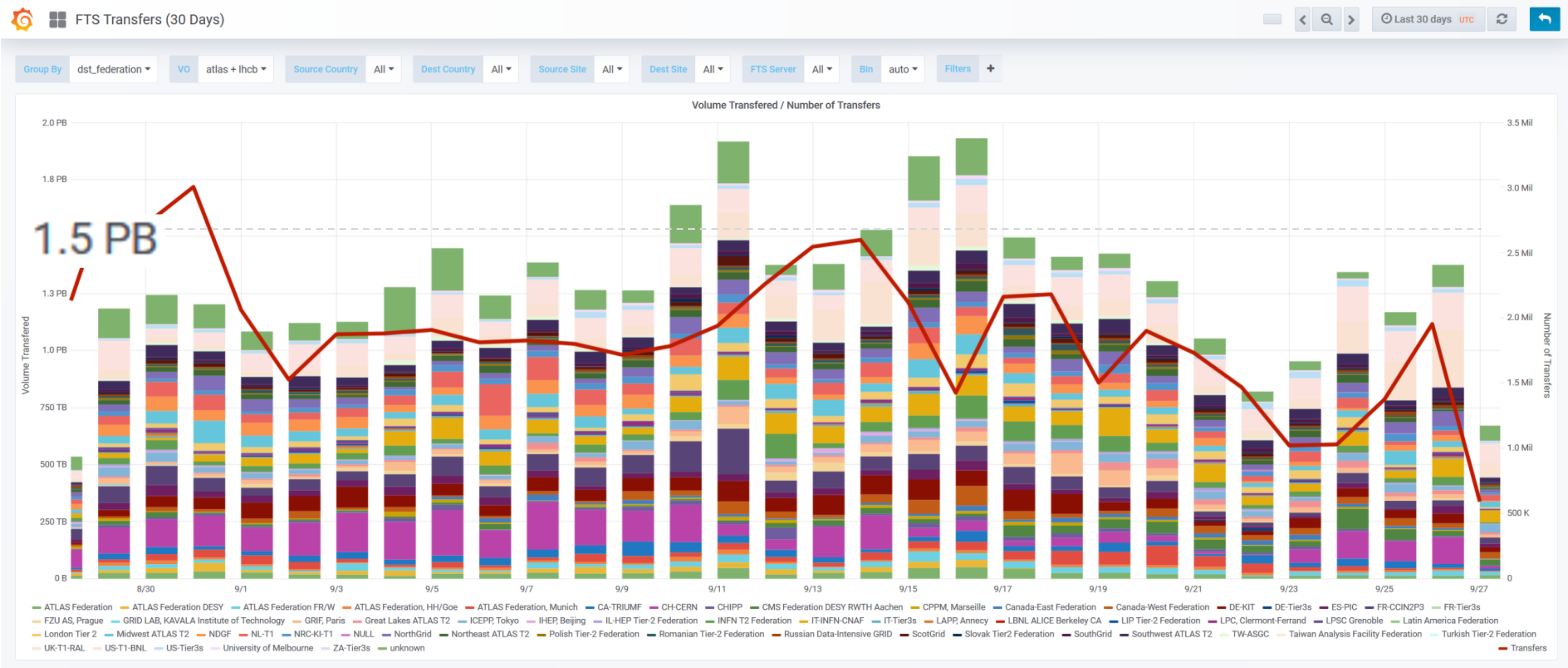


Waiting jobs (Week 40, 2022):



Source: NDPF Statistics overview, https://www.nikhef.nl/pdp/doc/stats/ - GRISview images: Jeff Templon for NDPF and STBC

High *throughput* computing is in the end about data



FTS Transfers (30 Days)

Group By: dst_federation · VO: atlas + lhcb · Source Country: All · Dest Country: All · Source Site: All · Dest Site: All · FTS Server: All · Bin: auto · Filters

source: https://monit-grafana.cern.ch/d/000000420/fts-transfers-30-day ; data: November 2020 ; CERN FTS instance WLCG: daily transfer volume ATLAS+LHCb

Can storage support your parallel processing

Basic storage properties

- throughput
- IOPS – I/O Operations per Second
- seek-time

but not many **file systems** support *concurrent parallel access* by many clients

- both data **and** (file system or index) meta-data must be scalably distributed
- typically sacrifice either instant consistency, or (POSIX) semantics, (or scalability) in a distributed storage system

Common commercial solutions: GPFS, … but also NetApp, HDS, Dell-EMC, have theirs
Common open source: BeeGFS, gluster, dCache, CephFS, Lustre, …

… likely do not use a file system if object storage does the job, *but then you need a catalogue/database*

Example: client-side managed GlusterFS

- scalable through independence of both clients and servers

- design is stateless: file system meta-data kept in each server's file system

- data itself can be replicated and protected but … inconsistencies in metadata linger around the corner in case of client failures (e.g. batch system worker nodes)

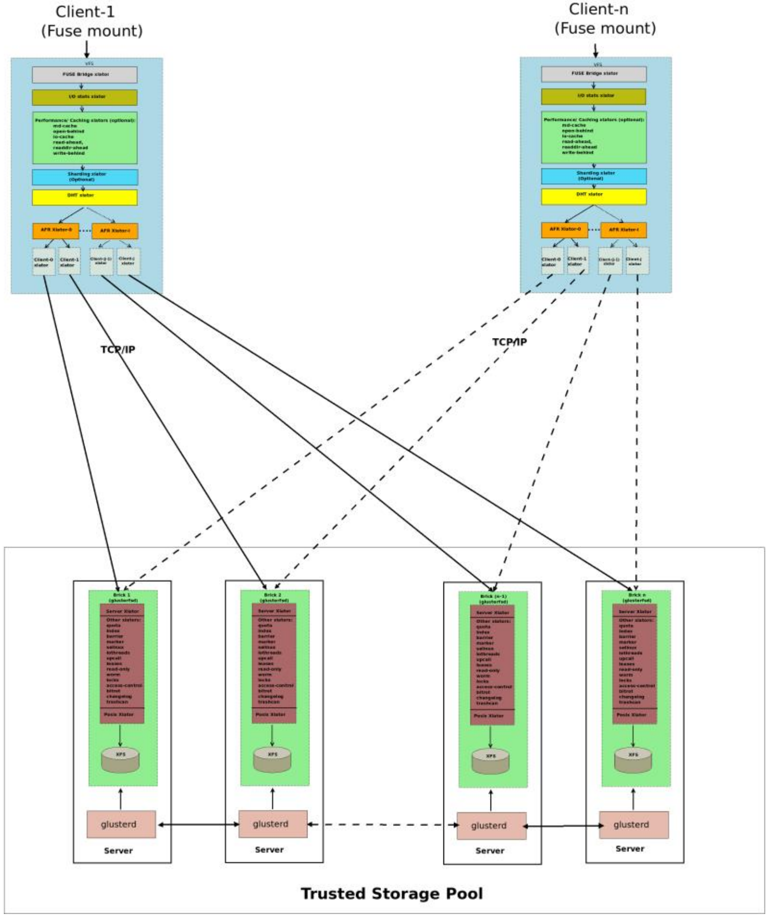
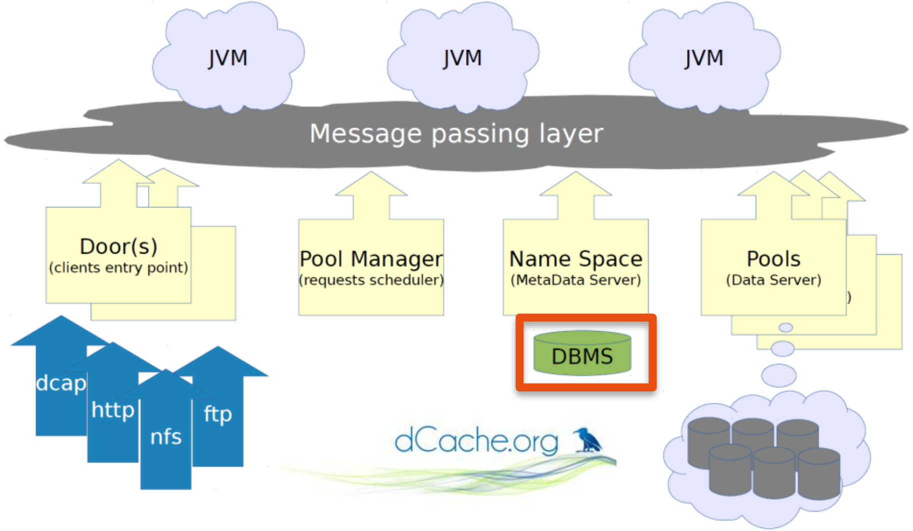
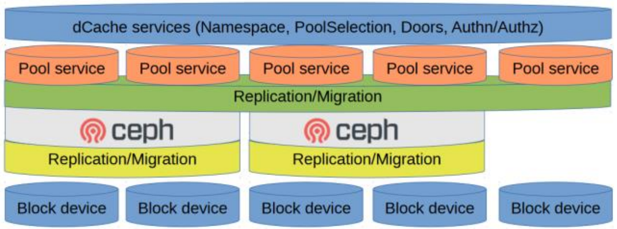


Image source Gluster community: https://docs.gluster.org/en/main/Quick-Start-Guide/Architecture/

Example: server-coherent distribution – dCache

- separate client entry points, storage access scheduling, filesystem meta-data (namespaces), and storage
- message layer for eventual consistency
- redirect-based access
  - doors and pools usually on all nodes
  - now also feature of standard NFSv4.1

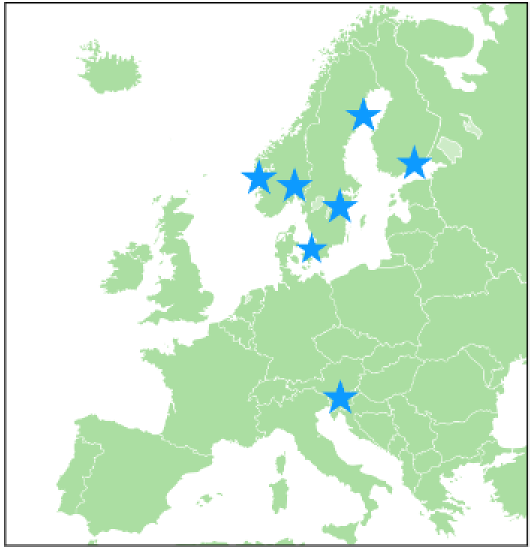




Images: Tigran Mkrtchyan (DESY, dCache.org), *dCache on steroids - delegated storage solutions*, ISGC 2016, https://dcache.org/manuals/publications.shtml

dCache: wide area distribution

- can be widely (long latency) distributed
  - Nordic Data Grid Facility: Sweden is quite long (~16ms RTT), and Ljubljana to Umeå is ~30ms RTT (~ 2900km)
- redirect-then-access model limits interactions with any single node across a long-distance links
- at ‘cost’ of POSIX features like *atime* or concurrent write
  - most distributed applications don’t need these anyway
  - but indeed it’s not a good backing store for databases ☺



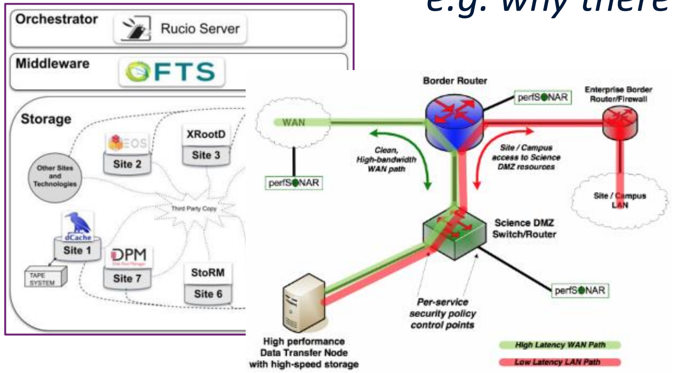
The NDGF dCache instance spans datacentres across Scandinavia and Slovenia, but is administered and used as a single instance.

Image NDGF instance: Jürgen Starek et al. (dCache team) at https://www.dcache.org/manuals/dCache-Whitepaper.pdf; https://dcache.org/manuals/Book-8.2

proceed to networking ▶

Structure of application data placement impacts storage (hardware) systems design

pre-staging all data locally allows for **latency hiding**,
posix-style access with lseek(2), and a fast, local, ‘$TMPDIR’
e.g. why there are Data Transfer Nodes (DTNs) in the ‘Science DMZ’ concept


Orchestrator
Rucio Server
Middleware
FTS
Storage
EOS
Site 2
XRootD
Site 3
Other Sites and Technologies
Third Party Copy
dCache
Site 1
DPM
Site 7
StoRM
Site 6
TAPE SYSTEM
Border Router
perfSONAR
Enterprise Border Router/Firewall
WAN
Clean, High-bandwidth WAN path
Site / Campus access to Science DMZ resources
perfSONAR
Site / Campus LAN
Science DMZ Switch/Router
perfSONAR
Per-service security policy control points
High performance Data Transfer Node with high-speed storage
High Latency WAN Path
Low Latency LAN Path

but, nowadays, pre-staging started coming at a cost, when using **SSDs** as local ‘scratch’ area … because of their hardware characteristic ‘endurance’

Photo HGST nVME from: Dmitry Nosachev on Wikimedia Commons CC-BY-SA; Image Science DMZ and Data Transfer Nodes: ESnet fasterdata.es.net

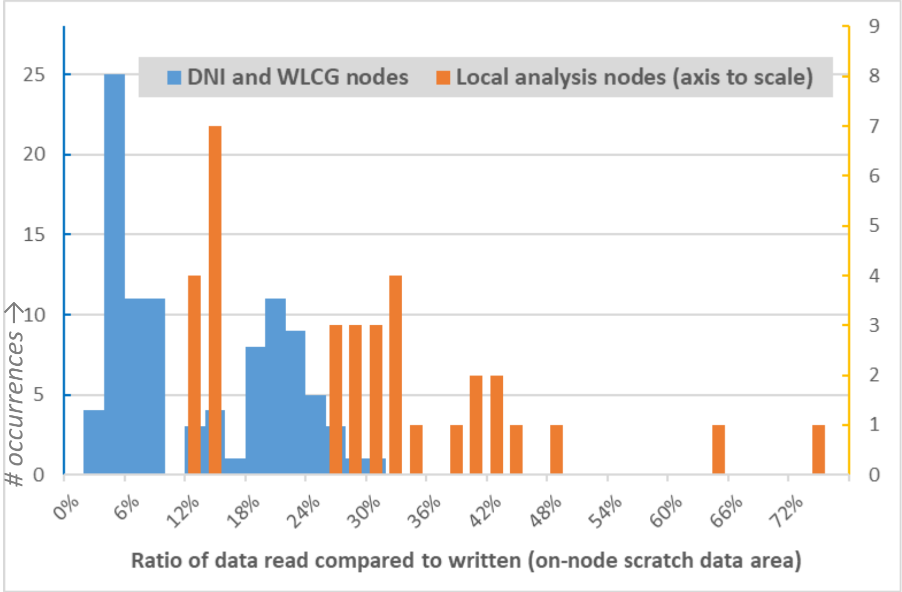
Especially with *WORN* storage: Write Once Read Never

Frequency distribution of **read-back vs. write** volume, observed on local scratch for NDPF execution nodes for ***outside ('grid') access (blue) vs local access (orange)***

Access pattern is rather different. But why?

- external users pre-stage, because it is built into data management frameworks (like DIRAC, Athena),
- 'local' users stream output data (dCache with NFSv4) and use $TMPDIR mainly for merging partial results

Different types of workload (here analysis vs processing) determine the choice of systems hardware



Data: NDPF execution nodes, based on SSD SMART data, integrated over total device lifetime; plot shows number of local analysis nodes scaled to DNI-WLCG count; collected using smartctl on 2020-10-28 – in total 97 'DNI' and 34 'STBC' SSDs were used in the analysis

As a service!

‘Cloud’ Services and
Service Management

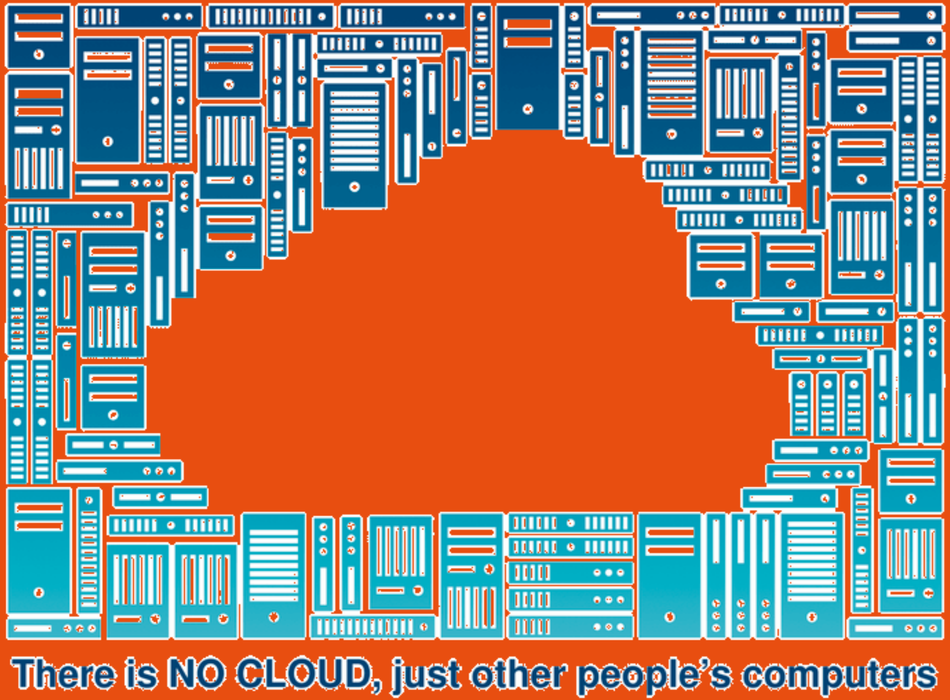


Image source: Free Software Foundation Europe - https://fsfe.org/

Scaling things '… as a service'

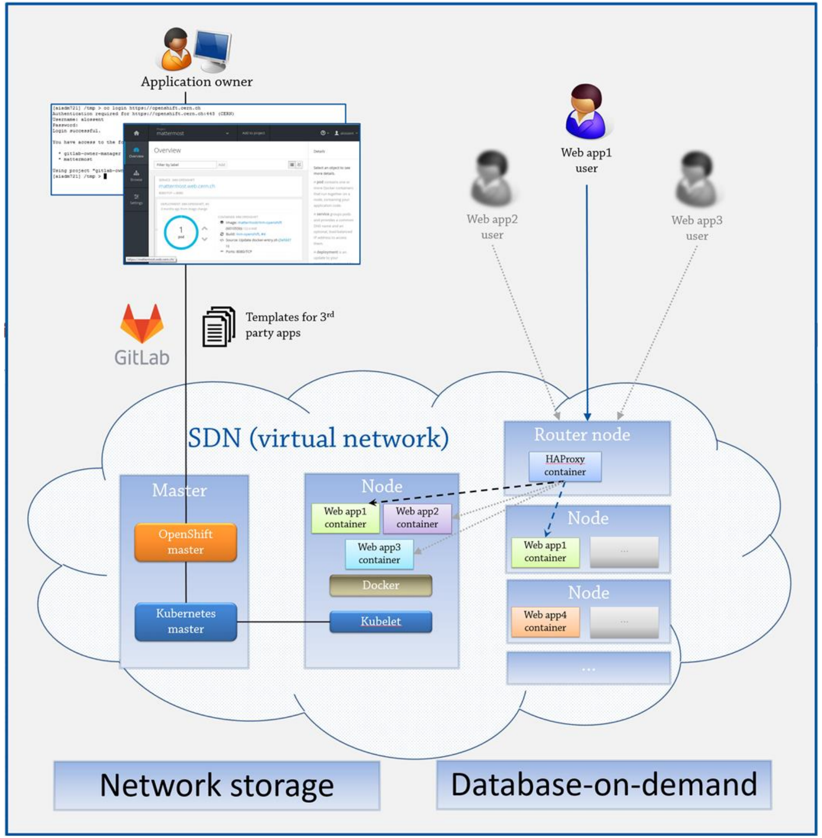
The managed servers usually are not physical

- although there is lots of 'fixed' virtualization of systems, network and (block) storage

When scale, or environment, must be flexible, you get ***software defined infrastructure***

- IaaS: Infrastructure as a Service
- PaaS: Platform as a Service (containers, but also a batch system …)
- SaaS: Software as a Service (like the science application portal like WeNMR)

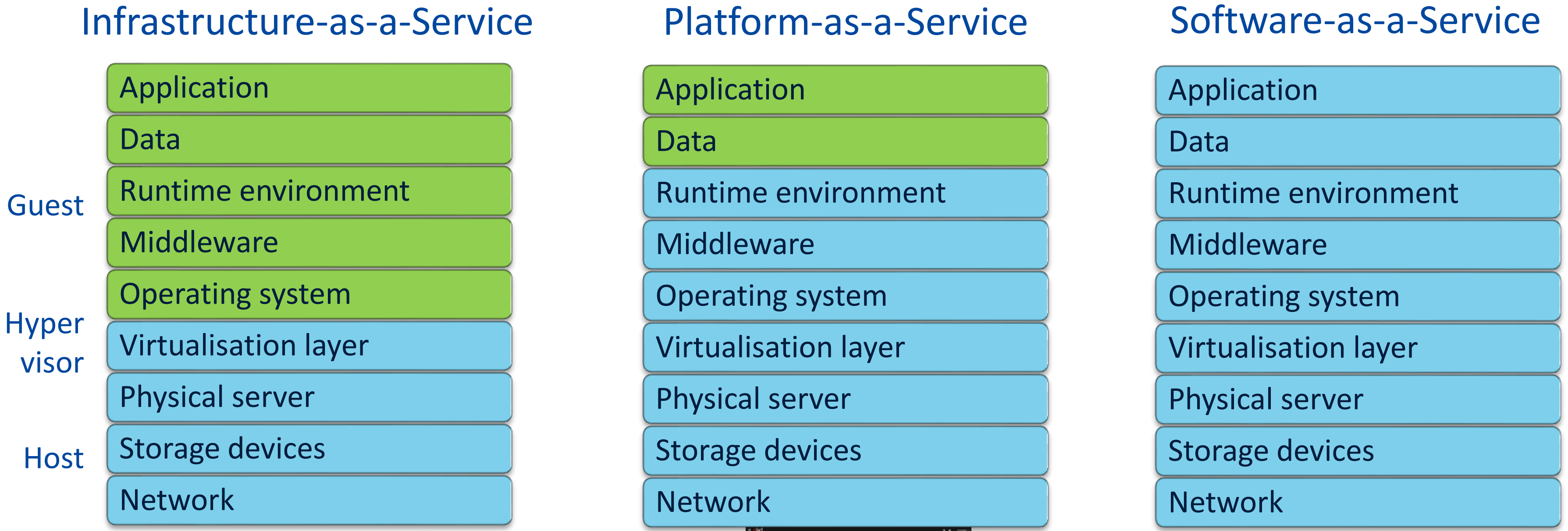
powerful tools, but also easy to get wrong (i.e. having plain-text secrets in the version control system to automate redeployment). And abstractions are *leaky*!



```
[aiadm721] /tmp > oc login https://openshift.cern.ch
Authentication required for https://openshift.cern.ch:443 (CERN)
Username: alossent
Password:
Login successful.

You have access to the fo

  * gitlab-owner-manager
  * mattermost

Using project "gitlab-own
[aiadm721] /tmp > 
```

Image from CERN's OpenShift, A Lossent *et al* 2017 *J. Phys.: Conf. Ser.* **898** 082037 https://doi.org/10.1088/1742-6596/898/8/082037

Moving the management boundary



Infrastructure-as-a-Service

Guest

Hyper visor

Host

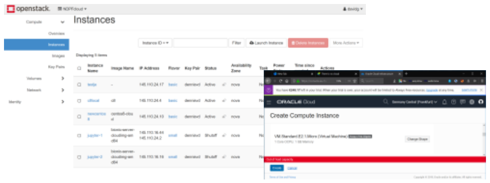
- Application
- Data
- Runtime environment
- Middleware
- Operating system
- Virtualisation layer
- Physical server
- Storage devices
- Network

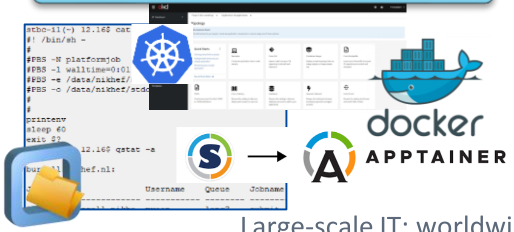
Platform-as-a-Service

- Application
- Data
- Runtime environment
- Middleware
- Operating system
- Virtualisation layer
- Physical server
- Storage devices
- Network

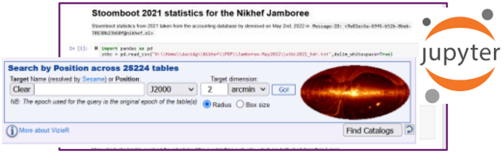
Software-as-a-Service

- Application
- Data
- Runtime environment
- Middleware
- Operating system
- Virtualisation layer
- Physical server
- Storage devices
- Network





```
stbc-11(~) 12.16$ cat
#! /bin/sh -
#
#PBS -N platformjob
#PBS -l walltime=0:01
#PBS -e /data/nikhef/
#PBS -o /data/nikhef/stdo
#
#
printenv
sleep 60
exit $?
            12.16$ qstat -a

bur      hef.nl:

                         Username    Queue    Jobname
------------  ----------  --------  -------
```



Stoomboot 2021 statistics for the Nikhef Jamboree

Search by Position across 25224 tables

Target Name (resolved by Sesame) or Position:

Target dimension:

Clear | J2000 | 2 | arcmin | Go!

NB: The epoch used for the query is the original epoch of the table(s)

● Radius ○ Box size

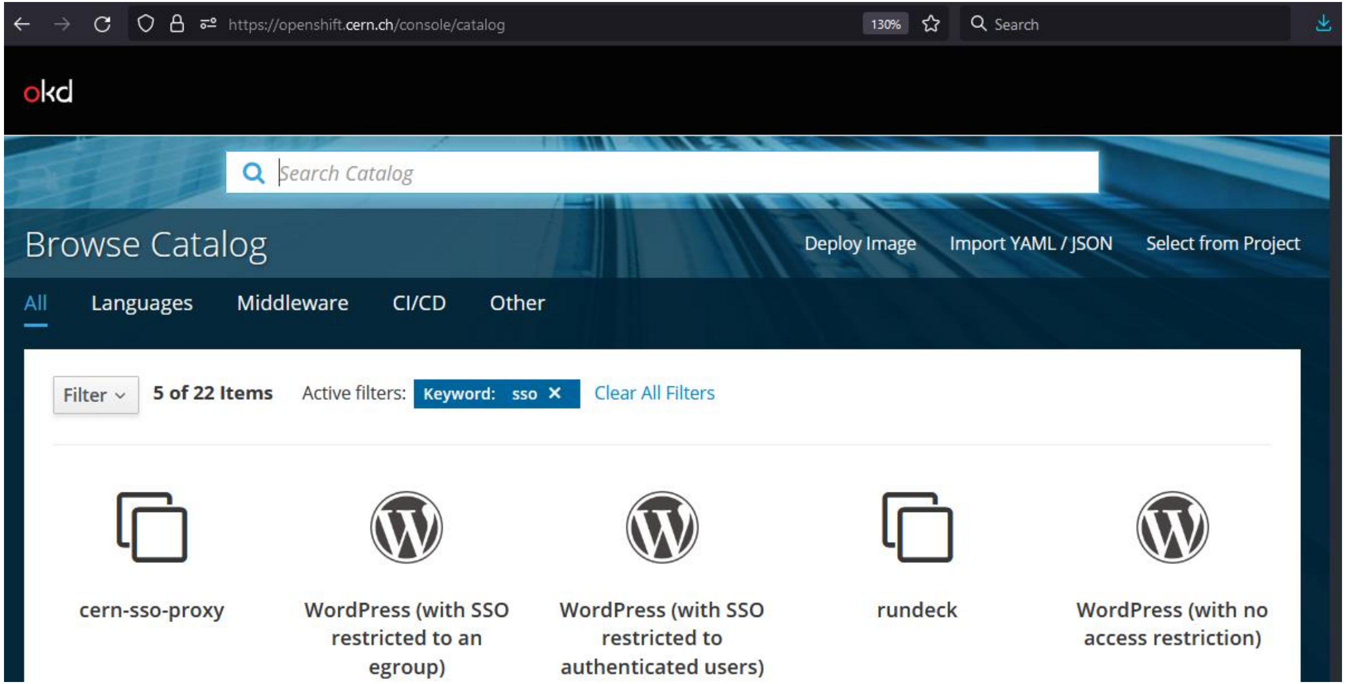
More about VizieR

Find Catalogs

Astronomy catalogue: https://vizier.cds.unistra.fr/

IaaS: openstack.com, Oracle OCI; PaaS: dsri.maastrichtuniversity.nl, apptainer.org, cvmfs.readthedocs.io, kubernetes.io, slurm.schedmd.com; SaaS: Jupyter.org

‘Cloudification’ eases systems management …



https://openshift.cern.ch/console/catalog

okd

Search Catalog

Browse Catalog

Deploy Image Import YAML / JSON Select from Project

All Languages Middleware CI/CD Other

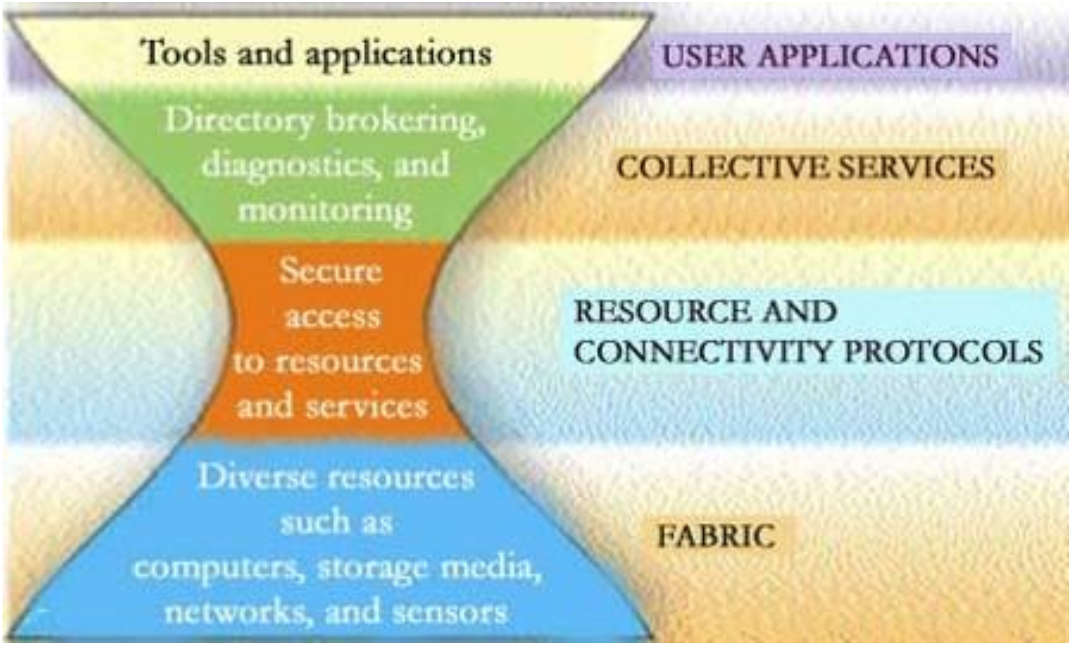
Filter 5 of 22 Items Active filters: Keyword: sso ✕ Clear All Filters

- cern-sso-proxy
- WordPress (with SSO restricted to an egroup)
- WordPress (with SSO restricted to authenticated users)
- rundeck
- WordPress (with no access restriction)

OpenShift (OKD) system at CERN (accessible for CERN users only) – at Maastricht use the DSRI infrastructure: https://dsri.maastrichtuniversity.nl/

Common interfaces to the different clouds?

‘protocol hourglass’


Tools and applications
USER APPLICATIONS
Directory brokering, diagnostics, and monitoring
COLLECTIVE SERVICES
Secure access to resources and services
RESOURCE AND CONNECTIVITY PROTOCOLS
Diverse resources such as computers, storage media, networks, and sensors
FABRIC

hourglass image: Alessio Merlo in The Condor on the Grid: state of art and open issues,

Standard interfaces for compute and data?

hourglass model 'kind-of' worked for IP and web with http as common standard

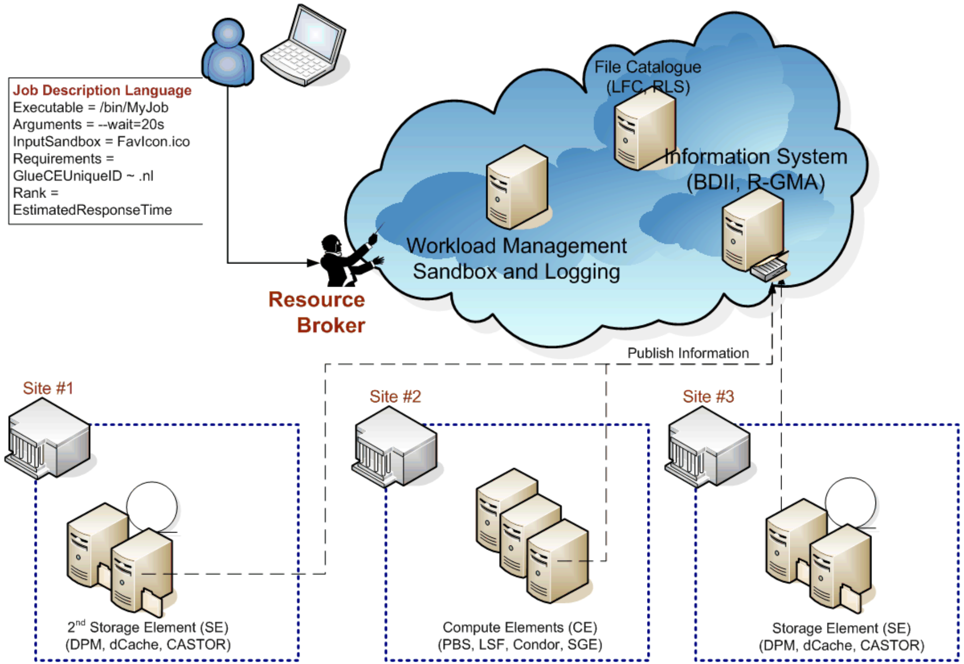
- a very simple stateless interface

protocols for higher-level services never quite reached this level of global interop

- requirements too complex and stateful
- use cases were usually scoped

slowly changing now but only for similarly simple things, like on-line object storage

Is distributed computing too bespoke …?



Job Description Language

```
Executable = /bin/MyJob
Arguments = --wait=20s
InputSandbox = FavIcon.ico
Requirements =
GlueCEUniqueID ~ .nl
Rank =
EstimatedResponseTime
```

Interoperable cloud? Compare OGF's OCCI WG GFD.221 (https://www.ogf.org/documents/GFD.221.pdf) with e.g. Amazon S3 API or the OwnCloud CS3 interfaces

DIRAC: spanning heterogeneous resource models

Add a scheduling layer!

‘any (IT) problem can be solved by adding an extra level of indirection’

DIRAC is just one example

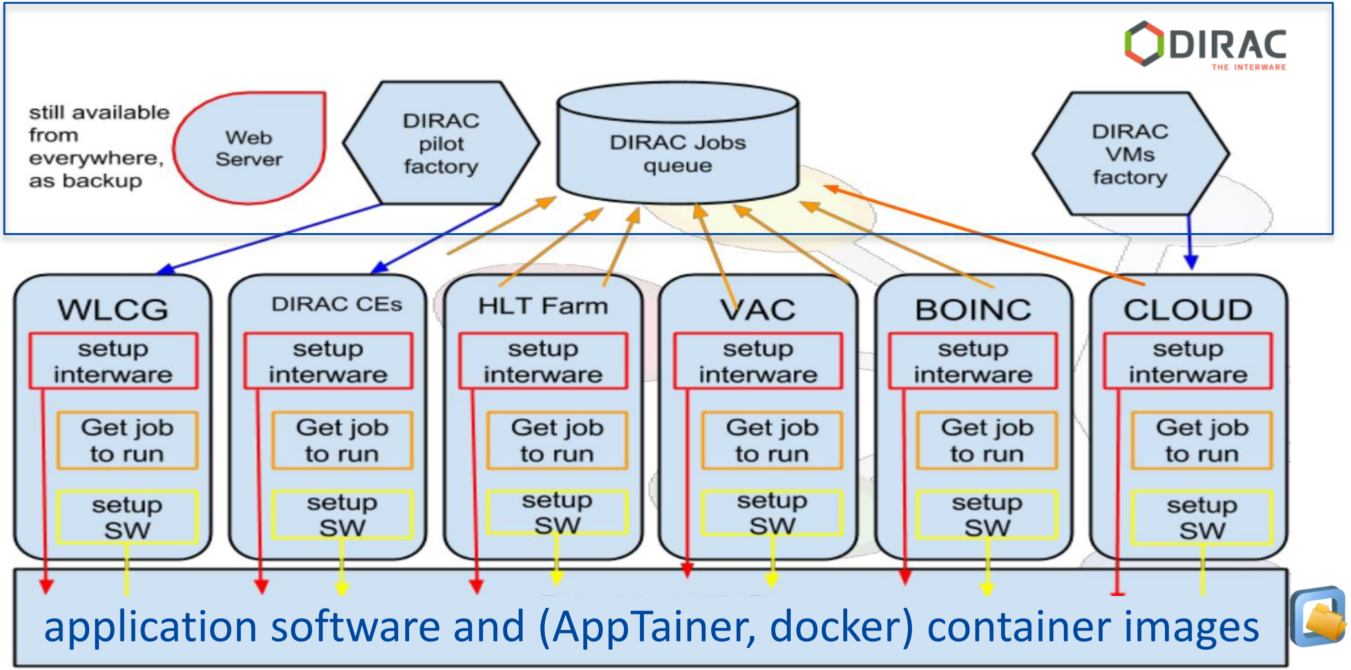
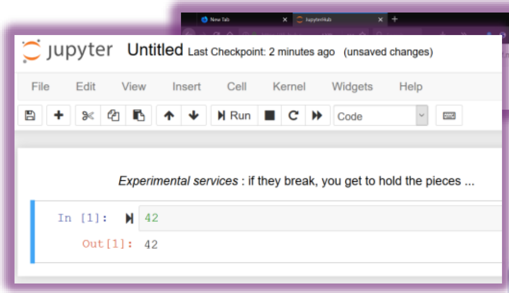


Image: DIRAC project, A. Tsaregorodtsev *et al.* CPPM Marseille, from https://dirac.readthedocs.io/ ; CVMFS (CERN VM File System) is a common software distribution platform using distributed signed data objects in a cached hierarchy using CDN techniques, see https://cernvm.cern.ch/fs/

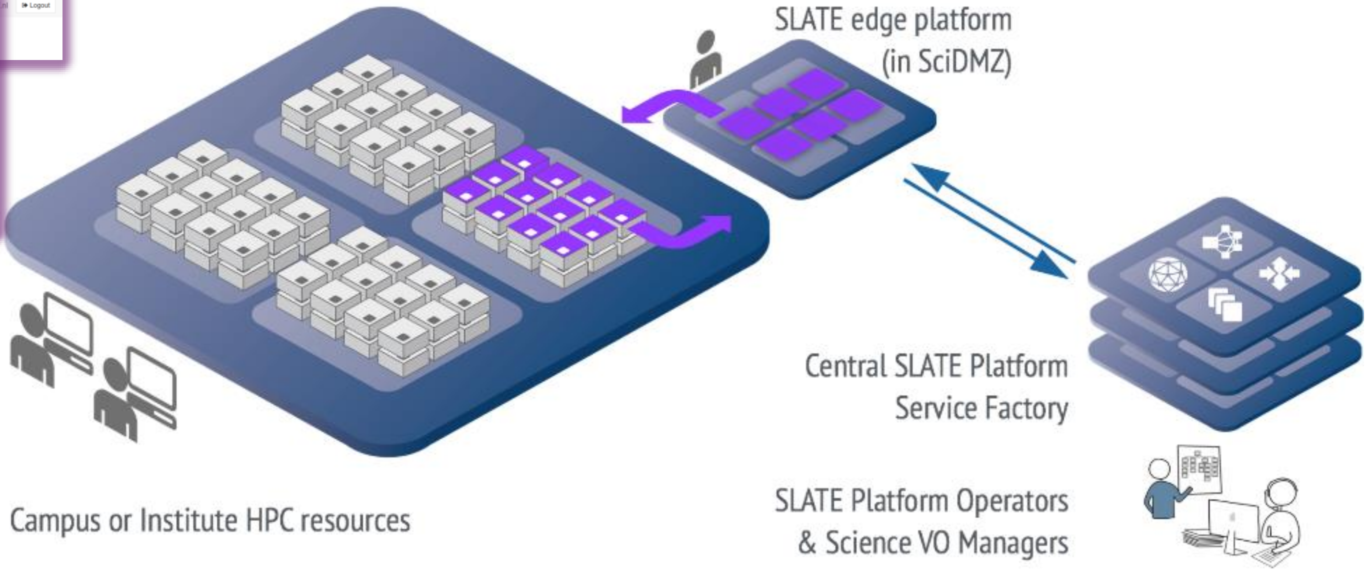
An overlay network of containers

Nobody wants a cloud per-se … what folk want is a solution …



Experimental services : if they break, you get to hold the pieces ...

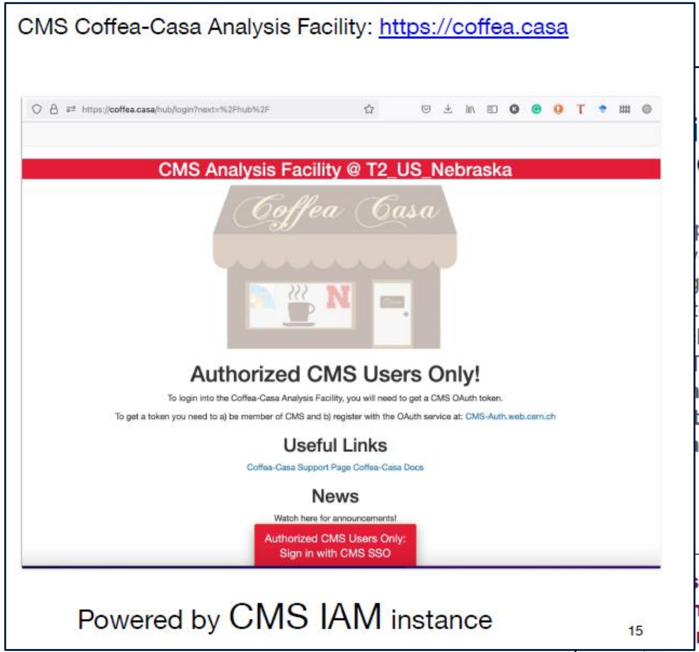
```
In [1]:  42
Out[1]:  42
```



'alien containers' HPC integration - container computing, using curated application images

Image sources: NDPF JupyterHub service "Callysto"; SLATE: Service Layer At The Edge – Rob Gartner (UChicago), Shawn KcMee (UMich) *et al.* – slateci.io

Containerised workloads: between 'PaaS' and 'SaaS'



CMS Coffea-Casa Analysis Facility: https://coffea.casa

CMS Analysis Facility @ T2_US_Nebraska

Authorized CMS Users Only!

To login into the Coffea-Casa Analysis Facility, you will need to get a CMS OAuth token.

To get a token you need to a) be member of CMS and b) register with the OAuth service at: CMS-Auth.web.cern.ch

Useful Links

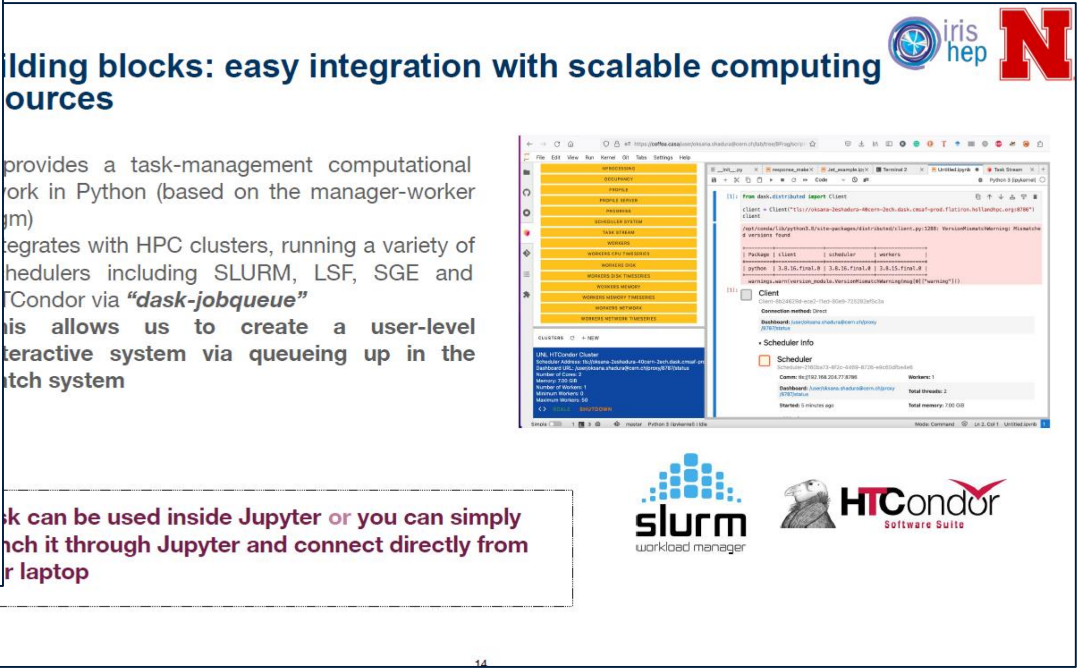
Coffea-Casa Support Page Coffea-Casa Docs

News

Watch here for announcements!

Powered by CMS IAM instance

15



ilding blocks: easy integration with scalable computing ources

provides a task-management computational ork in Python (based on the manager-worker gm)

tegrates with HPC clusters, running a variety of hedulers including SLURM, LSF, SGE and TCondor via ***"dask-jobqueue"***

is allows us to create a user-level teractive system via queueing up in the tch system

UNL HTCondor Cluster
Scheduler Address: tls://oksana-2eshadura-40cern-2ech.dask.cmsaf-pr
Dashboard URL: /user/oksana.shadura@cern.ch/proxy/8787/status
Number of Cores: 2
Memory: 7.00 GiB
Number of Workers: 1
Minimum Workers: 0
Maximum Workers: 50

```
[1]: from dask.distributed import Client

     client = Client("tls://oksana-2eshadura-40cern-2ech.dask.cmsaf-prod.flatiron.hollandhpc.org:8786")
     client
```

```
/opt/conda/lib/python3.8/site-packages/distributed/client.py:1288: VersionMismatchWarning: Mismatche
d versions found

+---------+-----------------+-----------------+-----------------+
| Package | client          | scheduler       | workers         |
+---------+-----------------+-----------------+-----------------+
| python  | 3.8.16.final.0  | 3.8.16.final.0  | 3.8.15.final.0  |
+---------+-----------------+-----------------+-----------------+
  warnings.warn(version_module.VersionMismatchWarning(msg[0]["warning"]))
```

[1]: Client
Client-8b24629d-ece2-11ed-80e9-725282ef5c3a

Connection method: Direct

Dashboard: /user/oksana.shadura@cern.ch/proxy/8787/status

- Scheduler Info

Scheduler
Scheduler-2160ba73-8f2c-4489-8726-e6c60dfbe4e6

Comm: tls://192.168.204.77:8786 **Workers:** 1

Dashboard: /user/oksana.shadura@cern.ch/proxy/8787/status **Total threads:** 2

Started: 5 minutes ago **Total memory:** 7.00 GiB

k can be used inside Jupyter or you can simply nch it through Jupyter and connect directly from r laptop

14

Images: Oksana Shadura et al (UNebraska Lincoln), Brian Bockelman (Morgridge Institute) at CHEP2023 https://indico.jlab.org/event/459/contributions/11610/

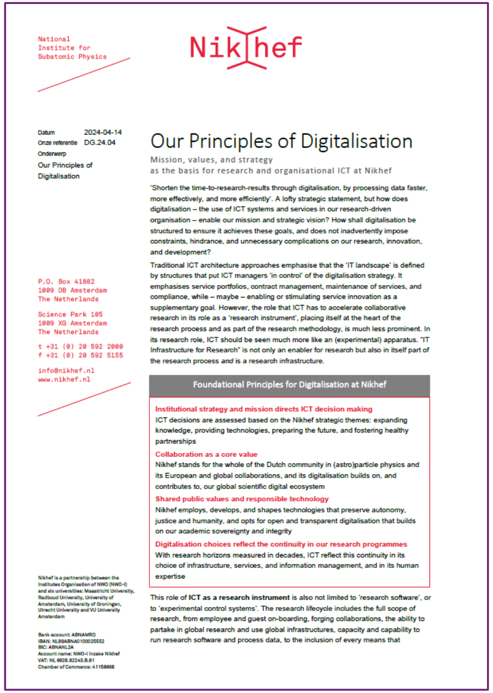
Services: serving the users, not the IT department

What actually drives your IT architecture and service management system?

- strategic requirements of your organisation?
- what are ‘appropriate’ service levels
  and impact in case things go differently?
- balancing stability, innovation, and engagement

Potentially separate ‘service classes’

- ‘enterprise’ services
- research computing services (the ‘primary business’)?
- experimental services (innovation, future strategy)?



National
Institute for
Subatomic Physics

Nikhef

Datum 2024-04-14
Onze referentie DG.24.04
Onderwerp
Our Principles of Digitalisation

Our Principles of Digitalisation

Mission, values, and strategy
as the basis for research and organisational ICT at Nikhef

'Shorten the time-to-research-results through digitalisation, by processing data faster, more effectively, and more efficiently'. A lofty strategic statement, but how does digitalisation – the use of ICT systems and services in our research-driven organisation – enable our mission and strategic vision? How shall digitalisation be structured to ensure it achieves these goals, and does not inadvertently impose constraints, hindrance, and unnecessary complications on our research, innovation, and development?

Traditional ICT architecture approaches emphasise that the 'IT landscape' is defined by structures that put ICT managers 'in control' of the digitalisation strategy. It emphasises service portfolios, contract management, maintenance of services, and compliance, while – maybe – enabling or stimulating service innovation as a supplementary goal. However, the role that ICT has to accelerate collaborative research in its role as a 'research instrument', placing itself at the heart of the research process and as part of the research methodology, is much less prominent. In its research role, ICT should be seen much more like an (experimental) apparatus. "IT Infrastructure for Research" is not only an enabler for research but also in itself part of the research process *and* is a research infrastructure.

Foundational Principles for Digitalisation at Nikhef

Institutional strategy and mission directs ICT decision making
ICT decisions are assessed based on the Nikhef strategic themes: expanding knowledge, providing technologies, preparing the future, and fostering healthy partnerships

Collaboration as a core value
Nikhef stands for the whole of the Dutch community in (astro)particle physics and its European and global collaborations, and its digitalisation builds on, and contributes to, our global scientific digital ecosystem

Shared public values and responsible technology
Nikhef employs, develops, and shapes technologies that preserve autonomy, justice and humanity, and opts for open and transparent digitalisation that builds on our academic sovereignty and integrity

Digitalisation choices reflect the continuity in our research programmes
With research horizons measured in decades, ICT reflect this continuity in its choice of infrastructure, services, and information management, and in its human expertise

This role of **ICT as a research instrument** is also not limited to 'research software', or to 'experimental control systems'. The research lifecycle includes the full scope of research, from employee and guest on-boarding, forging collaborations, the ability to partake in global research and use global infrastructures, capacity and capability to run research software and process data, to the inclusion of every means that

P.O. Box 41882
1009 DB Amsterdam
The Netherlands

Science Park 105
1098 XG Amsterdam
The Netherlands

t +31 (0) 20 592 2000
f +31 (0) 20 592 5155

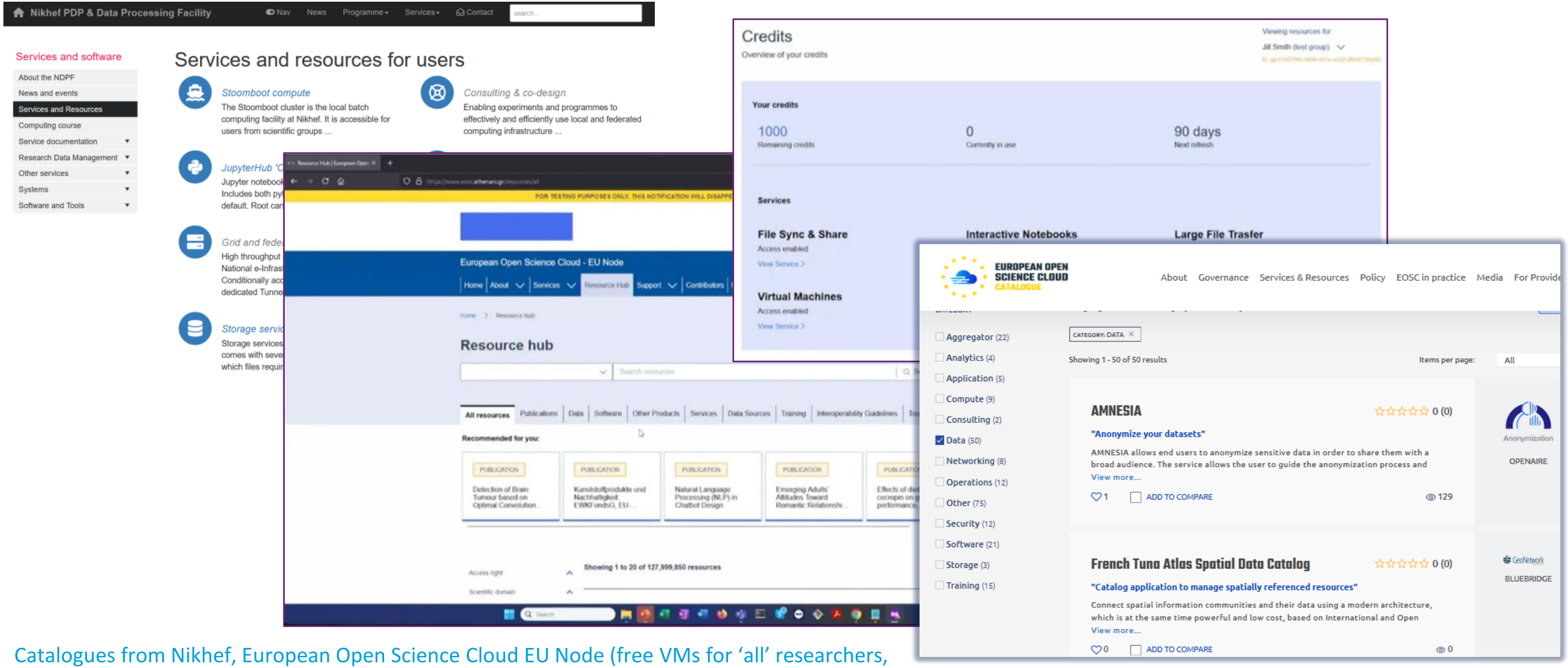
info@nikhef.nl
www.nikhef.nl

Nikhef is a partnership between the Institutes Organisation of NWO (NWO-I) and six universities: Maastricht University, Radboud University, University of Amsterdam, University of Groningen, Utrecht University and VU University Amsterdam

Bank account: ABNAMRO
IBAN: NL89ABNA0100025552
BIC: ABNANL2A
Account name: NWO-I inzake Nikhef
VAT: NL 0028.82243.B.01
Chamber of Commerce: 41150068

Alternative approach to ‘IT landscape’ architecture https://go.nikhef.nl/principles-of-digitalisation

Service portfolios – catalogues are nice, up to a point



Nikhef PDP & Data Processing Facility

Nav | News | Programme ▾ | Services ▾ | Contact | search...

Services and software

- About the NDPF
- News and events
- Services and Resources
- Computing course
- Service documentation ▾
- Research Data Management ▾
- Other services ▾
- Systems ▾
- Software and Tools ▾

Services and resources for users

Stoomboot compute
The Stoomboot cluster is the local batch computing facility at Nikhef. It is accessible for users from scientific groups ...

Consulting & co-design
Enabling experiments and programmes to effectively and efficiently use local and federated computing infrastructure ...

JupyterHub 'C
Jupyter notebook
Includes both py
default. Root can

Grid and fede
High throughput
National e-Infras
Conditionally acc
dedicated Tunne

Storage servic
Storage services
comes with seve
which files requir

Resource Hub | European Open

https://www.eosc.athenarc.gr/resources/all

FOR TESTING PURPOSES ONLY. THIS NOTIFICATION WILL DISAPPE

European Open Science Cloud - EU Node

Home | About ▾ | Services ▾ | Resource Hub | Support ▾ | Contributors

Home > Resource hub

Resource hub

Search resources

All resources | Publications | Data | Software | Other Products | Services | Data Sources | Training | Interoperability Guidelines | Too

Recommended for you:

- PUBLICATION: Detection of Brain Tumour based on Optimal Convolution...
- PUBLICATION: Kunststoffprodukte und Nachhaltigkeit: EWKFondsG, EU-...
- PUBLICATION: Natural Language Processing (NLP) in Chatbot Design
- PUBLICATION: Emerging Adults' Attitudes Toward Romantic Relationshi...
- PUBLICATIO: Effects of diet cecropin on g performance...

Access right ^
Scientific domain ^

Showing 1 to 20 of 127,999,850 resources

Credits

Overview of your credits

Viewing resources for:
Jill Smith (test group) ▾

Your credits

1000	0	90 days
Remaining credits	Currently in use	Next refresh

Services

File Sync & Share
Access enabled
View Service >

Interactive Notebooks

Large File Trasfer

Virtual Machines
Access enabled
View Service >

EUROPEAN OPEN SCIENCE CLOUD CATALOGUE

About | Governance | Services & Resources | Policy | EOSC in practice | Media | For Provide

- □ Aggregator (22)
- □ Analytics (4)
- □ Application (5)
- □ Compute (9)
- □ Consulting (2)
- ☑ Data (50)
- □ Networking (8)
- □ Operations (12)
- □ Other (75)
- □ Security (12)
- □ Software (21)
- □ Storage (3)
- □ Training (15)

CATEGORY: DATA ✕

Showing 1 - 50 of 50 results

Items per page: All

AMNESIA ☆☆☆☆☆ 0 (0)
"Anonymize your datasets"
AMNESIA allows end users to anonymize sensitive data in order to share them with a broad audience. The service allows the user to guide the anonymization process and
View more...
♡1 □ ADD TO COMPARE 👁 129
Anonymization
OPENAIRE

French Tuna Atlas Spatial Data Catalog ☆☆☆☆☆ 0 (0)
"Catalog application to manage spatially referenced resources"
Connect spatial information communities and their data using a modern architecture, which is at the same time powerful and low cost, based on International and Open
View more...
♡0 □ ADD TO COMPARE 👁 0
GeoNetwork
BLUEBRIDGE

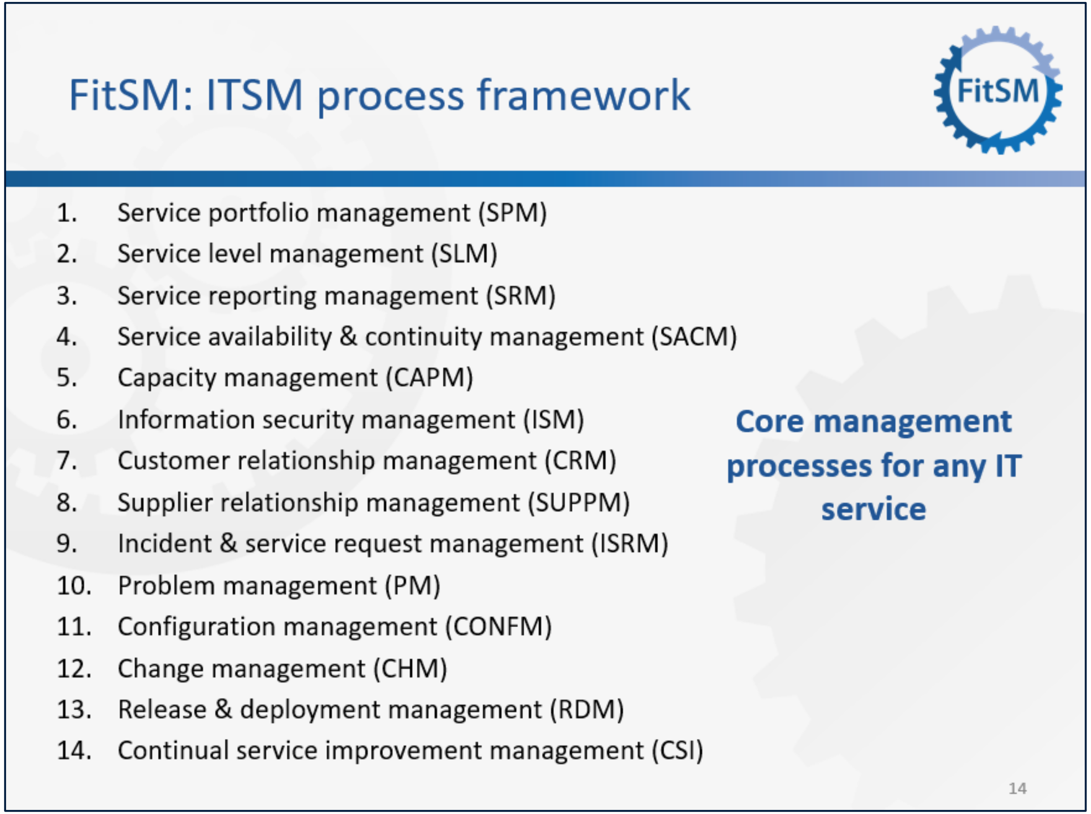
Catalogues from Nikhef, European Open Science Cloud EU Node (free VMs for 'all' researchers, subject to https://open-science-cloud.ec.europa.eu/system/files?file=2024-10/EOSC-EU-Node-User-Access-Policy-v1.0.pdf)

Example: FitSM – Federated IT Service Management

Structuring service management

- ISO 20k
- https://www.fitsm.eu/
- ITIL (now at ITIL v3)

and a whole bunch of others, like COBIT, AgileSM, …



FitSM: ITSM process framework

1. Service portfolio management (SPM)
2. Service level management (SLM)
3. Service reporting management (SRM)
4. Service availability & continuity management (SACM)
5. Capacity management (CAPM)
6. Information security management (ISM)
7. Customer relationship management (CRM)
8. Supplier relationship management (SUPPM)
9. Incident & service request management (ISRM)
10. Problem management (PM)
11. Configuration management (CONFM)
12. Change management (CHM)
13. Release & deployment management (RDM)
14. Continual service improvement management (CSI)

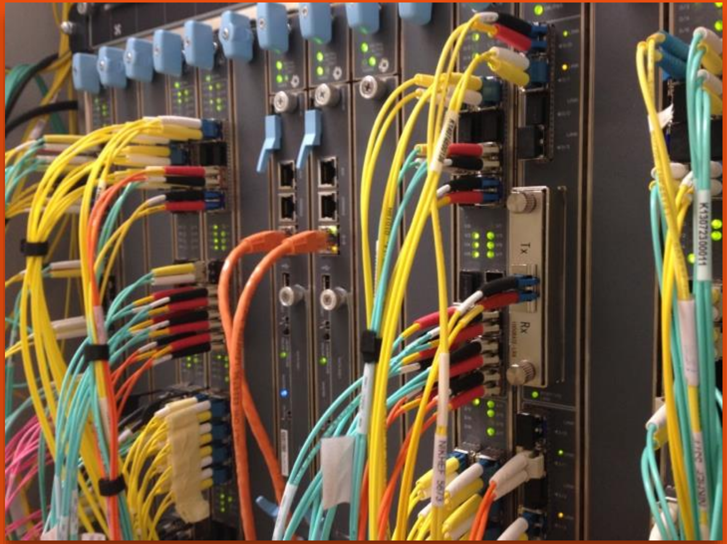
Core management processes for any IT service

14

Slide with PR list from https://www.fitsm.eu/

Putting ‘more than one’ thing together

Connecting the data:
The Internet Is Not Enough!



‘Elephant streams in a packet-switched internet’

*‘You may have plenty of shovels,
but where to leave the sand?’*

- wheelbarrow works fine in your garden
- want to send it to different places?
  Use waggons on a train,
  or ships with containers
- always from A-to-B?
  A conveyer belt will do much better!

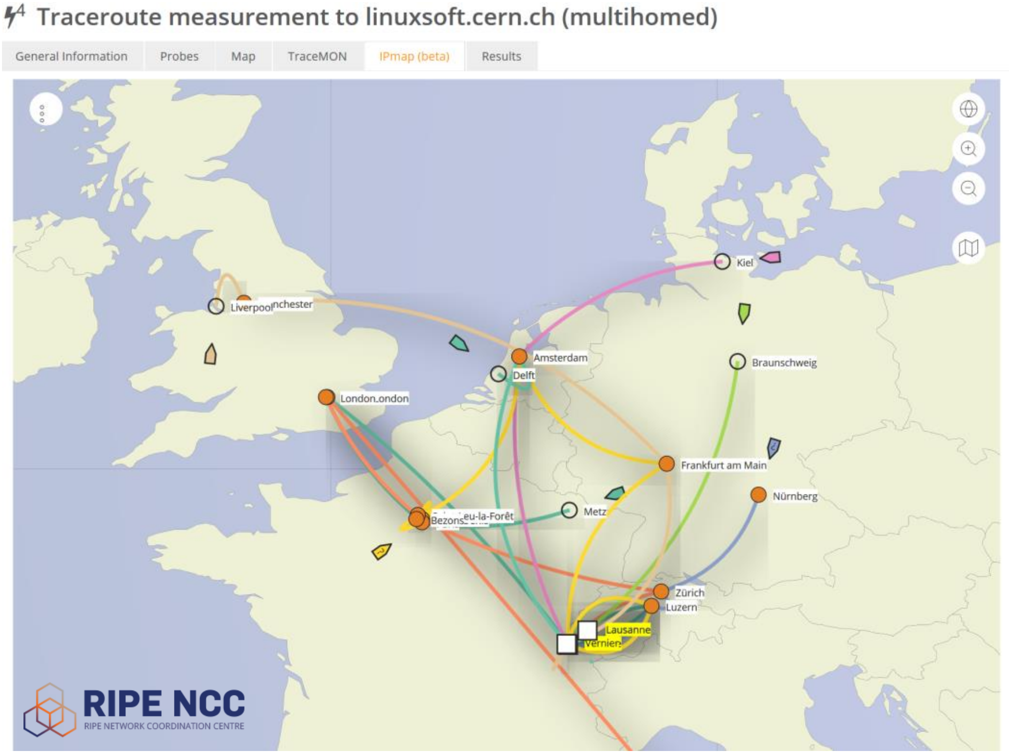
… although you still need
a hole to dump it in …

Image conveyor belt tunnel near Bluntisham, Cambridgeshire by Hugh Venables, CC-BY-SA-4.0 from https://www.geograph.org.uk/photo/4344525

A quick look at internet routing …

network paths
from various places
in Western Europe

towards an IP address at CERN

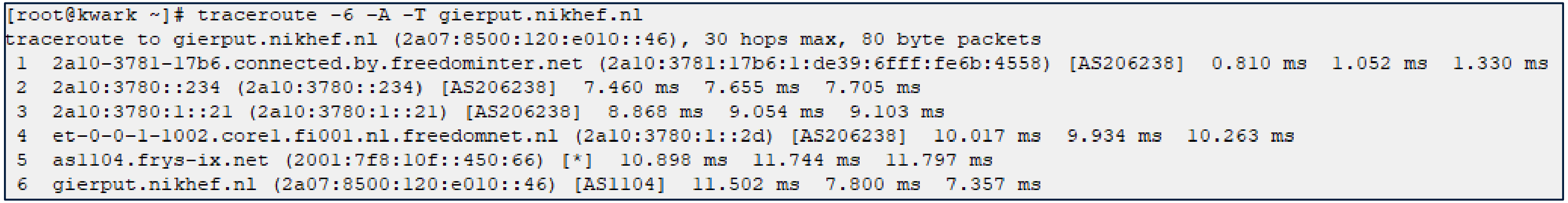


Traceroute measurement to linuxsoft.cern.ch (multihomed)

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

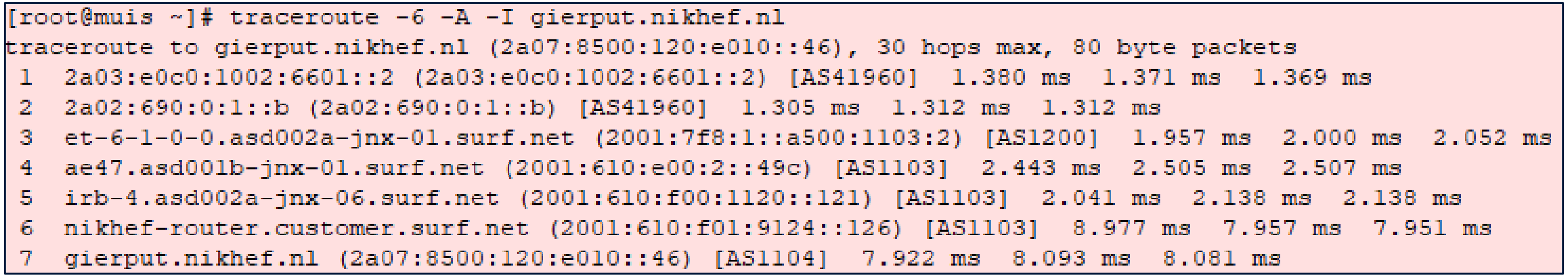
Many paths to Rome … i.e. to your server

- From a home connected to Freedom Internet to *spiegel.nikhef.nl*



```
[root@kwark ~]# traceroute -6 -A -T gierput.nikhef.nl
traceroute to gierput.nikhef.nl (2a07:8500:120:e010::46), 30 hops max, 80 byte packets
 1  2a10-3781-17b6.connected.by.freedominter.net (2a10:3781:17b6:1:de39:6fff:fe6b:4558) [AS206238]  0.810 ms  1.052 ms  1.330 ms
 2  2a10:3780::234 (2a10:3780::234) [AS206238]  7.460 ms  7.655 ms  7.705 ms
 3  2a10:3780:1::21 (2a10:3780:1::21) [AS206238]  8.868 ms  9.054 ms  9.103 ms
 4  et-0-0-1-1002.core1.fi001.nl.freedomnet.nl (2a10:3780:1::2d) [AS206238]  10.017 ms  9.934 ms  10.263 ms
 5  as1104.frys-ix.net (2001:7f8:10f::450:66) [*]  10.898 ms  11.744 ms  11.797 ms
 6  gierput.nikhef.nl (2a07:8500:120:e010::46) [AS1104]  11.502 ms  7.800 ms  7.357 ms
```

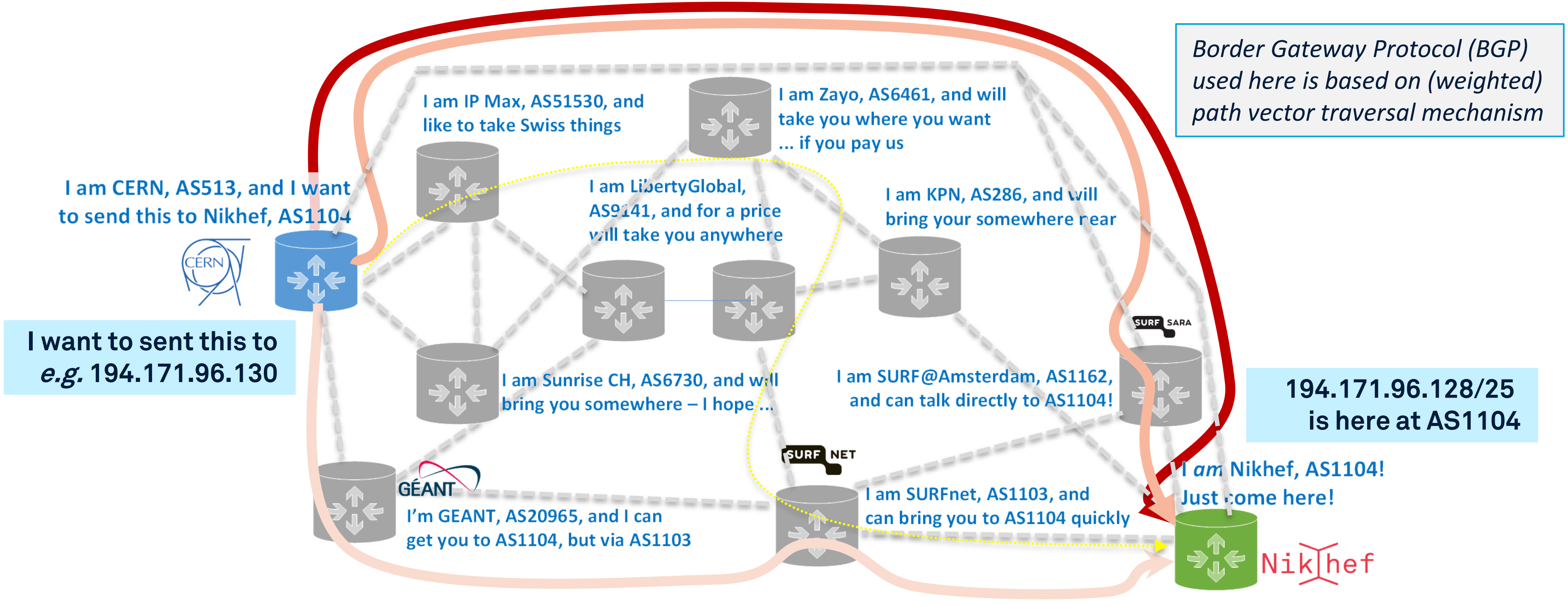
- but from Interparts in Lisse, NH:



```
[root@muis ~]# traceroute -6 -A -I gierput.nikhef.nl
traceroute to gierput.nikhef.nl (2a07:8500:120:e010::46), 30 hops max, 80 byte packets
 1  2a03:e0c0:1002:6601::2 (2a03:e0c0:1002:6601::2) [AS41960]  1.380 ms  1.371 ms  1.369 ms
 2  2a02:690:0:1::b (2a02:690:0:1::b) [AS41960]  1.305 ms  1.312 ms  1.312 ms
 3  et-6-1-0-0.asd002a-jnx-01.surf.net (2001:7f8:1::a500:1103:2) [AS1200]  1.957 ms  2.000 ms  2.052 ms
 4  ae47.asd001b-jnx-01.surf.net (2001:610:e00:2::49c) [AS1103]  2.443 ms  2.505 ms  2.507 ms
 5  irb-4.asd002a-jnx-06.surf.net (2001:610:f00:1120::121) [AS1103]  2.041 ms  2.138 ms  2.138 ms
 6  nikhef-router.customer.surf.net (2001:610:f01:9124::126) [AS1103]  8.977 ms  7.957 ms  7.951 ms
 7  gierput.nikhef.nl (2a07:8500:120:e010::46) [AS1104]  7.922 ms  8.093 ms  8.081 ms
```

AS41960: Interparts; AS1200: AMS-IX route reflector; AS1103: SURFnet; AS1104: Nikhef; AS206238: Freedom Internet – on the FrysIX there is direct L2 peering

Where do internet packets go anyway?


I am CERN, AS513, and I want to send this to Nikhef, AS1104
CERN
I am IP Max, AS51530, and like to take Swiss things
I am Zayo, AS6461, and will take you where you want ... if you pay us
I am LibertyGlobal, AS9141, and for a price will take you anywhere
I am KPN, AS286, and will bring your somewhere near
I am Sunrise CH, AS6730, and will bring you somewhere – I hope ...
I am SURF@Amsterdam, AS1162, and can talk directly to AS1104!
SURF SARA
SURF NET
GÉANT
I'm GEANT, AS20965, and I can get you to AS1104, but via AS1103
I am SURFnet, AS1103, and can bring you to AS1104 quickly
I *am* Nikhef, AS1104! Just come here!
Nikhef

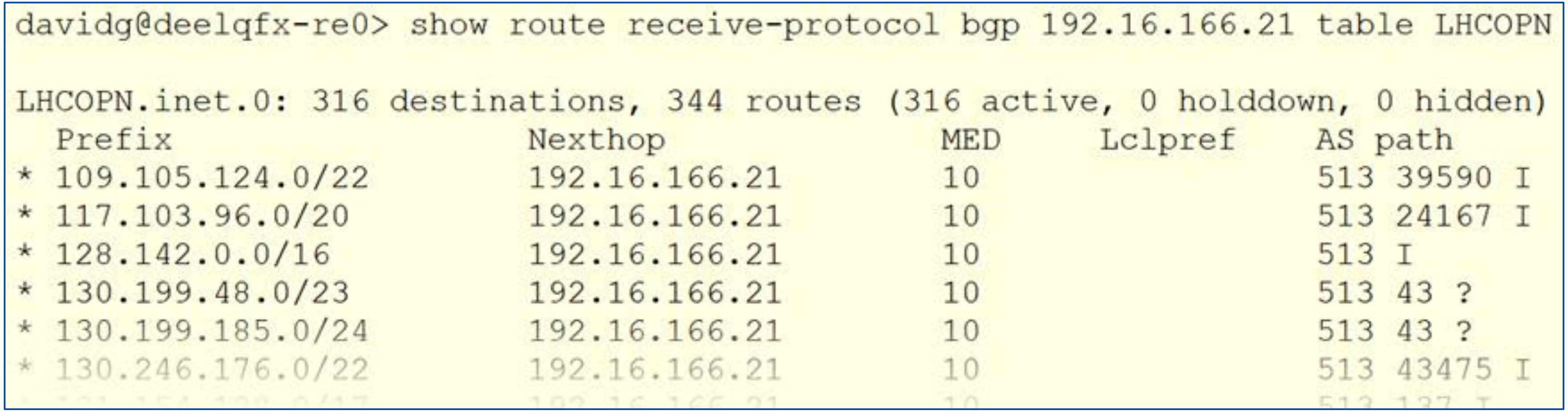
I want to sent this to *e.g.* 194.171.96.130

194.171.96.128/25 is here at AS1104

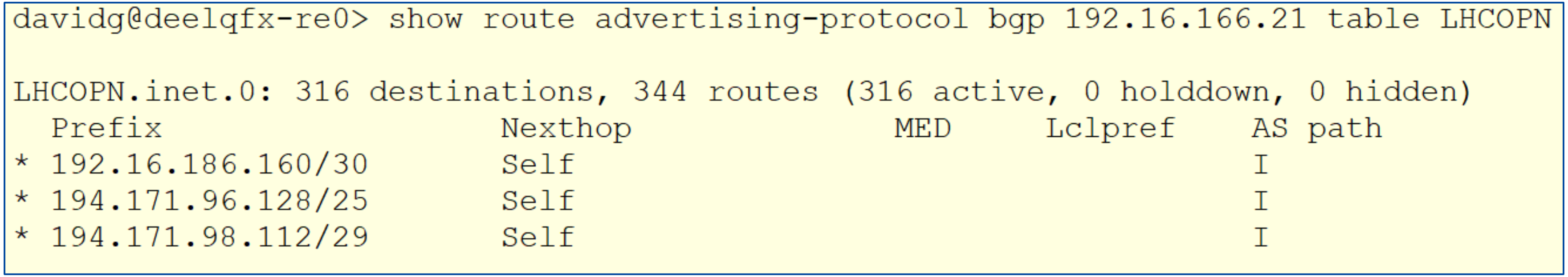
Border Gateway Protocol (BGP) used here is based on (weighted) path vector traversal mechanism

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)

Announcing routes: the Border Gateway Protocol



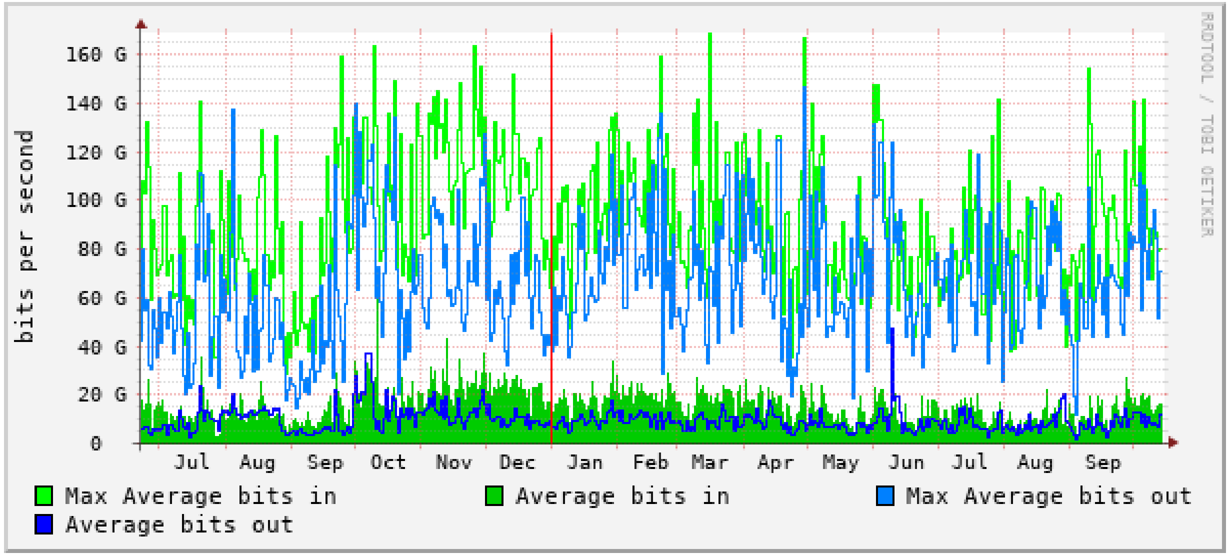
```
davidg@deelqfx-re0> show route receive-protocol bgp 192.16.166.21 table LHCOPN

LHCOPN.inet.0: 316 destinations, 344 routes (316 active, 0 holddown, 0 hidden)
  Prefix                  Nexthop              MED     Lclpref    AS path
* 109.105.124.0/22        192.16.166.21        10                 513 39590 I
* 117.103.96.0/20         192.16.166.21        10                 513 24167 I
* 128.142.0.0/16          192.16.166.21        10                 513 I
* 130.199.48.0/23         192.16.166.21        10                 513 43 ?
* 130.199.185.0/24        192.16.166.21        10                 513 43 ?
* 130.246.176.0/22        192.16.166.21        10                 513 43475 I
```



```
davidg@deelqfx-re0> show route advertising-protocol bgp 192.16.166.21 table LHCOPN

LHCOPN.inet.0: 316 destinations, 344 routes (316 active, 0 holddown, 0 hidden)
  Prefix                  Nexthop              MED     Lclpref    AS path
* 192.16.186.160/30       Self                                    I
* 194.171.96.128/25       Self                                    I
* 194.171.98.112/29       Self                                    I
```

IPv4 routes advertised from AS513/CERN (for all sites on LHCOPN) to AS1104/Nikhef (top), and the routes announced by AS1104/Nikhef to CERN, on 5 Nov 2022

Typical data traffic to and from the processing cluster



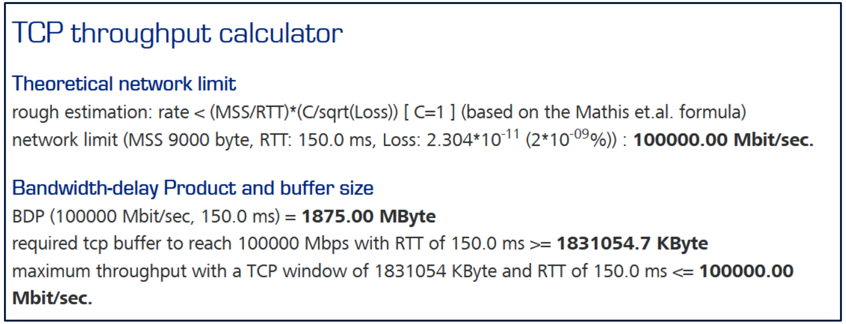
Source: Nikhef cricket graphs period June 2021 – October 2022 – aggregated (research) traffic to external peers from deelqfx – https://cricket.nikhef.nl/

Network is more than just what it says on the tin

More network bandwidth does not mean your *data* gets there faster

- memory requirements (since TCP needs a capability to re-transmit)

- tcp 'slow start'
- congestion control algorithms



TCP throughput calculator

Theoretical network limit

rough estimation: rate < (MSS/RTT)*(C/sqrt(Loss)) [C=1] (based on the Mathis et.al. formula)

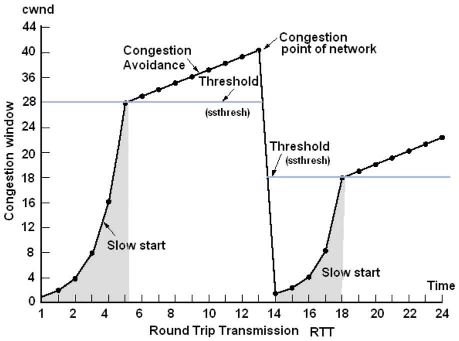
network limit (MSS 9000 byte, RTT: 150.0 ms, Loss: 2.304*10^{-11} (2*10^{-09}%)) : **100000.00 Mbit/sec.**

Bandwidth-delay Product and buffer size

BDP (100000 Mbit/sec, 150.0 ms) = **1875.00 MByte**

required tcp buffer to reach 100000 Mbps with RTT of 150.0 ms >= **1831054.7 KByte**

maximum throughput with a TCP window of 1831054 KByte and RTT of 150.0 ms <= **100000.00 Mbit/sec.**



Useful sources: https://www.switch.ch/network/tools/tcp_throughput/, https://fasterdata.es.net/

tcp slow-start graphic from Abed et al, *Improvement of TCP Congestion Window over LTE- Advanced Networks* IJoARiC&CE 2012

The cat video that destroyed it all …

latency AMS-GVA 17 ms
congestion event @20ms:
2 ms of UDP traffic to GVA

- TCP protocol sensitive to packet loss
    - 3 lost packets is enough to trigger this
- different congestion avoidance algorithms exists (~20 by now)
- loss severely impacts links w/large 'bandwidth-delay-product' (BDP)
- NL: ~3 ms, US East: 150ms

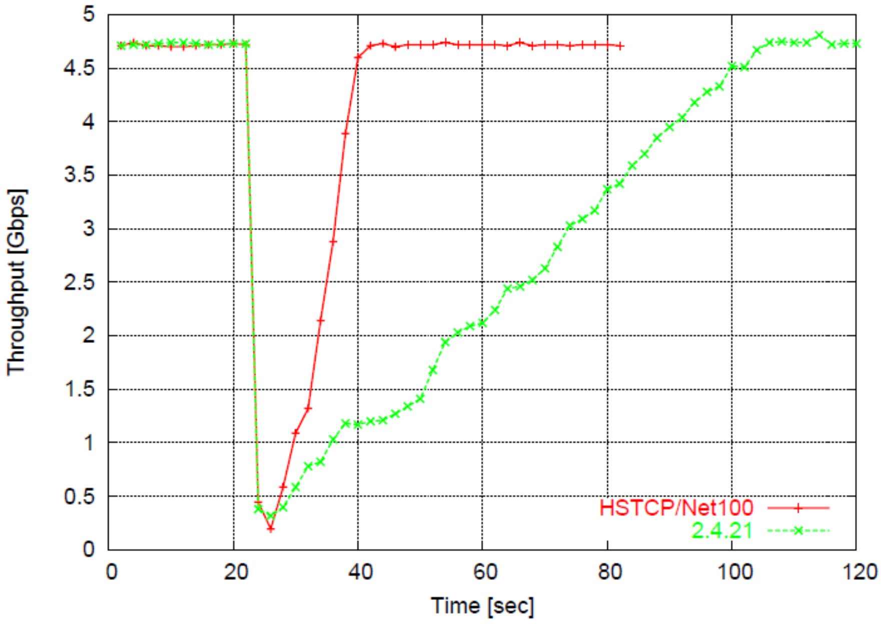


Figure 10: HSTCP versus stock TCP recovery time

source: Catalin Meirosu et al. *Native 10 Gigabit Ethernet experiments over long distances* in FGCS, doi:10.1016/j.future.2004.10.003 – aka. ATL-D-TN-0001

LHCOPN – distributing raw data

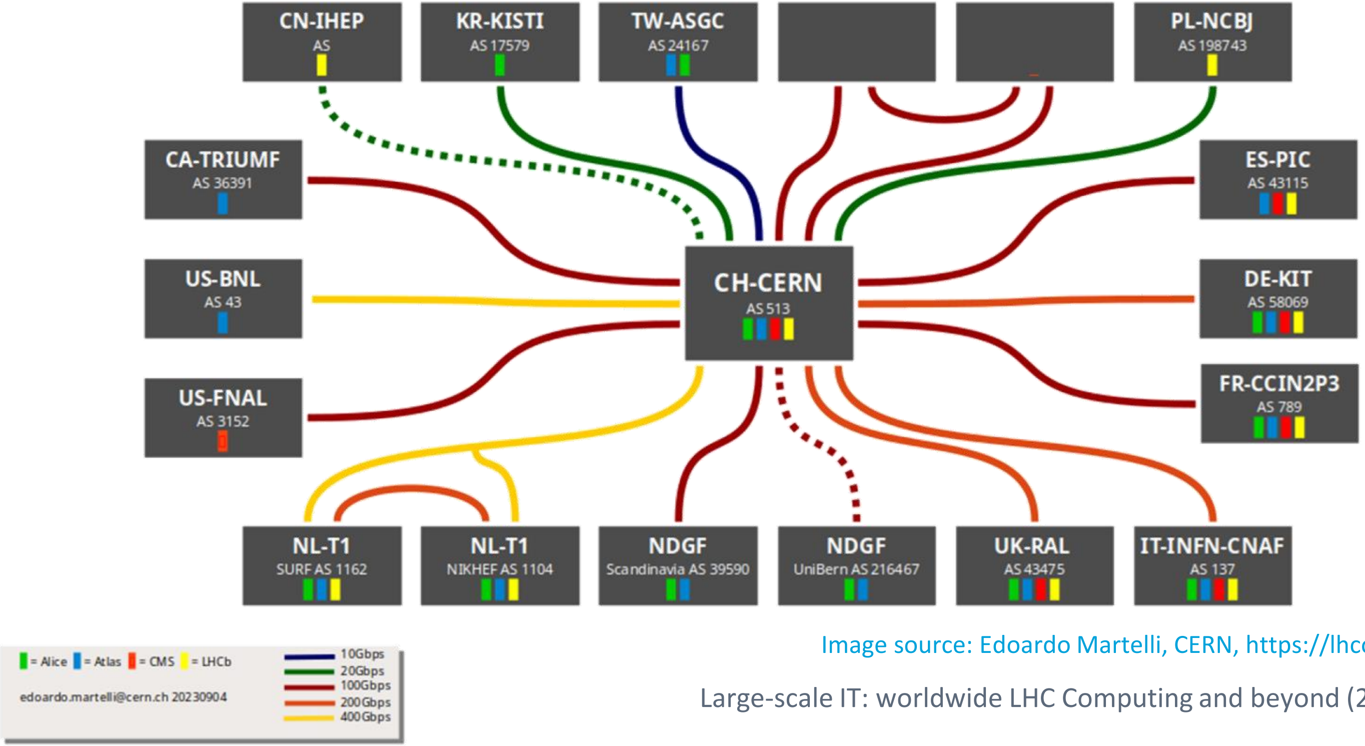
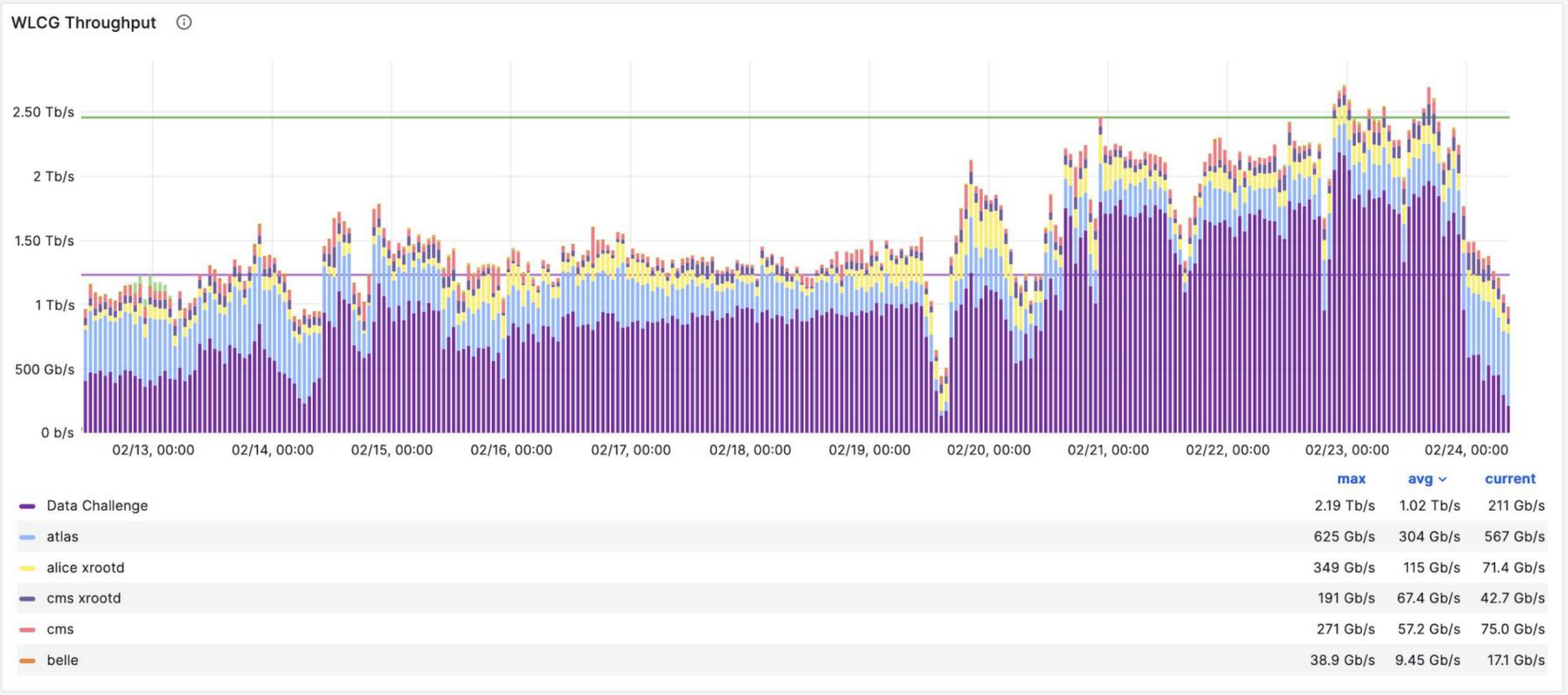


Image source: Edoardo Martelli, CERN, https://lhcopn.web.cern.ch/

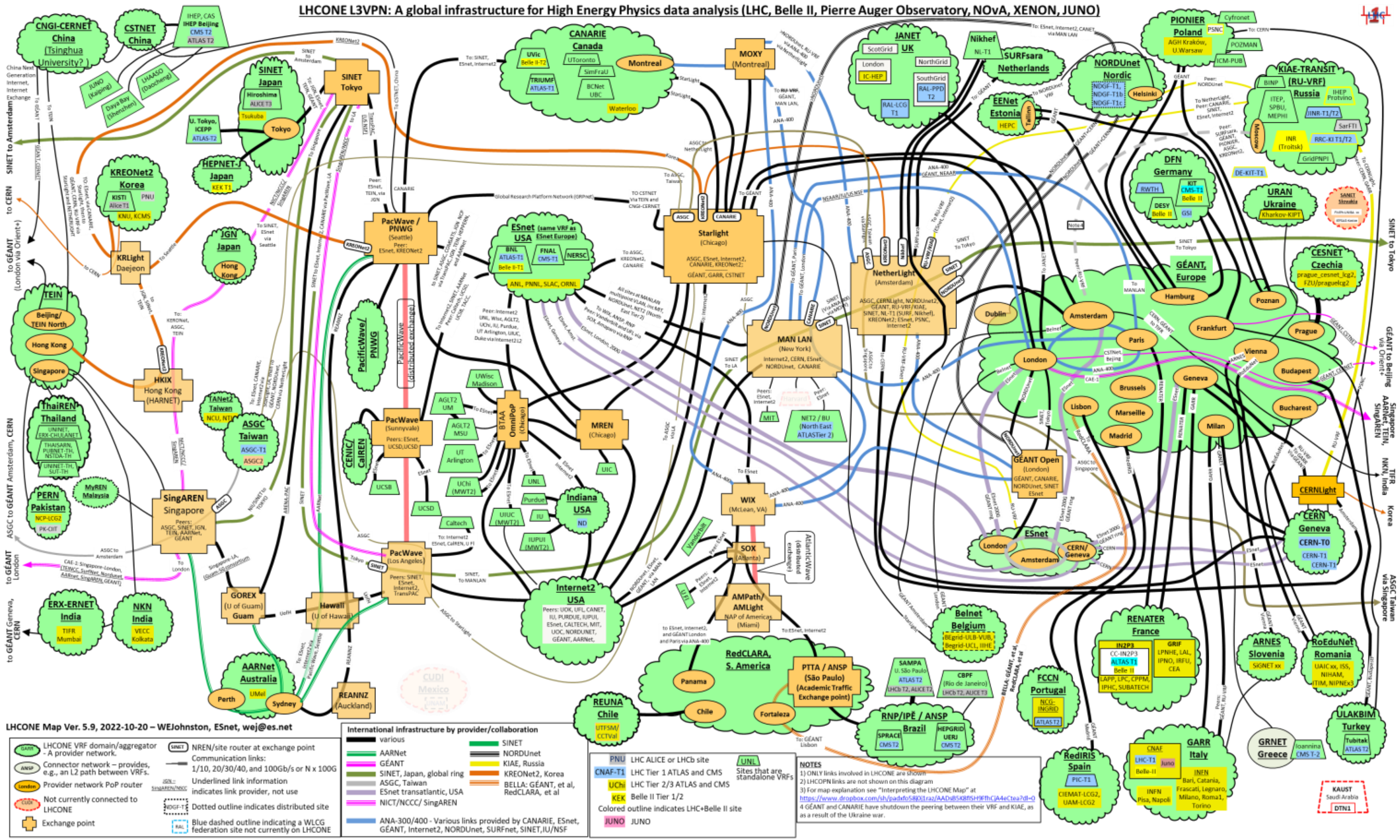
LHCOPN – traffic levels for data transfer (DC24)



	max	avg	current
Data Challenge	2.19 Tb/s	1.02 Tb/s	211 Gb/s
atlas	625 Gb/s	304 Gb/s	567 Gb/s
alice xrootd	349 Gb/s	115 Gb/s	71.4 Gb/s
cms xrootd	191 Gb/s	67.4 Gb/s	42.7 Gb/s
cms	271 Gb/s	57.2 Gb/s	75.0 Gb/s
belle	38.9 Gb/s	9.45 Gb/s	17.1 Gb/s

From Lassnig, M., & Wissing, C. et al. (2024). WLCG/DOMA Data Challenge 2024: Final Report. Zenodo. https://doi.org/10.5281/zenodo.11444180

LHCone



LHCONE L3VPN: A global infrastructure for High Energy Physics data analysis (LHC, Belle II, Pierre Auger Observatory, NOvA, XENON, JUNO)

LHCONE Map Ver. 5.9, 2022-10-20 – WEJohnston, ESnet, wej@es.net

GARR	LHCONE VRF domain/aggregator - A provider network.
ANSP	Connector network – provides, e.g., an L2 path between VRFs.
London	Provider network PoP router
CUDI	Not currently connected to LHCONE
✚	Exchange point
SINET	NREN/site router at exchange point
━	Communication links: 1/10, 20/30/40, and 100Gb/s or N x 100G
JGN – SingAREN/NICC	Underlined link information indicates link provider, not use
NDGF-T1	Dotted outline indicates distributed site
RAL	Blue dashed outline indicating a WLCG federation site not currently on LHCONE

International infrastructure by provider/collaboration

━	various	━	SINET
═	AARNet	═	NORDUnet
━	GÉANT	━	KIAE, Russia
━	SINET, Japan, global ring	━	KREONet2, Korea
━	ASGC, Taiwan	═	BELLA: GÉANT, et al, RedCLARA, et al
━	ESnet transatlantic, USA		
═	NICT/NCCC/ SingAREN		
━	ANA-300/400 - Various links provided by CANARIE, ESnet, GÉANT, Internet2, NORDUnet, SURFnet, SINET,IU/NSF		

PNU	LHC ALICE or LHCb site
CNAF-T1	LHC Tier 1 ATLAS and CMS
UChi	LHC Tier 2/3 ATLAS and CMS
KEK	Belle II Tier 1/2
UNL	Sites that are standalone VRFs
JUNO	JUNO

Colored outline indicates LHC+Belle II site

NOTES
1) ONLY links involved in LHCONE are shown
2) LHCOPN links are not shown on this diagram
3) For map explanation see "Interpreting the LHCONE Map" at https://www.dropbox.com/sh/padxfo58j0j1raz/AADsB5K8flSH9FfhCjA4eCtea?dl=0
4 GÉANT and CANARIE have shutdown the peering between their VRF and KIAE, as as a result of the Ukraine war.

LHCone ("LHC Open Network Environment") – visualization by Bill Johnston, ESnet version: October 2022 – updated with new AS1104 links

‘ScienceDMZ’

Predicable performance and data access for research

‘where research services, data, and researchers meet’

- latency hiding through caching
- **security zoning/segmentation** protects specific data sets
- **outside any enterprise perimeter**

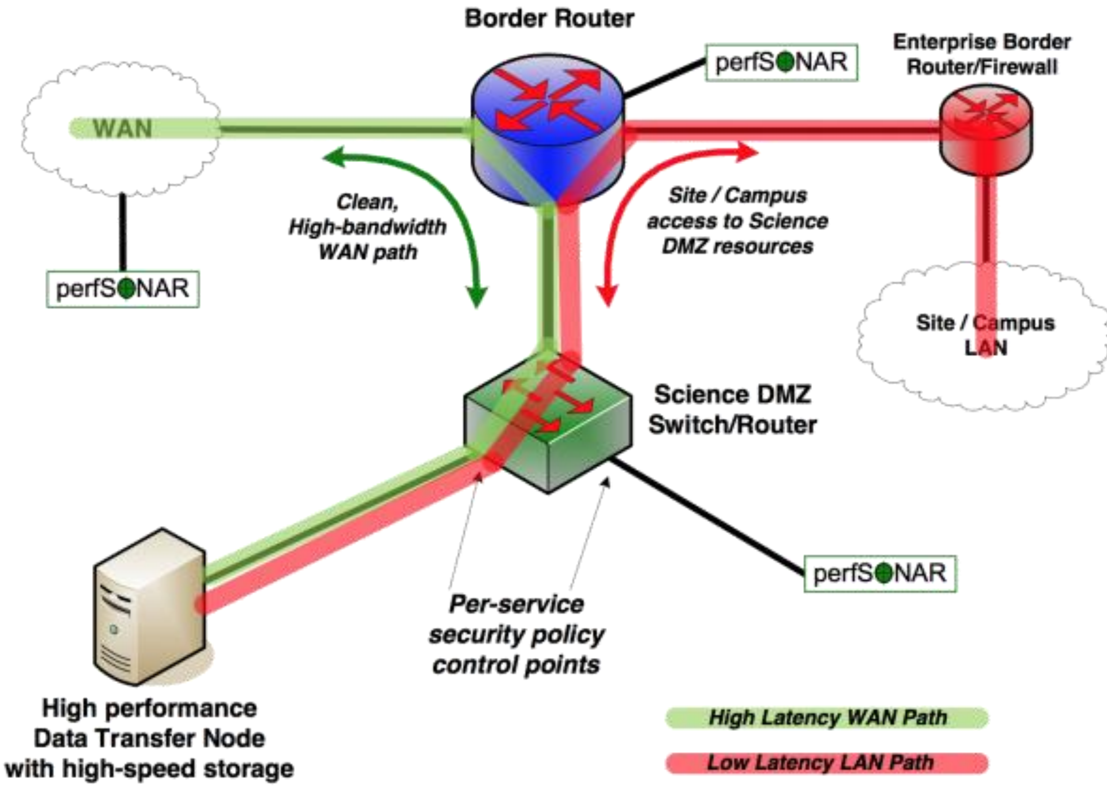
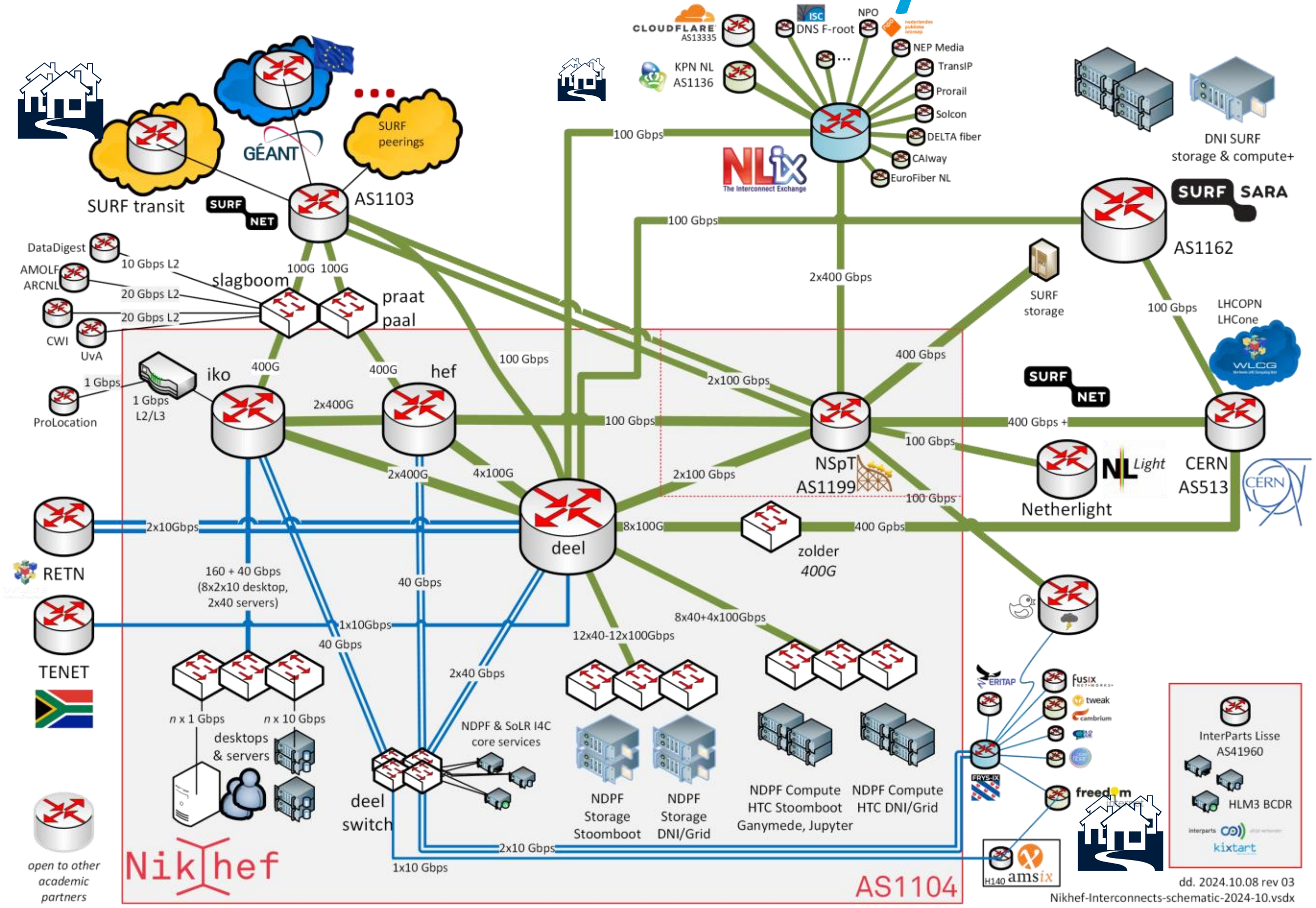

Border Router
perfSONAR
Enterprise Border Router/Firewall
WAN
perfSONAR
Clean, High-bandwidth WAN path
Site / Campus access to Science DMZ resources
Site / Campus LAN
Science DMZ Switch/Router
perfSONAR
Per-service security policy control points
High performance Data Transfer Node with high-speed storage
High Latency WAN Path
Low Latency LAN Path

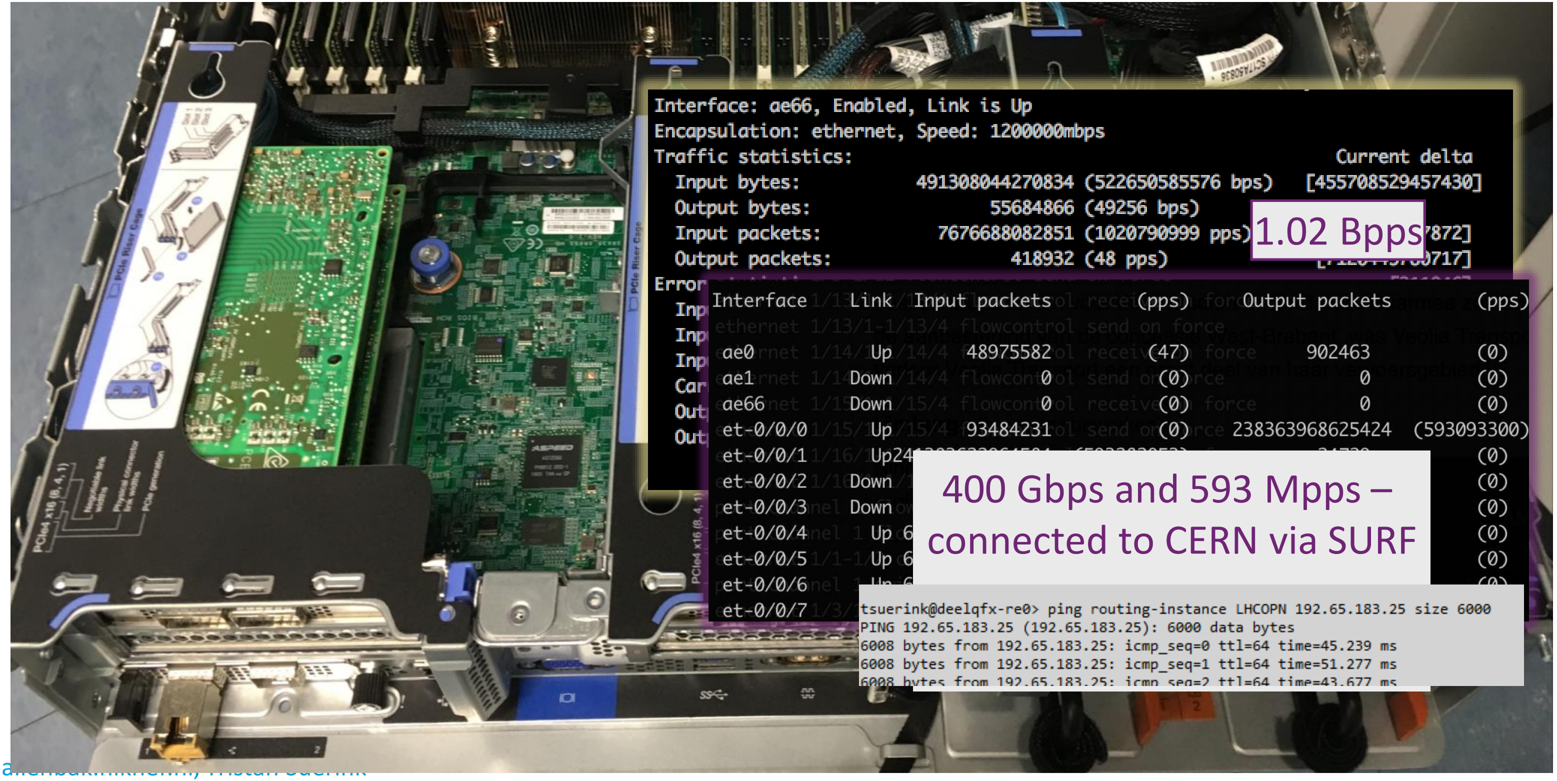
Image and ‘ScienceDMZ’ concept promulgated by ESnet (see fasterdata.es.net)

Just one random autonomous system: AS1104



AS1104
state as of Oct 2024

Exercising the network – sensor data and events



```
Interface: ae66, Enabled, Link is Up
Encapsulation: ethernet, Speed: 1200000mbps
Traffic statistics:                                          Current delta
  Input bytes:          491308044270834 (522650585576 bps)   [455708529457430]
  Output bytes:                55684866 (49256 bps)
  Input packets:          7676688082851 (1020790999 pps)
  Output packets:                418932 (48 pps)
Error
  Inp
  Inp
  Inp
  Car
  Out
  Out
```

1.02 Bpps

```
Interface     Link   Input packets      (pps)    Output packets         (pps)
 ae0          Up       48975582         (47)         902463             (0)
 ae1          Down            0          (0)              0             (0)
 ae66         Down            0          (0)              0             (0)
 et-0/0/0     Up       93484231          (0)  238363968625424   (593093300)
 et-0/0/1     Up                                                        (0)
 et-0/0/2     Down                                                      (0)
 et-0/0/3     Down                                                      (0)
 et-0/0/4     Up                                                        (0)
 et-0/0/5     Up                                                        (0)
 et-0/0/6
 et-0/0/7
```

400 Gbps and 593 Mpps – connected to CERN via SURF

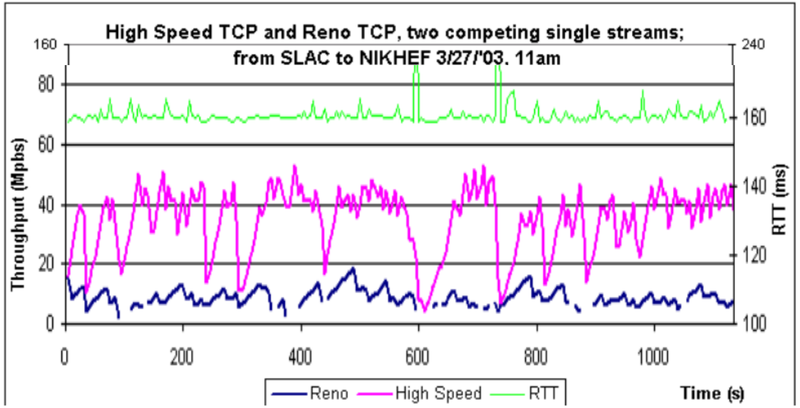
```
tsuerink@deelqfx-re0> ping routing-instance LHCOPN 192.65.183.25 size 6000
PING 192.65.183.25 (192.65.183.25): 6000 data bytes
6008 bytes from 192.65.183.25: icmp_seq=0 ttl=64 time=45.239 ms
6008 bytes from 192.65.183.25: icmp_seq=1 ttl=64 time=51.277 ms
6008 bytes from 192.65.183.25: icmp_seq=2 ttl=64 time=43.677 ms
```

Image: ba

Scaling data access: ‘system-aware design’ at application layer

Reading data ‘scattered’ in a file - simply using POSIX-like IO - when done over the network severely exposes latency

and TCP slow-start makes that even worse



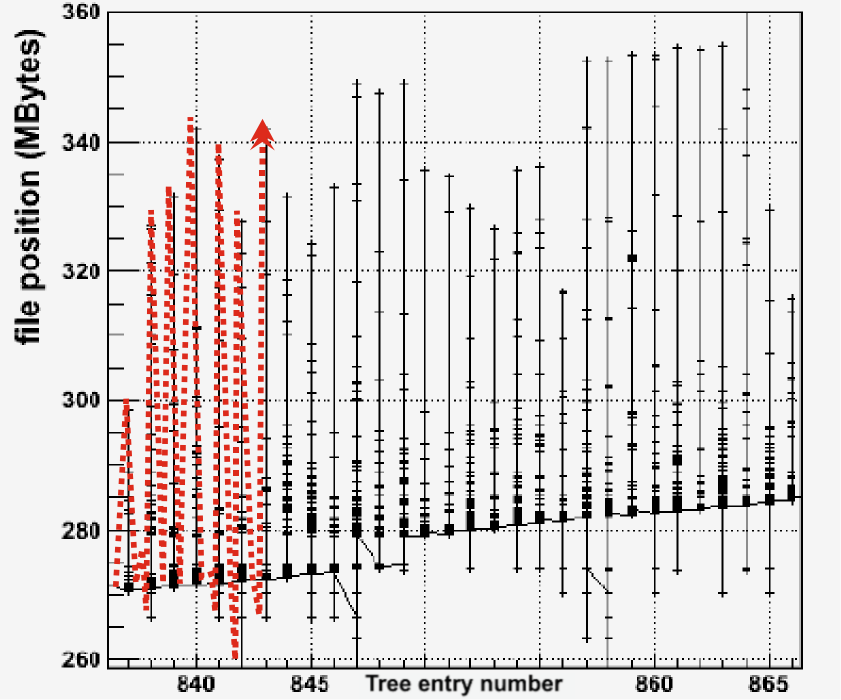
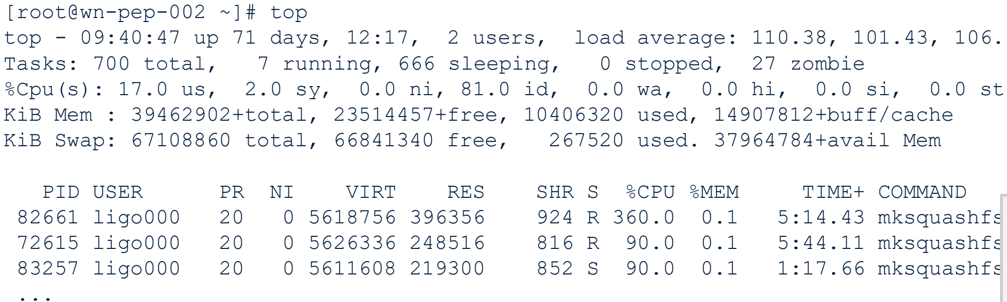


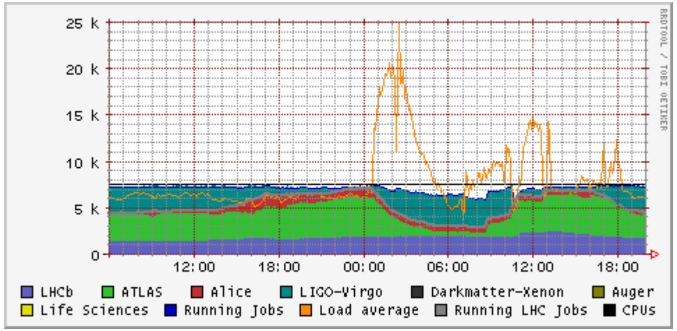
Image of TCP slow-start and packet loss impact (in Mpps): Antony Antony et al., Nikhef, for DataTAG, 2003(!)
Right: base graphic: Philippe Canal “Root I/O: the fast and the furious”, CHEP2010 Access pattern reflects Root versions < 5.28, before Ttree caching and ‘baskets’

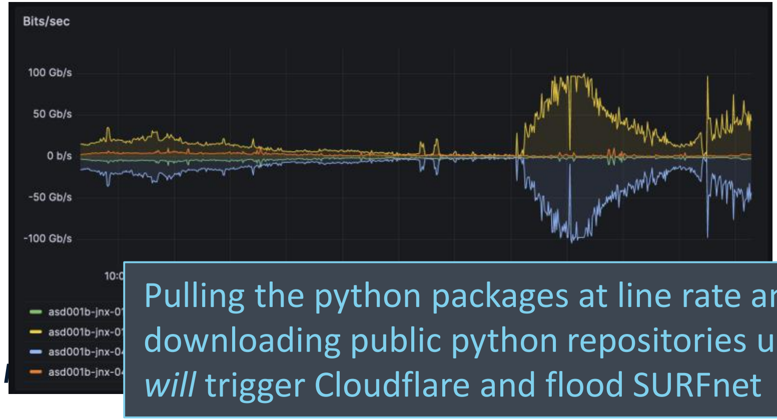
And sometimes traffic is triggered by researchers scaling up ‘accidentally’ from a laptop to a cluster without too much thought

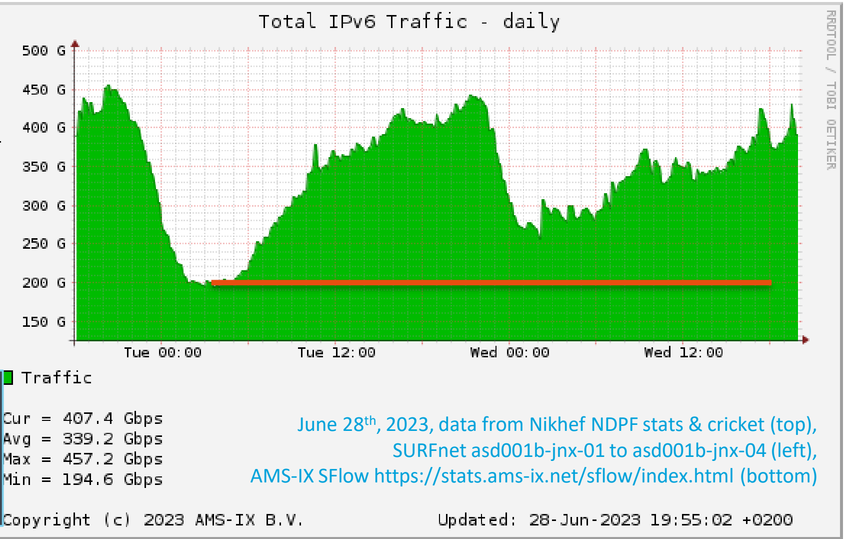
A researcher doing mass creation of containers, rebuilding their python ‘virtual env’ for each job, running on >> 4000 cores



```
[root@wn-pep-002 ~]# top
top - 09:40:47 up 71 days, 12:17,  2 users,  load average: 110.38, 101.43, 106.3
Tasks: 700 total,   7 running, 666 sleeping,   0 stopped,  27 zombie
%Cpu(s): 17.0 us,  2.0 sy,  0.0 ni, 81.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
KiB Mem : 39462902+total, 23514457+free, 10406320 used, 14907812+buff/cache
KiB Swap: 67108860 total, 66841340 free,   267520 used. 37964784+avail Mem

  PID USER      PR  NI    VIRT    RES    SHR S  %CPU %MEM     TIME+ COMMAND
82661 ligo000   20   0 5618756 396356    924 R 360.0  0.1   5:14.43 mksquashfs
72615 ligo000   20   0 5626336 248516    816 R  90.0  0.1   5:44.11 mksquashfs
83257 ligo000   20   0 5611608 219300    852 S  90.0  0.1   1:17.66 mksquashfs
...
```





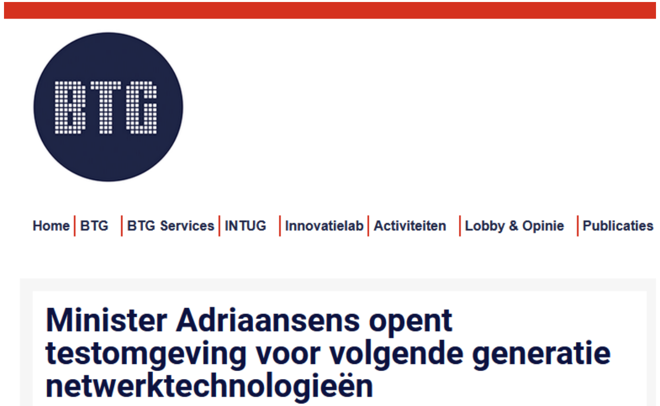


Pulling the python packages at line rate and downloading public python repositories ultimately *will* trigger Cloudflare and flood SURFnet

June 28th, 2023, data from Nikhef NDPF stats & cricket (top), SURFnet asd001b-jnx-01 to asd001b-jnx-04 (left), AMS-IX SFlow https://stats.ams-ix.net/sflow/index.html (bottom)

For example for HL-LHC, or SKA, more is needed > 2028 …

- ‘Typical’ network is now mixed 400G-100G
- Push experiments to 800Gbps in metro area, and a local (AMS) loop has been demonstrated
- next: 800 → 1600G AMS-GVA ☺



BTG
Home | BTG | BTG Services | INTUG | Innovatielab | Activiteiten | Lobby & Opinie | Publicaties

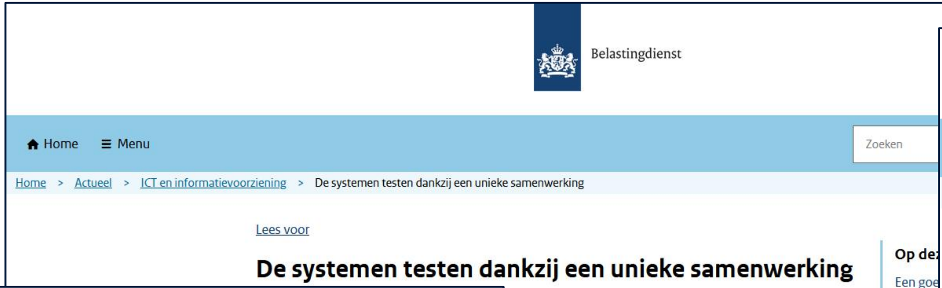
Minister Adriaansens opent testomgeving voor volgende generatie netwerktechnologieën

in Amsterdam is door minister Micky Adriaansens van Economische Zaken en Klimaat
tierotonde is een testomgeving waar SURF en Nikhef gaan experimenteren met nieuwe
ing beschikt over een internetsnelheid van 800 Gbit/s, wat meer dan 1000 keer sneller
n gemiddeld huishouden in Nederland. De innovatierotonde stelt Nederlandse
te doen naar de volgende generatie netwerktechnologieën.

en onderzoek naar bandbreedte op het internet groeit. Onderzoekers willen steeds meer
over de landsgrenzen heen met elkaar delen. De bandbreedte van het netwerk speelt
rote hoeveelheden data snel te kunnen verwerken, is de verwachting dat 800Gbit/s
n. De innovatierotonde maakt het mogelijk om te experimenteren met nieuwe

Web screenshot: btg.org,
Images Nokia 7750-SR1x in Nikhef AMS H234b: Tristan Suerink

Research data traffic looks like … a DDoS to others ☺



Belastingdienst

Home ≡ Menu Zoeken

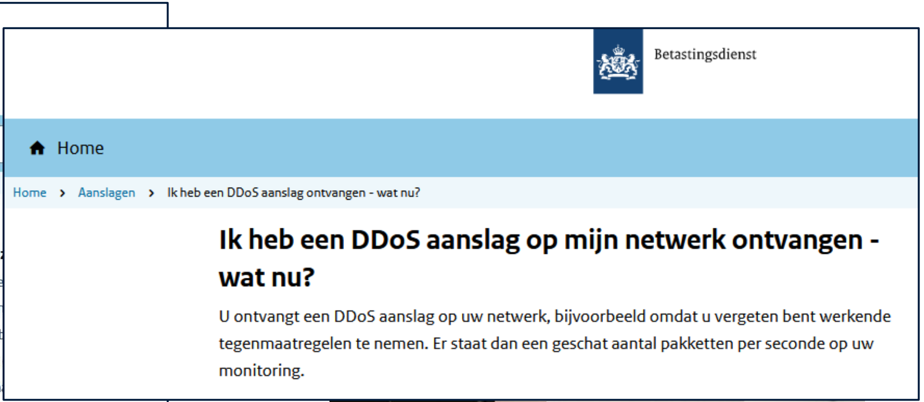
Home > Actueel > ICT en informatievoorziening > De systemen testen dankzij een unieke samenwerking

Lees voor

De systemen testen dankzij een unieke samenwerking

Het begon in 2018. Een bijzondere samenwerking tussen overheden, internetproviders- en exchanges, academische instanties, non- pr du sin n plat vapen alitie.

Op de Een goe Examen Wat geb Vragen Terug n



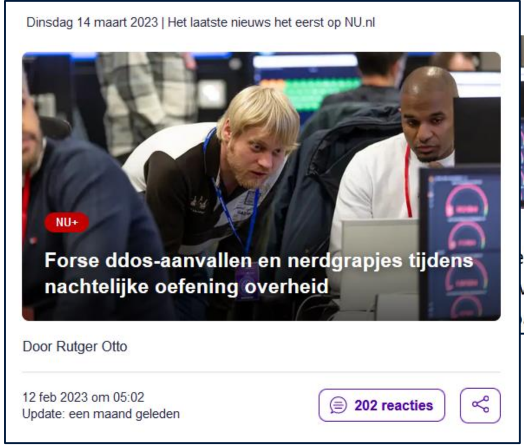
Betastingsdienst

Home

Home › Aanslagen › Ik heb een DDoS aanslag ontvangen - wat nu?

Ik heb een DDoS aanslag op mijn netwerk ontvangen - wat nu?

U ontvangt een DDoS aanslag op uw netwerk, bijvoorbeeld omdat u vergeten bent werkende tegenmaatregelen te nemen. Er staat dan een geschat aantal pakketten per seconde op uw monitoring.



Dinsdag 14 maart 2023 | Het laatste nieuws het eerst op NU.nl

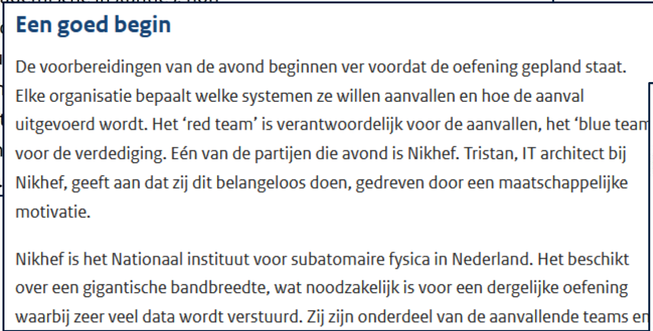
NU+

Forse ddos-aanvallen en nerdgrapjes tijdens nachtelijke oefening overheid

Door Rutger Otto

12 feb 2023 om 05:02
Update: een maand geleden

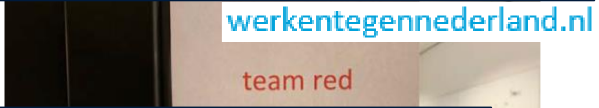
202 reacties



Een goed begin

De voorbereidingen van de avond beginnen ver voordat de oefening gepland staat. Elke organisatie bepaalt welke systemen ze willen aanvallen en hoe de aanval uitgevoerd wordt. Het 'red team' is verantwoordelijk voor de aanvallen, het 'blue team voor de verdediging. Eén van de partijen die avond is Nikhef. Tristan, IT architect bij Nikhef, geeft aan dat zij dit belangeloos doen, gedreven door een maatschappelijke motivatie.

Nikhef is het Nationaal instituut voor subatomaire fysica in Nederland. Het beschikt over een gigantische bandbreedte, wat noodzakelijk is voor een dergelijke oefening waarbij zeer veel data wordt verstuurd. Zij zijn onderdeel van de aanvallende teams en



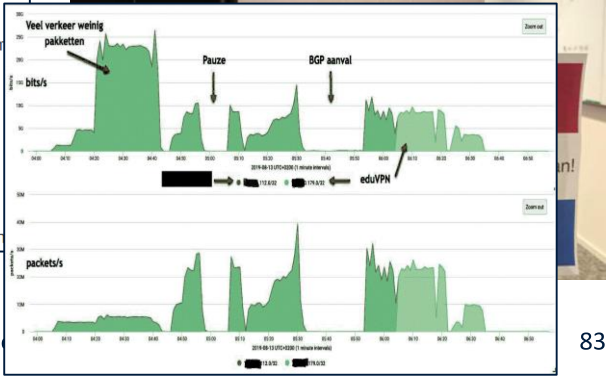


Image sources: belastingdienst.nl, rws.nl, nu.nl

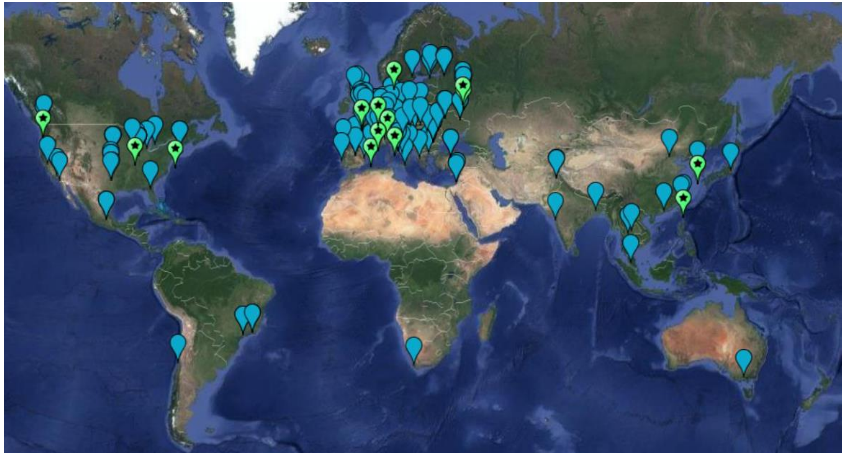
Access, Trust & Identity

More than one user, *from*
more than one organizational domain, *in*
more than one country

WLCG: when we met a global trust scaling issue

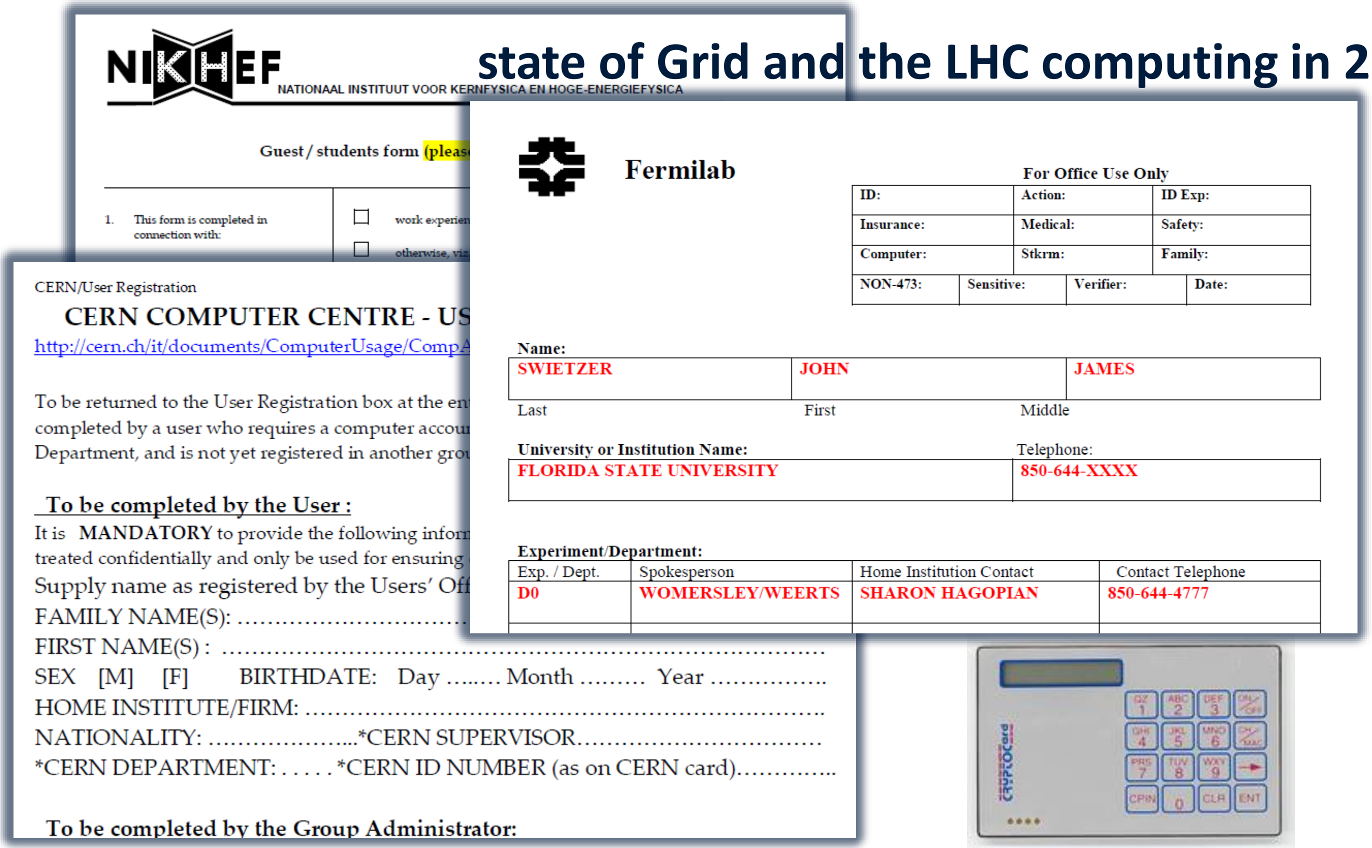
- 170 sites
- ~50 countries & regions
- ~20000 users

so … just *how* many interactions ??



people photo: a small part of the CMS collaboration in 2017, Credit: CMS-PHO-PUBLIC-2017-004-3; site map: WLCG sites from Maarten Litmaath (CERN) 2021

Scaling issues – credentials at each site does not work



state of Grid and the LHC computing in 2000

NIKHEF
NATIONAAL INSTITUUT VOOR KERNFYSICA EN HOGE-ENERGIEFYSICA

Guest / students form (pleas

1. This form is completed in connection with:

□ work experien

□ otherwise, vi

CERN/User Registration

CERN COMPUTER CENTRE - US

http://cern.ch/it/documents/ComputerUsage/CompA

To be returned to the User Registration box at the en
completed by a user who requires a computer accou
Department, and is not yet registered in another gro

To be completed by the User :

It is **MANDATORY** to provide the following infor
treated confidentially and only be used for ensuring
Supply name as registered by the Users' Off
FAMILY NAME(S): ……………………………
FIRST NAME(S) : ………………………………………………………………………
SEX [M] [F] BIRTHDATE: Day …..…. Month ……… Year …………….
HOME INSTITUTE/FIRM: ………………………………………………………………….
NATIONALITY: ………………...*CERN SUPERVISOR……………………………
*CERN DEPARTMENT: . . . . . *CERN ID NUMBER (as on CERN card)…………..

To be completed by the Group Administrator:

Fermilab

For Office Use Only

ID:	**Action:**		**ID Exp:**
Insurance:	**Medical:**		**Safety:**
Computer:	**Stkrm:**		**Family:**
NON-473:	**Sensitive:**	**Verifier:**	**Date:**

Name:

SWIETZER	JOHN	JAMES
Last	First	Middle

University or Institution Name:	Telephone:
FLORIDA STATE UNIVERSITY	850-644-XXXX

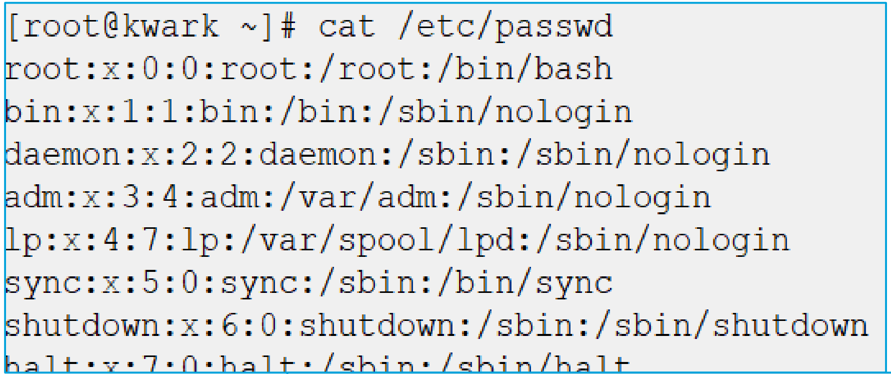
Experiment/Department:

Exp. / Dept.	Spokesperson	Home Institution Contact	Contact Telephone
D0	WOMERSLEY/WEERTS	SHARON HAGOPIAN	850-644-4777

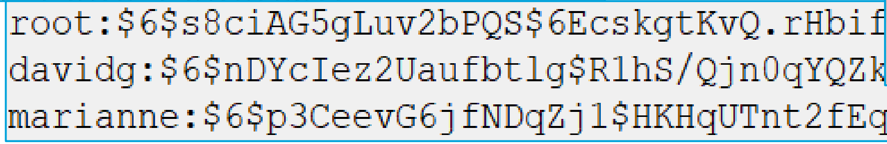
Authentication – proving *who* are you

Authenticating to a *single service* is relatively simple

- per-service identity (username) and secrets (e.g. password or TOTP token)
- server-side: list of valid users and (hashed and hopefully salted) secrets



```
[root@kwark ~]# cat /etc/passwd
root:x:0:0:root:/root:/bin/bash
bin:x:1:1:bin:/bin:/sbin/nologin
daemon:x:2:2:daemon:/sbin:/sbin/nologin
adm:x:3:4:adm:/var/adm:/sbin/nologin
lp:x:4:7:lp:/var/spool/lpd:/sbin/nologin
sync:x:5:0:sync:/sbin:/bin/sync
shutdown:x:6:0:shutdown:/sbin:/sbin/shutdown
halt:x:7:0:halt:/sbin:/sbin/halt
```



```
root:$6$s8ciAG5gLuv2bPQS$6EcskgtKvQ.rHbif
davidg:$6$nDYcIez2Uaufbtlg$R1hS/Qjn0qYQZk
marianne:$6$p3CeevG6jfNDqZj1$HKHqUTnt2fEqQfkA/m5J3oAOA0zSvgLCKOSQhPS
```

Passport image: cropped from original by Jon Tyson on Unsplash https://unsplash.com/photos/Hid-yhommOg

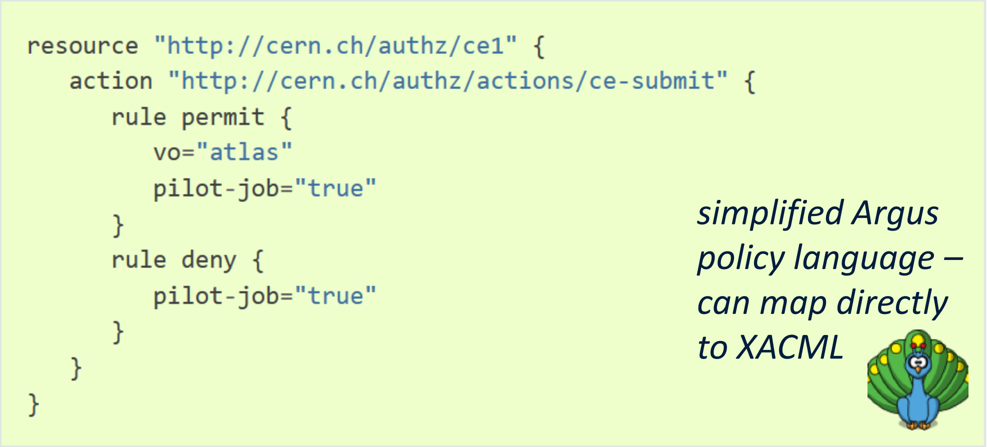
Authorization – what you are allowed to do

soon needs specifying **access rights** to resources, based on an access **policy**

- might be implicit or ad-hoc

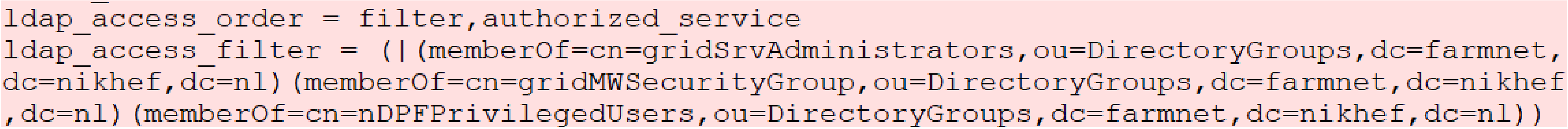
- be in formal policy language
  like XACML (*example: Argus PDP)*

- or be service-specific
  example: Linux sssd config



```
resource "http://cern.ch/authz/ce1" {
   action "http://cern.ch/authz/actions/ce-submit" {
      rule permit {
         vo="atlas"
         pilot-job="true"
      }
      rule deny {
         pilot-job="true"
      }
   }
}
```

simplified Argus policy language – can map directly to XACML



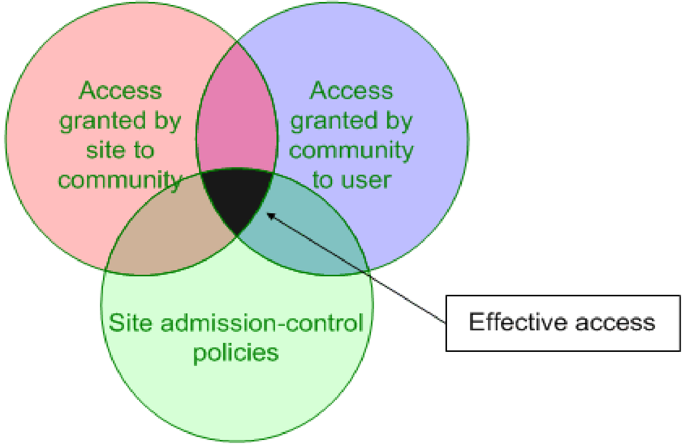
```
ldap_access_order = filter,authorized_service
ldap_access_filter = (|(memberOf=cn=gridSrvAdministrators,ou=DirectoryGroups,dc=farmnet,
dc=nikhef,dc=nl)(memberOf=cn=gridMWSecurityGroup,ou=DirectoryGroups,dc=farmnet,dc=nikhef
,dc=nl)(memberOf=cn=nDPFPrivilegedUsers,ou=DirectoryGroups,dc=farmnet,dc=nikhef,dc=nl))
```

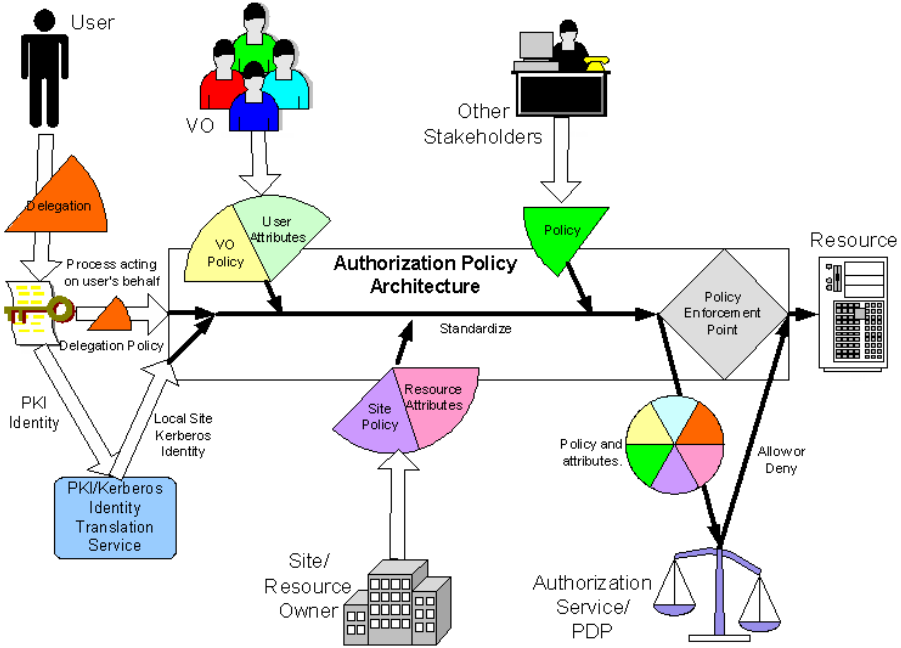
Policy example: Argus system, https://argus-documentation.readthedocs.io/en/stable/misc/examples.html; service-specific: sssd.conf ldap auth_provider

Authorization and access control

Access control is ultimately enforced by the service provider

(unless data-level encryption is used, where the data owner retains some control)


Access granted by site to community
Access granted by community to user
Site admission-control policies
Effective access



policy overlap diagram by Olle Mulmo, KTH for EGEE-I JRA3, policy pie: OpenGrod Forum OGSA working group and Globus Alliance

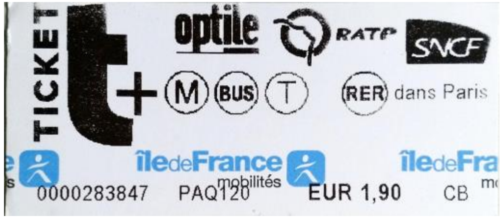
Authorization policy subjects

AuthZ policies need subject attributes (‘claims’)

- **bound to an verifiable identity** statement
  - e.g. visa are strongly linked to a specific entity, and asserted by a trusted party (by the service)
- be a **bearer token**
  - scoped to a relying party, a service, or an action
- **self-asserted**
  - quite useless unless backed by verifiable evidence, like in self-sovereign identity schemes

Transport mechanisms (see also RFC2903)

- pushed alongside the service access,
- pulled from the source as needed, or
- pushed by the attribute source as an agent



USA visa image source: https://2009-2017.state.gov/m/ds/rls/rpt/79785.htm ; RATP bearer token, issued for the Paris public transport system

Access control in a single domain

- Dedicated to each service where you need access

- Usually strongly linked to authorization: at times even different accounts for different roles

- In a multi-organizational system becomes

$$\mathcal{O}(n_{sites} * n_{services}) * \mathcal{O}(n_{users})$$

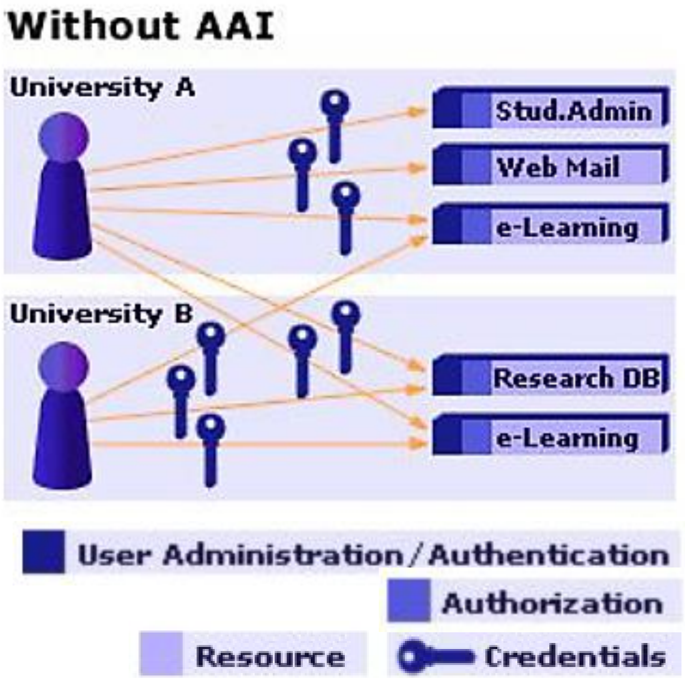


Image: AARC NA2 training module “Authentication and Authorisation 101” - https://aarc-community.org/training/aai-101/

Authentication and Authorization Infrastructure

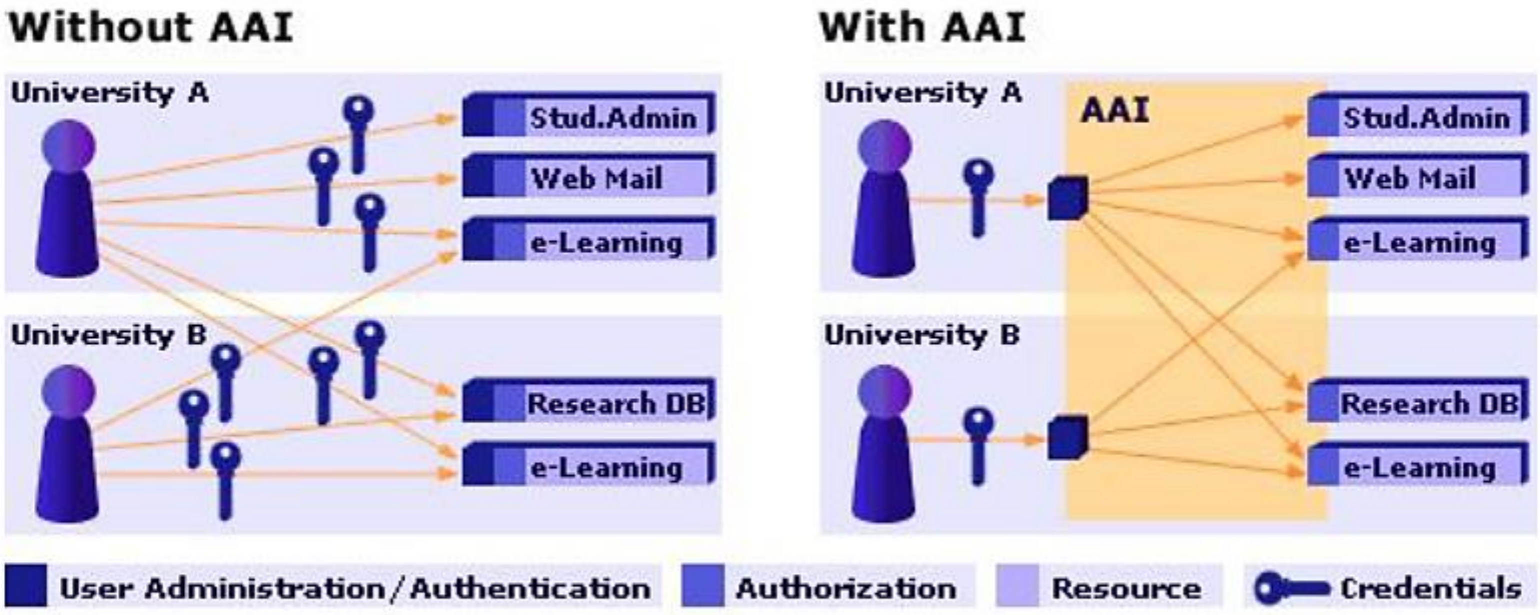
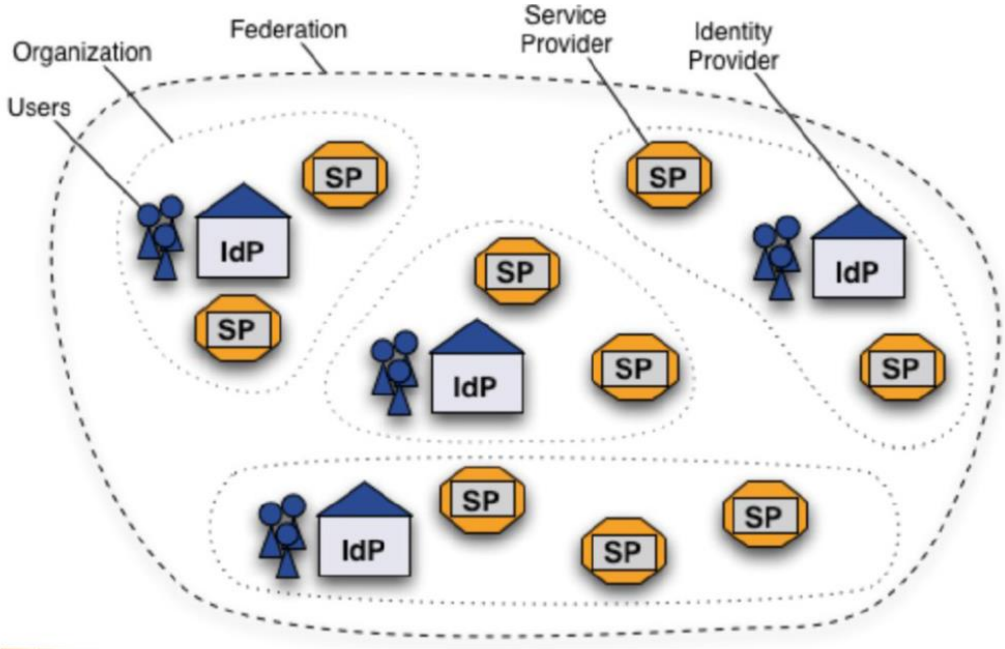


Image: AARC NA2 training module "Authentication and Authorisation 101" - https://aarc-community.org/training/aai-101/

Federation

portability of identity information across otherwise autonomous administrative domains

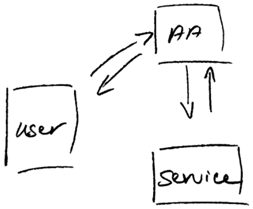


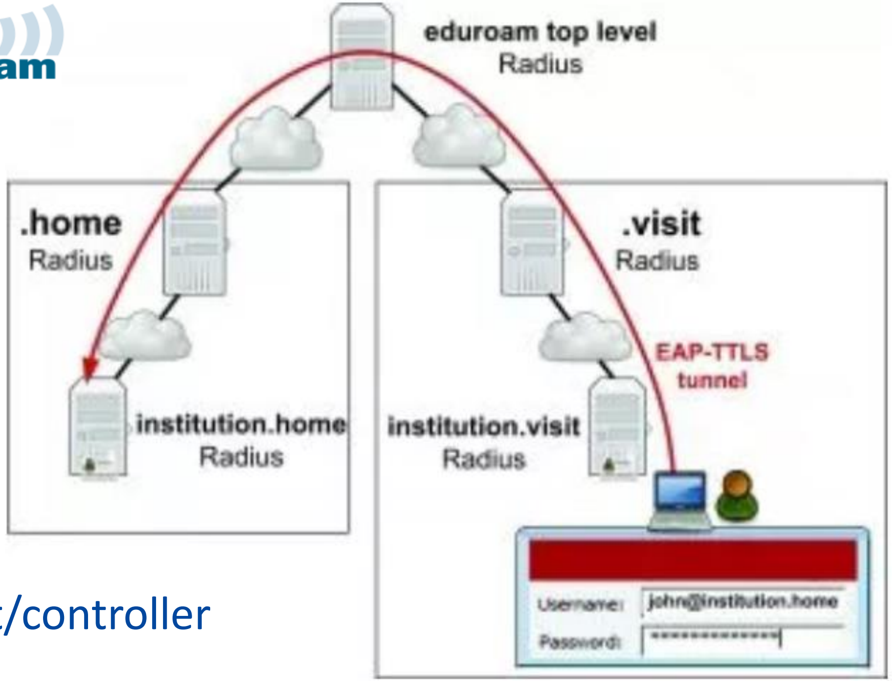
Shibboleth IdP image and SAML2 auth flow by SWITCH (CH) – see also https://refeds.org/ on federation structure and (assurance and security) guidelines

One simple federation you know: eduroam

service-specific trust between organisations globally

hierarchical RADIUS servers based an 802.1x secure exchange over TLS or EAP-TTLS tunneling your credentials back to your home institution



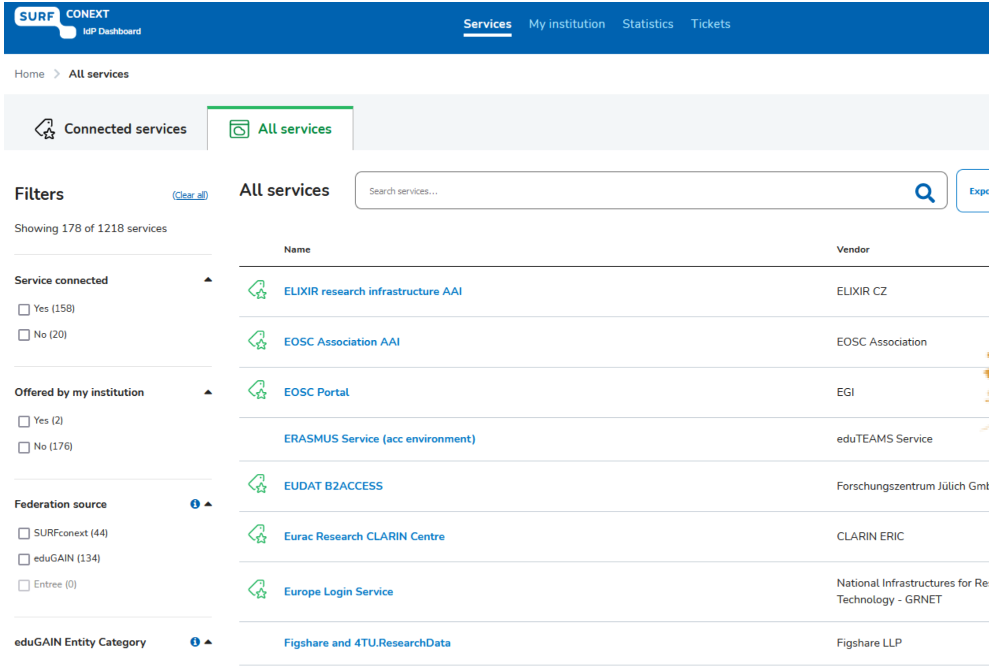


Username: john@institution.home
Password: *************

RADIUS server then instructs WiFi access point/controller

eduroam: Klaas Wieringa et al., image from https://eduroam.org/how/, GEANT ; RADIUS: RC2865 https://www.rfc-editor.org/rfc/rfc2865; see also freeradius.org

Multipurpose federation with SAML: SURFconext & eduGAIN



SURF CONEXT IdP Dashboard

Services | My institution | Statistics | Tickets | DG

Home > All services

Connected services | All services

Filters (Clear all)

Showing 178 of 1218 services

Service connected

□ Yes (158)

□ No (20)

Offered by my institution

□ Yes (2)

□ No (176)

Federation source

□ SURFconext (44)

□ eduGAIN (134)

□ Entree (0)

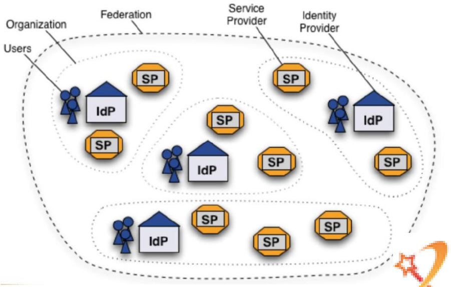
eduGAIN Entity Category

All services

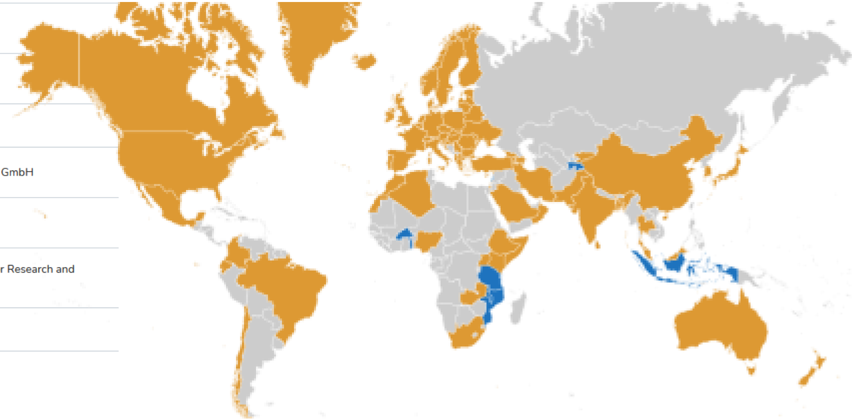
Search services...

Export overview as csv

Name	Vendor
ELIXIR research infrastructure AAI	ELIXIR CZ
EOSC Association AAI	EOSC Association
EOSC Portal	EGI
ERASMUS Service (acc environment)	eduTEAMS Service
EUDAT B2ACCESS	Forschungszentrum Jülich GmbH
Eurac Research CLARIN Centre	CLARIN ERIC
Europe Login Service	National Infrastructures for Research and Technology - GRNET
Figshare and 4TU.ResearchData	Figshare LLP

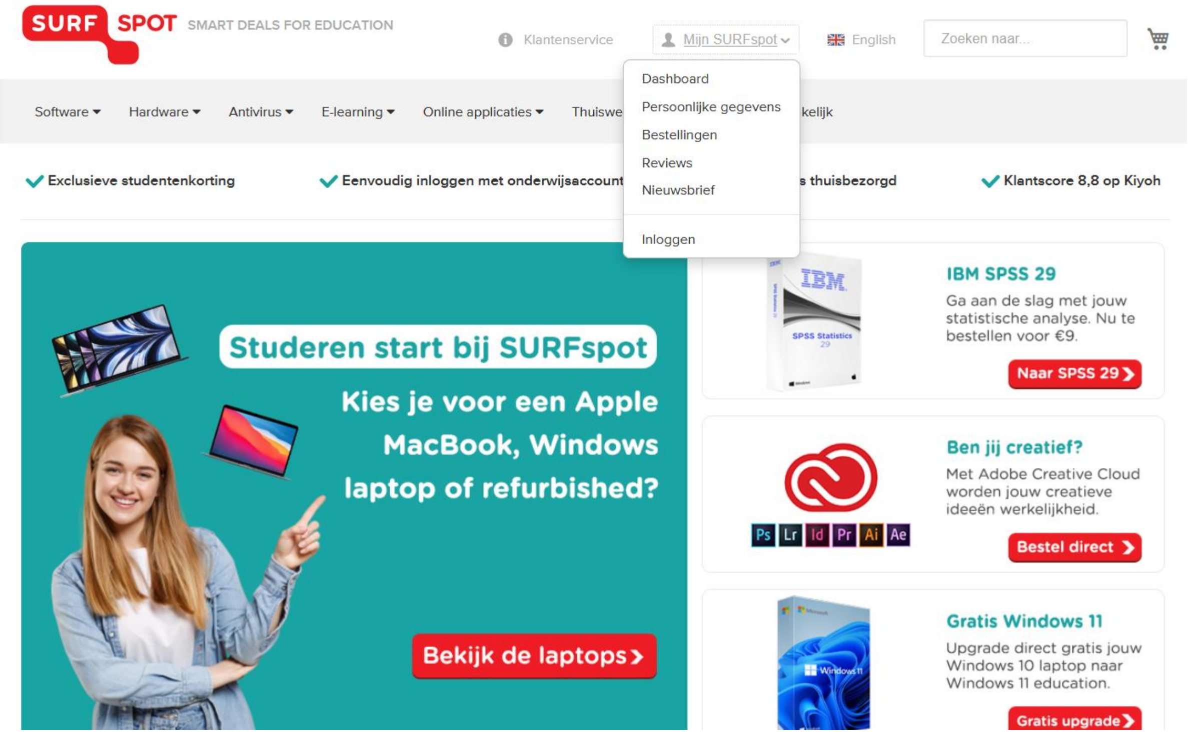


eduGAIN



Images: SURFconext IdP dashboard by SURF, showing some services tagged with REFEDS R&S; eduGAIN map: GEANT, https://technical.edugain.org/status

Your favourite federated service?



✓ Exclusieve studentenkorting ✓ Eenvoudig inloggen met onderwijsaccount s thuisbezorgd ✓ Klantscore 8,8 op Kiyoh

Studeren start bij SURFspot

Kies je voor een Apple MacBook, Windows laptop of refurbished?

IBM SPSS 29

Ga aan de slag met jouw statistische analyse. Nu te bestellen voor €9.

Ben jij creatief?

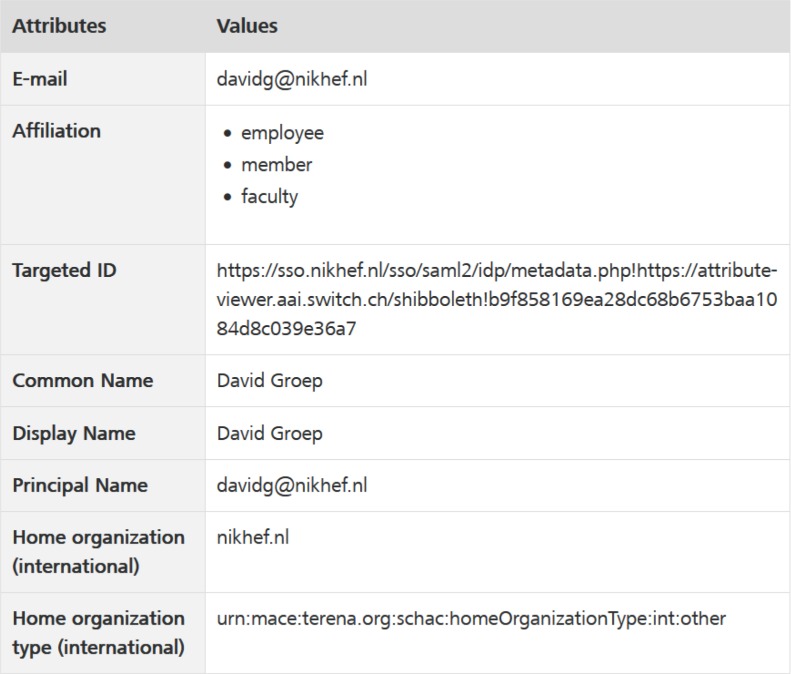
Met Adobe Creative Cloud worden jouw creatieve ideeën werkelijkheid.

Gratis Windows 11

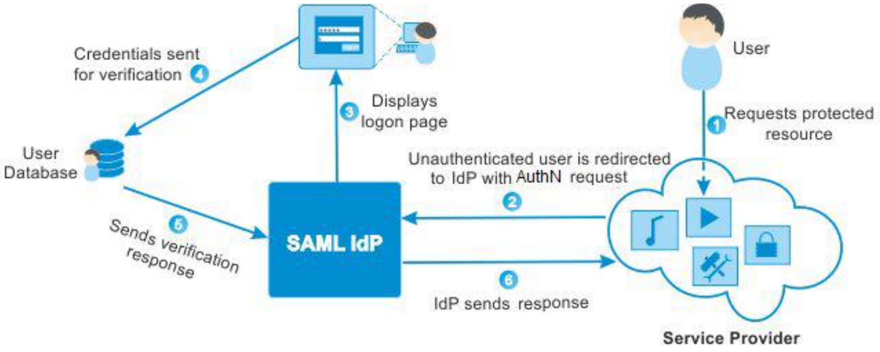
Upgrade direct gratis jouw Windows 10 laptop naar Windows 11 education.

https://surfspot.nl/

SAML federation

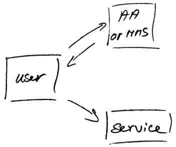


Attributes	Values
E-mail	davidg@nikhef.nl
Affiliation	• employee • member • faculty
Targeted ID	https://sso.nikhef.nl/sso/saml2/idp/metadata.php!https://attribute-viewer.aai.switch.ch/shibboleth!b9f858169ea28dc68b6753baa1084d8c039e36a7
Common Name	David Groep
Display Name	David Groep
Principal Name	davidg@nikhef.nl
Home organization (international)	nikhef.nl
Home organization type (international)	urn:mace:terena.org:schac:homeOrganizationType:int:other



Credentials sent for verification 4
Displays logon page 3
User
Requests protected resource 1
User Database
Unauthenticated user is redirected to IdP with AuthN request
2
Sends verification response 5
SAML IdP
6
IdP sends response
Service Provider

SAML2.0 auth flow



Try at https://attribute-viewer.nikhef.nl/ and select “Login via a global authentication SAML source”

Firefox: use F12, and SAML tracer https://addons.mozilla.org/nl/firefox/addon/saml-tracer/ (by Tim van Dijen of SimpleSAMLphp fame)

SAML WebSSO flow image: SWITCH, CH

Under the hood, sends a (signed) XML document

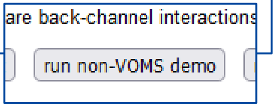
```
<saml:Subject>
     <saml:NameID Format="urn:oasis:names:tc:SAML:2.0:nameid-format:persistent">xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx</saml:NameID>
     <saml:SubjectConfirmation Method="urn:oasis:names:tc:SAML:2.0:cm:bearer">
       <saml:SubjectConfirmationData NotOnOrAfter="2022-10-21T18:16:40Z"
         Recipient="https://attribute-viewer.aai.switch.ch/Shibboleth.sso/SAML2/POST"
         InResponseTo="_64c10a60c382bdaeb328653d9d25951c" /></saml:SubjectConfirmation>
   </saml:Subject>
   <saml:Conditions NotBefore="2022-10-21T18:11:39Z"
                    NotOnOrAfter="2022-10-21T18:16:40Z">
     <saml:AudienceRestriction>
       <saml:Audience>https://attribute-viewer.aai.switch.ch/shibboleth</saml:Audience>
     </saml:AudienceRestriction>
   </saml:Conditions>
   <saml:AuthnStatement AuthnInstant="2022-10-21T17:33:29
                        SessionNotOnOrAfter="2022-10-22T0
                        SessionIndex="_90f745f18f712b6a56
    <saml:AuthnContext>
        <saml:AuthnContextClassRef>urn:oasis:names:tc:SAM
        <saml:AuthenticatingAuthority>https://sso.nikhef.
    </saml:AuthnContext>
   </saml:AuthnStatement>
```

```
<saml:AttributeStatement>
      <saml:Attribute Name="urn:mace:dir:attribute-def:cn"
                      NameFormat="urn:oasis:names:tc:SAML:2.0:attrname-format:uri">
        <saml:AttributeValue xsi:type="xs:string">David Groep</saml:AttributeValue>
      </saml:Attribute>
      <saml:Attribute Name="urn:oid:2.5.4.3"
                      NameFormat="urn:oasis:names:tc:SAML:2.0:attrname-format:uri">
        <saml:AttributeValue xsi:type="xs:string">David Groep</saml:AttributeValue>
      </saml:Attribute>
      <saml:Attribute Name="urn:mace:dir:attribute-def:eduPersonAffiliation"
                      NameFormat="urn:oasis:names:tc:SAML:2.0:attrname-format:uri">
        <saml:AttributeValue xsi:type="xs:string">employee</saml:AttributeValue>
        <saml:AttributeValue xsi:type="xs:string">member</saml:AttributeValue>
        <saml:AttributeValue xsi:type="xs:string">faculty</saml:AttributeValue>
      </saml:Attribute>
      <saml:Attribute Name="urn:oid:1.3.6.1.4.1.5923.1.1.1.1"
      ...
```

Different tech, similar concept: X.509 RFC5280 client certificates

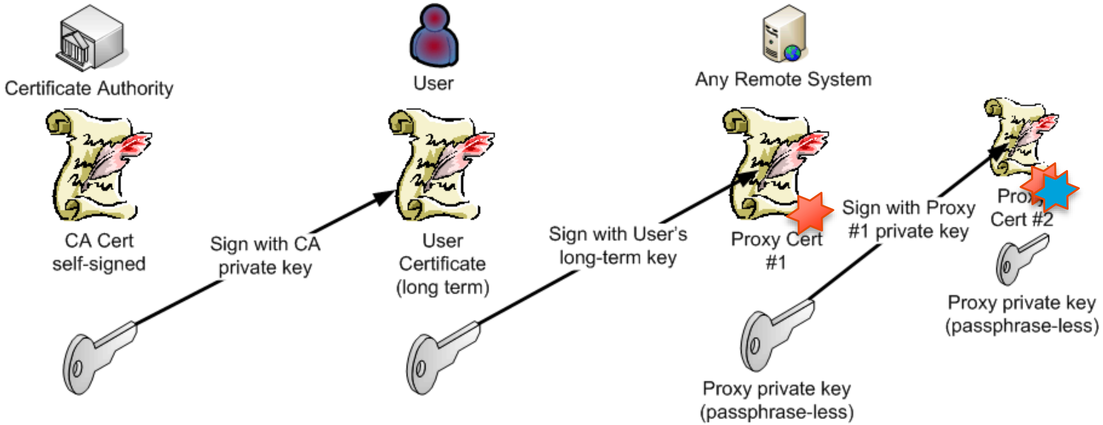
```
Version: 3 (0x2)
Serial Number:
  34:f3:e3:5f:c0:53:0b:a6:ef:2b:4a:79:01:b5:50:3b
Signature Algorithm: sha384WithRSAEncryption
Issuer: C = NL, O = GEANT Vereniging, CN = GEANT eScience Personal CA 4
Validity
  Not Before: Apr  2 00:00:00 2022 GMT
  Not After : May  2 23:59:59 2023 GMT
Subject: DC = org, DC = terena, DC = tcs, C = NL, O = Nikhef, CN = David Groep davidg@nikhef.nl
Subject Public Key Info:
  Public Key Algorithm: rsaEncryption
    RSA Public-Key: (4096 bit)
    Modulus:
      00:f0:0d:c0:ff:ee:f0:0d:f0:0d:c0:ff:ee:f0:0d:
      ...
      ff:50:6d
    Exponent: 65537 (0x10001)
X509v3 extensions:
  X509v3 Key Usage: critical
    Digital Signature, Key Encipherment
  X509v3 Basic Constraints: critical
    CA:FALSE
  X509v3 Extended Key Usage:
    E-mail Protection, TLS Web Client Authentication
  X509v3 Certificate Policies:
    Policy: 1.2.840.113612.5.2.2.5
```

You should be able to get an ‘IGTF-DOGWOOD’ assurance certificate from RCauth.eu.
Go to https://rcdemo.nikhef.nl/ and select the ‘Basic demo’ and use ‘run non-VOMS’ to get and view your short-lived certificate


are back-channel interactions

Certificates chains & constraint proxy identity delegation

- PKIX certificates are ASN.1 structures in a distinguished binary encoding (DER format)
- contains the tuple (issuer, subject, serial) + validity period + key material + **extensions**
- within it is the message digest (hash), signed with private key of the issuer
- Verifiable using the issuer’s public key


Certificate Authority
User
Any Remote System
CA Cert self-signed
Sign with CA private key
User Certificate (long term)
Sign with User's long-term key
Proxy Cert #1
Sign with Proxy #1 private key
Proxy Cert #2
Proxy private key (passphrase-less)
Proxy private key (passphrase-less)

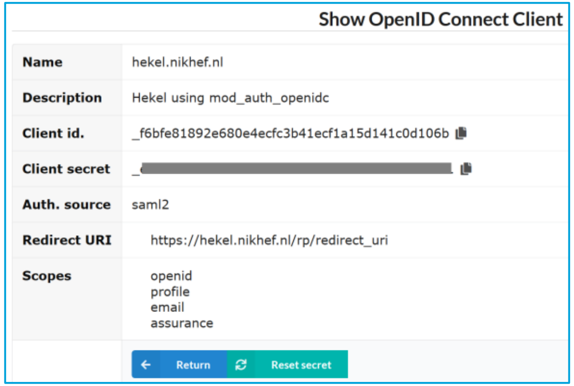
RFC3820 ‘proxy’ certificates extend this concept to (policy-constraint) identity delegation

To get an RFC3820 proxy certificate using your own federated identity, use RCauth.eu – see https://rcdemo.nikhef.nl/ and use the “Basic Demo” option

OpenID Connect and OAuth2

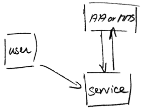
- Quite .well-known
  (used by lots modern ‘non-enterprise’ SSO)

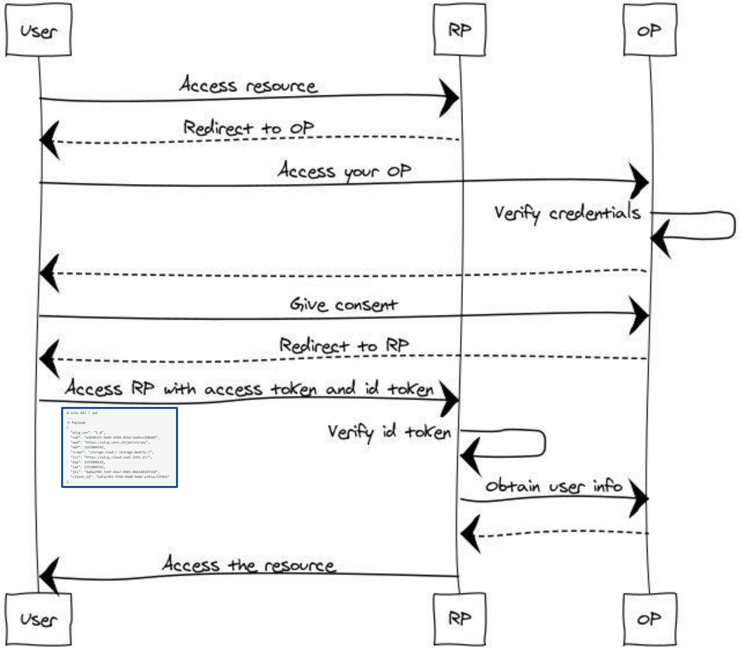
- shows in its initial design: one source of identity (Openid Provider, ‘OP’), and many services (Relaying Parties, ‘RP’)



Show OpenID Connect Client

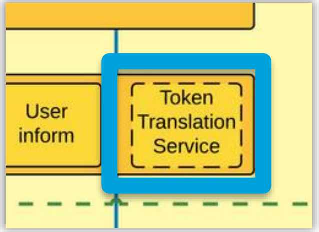
Name	hekel.nikhef.nl
Description	Hekel using mod_auth_openidc
Client id.	_f6bfe81892e680e4ecfc3b41ecf1a15d141c0d106b
Client secret	_[redacted]
Auth. source	saml2
Redirect URI	https://hekel.nikhef.nl/rp/redirect_uri
Scopes	openid profile email assurance





Shown is the ‘implicit flow’, other flows possible. Image source: AARC NA2 training on AAI 101
See https://openid.net/ for protocols and standardization work

Federation: different technologies, same idea



SAML - Security Assertion Markup Language and WebSSO ('SAML2Int')

- XML-formatted 'attribute statements' over web transport (usually POST)
- SAML-Metadata: list of entities with description of bindings with entityAttributes

PKI - Public Key Infrastructures

- trusted third party (a *certification authority* a.k.a. *CA*)
  signs X.509 formatted certificates with name, issuer, serial number, and extensions
- CAs can sign end-entities as well as other CAs (hierarchically or by cross-signing)
- *bridge CAs* render a technical implementation of a shared policy (assurance)
- *policy-bridges* don't sign anything, but curate *distribution*
  (like browsers and operating systems based on CA/BF requirements, IGTF for research infras)

OpenID Federation – Federating OpenID Connect parties

- federate end-points for OIDC Providers and Relying Parties (or OAuth2), with similar models

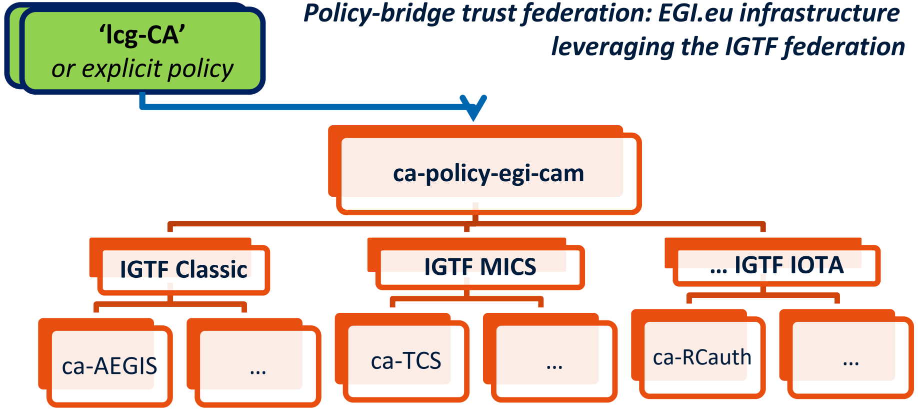
note federation based on 'ultimate trust' domains (e.g. cross-realm Kerberos) also exists …

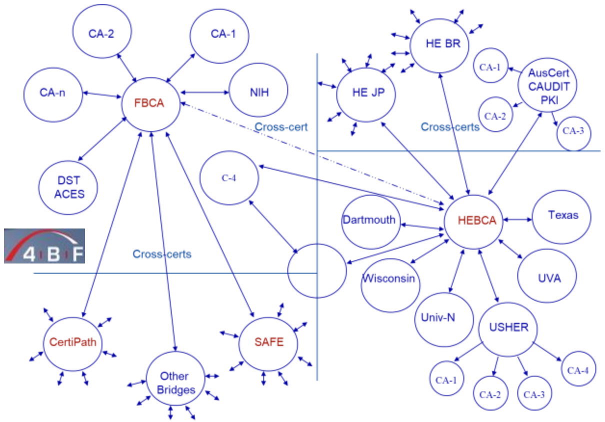
See www.oasis.org for SAML; RFC5280 (tech) & RFC3247 (policy) for PKIX, https://igtf.net/ and https://cabforum.org;
OpenID Connect Federation: https://openid.net/specs/openid-connect-federation-1_0.html

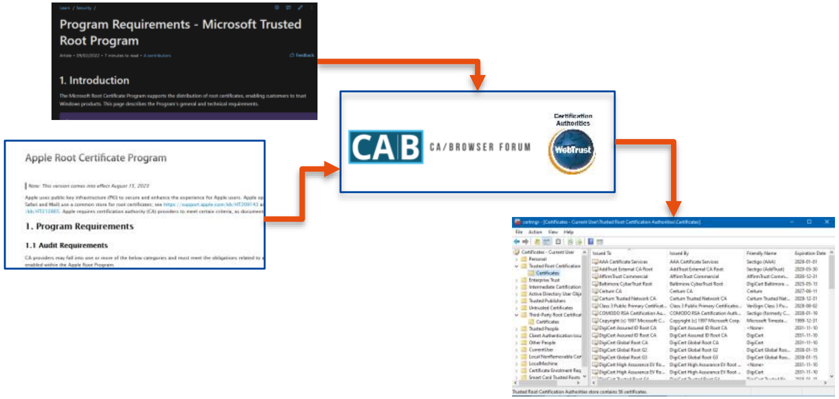
Federation: technological …
… or policy bridge

trust remains with the relying party
can be *bridged* by either cross-signing (left)
or by policy agreements (right)

Policy-bridge trust federation: EGI.eu infrastructure leveraging the IGTF federation







Program Requirements - Microsoft Trusted Root Program

Article • 09/02/2022 • 7 minutes to read • 4 contributors

1. Introduction

The Microsoft Root Certificate Program supports the distribution of root certificates, enabling customers to trust Windows products. This page describes the Program's general and technical requirements.

Apple Root Certificate Program

Note: This version comes into effect August 15, 2023

Apple uses public key infrastructure (PKI) to secure and enhance the experience for Apple users. Apple op… Safari and Mail) use a common store for root certificates; see https://support.apple.com/kb/HT209143 a… /kb/HT212865. Apple requires certification authority (CA) providers to meet certain criteria, as document…

1. Program Requirements

1.1 Audit Requirements

CA providers may fall into one or more of the below categories and must meet the obligations related to … enabled within the Apple Root Program.

certmgr - [Certificates - Current User\Trusted Root Certification Authorities\Certificates]

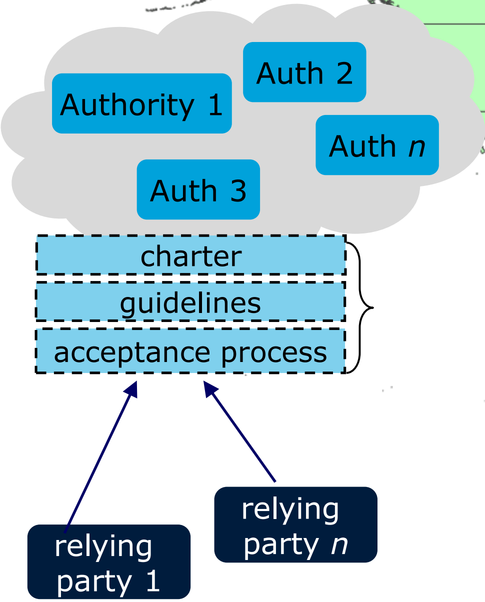
Certificates - Current User
- Personal
- Trusted Root Certification
  - Certificates
- Enterprise Trust
- Intermediate Certification
- Active Directory User Obje
- Trusted Publishers
- Untrusted Certificates
- Third-Party Root Certificat
  - Certificates
- Trusted People
- Client Authentication Issu
- Other People
- CurrentUser
- Local NonRemovable Cer
- LocalMachine
- Certificate Enrolment Req
- Smart Card Trusted Roots

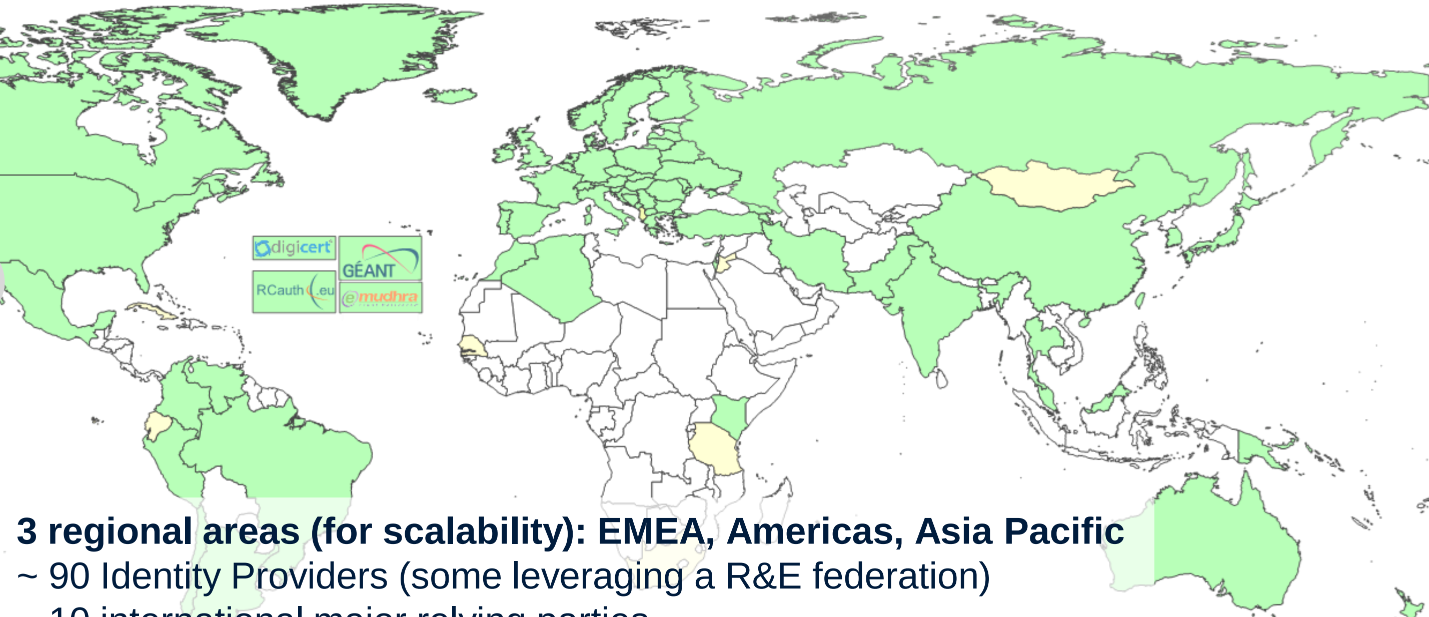
Issued To	Issued By	Friendly Name	Expiration Date
AAA Certificate Services	AAA Certificate Services	Sectigo (AAA)	2028-01-01
AddTrust External CA Root	AddTrust External CA Root	Sectigo (AddTrust)	2020-05-30
AffirmTrust Commercial	AffirmTrust Commercial	AffirmTrust Comm...	2030-12-31
Baltimore CyberTrust Root	Baltimore CyberTrust Root	DigiCert Baltimore ...	2025-05-13
Certum CA	Certum CA	Certum	2027-06-11
Certum Trusted Network CA	Certum Trusted Network CA	Certum Trusted Net...	2029-12-31
Class 3 Public Primary Certificat...	Class 3 Public Primary Certificatio...	VeriSign Class 3 Pu...	2028-08-02
COMODO RSA Certification Au...	COMODO RSA Certification Auth...	Sectigo (formerly C...	2038-01-19
Copyright (c) 1997 Microsoft C...	Copyright (c) 1997 Microsoft Corp.	Microsoft Timesta...	1999-12-31
DigiCert Assured ID Root CA	DigiCert Assured ID Root CA	<None>	2031-11-10
DigiCert Assured ID Root CA	DigiCert Assured ID Root CA	DigiCert	2031-11-10
DigiCert Global Root CA	DigiCert Global Root CA	DigiCert	2031-11-10
DigiCert Global Root G2	DigiCert Global Root G2	DigiCert Global Roo...	2038-01-15
DigiCert Global Root G3	DigiCert Global Root G3	DigiCert Global Roo...	2038-01-15
DigiCert High Assurance EV Ro...	DigiCert High Assurance EV Root ...	<None>	2031-11-10
DigiCert High Assurance EV Ro...	DigiCert High Assurance EV Root ...	DigiCert	2031-11-10

Trusted Root Certification Authorities store contains 56 certificates.

Left-hand image: 4 Bridges Forum, source: Scott Rea (then: Dartmouth University)
Images: cabforum.org, WebTrust logo: from DigiCert.com; image MS root store, https://learn.microsoft.com/en-us/security/trusted-root/program-requirements

Policy-bridged global federations for research computing





3 regional areas (for scalability): EMEA, Americas, Asia Pacific

~ 90 Identity Providers (some leveraging a R&E federation)

~ 10 international major relying parties

~ 60 countries / economic areas / international treaty orgs

> 1000 relying service provider collaborations

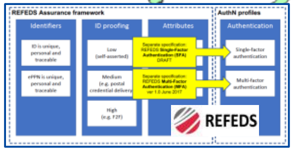
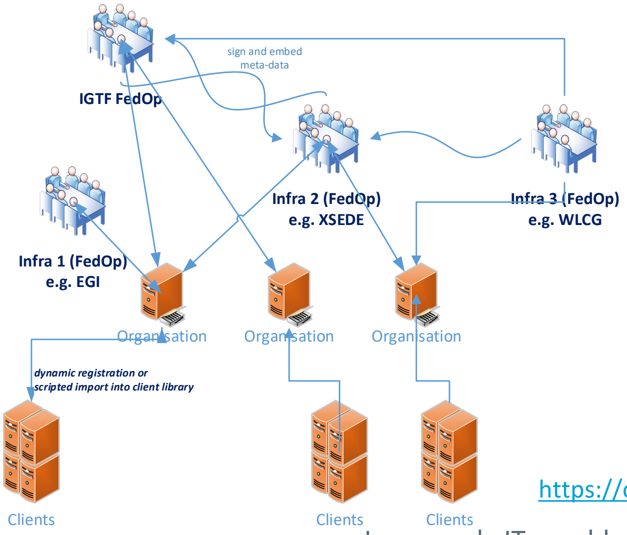

REFEDS Assurance framework
AuthN profiles
Identifiers
ID proofing
Attributes
Authentication
Low
Medium
High
Single-factor authentication
Multi-factor authentication
REFEDS

Image: Interoperable Global Trust Federation IGTF, https://igtf.net/; REFEDS Assurance Framework RAF: http://refeds.org/assurance, https://refeds.org/profile/mfa

OpenID Federation

OIDC endpoints + trust policy data for registration can be federated in a meta-data feed

- makes OIDC ‘federatable’ (plain oidc is single OP)
- as for PKIX, can be technical or policy bridge
- delegated metadata makes ‘OIDC-fed’ scale in webscale scenarios



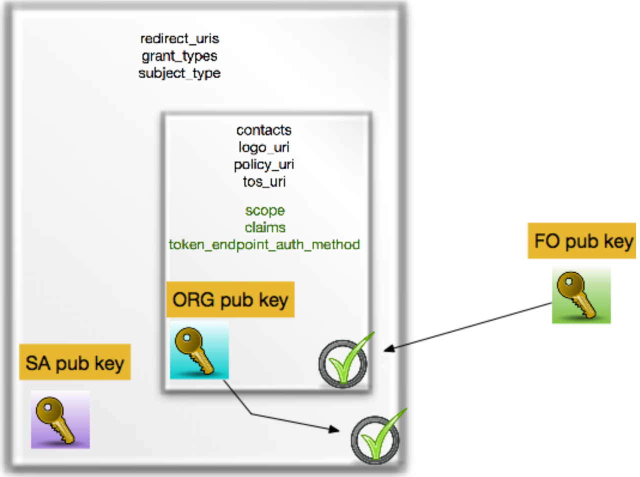


Image: Roland Hedberg, University of Umeå
OpenID Connect Fedrration:
https://openid.net/specs/openid-connect-federation-1_0.html

Federation: technology, interoperability, policy

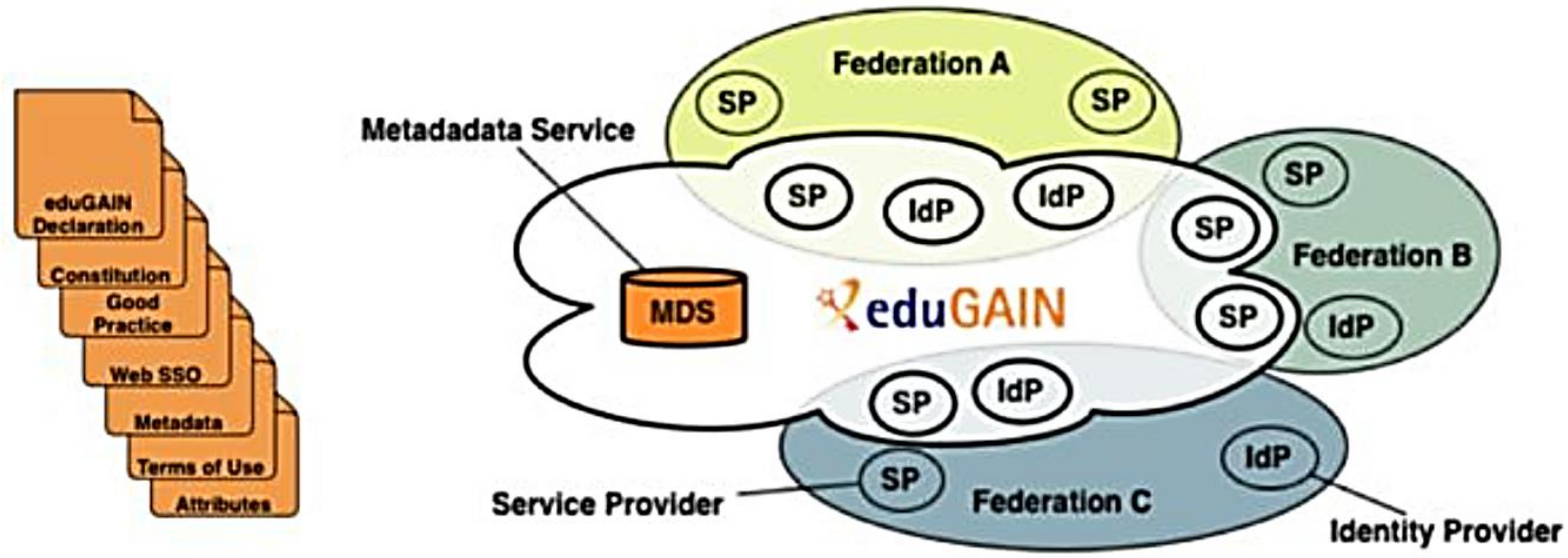
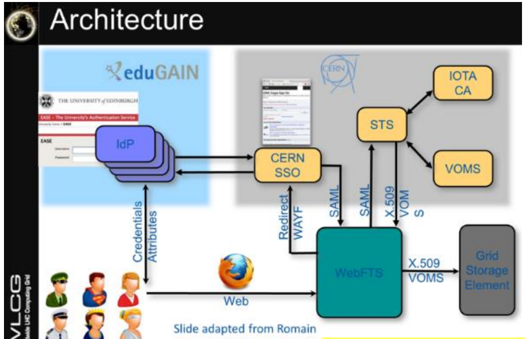


Image from SWITCH (CH) and edugain.org

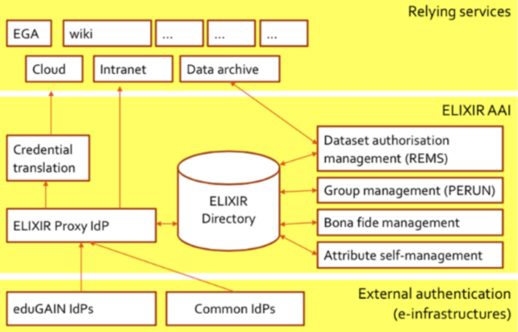
Managing complexities of federation & identity

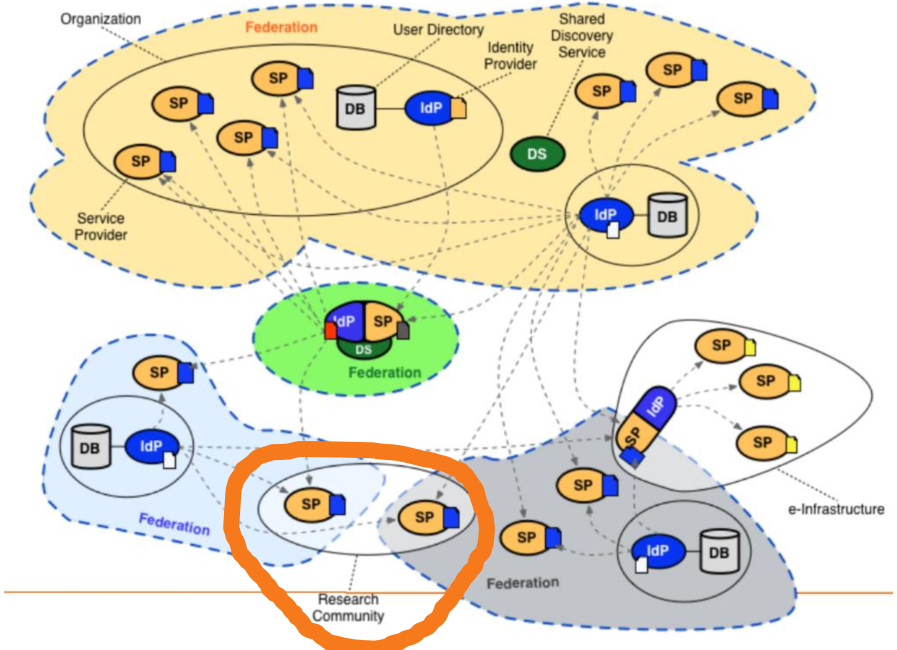


Slide adapted from Romain Wartel, GDB Sept 2014

WebFTS prototype
'FIM4R' in wLCG
Romain Wartel et al.

ELIXIR reference architecture 2016
Mikael Linden et al.





communities had either invented
their own 'proxy' model to abstract complexity

or they were composed of many services
each of which had to manage federation complexity

Community images: Romain Wartel, CERN; Mikael Linden, CSC; Lukas Hammerle, SWITCH

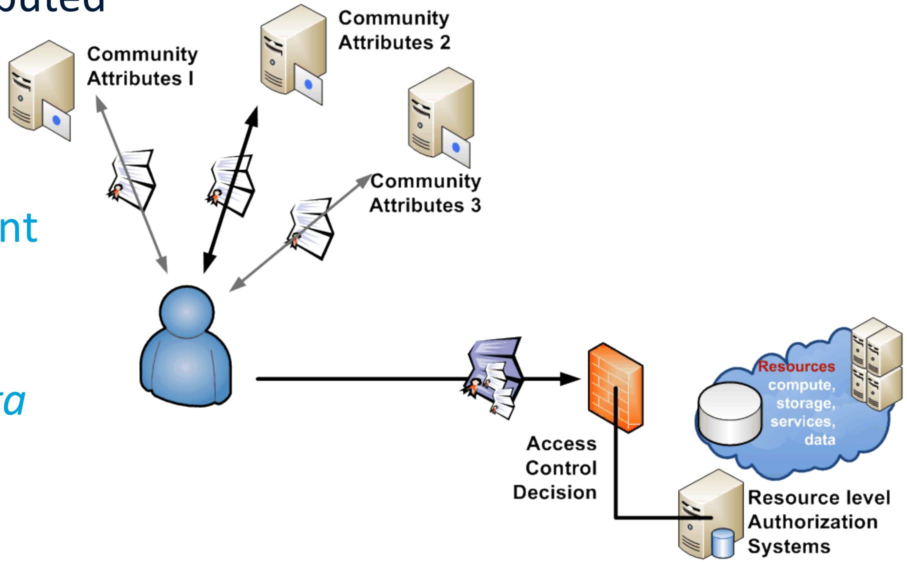
Multiple sources of authority: the community

- authorization assertion providers (attribute authorities) use the identifier(s) from authentication in their membership services

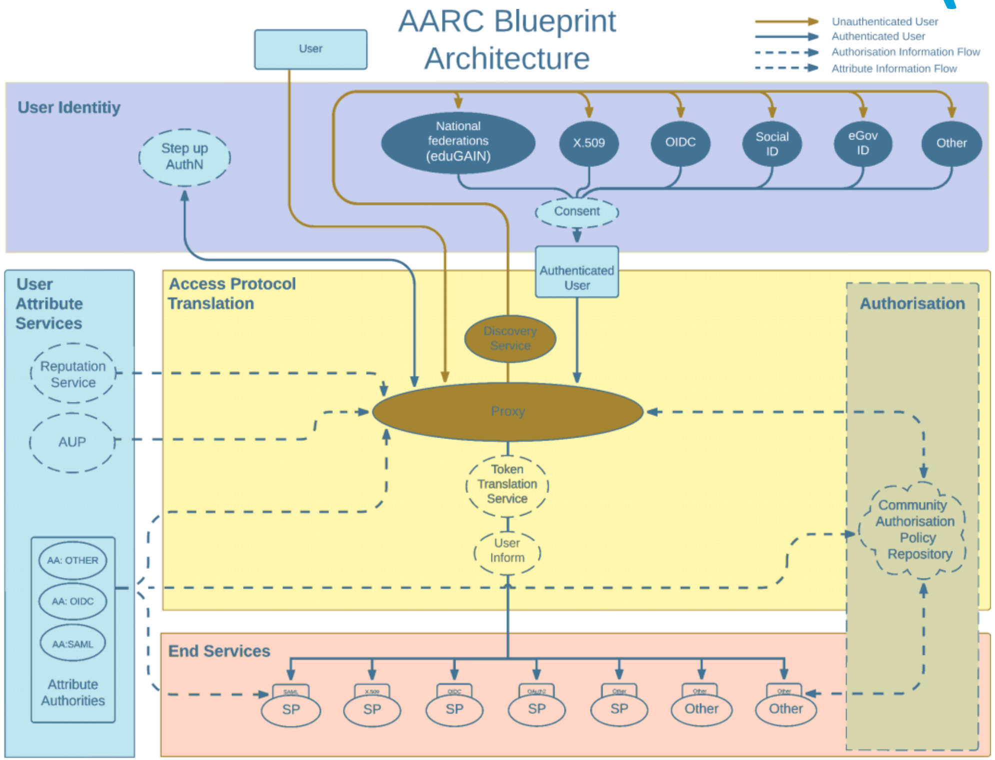
- *source of authority* for attributes is distributed

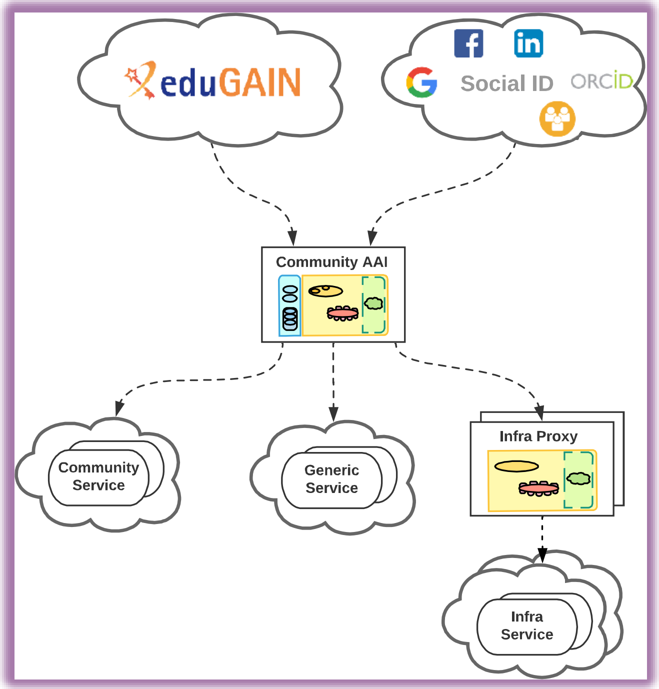
for example:
- community membership from an experiment
- affiliation status from home organisation

may be jointly needed to access sensitive data that is subject to medical-ethical clearance



Most trust flows from the (research) community



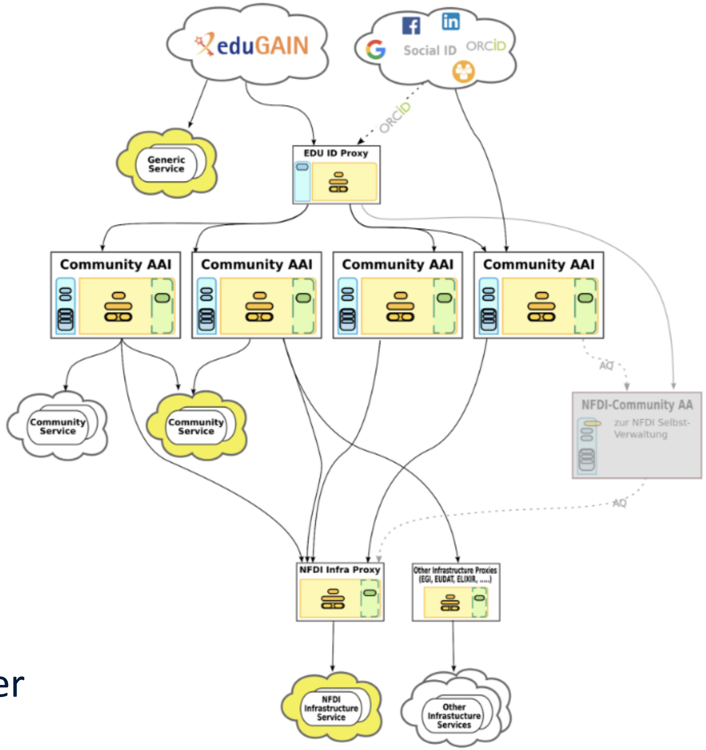


AARC Blueprint Architecture (2019) AARC-G045 https://aarc-community.org/guidelines/aarc-g045/; stacked proxies: EOSC AAI Architecture
EOSC Authentication and Authorization Infrastructure (AAI), ISBN 978-92-76-28113-9, http://doi.org/10.2777/8702

Composite AAIs: proxies beyond just the research infrastructures

Proxy model harmonizes IdPs from many sources

- **eduID**-style identifiers
  - ‘life-long learning’ identifiers
  - independent student identifier (the ESI) for mobility & Erasmus-without-papers
  - eduGAIN-alignment, but also a ‘provider of last resort’
- **eIDAS** and government eID (e.g. DigID)
  - identity assurance step-up
- **ORCID** provides identifier portability through linking
  - provides name linking and persistent attribution
  - since it persists, also very useful to allow access *independent of home organisation* throughout a carreer



Composite AAI image source: Christos Kanellopoulos (GEANT), Marcus Hardt (KIT)

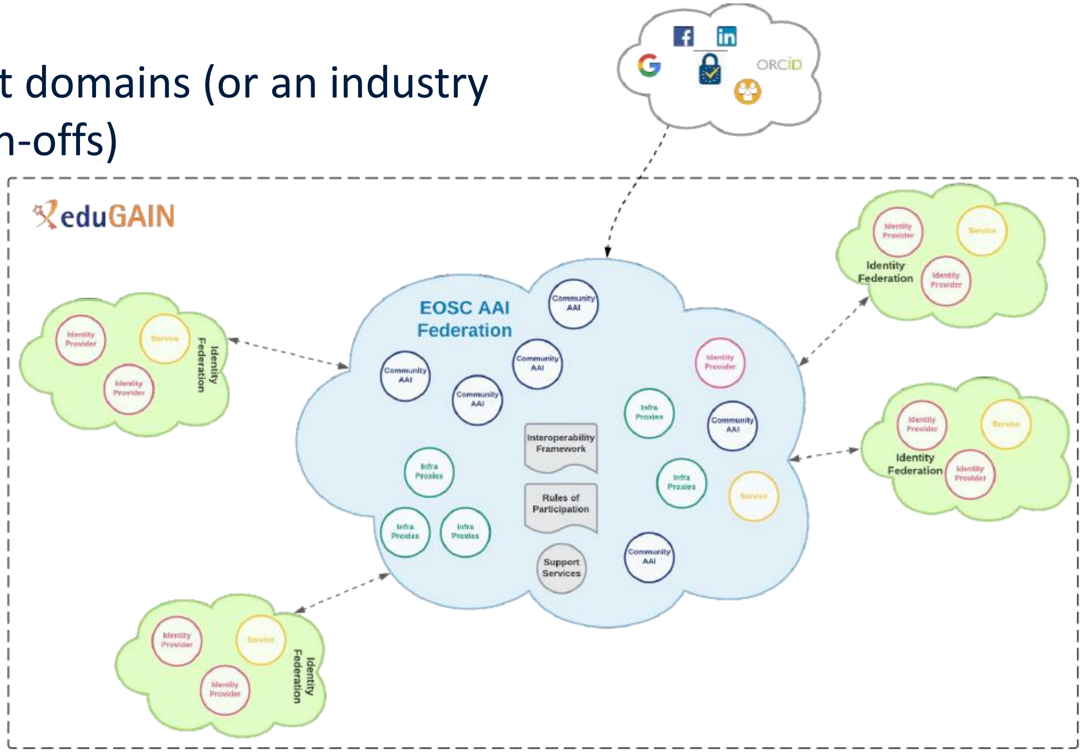
When many proxies from different groups come together

When collaborations cross different domains (or an industry sector with lots of mergers and spin-offs)

- proxies with each group
- inter-federate SP/IdP interfaces
- each federation can add own policy and entity filtering

Example

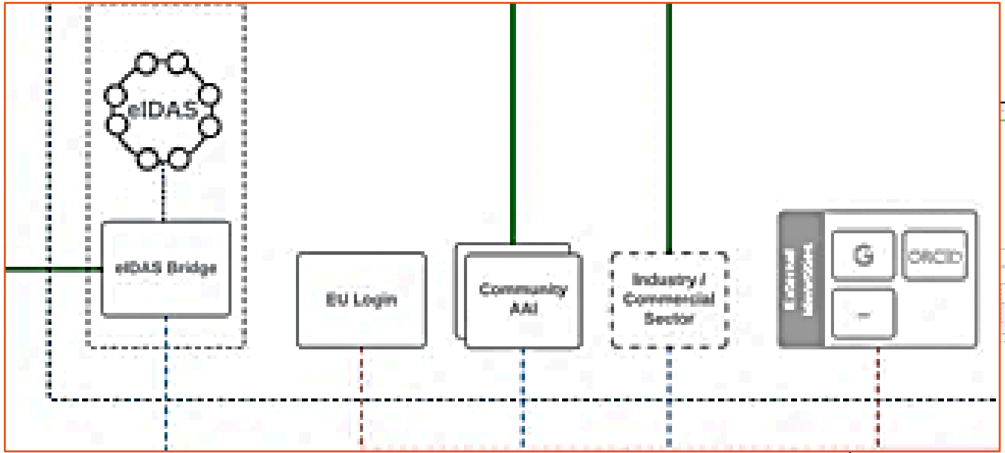
European Open Science Cloud (EOSC) AAI based on federations and proxies

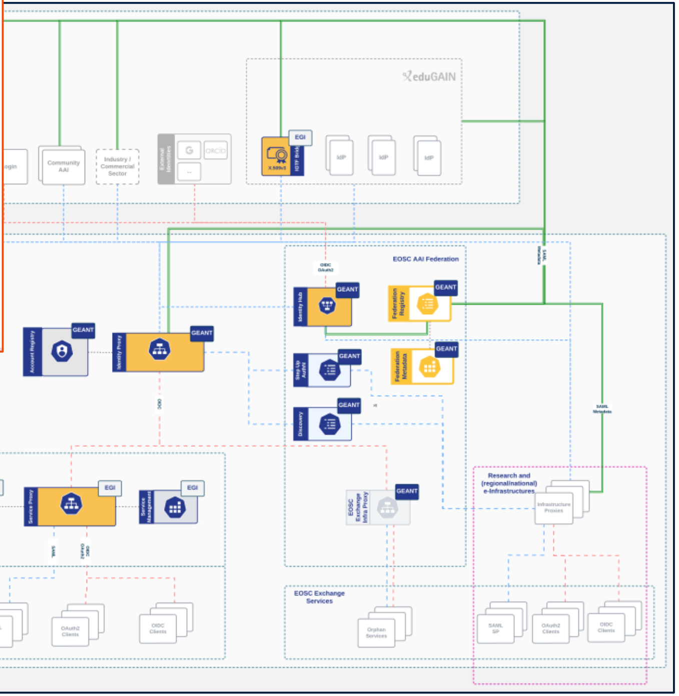


Christos Kanellopoulos (GEANT) for the EOSC AAI Federation in “The EOSC Core”, https://eoscfuture.eu/wp-content/uploads/2022/04/EOSC-Core.pdf

European Open Science Cloud (EOSC) AAI Federation

***Identity assurance** brings the true value: authenticators are aplenty, and 'MFA' far less interesting than vetted identities. But HEI home IdPs seem reluctant to provide it …*



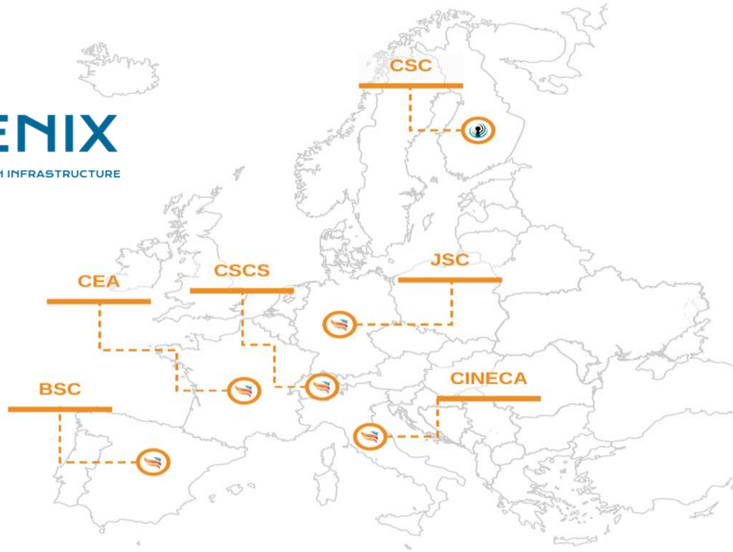


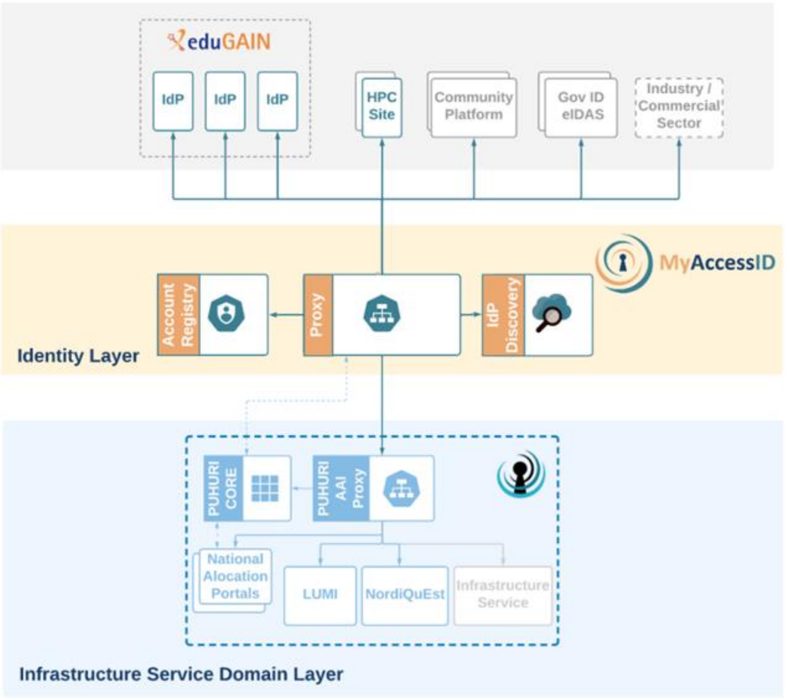
user identity comes 'with the user' from outside, mediated by the research community, ORCID, or from the home member state involved

Image: EOSC AAI for the EOSC Core and Exchange Federation for the EOSC European Node by Christos Kanellopoulos, Nicolas Liampotis, David Groep (June 2023)

Same blocks underlie e.g. the Fenix and Puhuri HPC ecosystem

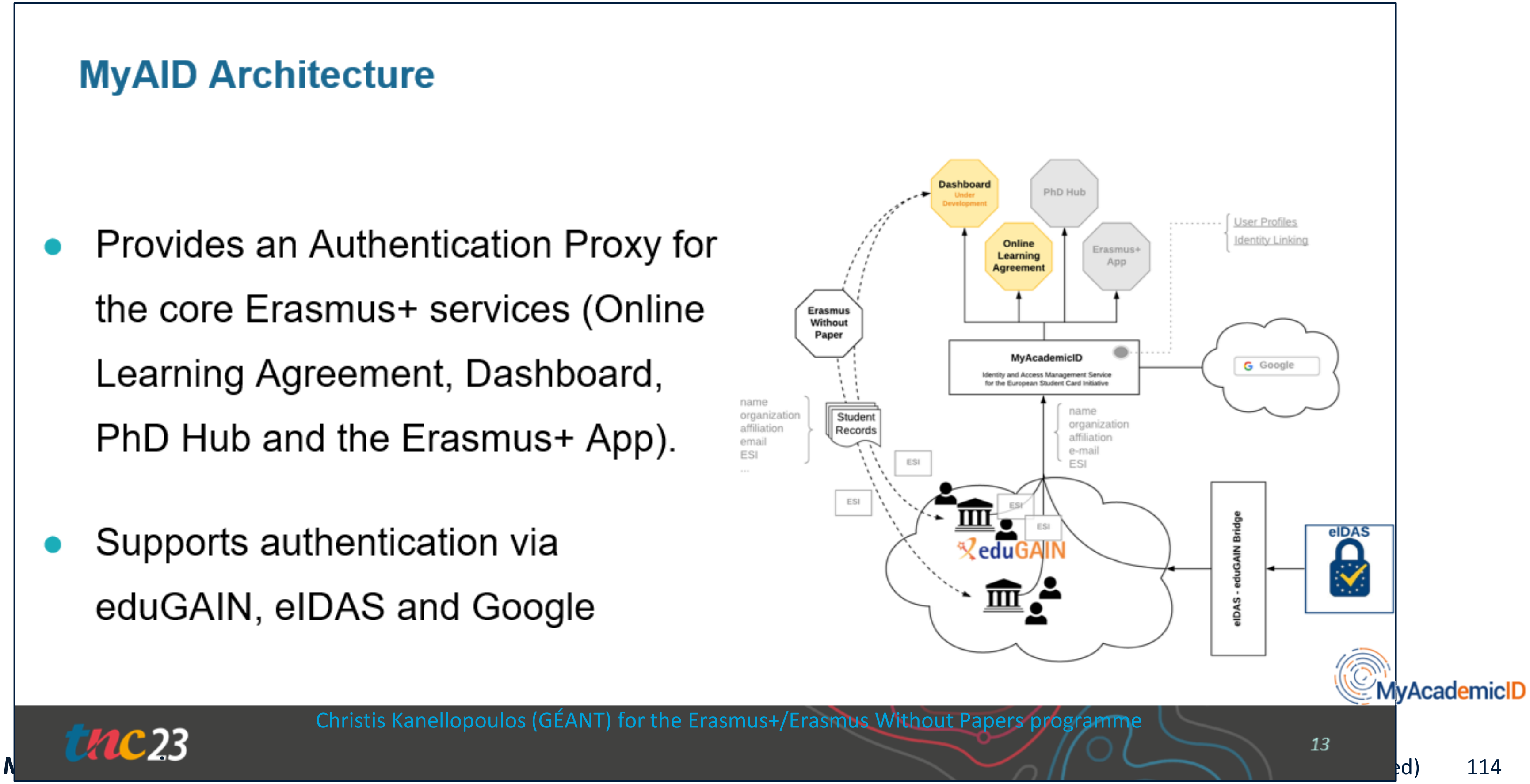
FENIX
RESEARCH INFRASTRUCTURE





Fenix image via Christos Kanellopoulos, diagram via Anders Sjöström (NeIC, Puhuri) at the TNC23 workshop

Also the basic blocks for your student identity& Erasmus+



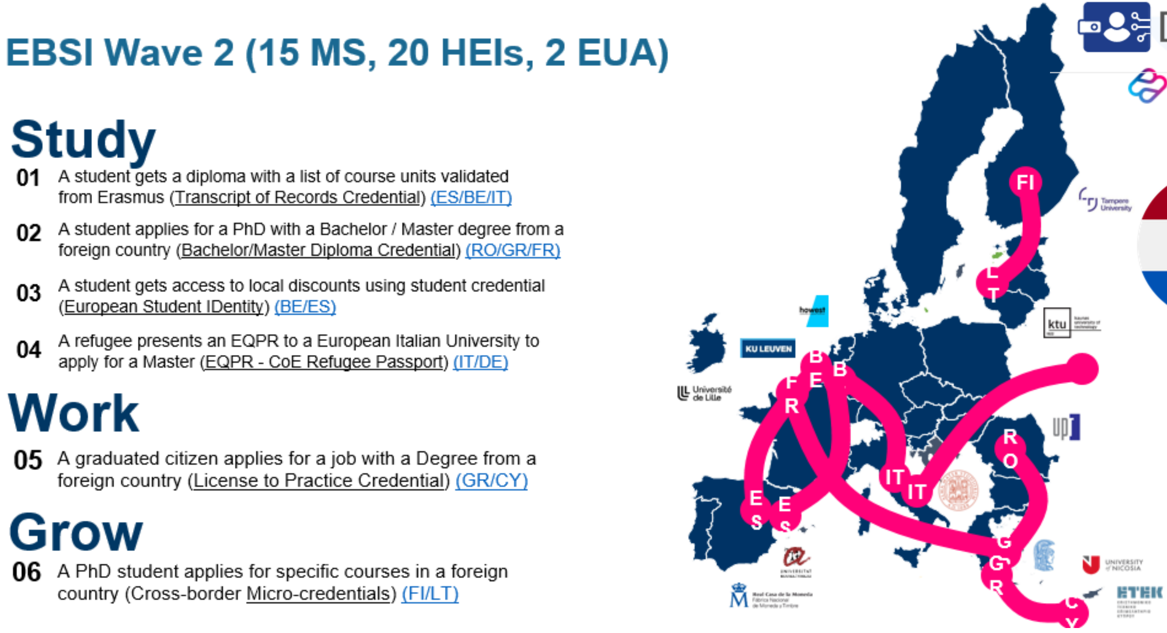
MyAID Architecture

- Provides an Authentication Proxy for the core Erasmus+ services (Online Learning Agreement, Dashboard, PhD Hub and the Erasmus+ App).
- Supports authentication via eduGAIN, eIDAS and Google

Christis Kanellopoulos (GÉANT) for the Erasmus+/Erasmus Without Papers programme

13

What value does our university ID bring in a life-long learning environment? Time to think less institution-centric?



EBSI Wave 2 (15 MS, 20 HEIs, 2 EUA)

Study

01 A student gets a diploma with a list of course units validated from Erasmus (Transcript of Records Credential) (ES/BE/IT)

02 A student applies for a PhD with a Bachelor / Master degree from a foreign country (Bachelor/Master Diploma Credential) (RO/GR/FR)

03 A student gets access to local discounts using student credential (European Student IDentity) (BE/ES)

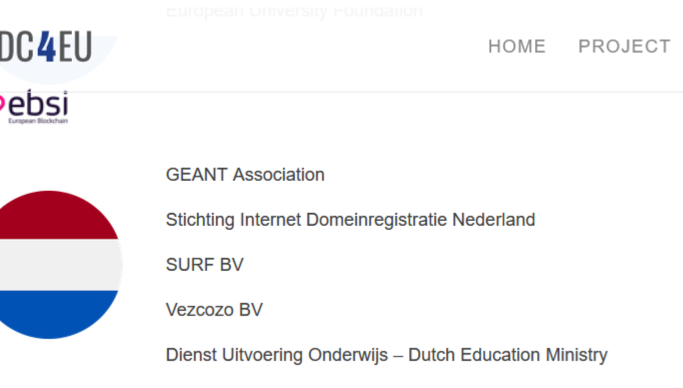
04 A refugee presents an EQPR to a European Italian University to apply for a Master (EQPR - CoE Refugee Passport) (IT/DE)

Work

05 A graduated citizen applies for a job with a Degree from a foreign country (License to Practice Credential) (GR/CY)

Grow

06 A PhD student applies for specific courses in a foreign country (Cross-border Micro-credentials) (FI/LT)



European University Foundation

HOME PROJECT

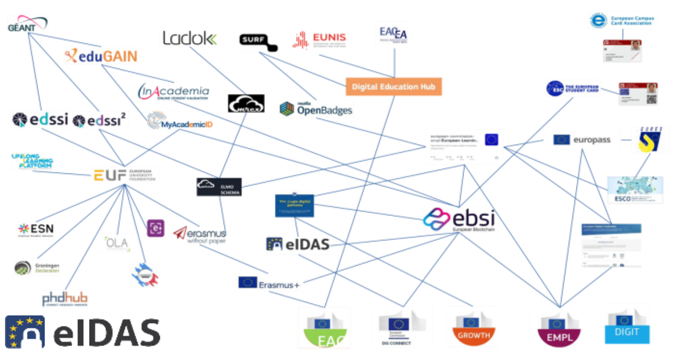
GEANT Association

Stichting Internet Domeinregistratie Nederland

SURF BV

Vezcozo BV

Dienst Uitvoering Onderwijs – Dutch Education Ministry

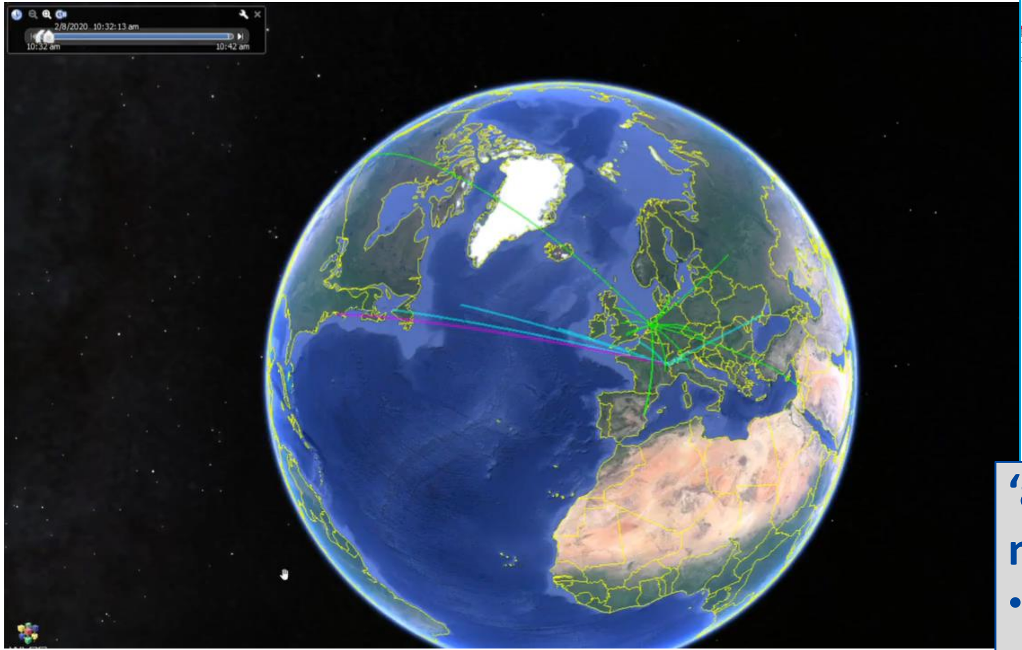


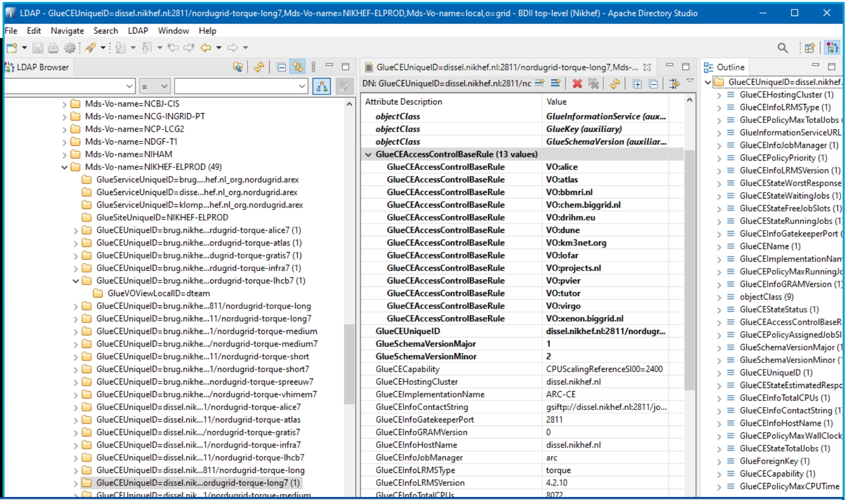
Images from Lluís Ariño, for the DC4EU project. See e.g. https://www.dc4eu.eu/consortium/#netherlands

Putting it back together again

Common patterns in scalability

A global infrastructure of EGI, OSG and WLCG, …




LDAP - GlueCEUniqueID=dissel.nikhef.nl:2811/nordugrid-torque-long7,Mds-Vo-name=NIKHEF-ELPROD,Mds-Vo-name=local,o=grid - BDII top-level (Nikhef) - Apache Directory Studio
File
Edit
Navigate
Search
LDAP
Window
Help
LDAP Browser
GlueCEUniqueID=dissel.nikhef.nl:2811/nordugrid-torque-long7,Mds-...
DN: GlueCEUniqueID=dissel.nikhef.nl:2811/nc
Outline

```
> Mds-Vo-name=NCBJ-CIS
> Mds-Vo-name=NCG-INGRID-PT
> Mds-Vo-name=NCP-LCG2
> Mds-Vo-name=NDGF-T1
> Mds-Vo-name=NIHAM
v Mds-Vo-name=NIKHEF-ELPROD (49)
    GlueServiceUniqueID=brug....hef.nl_org.nordugrid.arex
    GlueServiceUniqueID=disse...hef.nl_org.nordugrid.arex
    GlueServiceUniqueID=klomp...hef.nl_org.nordugrid.arex
    GlueSiteUniqueID=NIKHEF-ELPROD
  > GlueCEUniqueID=brug.nikhe...rdugrid-torque-alice7 (1)
  > GlueCEUniqueID=brug.nikhe...ordugrid-torque-atlas (1)
  > GlueCEUniqueID=brug.nikhe...dugrid-torque-gratis7 (1)
  > GlueCEUniqueID=brug.nikhe...rdugrid-torque-infra7 (1)
  v GlueCEUniqueID=brug.nikhe...ordugrid-torque-lhcb7 (1)
      GlueVOViewLocalID=dteam
  > GlueCEUniqueID=brug.nikhe...811/nordugrid-torque-long
  > GlueCEUniqueID=brug.nikhe...11/nordugrid-torque-long7
  > GlueCEUniqueID=brug.nikhe...1/nordugrid-torque-medium
  > GlueCEUniqueID=brug.nikhe.../nordugrid-torque-medium7
  > GlueCEUniqueID=brug.nikhe...11/nordugrid-torque-short
  > GlueCEUniqueID=brug.nikhe...1/nordugrid-torque-short7
  > GlueCEUniqueID=brug.nikhe...nordugrid-torque-spreeuw7
  > GlueCEUniqueID=brug.nikhe.../nordugrid-torque-vhimem7
  > GlueCEUniqueID=dissel.nik...1/nordugrid-torque-alice7
  > GlueCEUniqueID=dissel.nik...11/nordugrid-torque-atlas
  > GlueCEUniqueID=dissel.nik.../nordugrid-torque-gratis7
  > GlueCEUniqueID=dissel.nik...1/nordugrid-torque-infra7
  > GlueCEUniqueID=dissel.nik...11/nordugrid-torque-lhcb7
  > GlueCEUniqueID=dissel.nik...811/nordugrid-torque-long
  > GlueCEUniqueID=dissel.nik...ordugrid-torque-long7 (1)
  > GlueCEUniqueID=dissel.nik...1/nordugrid-torque-medium
```

Attribute Description	Value
objectClass	*GlueInformationService (aux...*
objectClass	*GlueKey (auxiliary)*
objectClass	*GlueSchemaVersion (auxiliar...*
GlueCEAccessControlBaseRule (13 values)	
GlueCEAccessControlBaseRule	**VO:alice**
GlueCEAccessControlBaseRule	**VO:atlas**
GlueCEAccessControlBaseRule	**VO:bbmri.nl**
GlueCEAccessControlBaseRule	**VO:chem.biggrid.nl**
GlueCEAccessControlBaseRule	**VO:drihm.eu**
GlueCEAccessControlBaseRule	**VO:dune**
GlueCEAccessControlBaseRule	**VO:km3net.org**
GlueCEAccessControlBaseRule	**VO:lofar**
GlueCEAccessControlBaseRule	**VO:projects.nl**
GlueCEAccessControlBaseRule	**VO:pvier**
GlueCEAccessControlBaseRule	**VO:tutor**
GlueCEAccessControlBaseRule	**VO:virgo**
GlueCEAccessControlBaseRule	**VO:xenon.biggrid.nl**
GlueCEUniqueID	**dissel.nikhef.nl:2811/nordugr...**
GlueSchemaVersionMajor	**1**
GlueSchemaVersionMinor	**2**
GlueCECapability	CPUScalingReferenceSI00=2400
GlueCEHostingCluster	dissel.nikhef.nl
GlueCEImplementationName	ARC-CE
GlueCEInfoContactString	gsiftp://dissel.nikhef.nl:2811/jo...
GlueCEInfoGatekeeperPort	2811
GlueCEInfoGRAMVersion	0
GlueCEInfoHostName	dissel.nikhef.nl
GlueCEInfoJobManager	arc
GlueCEInfoLRMSType	torque
GlueCEInfoLRMSVersion	4.2.10
GlueCEInfoTotalCPUs	8072

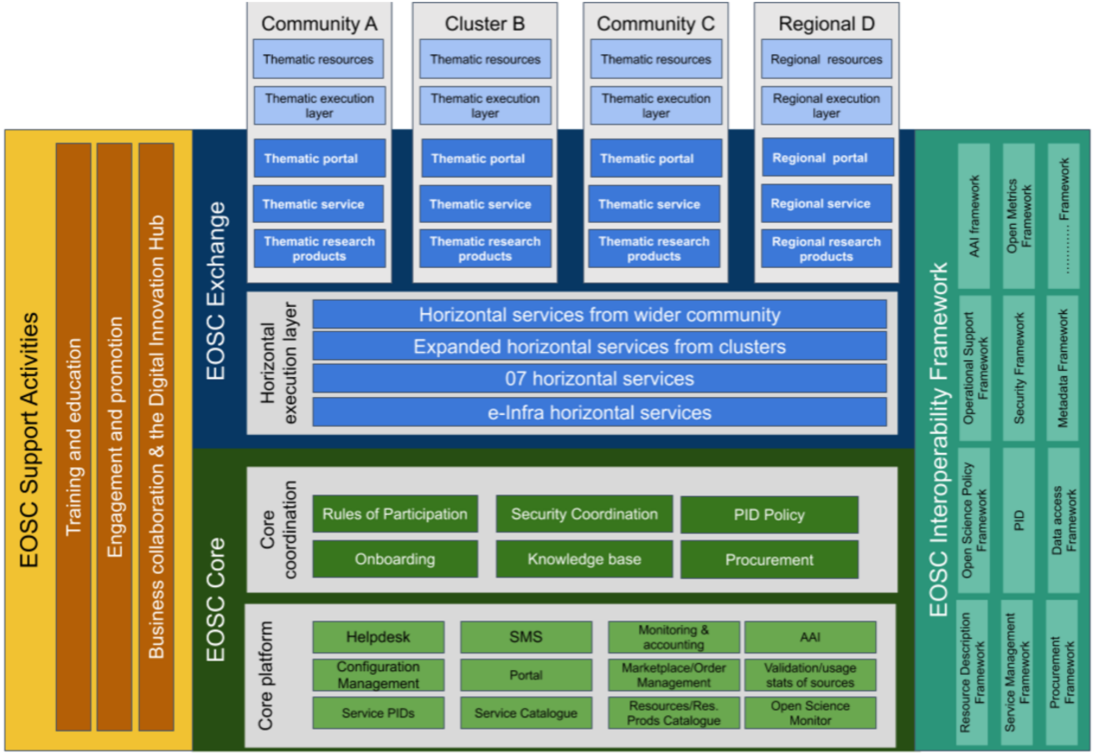
```
v GlueCEUniqueID=dissel.nikhef.
  > GlueCEHostingCluster (1)
  > GlueCEInfoLRMSType (1)
  > GlueCEPolicyMaxTotalJobs (
  > GlueInformationServiceURL
  > GlueCEInfoJobManager (1)
  > GlueCEPolicyPriority (1)
  > GlueCEInfoLRMSVersion (1)
  > GlueCEStateWorstResponse
  > GlueCEStateWaitingJobs (1)
  > GlueCEStateFreeJobSlots (1)
  > GlueCEStateRunningJobs (1
  > GlueCEInfoGatekeeperPort (
  > GlueCEName (1)
  > GlueCEImplementationNam
  > GlueCEPolicyMaxRunningJo
  > GlueCEInfoGRAMVersion (1)
  > objectClass (9)
  > GlueCEStateStatus (1)
  > GlueCEAccessControlBaseR
  > GlueCEPolicyAssignedJobSl
  > GlueSchemaVersionMajor (1
  > GlueSchemaVersionMinor (1
  > GlueCEUniqueID (1)
  > GlueCEStateEstimatedRespo
  > GlueCEInfoTotalCPUs (1)
  > GlueCEInfoContactString (1
  > GlueCEInfoHostName (1)
  > GlueCEPolicyMaxWallClock
  > GlueCEStateTotalJobs (1)
  > GlueForeignKey (1)
  > GlueCECapability (1)
  > GlueCEPolicyMaxCPUTime
```

‘an infrastructure with components matched to application need’

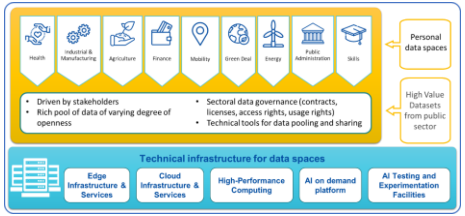
- systems architecture: compute (HTC clusters), networking, storage, and application structure
- in a balanced and {energy,cost}-efficient setup

BerkeleyDB Information System for EGI, from top-level BDII at ldap://bdii03.nikhef.nl:2170/o=grid; Earth visualization: https://dashb-earth.cern.ch/, Google Earth

European Open Science Cloud (EOSC) ecosystem example



and many more systems and ‘data spaces’ besides EOSC: *e.g.* Copernicus EO data, GAIA-X, sectoral spaces, …


Health
Industrial & Manufacturing
Agriculture
Finance
Mobility
Green Deal
Energy
Public Administration
Skills
Personal data spaces
High Value Datasets from public sector
Technical infrastructure for data spaces
Edge Infrastructure & Services
Cloud Infrastructure & Services
High-Performance Computing
AI on demand platform
AI Testing and Experimentation Facilities

- Driven by stakeholders
- Rich pool of data of varying degree of openness
- Sectoral data governance (contracts, licenses, access rights, usage rights)
- Technical tools for data pooling and sharing

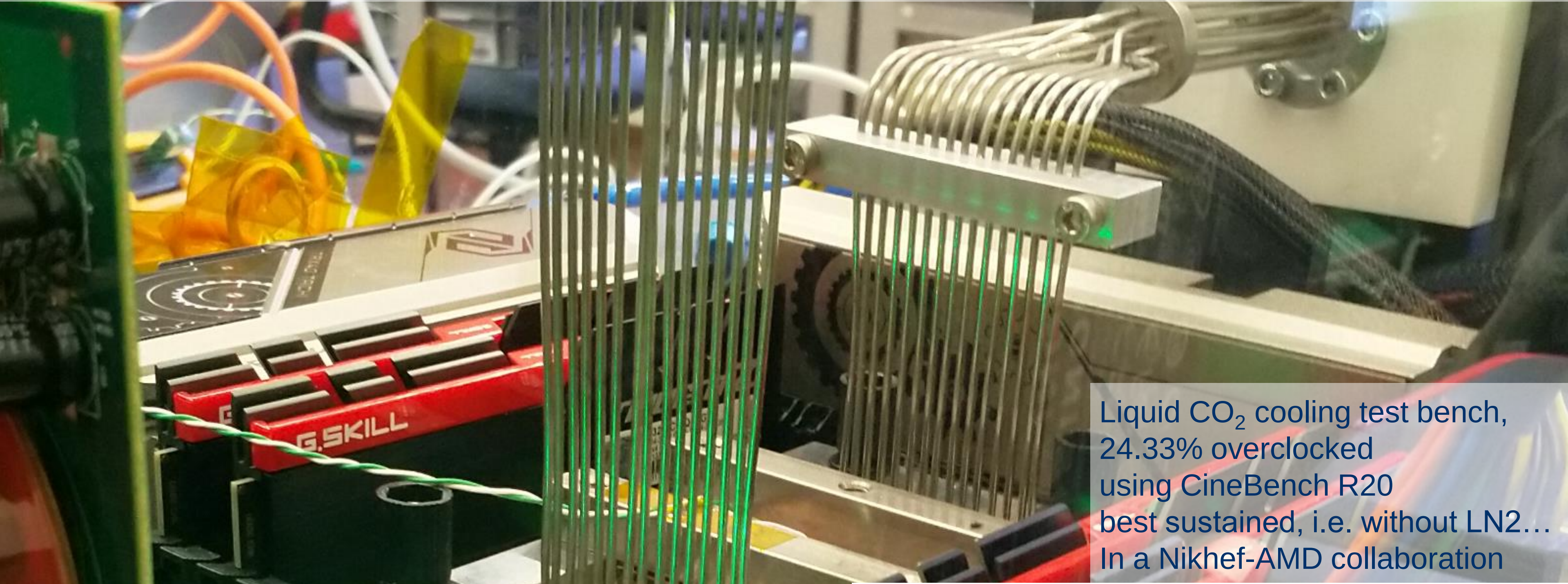
EOSC Future

EOSC: https://eoscfuture.eu/wp-content/uploads/2022/04/EOSC-Core.pdf; data spaces image: https://digital-strategy.ec.europa.eu/en/library/building-data-economy-brochure

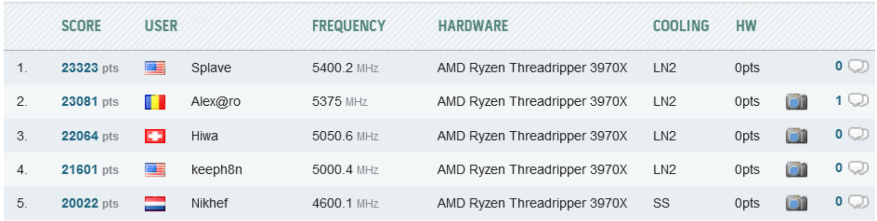
Looking for a common pattern?

- It’s all about *balanced* systems
  - systems are like congested highways: no use solving just *one* bottleneck
  - and the bottlenecks may be inside the system as well as in interconnects
- Horizontal scaling, and be as stateless as possible
  - although persistent storage obviously has to retain some state ☺
  - edge scales horizontally, and scaling from 2+ is much easier than from 1→ 2
- You can move problems around, but it’s hard to actually *solve* them
  - e.g. lack of a single common interface implies one needs adaptors and plugins
- Scaling *collaboration and trust* federation is as complex as scaling systems
  - composing services across administrative domains is ubiquitous
  - but beyond a certain size, $\mathcal{O}(100)$, you will find need for some policy and review

… since some things are fun, but not quite *that* scalable …

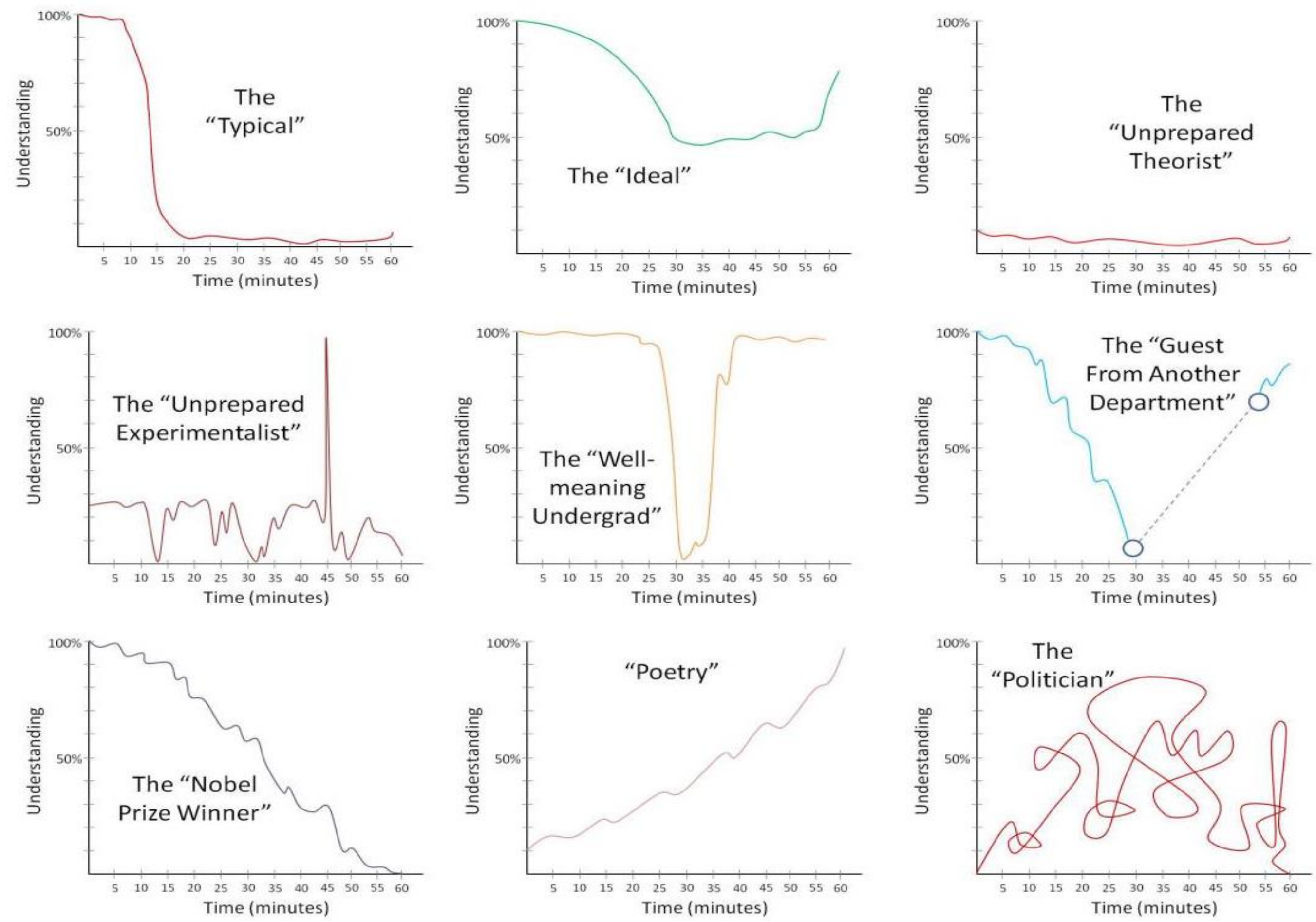


Liquid CO_2 cooling test bench,
24.33% overclocked
using CineBench R20
best sustained, i.e. without LN2…
In a Nikhef-AMD collaboration



	SCORE	USER		FREQUENCY	HARDWARE	COOLING	HW		
1.	23323 pts	🇺🇸	Splave	5400.2 MHz	AMD Ryzen Threadripper 3970X	LN2	0pts		0 💬
2.	23081 pts	🇷🇴	Alex@ro	5375 MHz	AMD Ryzen Threadripper 3970X	LN2	0pts	📷	1 💬
3.	22064 pts	🇨🇭	Hiwa	5050.6 MHz	AMD Ryzen Threadripper 3970X	LN2	0pts	📷	0 💬
4.	21601 pts	🇺🇸	keeph8n	5000.4 MHz	AMD Ryzen Threadripper 3970X	LN2	0pts	📷	0 💬
5.	20022 pts	🇳🇱	Nikhef	4600.1 MHz	AMD Ryzen Threadripper 3970X	SS	0pts	📷	0 💬

T Suerink, K de Roo: https://hwbot.org/submission/4539341_nikhef_cinebench___r20_with_benchmate_ryzen_threadripper_3970x_20022_pts



http://manyworldstheory.com/2013/10/03/the-9-kinds-of-physics-seminar/

More Q&A time!

David Groep, davidg@nikhef.nl

https://www.nikhef.nl/~davidg/presentations/

https://orcid.org/0000-0003-1026-6606

Nikhef

Maastricht University | Department of Advanced Computing Sciences

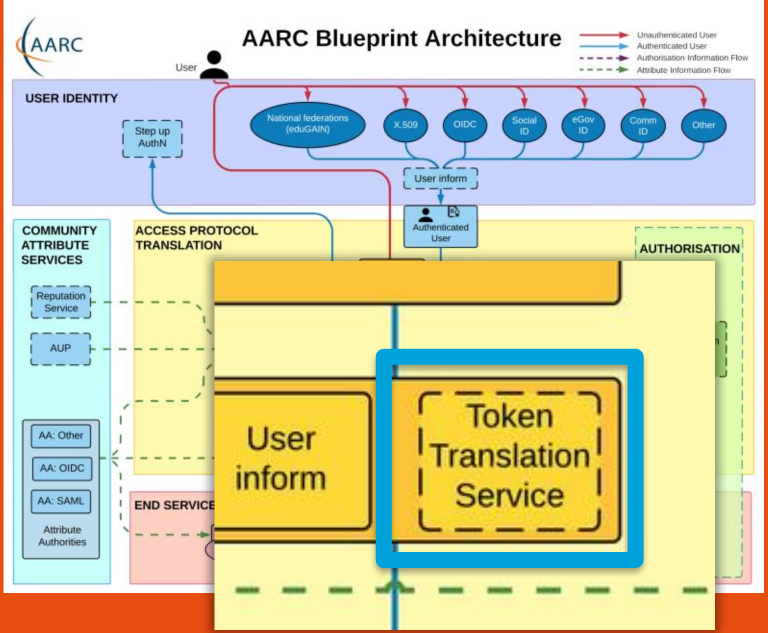
Distributed collaborative services
a more technical example with RCauth.eu

Credential translation in the AARC BPA

… building RCauth.eu

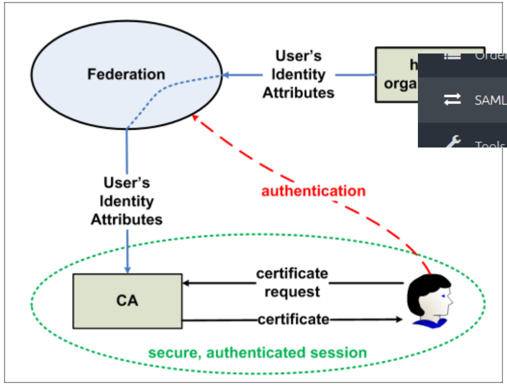
Leveraging federation and collaboration

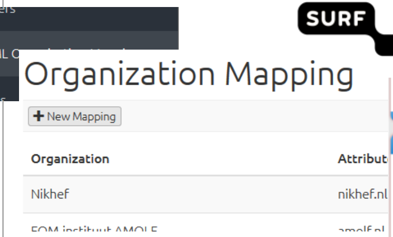
for ubiquitous research credentials



Bridges and Token Translation Services

TCS - for users that manage to grasp the idea


Federation
User's Identity Attributes
h orga
User's Identity Attributes
authentication
CA
certificate request
certificate
secure, authenticated session

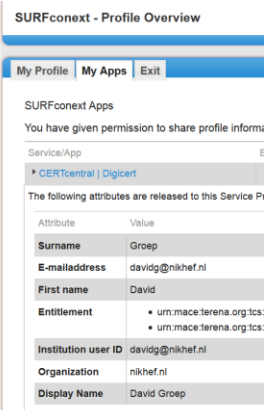


Organization Mapping

+ New Mapping

Organization	Attribut
Nikhef	nikhef.nl
FOM instituut AMOLF	amolf.nl

SURF GÉANT


SURFconext - Profile Overview

SURFconext Apps

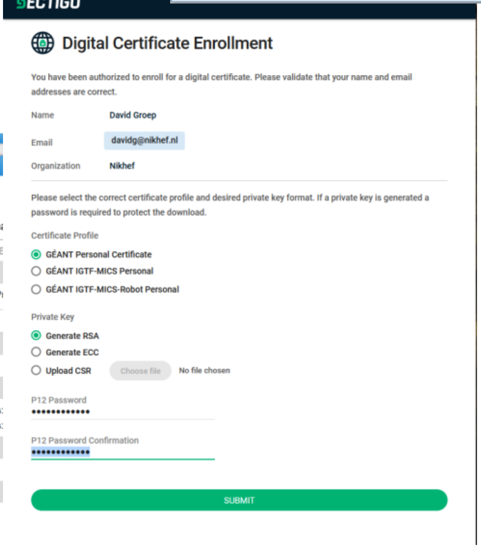
You have given permission to share profile informa

Service/App

‣ CERTcentral | Digicert

The following attributes are released to this Service Pr

Attribute	Value
Surname	Groep
E-mailaddress	davidg@nikhef.nl
First name	David
Entitlement	• urn:mace:terena.org:tcs: • urn:mace:terena.org:tcs:
Institution user ID	davidg@nikhef.nl
Organization	nikhef.nl
Display Name	David Groep



Digital Certificate Enrollment

You have been authorized to enroll for a digital certificate. Please validate that your name and email addresses are correct.

Name David Groep

Email davidg@nikhef.nl

Organization Nikhef

Please select the correct certificate profile and desired private key format. If a private key is generated a password is required to protect the download.

Certificate Profile

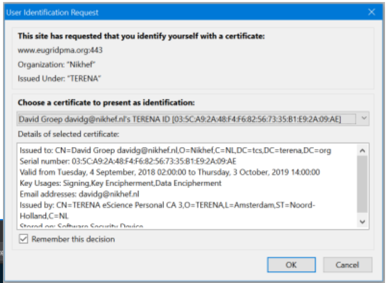
● GÉANT Personal Certificate
○ GÉANT IGTF-MICS Personal
○ GÉANT IGTF-MICS-Robot Personal

Private Key

● Generate RSA
○ Generate ECC
○ Upload CSR Choose file No file chosen

P12 Password ••••••••••••

P12 Password Confirmation ••••••••••••


User Identification Request

This site has requested that you identify yourself with a certificate:

www.eugridpma.org:443

Organization: "Nikhef"

Issued Under: "TERENA"

Choose a certificate to present as identification:

David Groep davidg@nikhef.nl's TERENA ID [03:5C:A9:2A:48:F4:F6:82:56:73:35:B1:E9:2A:09:AE]

Details of selected certificate:

Issued to: CN=David Groep davidg@nikhef.nl,O=Nikhef,C=NL,DC=tcs,DC=terena,DC=org
Serial number: 03:5C:A9:2A:48:F4:F6:82:56:73:35:B1:E9:2A:09:AE
Valid from Tuesday, 4 September, 2018 02:00:00 to Thursday, 3 October, 2019 14:00:00
Key Usages: Signing,Key Encipherment,Data Encipherment
Email addresses: davidg@nikhef.nl
Issued by: CN=TERENA eScience Personal CA 3,O=TERENA,L=Amsterdam,ST=Noord-Holland,C=NL
Stored on: Software Security Device

☑ Remember this decision

TCS is a SAML Service Provider (today by Sectigo) to eduGAIN: where eligible authenticated users obtain client certificates for access to many research services

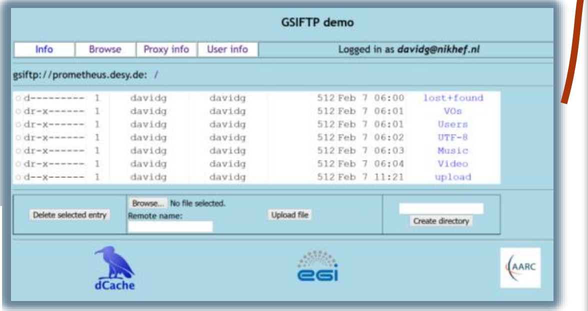
A globally recognized identity for all employees & students (they are automatically eligible!).

GEANT Trusted Certificate Service - https://ca.dutchgrid.nl/tcs/, https://cert-manager.com/customer/surfnet/idp/clientgeant, https://www.geant.org/Services/Trust_identity_and_security/Pages/TCS.aspx

Seamless in-line token translation services from ‘SAML’ to PKIX

user facing | hidden back-end

Community Science Portal



GSIFTP demo

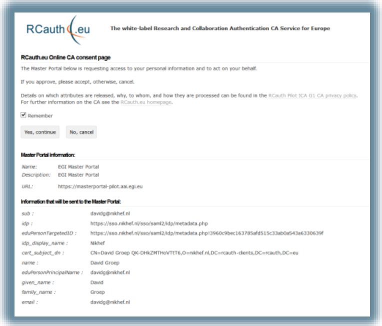
Logged in as ***davidg@nikhef.nl***

gsiftp://prometheus.desy.de: /

```
○ d---------  1   davidg   davidg   512 Feb 7 06:00  lost+found
○ dr-x------  1   davidg   davidg   512 Feb 7 06:01  VOs
○ dr-x------  1   davidg   davidg   512 Feb 7 06:01  Users
○ dr-x------  1   davidg   davidg   512 Feb 7 06:02  UTF-8
○ dr-x------  1   davidg   davidg   512 Feb 7 06:03  Music
○ dr-x------  1   davidg   davidg   512 Feb 7 06:04  Video
○ d--x------  1   davidg   davidg   512 Feb 7 11:21  upload
```

Browse... No file selected.

Remote name:



RCauth.eu — The white-label Research and Collaboration Authentication CA Service for Europe

RCauth.eu Online CA consent page

The Master Portal below is requesting access to your personal information and to act on your behalf.

If you approve, please accept, otherwise, cancel.

Details on which attributes are released, why, to whom, and how they are processed can be found in the RCauth Pilot ICA G1 CA privacy policy. For further information on the CA see the RCauth.eu homepage.

☑ Remember

Master Portal information:

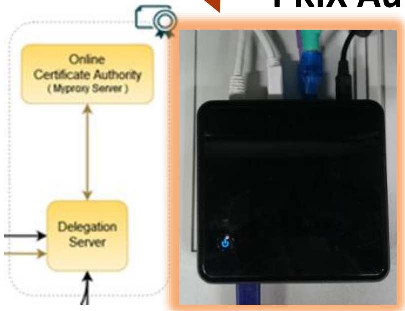
Name: EGI Master Portal
Description: EGI Master Portal
URL: https://masterportal-pilot.aai.egi.eu

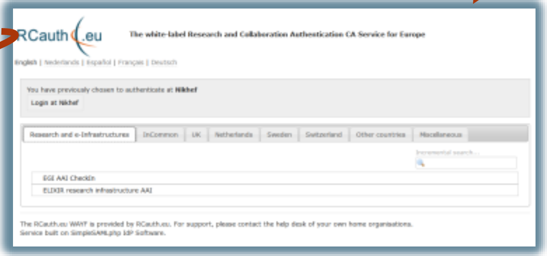
Information that will be sent to the Master Portal:

sub : davidg@nikhef.nl
idp : https://sso.nikhef.nl/sso/saml2/idp/metadata.php
eduPersonTargetedID : https://sso.nikhef.nl/sso/saml2/idp/metadata.php!3960c9bec163785afd515c33ab0a543a6330639f
idp_display_name : Nikhef
cert_subject_dn : CN=David Groep QK-DHkZMTHoVTtT6,O=nikhef.nl,DC=rcauth-clients,DC=rcauth,DC=eu
name : David Groep
eduPersonPrincipalName : davidg@nikhef.nl
given_name : David
family_name : Groep
email : davidg@nikhef.nl

Infrastructure Master Portal Credential Store

IGTF accredited PKIX Authority





RCauth.eu — The white-label Research and Collaboration Authentication CA Service for Europe

English | Nederlands | Español | Français | Deutsch

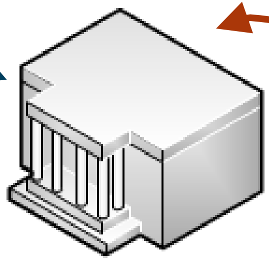
You have previously chosen to authenticate at **Nikhef**

EGI AAI CheckIn
ELIXIR research infrastructure AAI

The RCauth.eu WAYF is provided by RCauth.eu. For support, please contact the help desk of your own home organisations.
Service built on SimpleSAMLphp IdP Software.

Policy Filtering WAYF to eduGAIN

User Home Org
or Infrastructure IdP



REFEDS R&S
Sirtfi Trust

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

Built on CILogon and MyProxy, see www.cilogon.org CILogon

Unique certificated from FIM via eduPerson and REFEDS R&S

Sources of naming and uniqueness, that work *today*

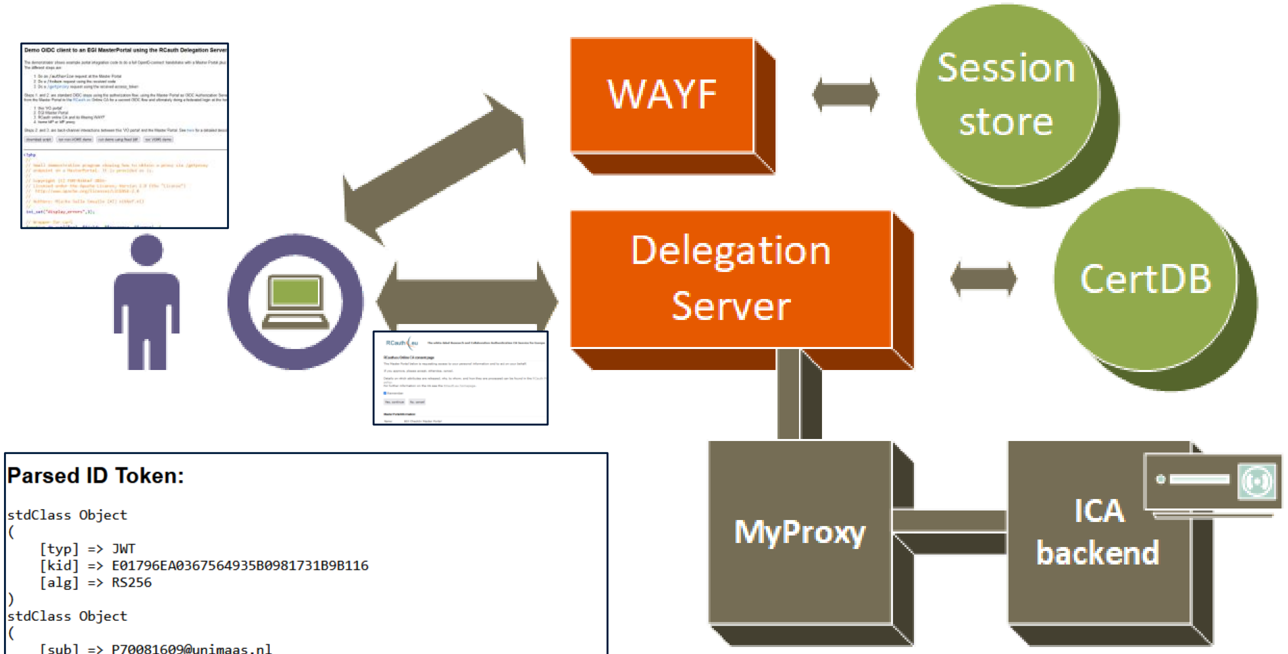
- **eduPersonPrincipalName** – scoped point-in-time unique identifier, which could be, but usually is not, privacy preserving: “davidg@nikhef.nl”, “P70081609@maastrichtuniversity.nl”
- **eduPersonTargetedID** – scoped transient non-reassigned identifier, like urn:geant:nikhef.nl:nikidm:idp:sso!*27c8d63ed42c84af2875e2984*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

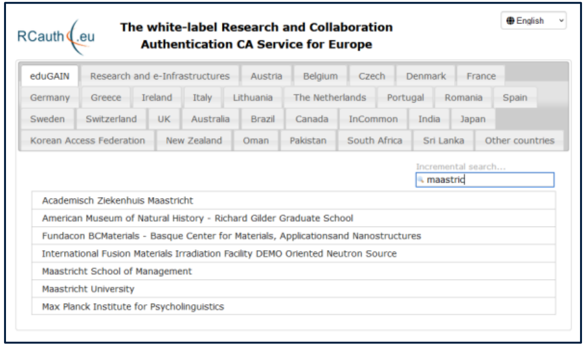
Plus the (domain-name based) schacHomeOrganisation and a ‘**representation of the real name**’

/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”

The ‘back side’ of a typical RCauth portal data flow


WAYF
Session store
Delegation Server
CertDB
MyProxy
ICA backend
Demo OIDC client to an EGI MasterPortal using the RCauth Delegation Server
RCauth.eu



The white-label Research and Collaboration Authentication CA Service for Europe

- Academisch Ziekenhuis Maastricht
- American Museum of Natural History - Richard Gilder Graduate School
- Fundacon BCMaterials - Basque Center for Materials, Applicationsand Nanostructures
- International Fusion Materials Irradiation Facility DEMO Oriented Neutron Source
- Maastricht School of Management
- Maastricht University
- Max Planck Institute for Psycholinguistics

Parsed ID Token:

```
stdClass Object
(
    [typ] => JWT
    [kid] => E01796EA0367564935B0981731B9B116
    [alg] => RS256
)
stdClass Object
(
    [sub] => P70081609@unimaas.nl
    [idp] => http://login.maastrichtuniversity.nl/adfs/services/trust
    [eduPersonTargetedID] => http://login.maastrichtuniversity.nl/adfs
    [idp_display_name] => Maastricht University
    [cert_subject_dn] => CN=Groep\, David (DACS) KWwWAnhI4psmiGTw 1,O=
    [name] => Groep, David (DACS)
    [eduPersonPrincipalName] => P70081609@unimaas.nl
    [given_name] => David
    [family_name] => Groep
    [email] => david.groep@maastrichtuniversity.nl
    [iss] => https://aai.egi.eu/mp-oa2-server
```

Proxy information:

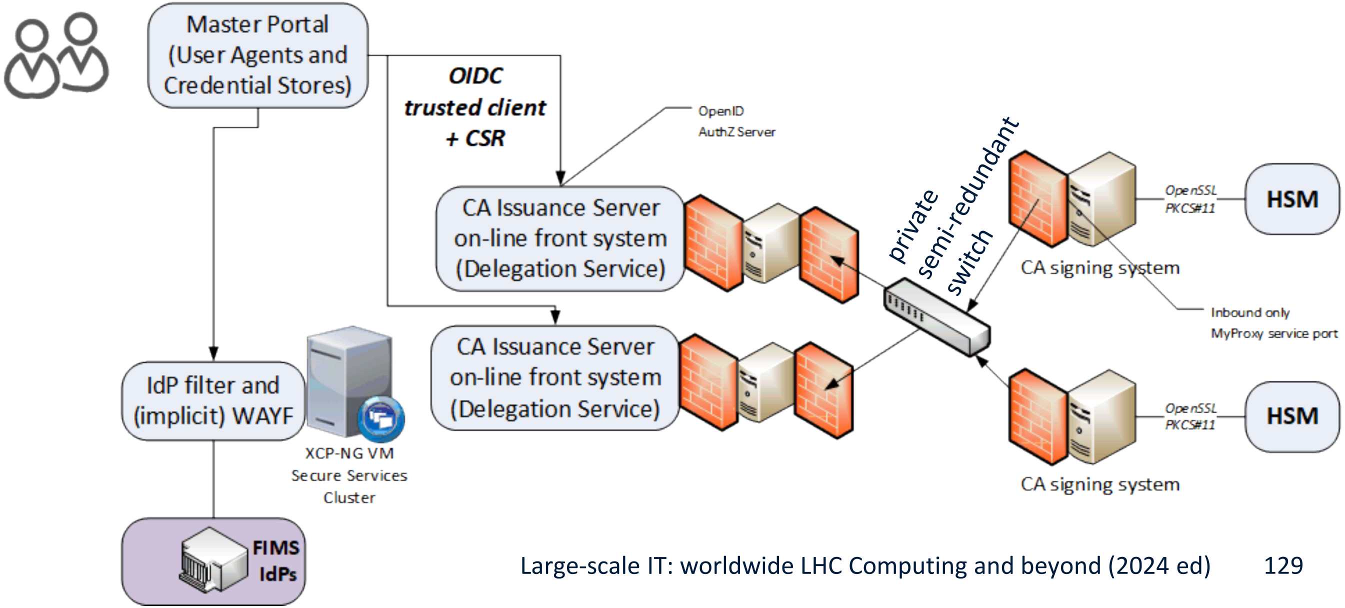
```
subject   : /DC=eu/DC=rcauth/DC=rcauth-clients/O=maastrichtuniversity.nl/CN=Groep, David (DACS) KWwWAnhI4psmiGTw 1/CN=208760481/CN=466908503
issuer    : /DC=eu/DC=rcauth/DC=rcauth-clients/O=maastrichtuniversity.nl/CN=Groep, David (DACS) KWwWAnhI4psmiGTw 1/CN=208760481
identity  : /DC=eu/DC=rcauth/DC=rcauth-clients/O=maastrichtuniversity.nl/CN=Groep, David (DACS) KWwWAnhI4psmiGTw 1/CN=208760481
type      : RFC compliant proxy
strength  : 2048 bits
path      : /tmp/x509up_uiI4TkF
```

With a single, yet fully compliant, 'Heath Robinson' CA

A single-site locally-highly-available RCauth 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 (physical hardware, albeit with redundant disks and powersupplies)

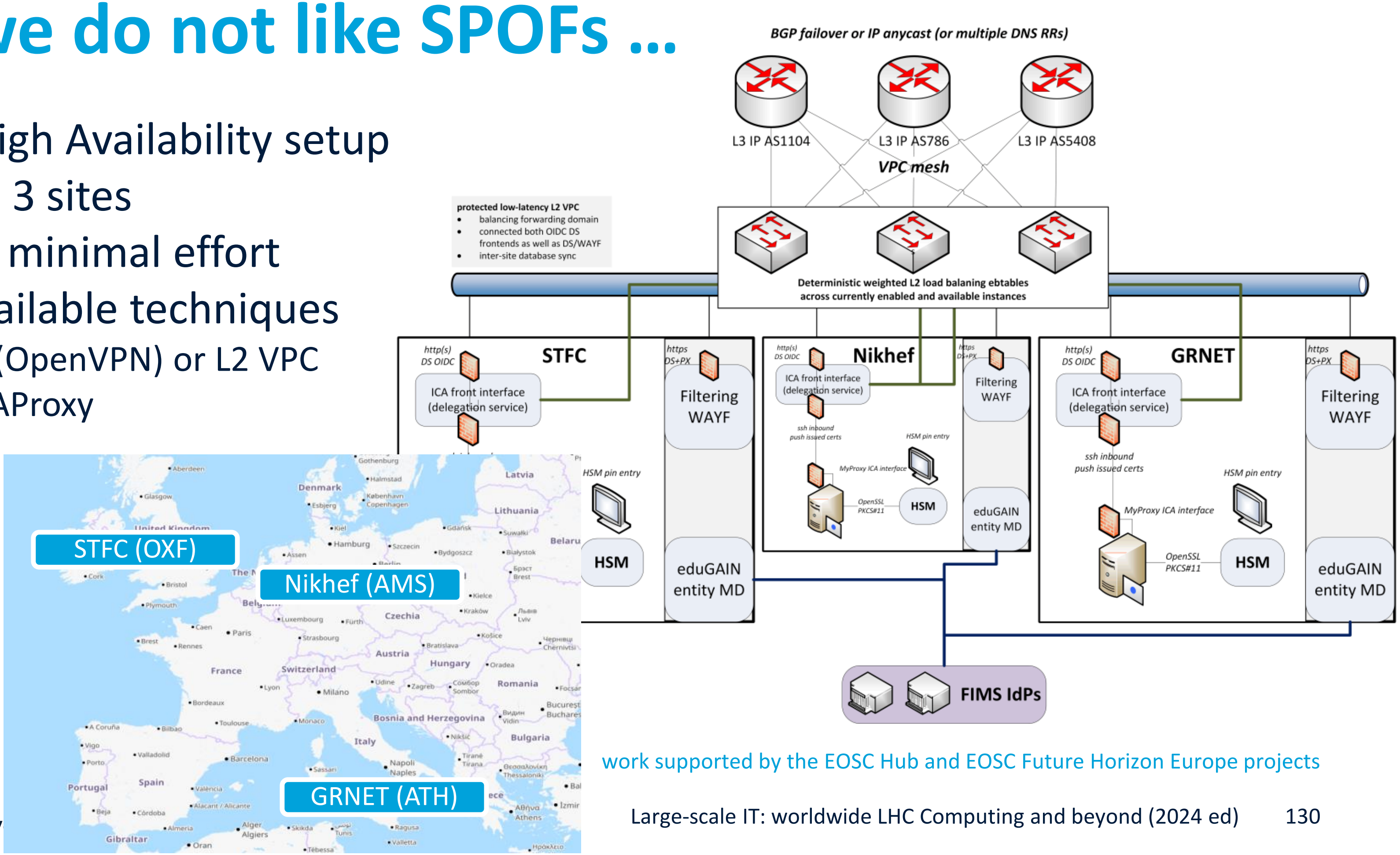
Eliminated SPOFs first using ‘local HA’


Master Portal (User Agents and Credential Stores)
OIDC trusted client + CSR
OpenID AuthZ Server
CA Issuance Server on-line front system (Delegation Service)
CA Issuance Server on-line front system (Delegation Service)
private semi-redundant switch
OpenSSL PKCS#11
HSM
CA signing system
Inbound only MyProxy service port
OpenSSL PKCS#11
HSM
CA signing system
IdP filter and (implicit) WAYF
XCP-NG VM Secure Services Cluster
FIMS IdPs

Since we do not like SPOFs …

Distributed High Availability setup

- across the 3 sites
- design for minimal effort
- readily-available techniques
    - L3 VPN (OpenVPN) or L2 VPC
    - Linux HAProxy

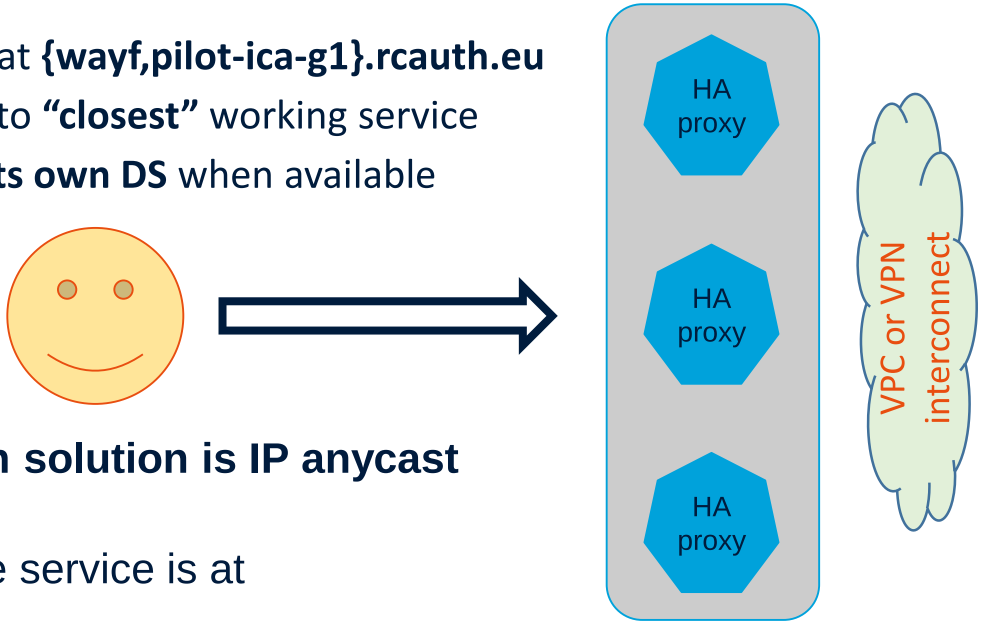

BGP failover or IP anycast (or multiple DNS RRs)
L3 IP AS1104
L3 IP AS786
L3 IP AS5408
VPC mesh
protected low-latency L2 VPC
balancing forwarding domain
connected both OIDC DS frontends as well as DS/WAYF
inter-site database sync
Deterministic weighted L2 load balaning ebtables across currently enabled and available instances
STFC
http(s) DS OIDC
ICA front interface (delegation service)
HSM pin entry
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
STFC (OXF)
Nikhef (AMS)
GRNET (ATH)

work supported by the EOSC Hub and EOSC Future Horizon Europe projects

A *transparent* multi-site setup is needed for the user

User

- connects to HA proxy at **{wayf,pilot-ica-g1}.rcauth.eu**
- HA proxy sends users to **“closest”** working service
- primarily **forward to its own DS** when available



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

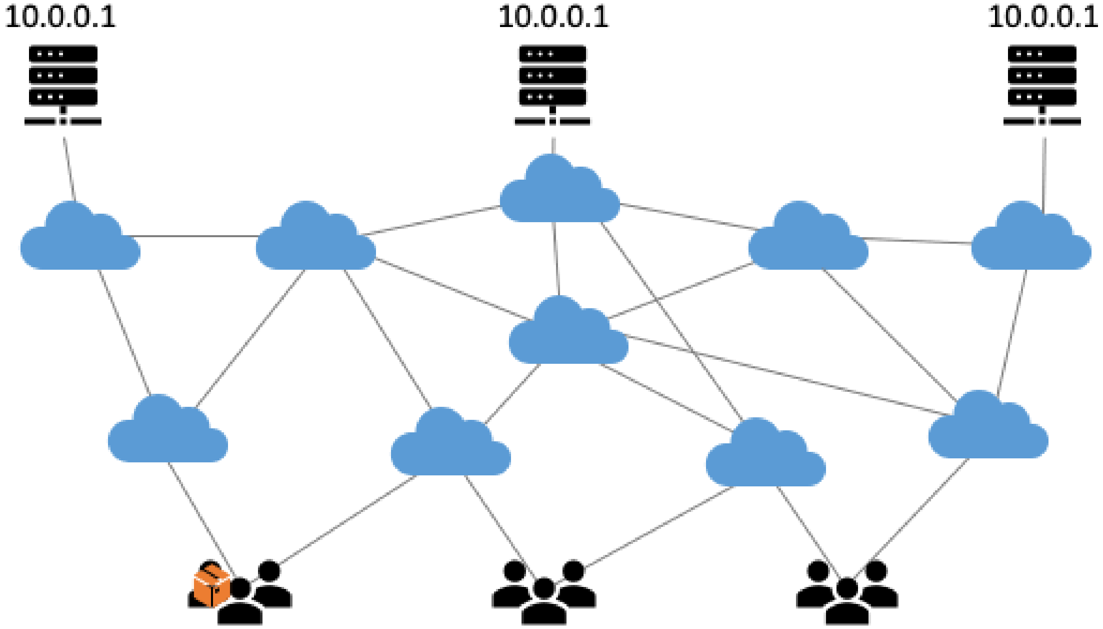
Straightforward proven solution is IP anycast

wherever the user is, the service is at

- **2a07:8504:01a0::1**
- or for legacy IP users at 145.116.216.1

selected imagery: Mischa Sallé, Jens Jensen, Nicolas Liampotis

Anycast: when the same place exists many times



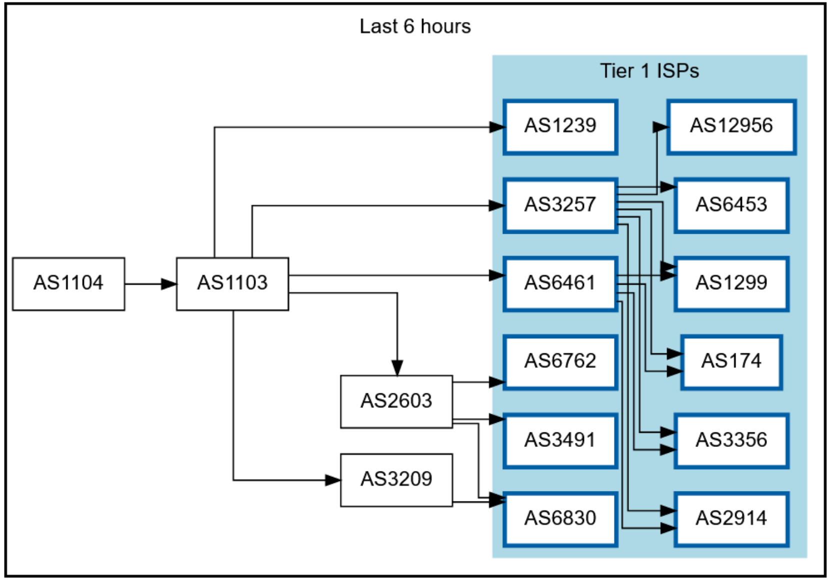
So we used

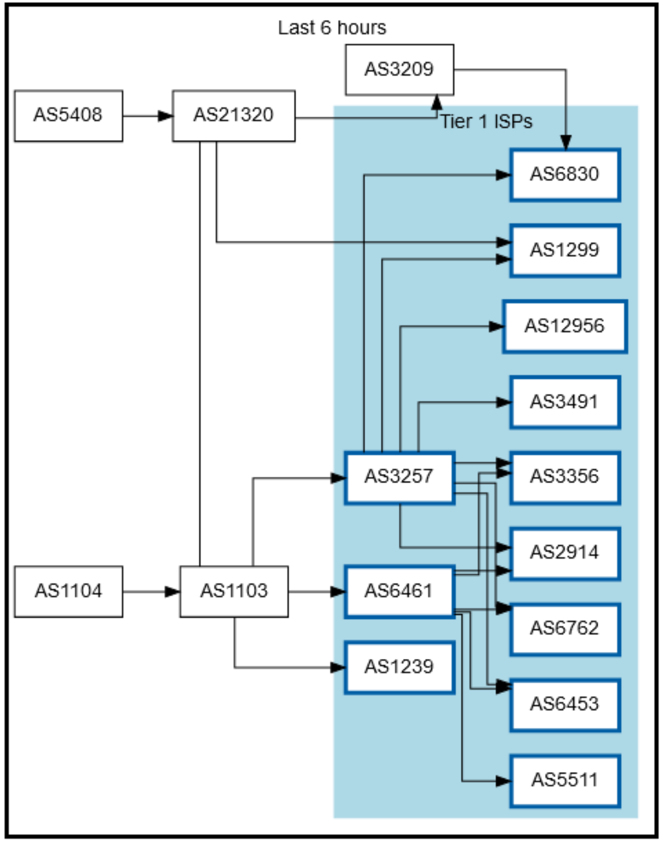
- 3 (for now: 2) sites
- one VM at each site exposing 2a07:8504:01a0::1
- smallest v6 subnet (/48)
- bird + a service probe
- each site's own ASN
- some IRR DB editing
- IPv4 is similar, with a /24

and some monitoring

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

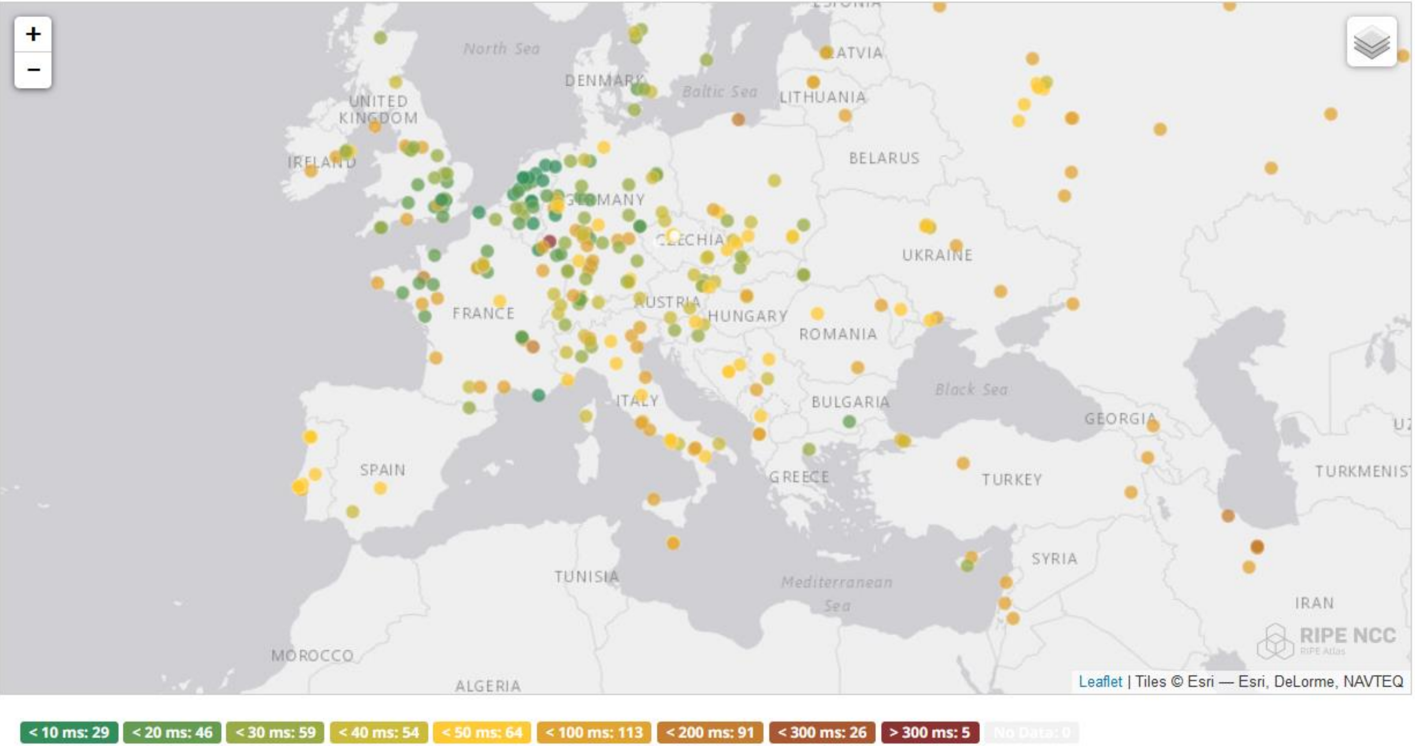
Getting 2a07:8504:1a0::/48 out there





route maps: bgp.tools for 2a07:8504:1a0::/48 – IPv4 for 145.116.216.0/24 is similar – imagery from November 2022

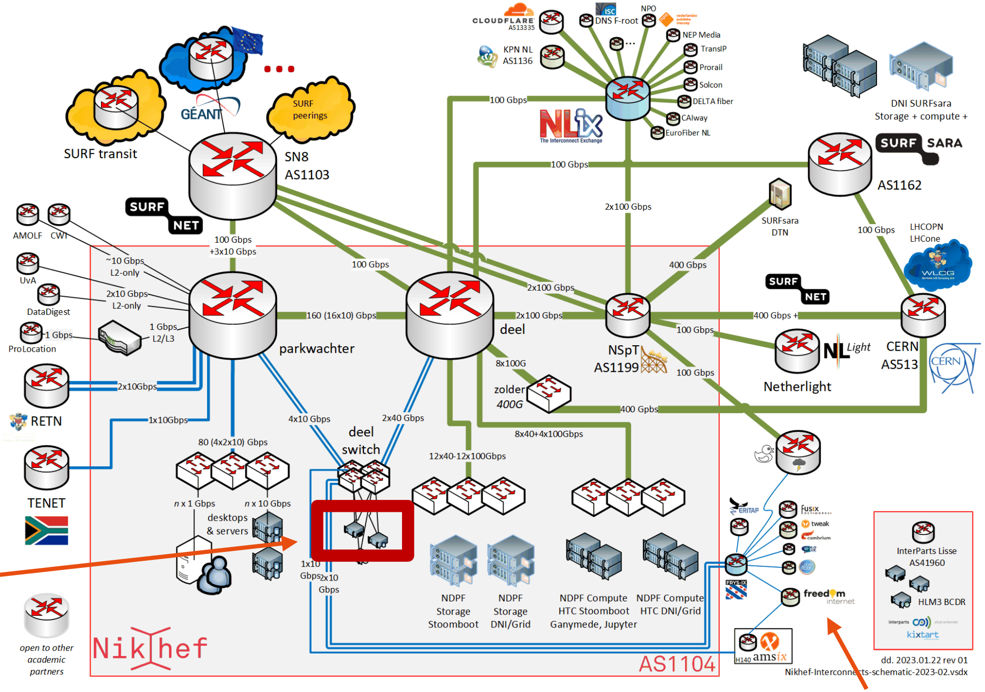
And you get reasonable load balancing in Europe for free



map: RIPE NCC RIPE Atlas - 500 probes, distributed across Europe (https://atlas.ripe.net/measurements/50949024/)

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

me, at home

Route from home to RCauth.eu, from my home Freedom Internet ISP

RSA Crypto

Just in case … you cannot factor ‘55’

Establishing trust at a distance

Remote trust needs cryptography in some way

Client authentication

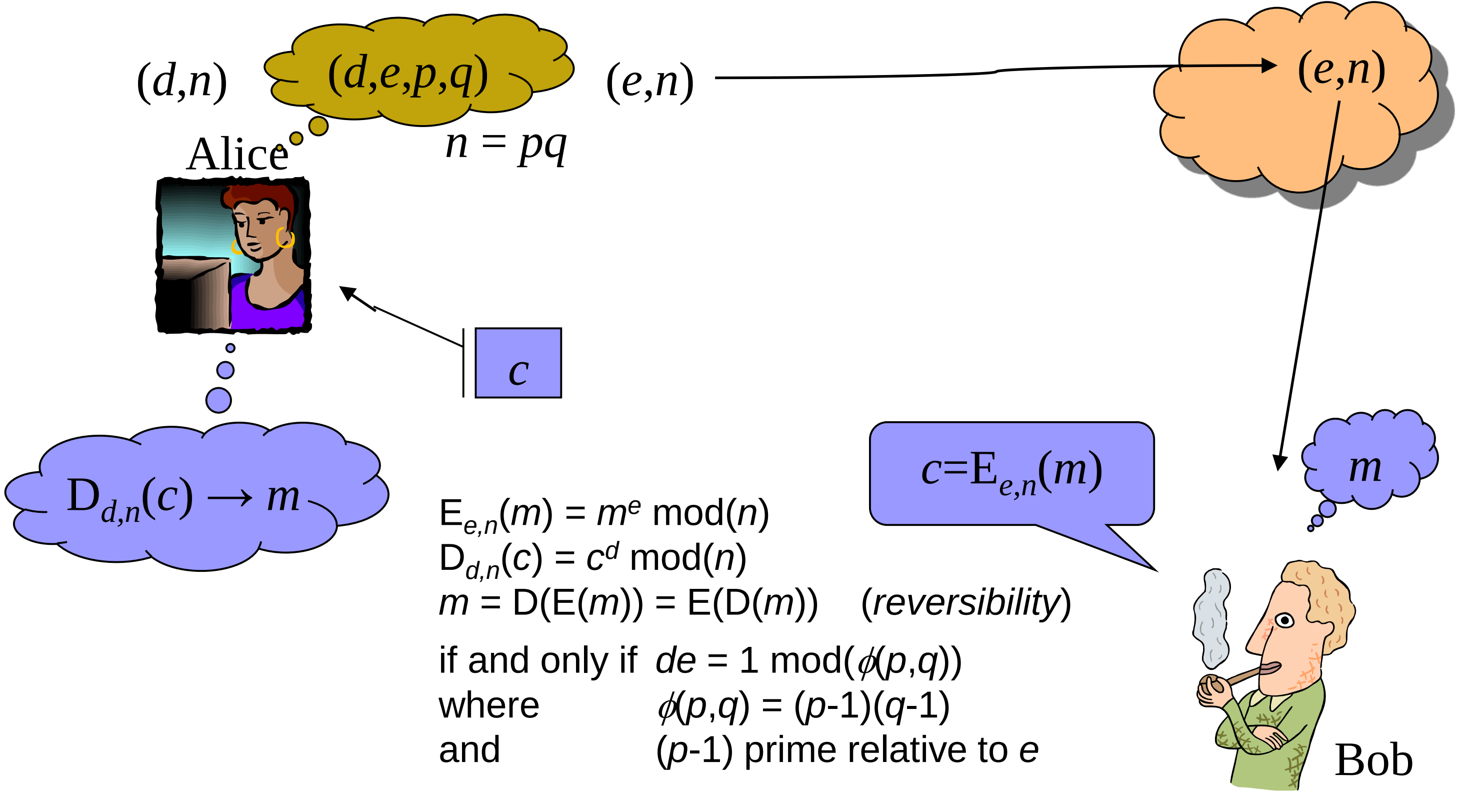
- pre-shared secrets, may be salted hashed on service side
- required: secure one-way hash function
- need a **protected channel** between identifiable end-points

Mutual authentication

- eithers need a lot of shared keys, a trusted third party (TTP), or mesh validation (WoT)
- with the TTP and multiple services comes the need for crypto
- across administrative domains, *key distribution* is the larger challenge

The cryptography used can be either *symmetric* or *asymmetric*, ‘public key’

Asymmetric crypto: RSA interlude needed?



(d,n) (d,e,p,q) (e,n)

$n = pq$

Alice

(e,n)

c

$D_{d,n}(c) \rightarrow m$

$c = E_{e,n}(m)$

m

Bob

$E_{e,n}(m) = m^e \bmod(n)$
$D_{d,n}(c) = c^d \bmod(n)$
$m = D(E(m)) = E(D(m))$ *(reversibility)*

if and only if $de = 1 \bmod(\phi(p,q))$
where $\phi(p,q) = (p\text{-}1)(q\text{-}1)$
and (p-1) prime relative to e

Rivest, Shamir and Adleman, Communications of the ACM 21 (2), 120-126

6-bit RSA (note: this might be broken quickly …)

- Take a (small) value e = **3**
- Generate a set of primes (p,q), each with a length of $k/2$ bits, with (p-1) prime relative to e.
  (p,q) = **(11,5)**
- $\phi(p,q)$ = (11-1)(5-1) = **40**; $n=pq=$**55**
- find d, in this case **27** [3*27 = 81 = 1 mod(40)]

- Public Key: **(3,55)**
- Private Key: **(27,55)**

$E_{e,n}(m) = m^e \bmod(n)$
$D_{d,n}(c) = c^d \bmod(n)$
$m = D(E(m)) = E(D(m))$
if a.o. if $de = 1 \bmod(\phi(p,q))$
where $\phi(p,q) = (p-1)(q-1)$

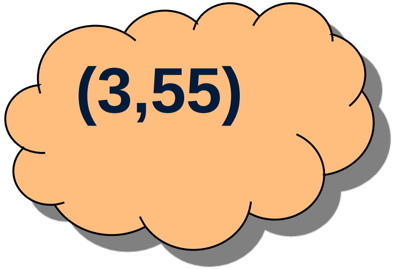
Message exchange

Encryption:

- Bob thinks of a plaintext $m(<n)$ = **18**
- Encrypt with Alice's public key **(3,55)**
- $c=\mathrm{E}_{\mathbf{3;55}}(18)=18^3$ mod(55) = 5832 mod(55) = **2**
- send message **"2"**

Decryption:

- Alice gets **"2"**
- she knows private key **(27,55)**
- $\mathrm{E}_{\mathbf{27;55}}(2) = 2^{27}$ mod(55) = **18** !



(3,55)

$\mathrm{E}_{e,n}(m) = m^e \bmod(n)$

$\mathrm{D}_{d,n}(c) = c^d \bmod(n)$

$m = \mathrm{D}(\mathrm{E}(m)) = \mathrm{E}(\mathrm{D}(m))$

if a.o. if $de = 1 \bmod(\phi(p,q))$

where $\phi(p,q) = (p\text{-}1)(q\text{-}1)$

If you just have (3,55), it's hard to get the 27…

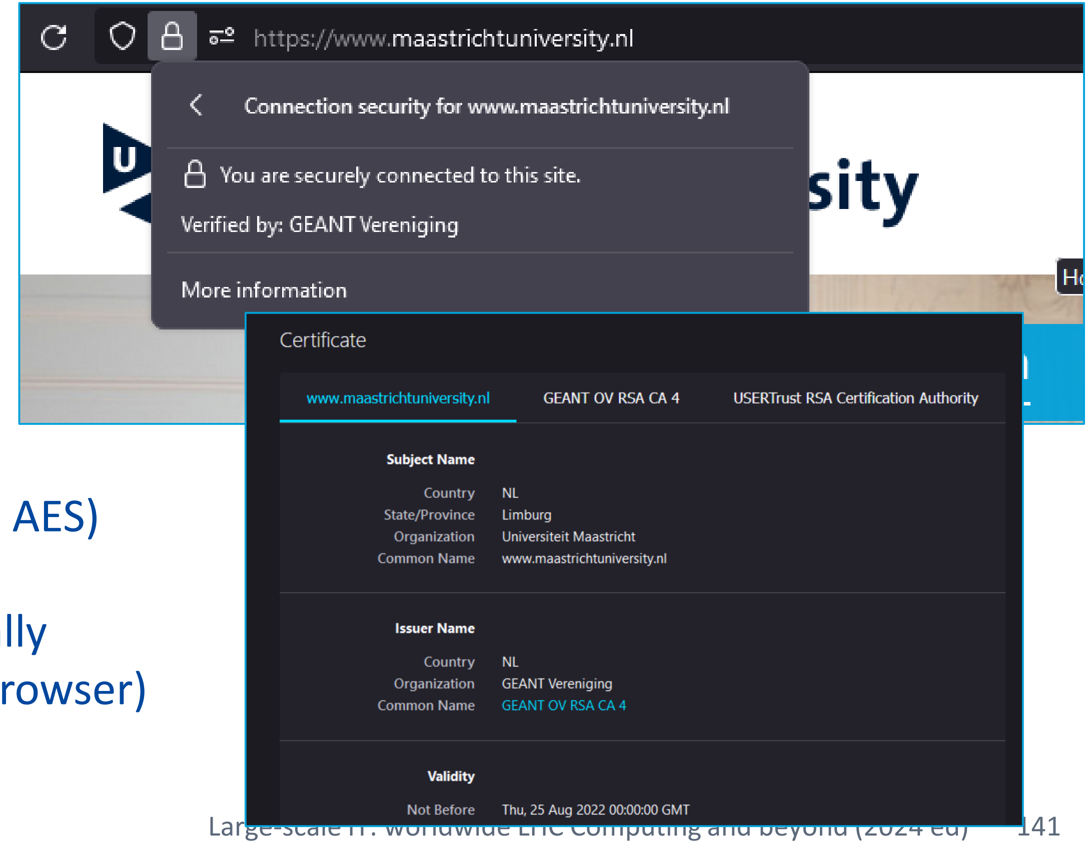
but also: the maximum plaintext is limited by the modulus length

The most used asymmetric crypto application

Asymmetric crypto underpins the transport layer security of all of the web today

- ASN.1 syntax data with X.509 (RFC5280) structure
- mostly RSA or Elliptic Curves (EC)
- used to negotiate a (symmetric) bulk cipher (typically AES)

then used to protect channel to usually *unauthenticated* client application (browser)



https://www.maastrichtuniversity.nl

Connection security for www.maastrichtuniversity.nl

You are securely connected to this site.

Verified by: GEANT Vereniging

More information

Certificate

www.maastrichtuniversity.nl | GEANT OV RSA CA 4 | USERTrust RSA Certification Authority

Subject Name

Country	NL
State/Province	Limburg
Organization	Universiteit Maastricht
Common Name	www.maastrichtuniversity.nl

Issuer Name

Country	NL
Organization	GEANT Vereniging
Common Name	GEANT OV RSA CA 4

Validity

Not Before	Thu, 25 Aug 2022 00:00:00 GMT

Other ancillary materials

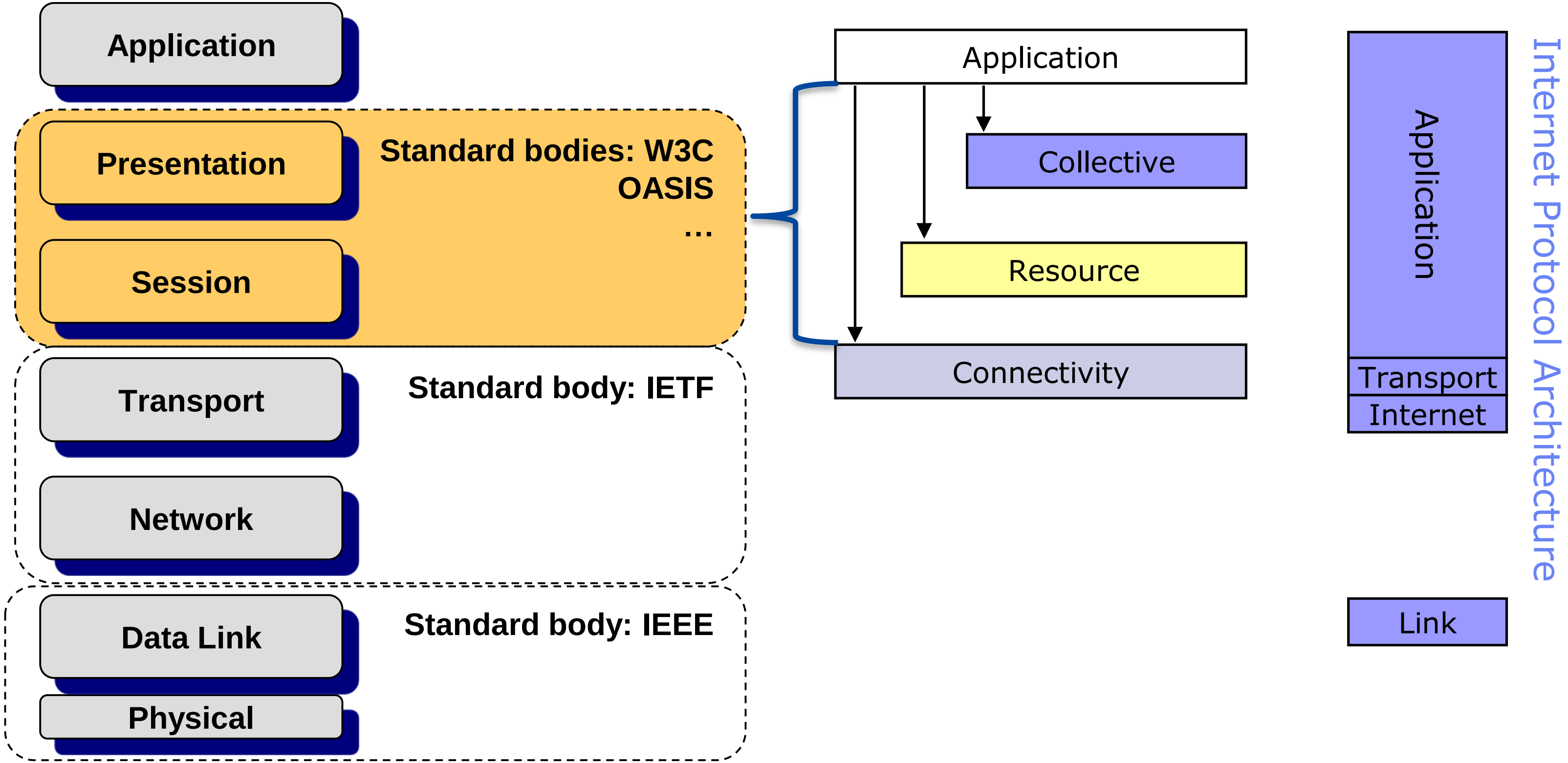
these generic slides do not form part of the module, but are just general background knowledge and example

Open Systems Interconnection model (OSI model)

Layer			Function
Host layers	7	Application	High-level protocols (resource sharing, remote file access)
	6	Presentation	Translation of data between a networking service and an application
	5	Session	Managing communication sessions, i.e., continuous exchange of information in the form of multiple back-and-forth transmissions between two nodes
	4	Transport	Reliable transmission of data segments between points on a network
Media layers	3	Network	Addressing, routing and traffic control
	2	Data link	Transmission of data frames between two nodes connected by a physical layer
	1	Physical	Transmission and reception of raw bit streams over a physical medium

OSI X.200 layering model, ITU-T (CCITT), https://www.itu.int/rec/T-REC-X.200; image adapted from https://en.wikipedia.org/wiki/OSI_model

OSI vs Internet Protocol Architecture model



Private (direct) peerings to distribute traffic load

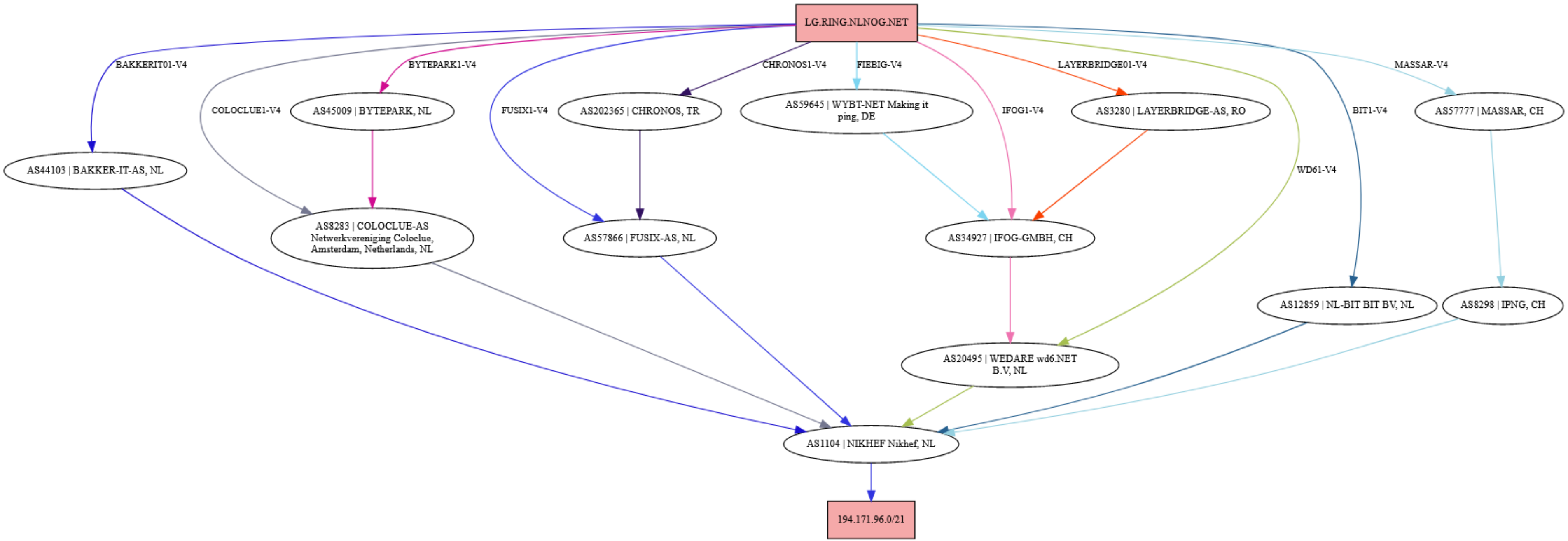
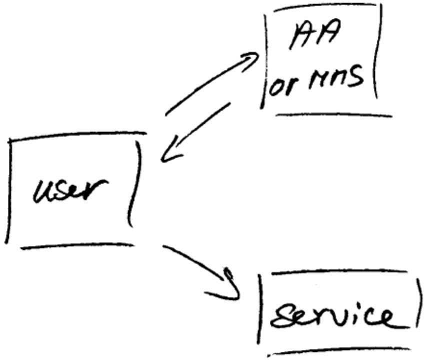
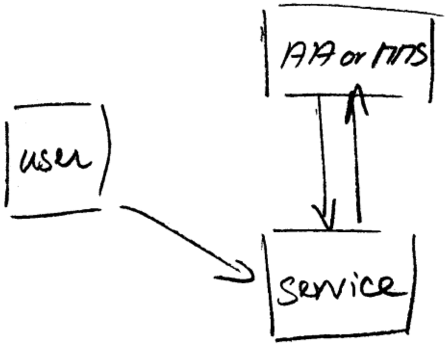


Image sources: NLNOG RING map https://lg.ring.nlnog.net/

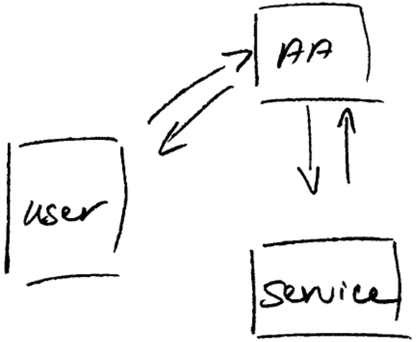
RFC2904 authorization models: three AuthZ flows



‘push’



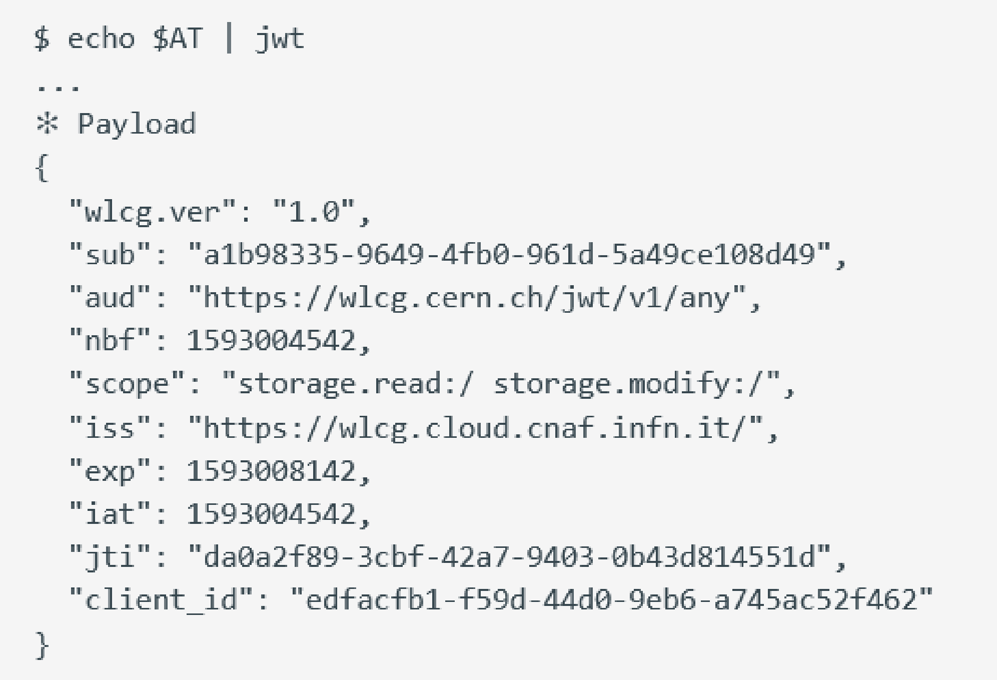
‘pull’



‘agent’

Authorization models: AAA Authorization Framework, RFC2904, Vollbrecht et al.

OAuth2 & JWTs: assertions can be quite detailed

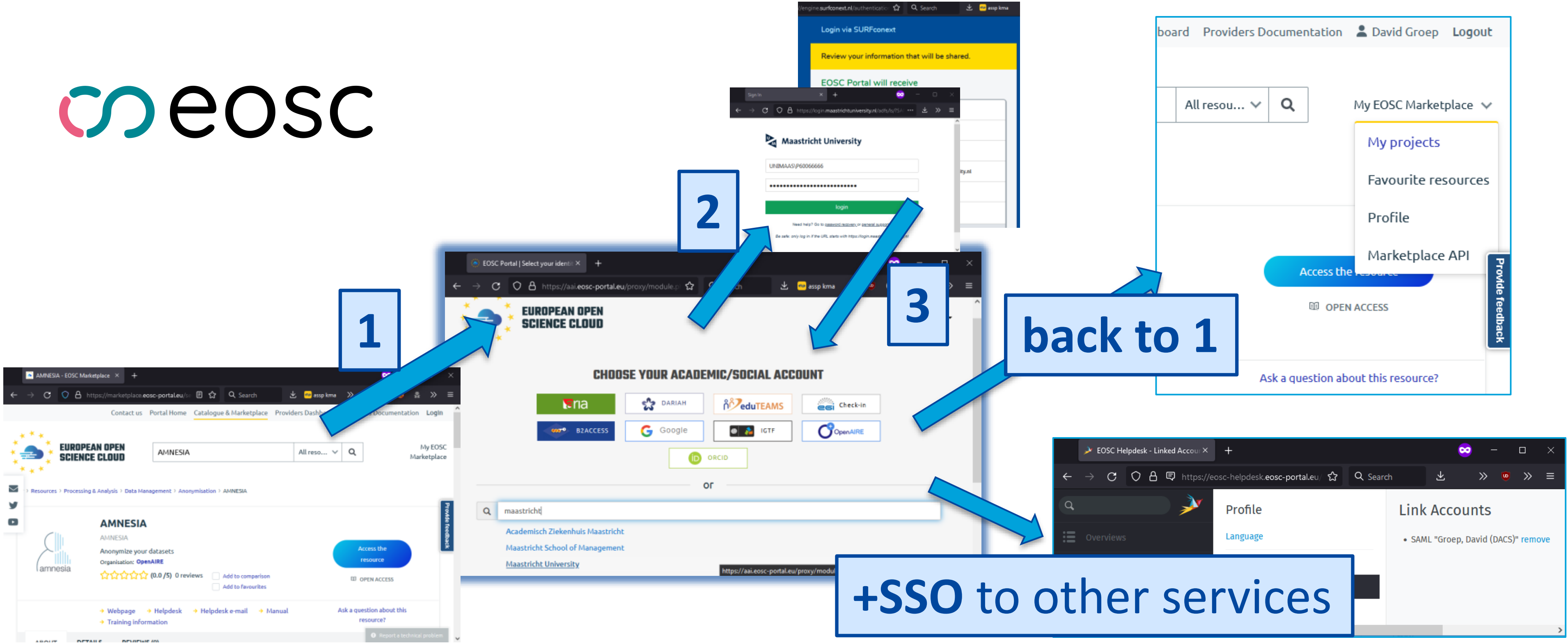


```
$ echo $AT | jwt
...
✻ Payload
{
  "wlcg.ver": "1.0",
  "sub": "a1b98335-9649-4fb0-961d-5a49ce108d49",
  "aud": "https://wlcg.cern.ch/jwt/v1/any",
  "nbf": 1593004542,
  "scope": "storage.read:/ storage.modify:/",
  "iss": "https://wlcg.cloud.cnaf.infn.it/",
  "exp": 1593008142,
  "iat": 1593004542,
  "jti": "da0a2f89-3cbf-42a7-9403-0b43d814551d",
  "client_id": "edfacfb1-f59d-44d0-9eb6-a745ac52f462"
}
```

OAuth2 Access Token following the WLCG AuthZ WG Profile, from: https://wlcg-authz-wg.github.io/wlcg-authz-docs/token-based-authorization/

Example flow in the European Open Science Cloud

eosc



EOSC Portal & Marketplace Amnesia service by the OpenAIRE e-infrastructure, EOSC Helpdesk: Zammad hosted by KIT https://eosc-helpdesk.eosc-portal.eu

RCauth demonstrator

RCauth is an AARC BPA token translation service that forges X.509 end-user certificates that are managed in a central portal for you (the portal is ‘elevator.nikhef.nl’)

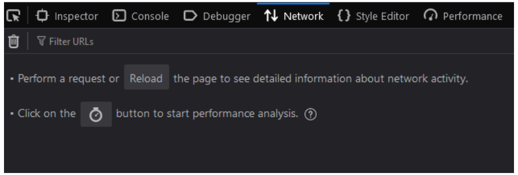
Qualified users are all those in eduGAIN with basic assurance (Sirtfi version 1 + Research & Scholarship entity categories), and everyone in a Dutch SURF ‘Annex IX’ institution – such as UM

Your end-entity certificate is globally IGTF trusted under the ‘Identifier Only Trust Assurance’ (IOTA) profile

RCauth do-it-yourself demo

- Go to https://rcdemo.nikhef.nl/
- select “Basic Demo”

- Enable browser Inspector (F12) on the network tab, (and start the SAML tracer extension if you have it)

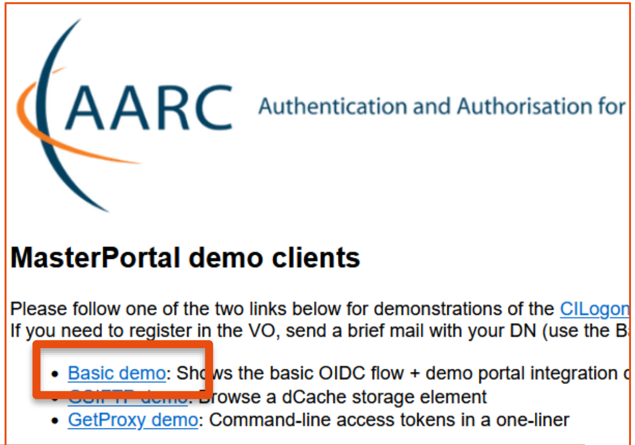


• Perform a request or Reload the page to see detailed information about network activity.

• Click on the button to start performance analysis.

- Run the “non-VOMS demo”

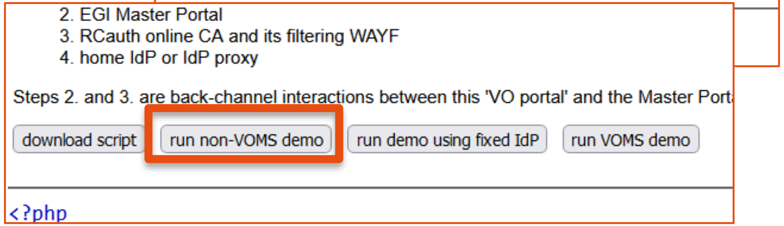
- From eduGAIN, select “Maastricht University”



MasterPortal demo clients

Please follow one of the two links below for demonstrations of the CILogon
If you need to register in the VO, send a brief mail with your DN (use the B

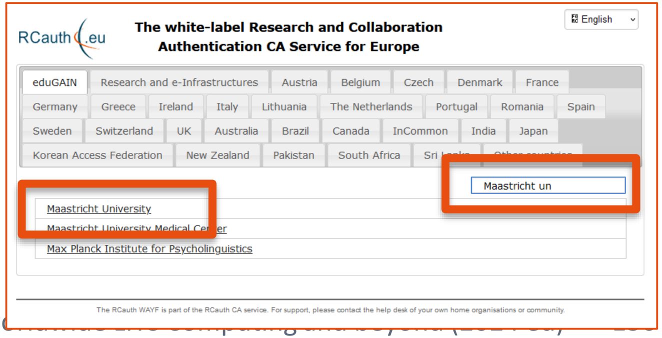
- Basic demo: Shows the basic OIDC flow + demo portal integration o
- GSIFTP demo: Browse a dCache storage element
- GetProxy demo: Command-line access tokens in a one-liner



2. EGI Master Portal
3. RCauth online CA and its filtering WAYF
4. home IdP or IdP proxy

Steps 2. and 3. are back-channel interactions between this 'VO portal' and the Master Port

<?php



The white-label Research and Collaboration Authentication CA Service for Europe

Maastricht un

Maastricht University
Maastricht University Medical Center
Max Planck Institute for Psycholinguistics

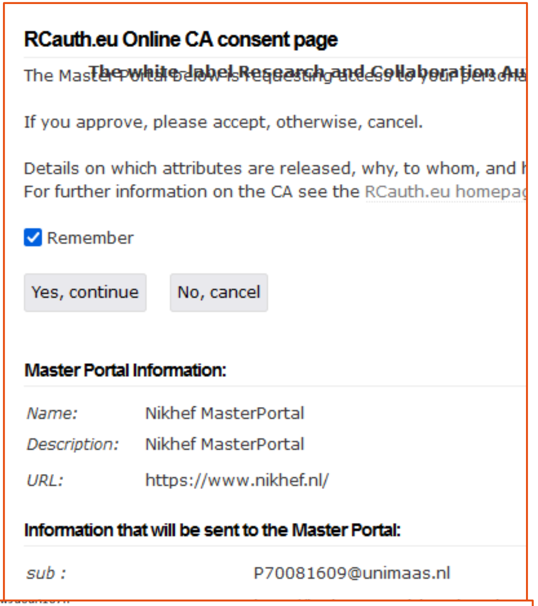
The RCauth WAYF is part of the RCauth CA service. For support, please contact the help desk of your own home organisations or community.

RCauth: SAML to UM, but OIDC for your credential management service

- Approve transfer in OIDC flow & see your PKI X.509 user cert!

- Review the network interactions with
  - engine.surfconext.nl
  - login.maastrichtuniversity.nl
  - pilot-ca1.rcauth.eu
  - elevator.nikhef.nl (this is the credential management service where your long-term private key is)

- What is the difference in the POSTs?
- Can you see the difference in the SAML and the OIDC flow?



RCauth.eu Online CA consent page

The white-label Research and Collaboration Au

The Master Portal below is requesting access to your persona

If you approve, please accept, otherwise, cancel.

Details on which attributes are released, why, to whom, and h
For further information on the CA see the RCauth.eu homepag

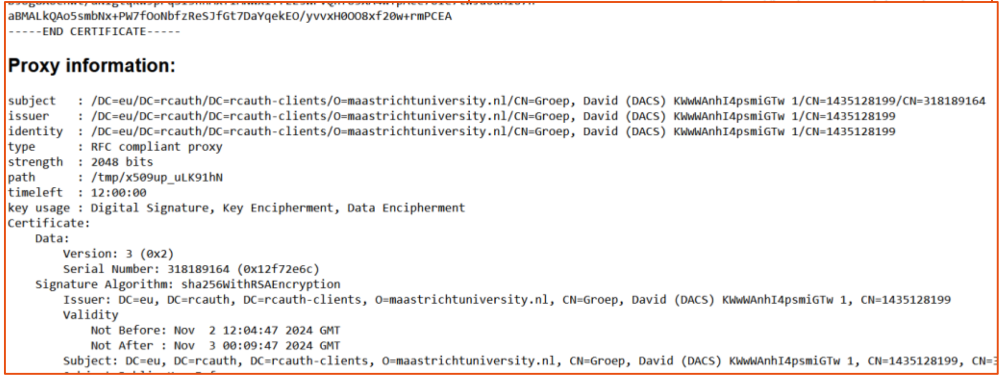
☑ Remember

Master Portal Information:

Name:	Nikhef MasterPortal
Description:	Nikhef MasterPortal
URL:	https://www.nikhef.nl/

Information that will be sent to the Master Portal:

sub :	P70081609@unimaas.nl



```
aBMALkQAo5smbNx+PW7fOoNbfzReSJfGt7DaYqekEO/yvvxH0OO8xf20w+rmPCEA
-----END CERTIFICATE-----
```

Proxy information:

```
subject   : /DC=eu/DC=rcauth/DC=rcauth-clients/O=maastrichtuniversity.nl/CN=Groep, David (DACS) KWwWAnhI4psmiGTw 1/CN=1435128199/CN=318189164
issuer    : /DC=eu/DC=rcauth/DC=rcauth-clients/O=maastrichtuniversity.nl/CN=Groep, David (DACS) KWwWAnhI4psmiGTw 1/CN=1435128199
identity  : /DC=eu/DC=rcauth/DC=rcauth-clients/O=maastrichtuniversity.nl/CN=Groep, David (DACS) KWwWAnhI4psmiGTw 1/CN=1435128199
type      : RFC compliant proxy
strength  : 2048 bits
path      : /tmp/x509up_uLK91hN
timeleft  : 12:00:00
key usage : Digital Signature, Key Encipherment, Data Encipherment
Certificate:
    Data:
        Version: 3 (0x2)
        Serial Number: 318189164 (0x12f72e6c)
    Signature Algorithm: sha256WithRSAEncryption
        Issuer: DC=eu, DC=rcauth, DC=rcauth-clients, O=maastrichtuniversity.nl, CN=Groep, David (DACS) KWwWAnhI4psmiGTw 1, CN=1435128199
        Validity
            Not Before: Nov  2 12:04:47 2024 GMT
            Not After : Nov  3 00:09:47 2024 GMT
        Subject: DC=eu, DC=rcauth, DC=rcauth-clients, O=maastrichtuniversity.nl, CN=Groep, David (DACS) KWwWAnhI4psmiGTw 1, CN=1435128199, CN=3
```

Nulla folia post hoc sunt

Thanks for watching!

“En daarmee, geachte luisteraars, laat ik u over aan
de verpozing die uw babbelklant u gemeenlijk pleegt te bieden.”