Computing for Research & the Worldwide LHC Computing Grid

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

Maastricht University | Department of Advanced Computing Sciences

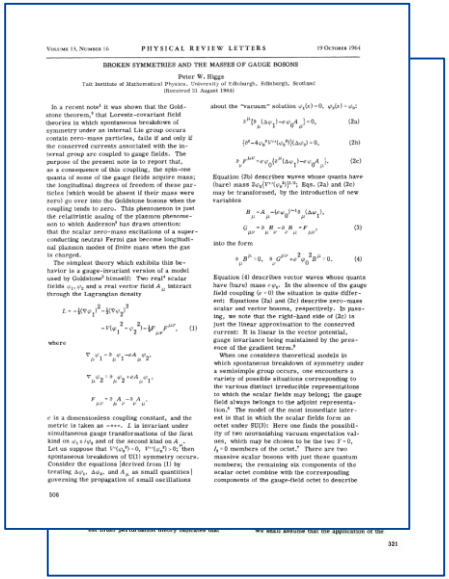
Nikhef

David Groep
DACS & Nikhef
7 November 2023
KEN3239 rev 1.1 RC2

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 plasma 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

$$L = -\tfrac{1}{2}(\nabla\varphi_1)^2 - \tfrac{1}{2}(\nabla\varphi_2)^2 - V(\varphi_1^2+\varphi_2^2) - \tfrac{1}{2}(\nabla\wedge A)^2 + \tfrac{1}{2}e^2\varphi_1^2 A_\mu A^\mu + \ldots \quad (1)$$

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 φ_1, φ_2, and of the second kind on A_μ. Let us suppose that $V(\varphi_1^2) > 0$, $V'(\varphi_1^2) > 0$, ^{5} then spontaneous breakdown of $U(1)$ symmetry occurs. Consider the equations derived from (1) by treating $\delta\varphi_1$, $\delta\varphi_2$, and δA_μ as small quantities governing the propagation of small oscillations about the "vacuum" solution $\varphi_1(x) = 0$, $\varphi_2(x) = \varphi_0$:

$$\partial^\mu[\partial_\mu\delta\varphi_1 + eA_\mu\varphi_0] = 0, \quad (2a)$$

$$(\partial^\mu - e^2\varphi_0^2)(\partial_\mu A_\nu - \partial_\nu A_\mu) = 0, \quad (2b)$$

$$\partial_\mu[\partial^\mu A_\nu - \partial_\nu A_\mu] - e^2\varphi_0 A_\nu. \quad (2c)$$

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

$$B_\mu = A_\mu - (e\varphi_0)^{-1}\partial_\mu(\delta\varphi_2),$$

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

into the form

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

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.[6]

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.[7] 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 $T=0$, $I_3=0$ 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

about the "vacuum" solution $\varphi_1(x) = 0$, $\varphi_2(x) = \varphi_0$:

$$\partial^\mu[\partial_\mu\delta\varphi_1 + eA_\mu\varphi_0] = 0, \quad (2a)$$

$$(\partial^\mu - e^2\varphi_0^2)(\partial_\mu A_\nu - \partial_\nu A_\mu) = 0, \quad (2b)$$

$$\partial_\mu[\partial^\mu A_\nu - \partial_\nu A_\mu] - e^2\varphi_0 A_\nu. \quad (2c)$$

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

$$B_\mu = A_\mu - (e\varphi_0)^{-1}\partial_\mu(\delta\varphi_2),$$

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

into the form

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

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.[6]

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.[7] 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 $T=0$, $I_3=0$ 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

321

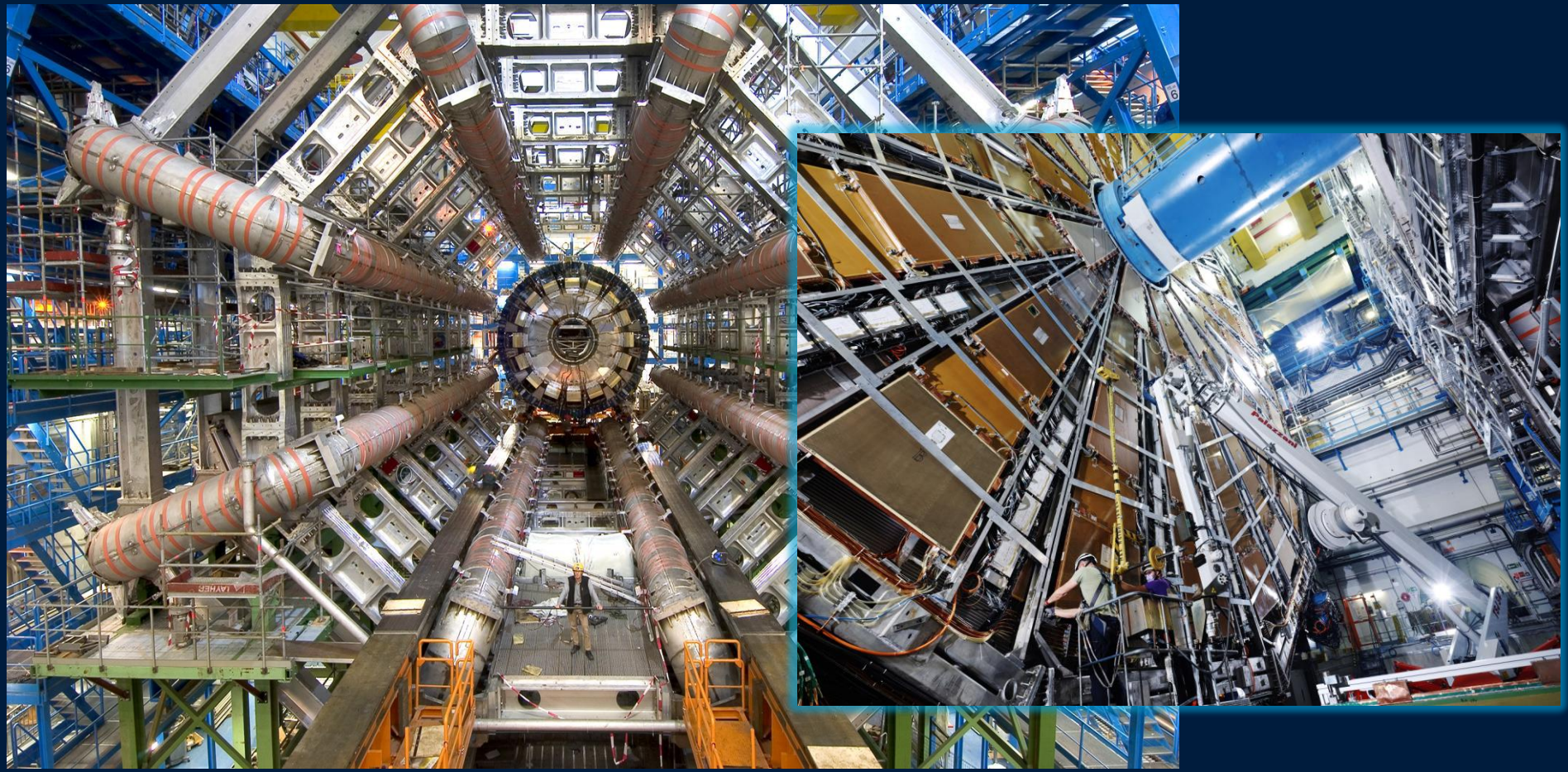
P. Higgs, Phys. Rev. Lett. 13, 508:

16823 characters, 165 kByte PDF

1998 - 2012 … 2028: HL-LHC … 2035+

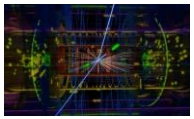
~50 PiB/year primary data

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 – 40M events/sec



ATLAS RAW single event
ROD File
1.60 MB

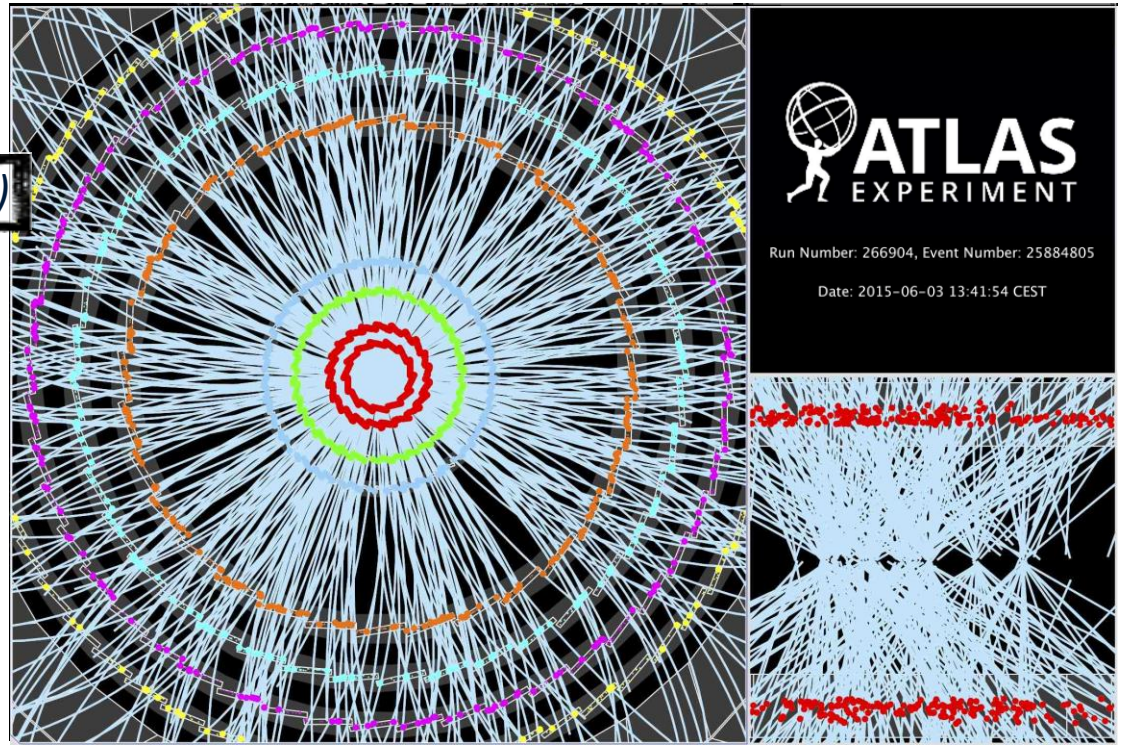
***~60 TByte/s** (compressed)*

Trigger system selects 600 Hz ~ 1 GB/s data

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

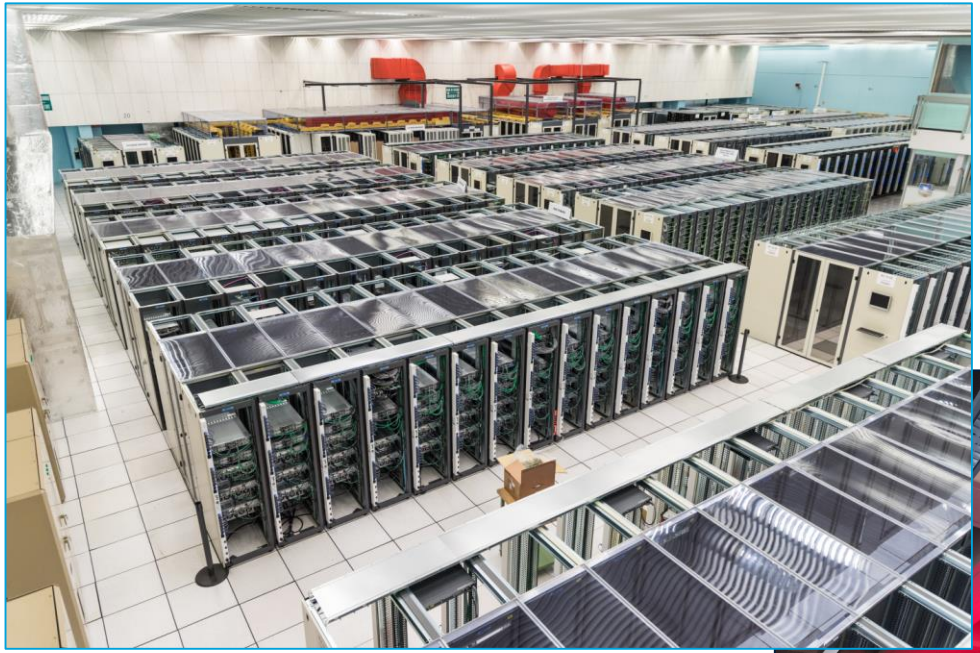
~10k researchers

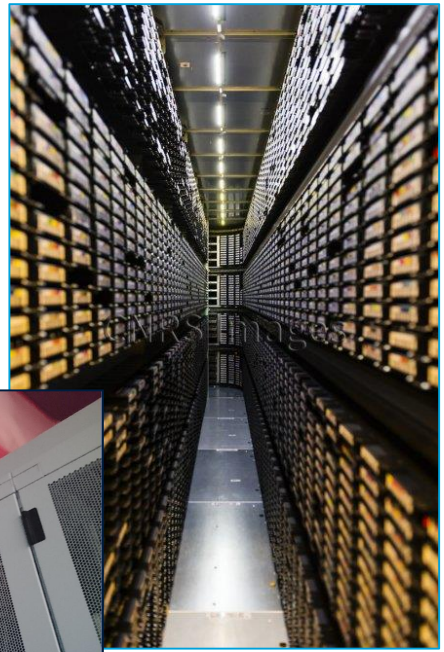
CERN and 170 institutes

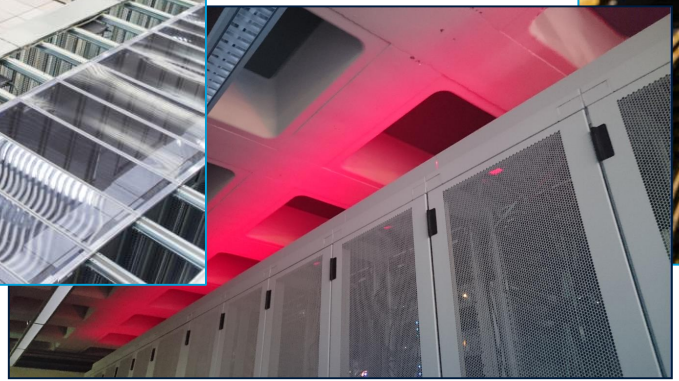


Display of a proton-proton collision event recorded by ATLAS on 3 June 2015, with the first LHC stable beams at a collision energy of 13 TeV; Event processing time: v19.0.1.1 as per Jovan Mitrevski and 2015 J. Phys.: Conf. Ser. 664 072034 (CHEP2015)

‘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

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 …*

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 Computing**, herding systems, cloud, and containers
- **distributed computing**, scalable storage and distributed data placement

Networking the systems: linking 'more than one' globally

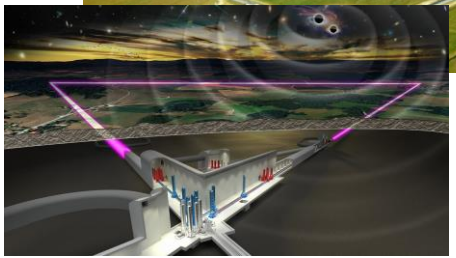
- **network design**: elephants vs. mice in shipping large quantities of data … and on cat videos
- *Optical Private Networks* and the *Open Networking Environment LHCone*

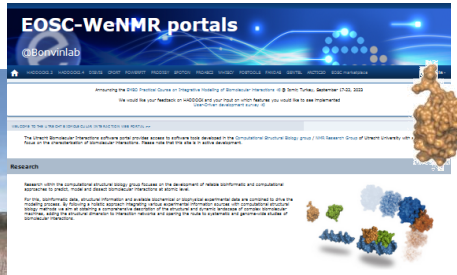
Networking the people

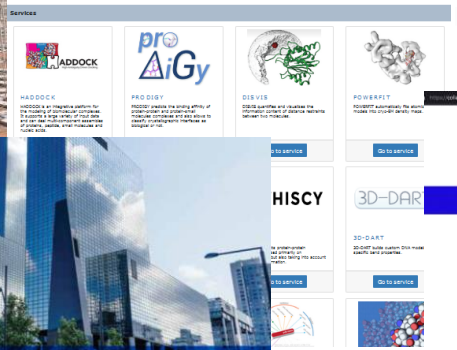
- **authentication and authorization** technologies
- **multilateral federation:** identity, community management & global trust

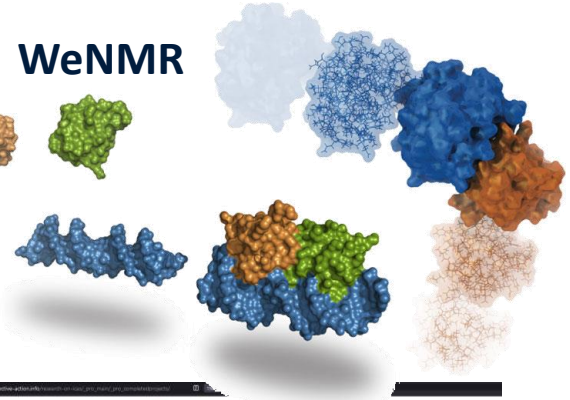
Putting it all together again *(and maybe an example of a federated anycasted authentication service)*

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



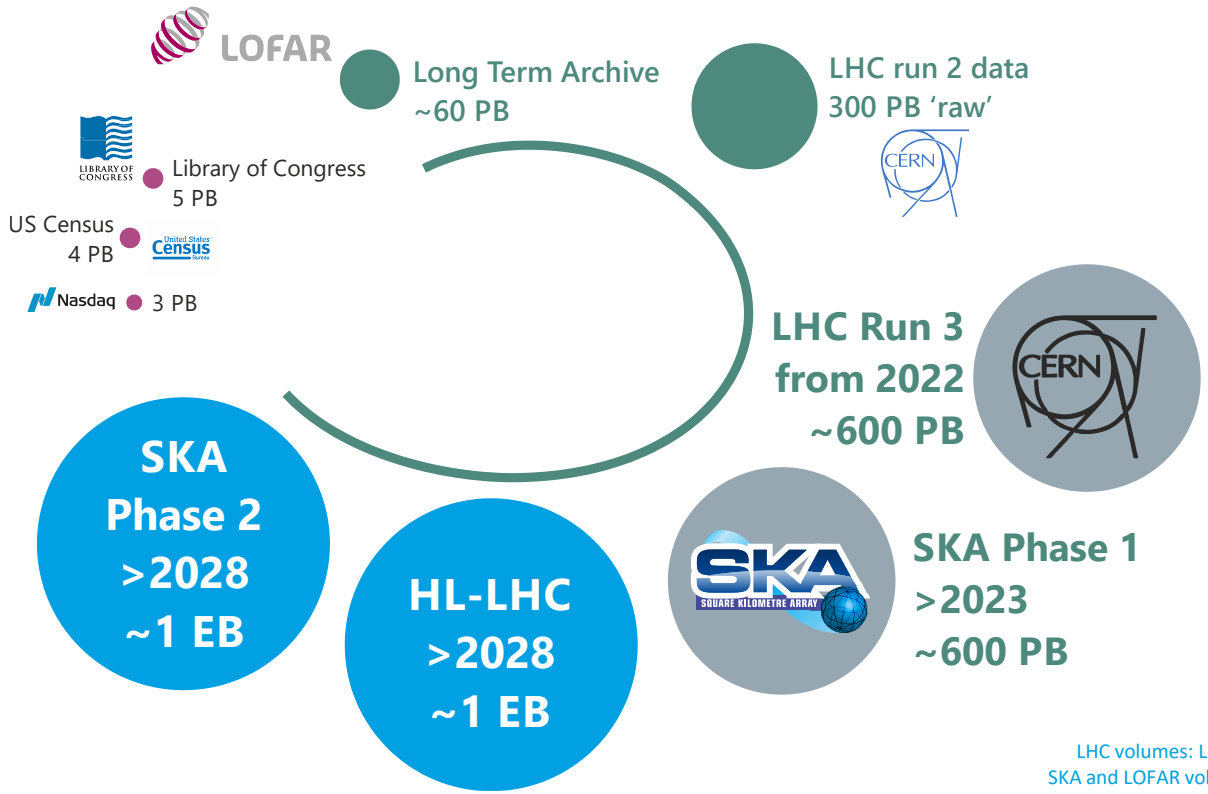






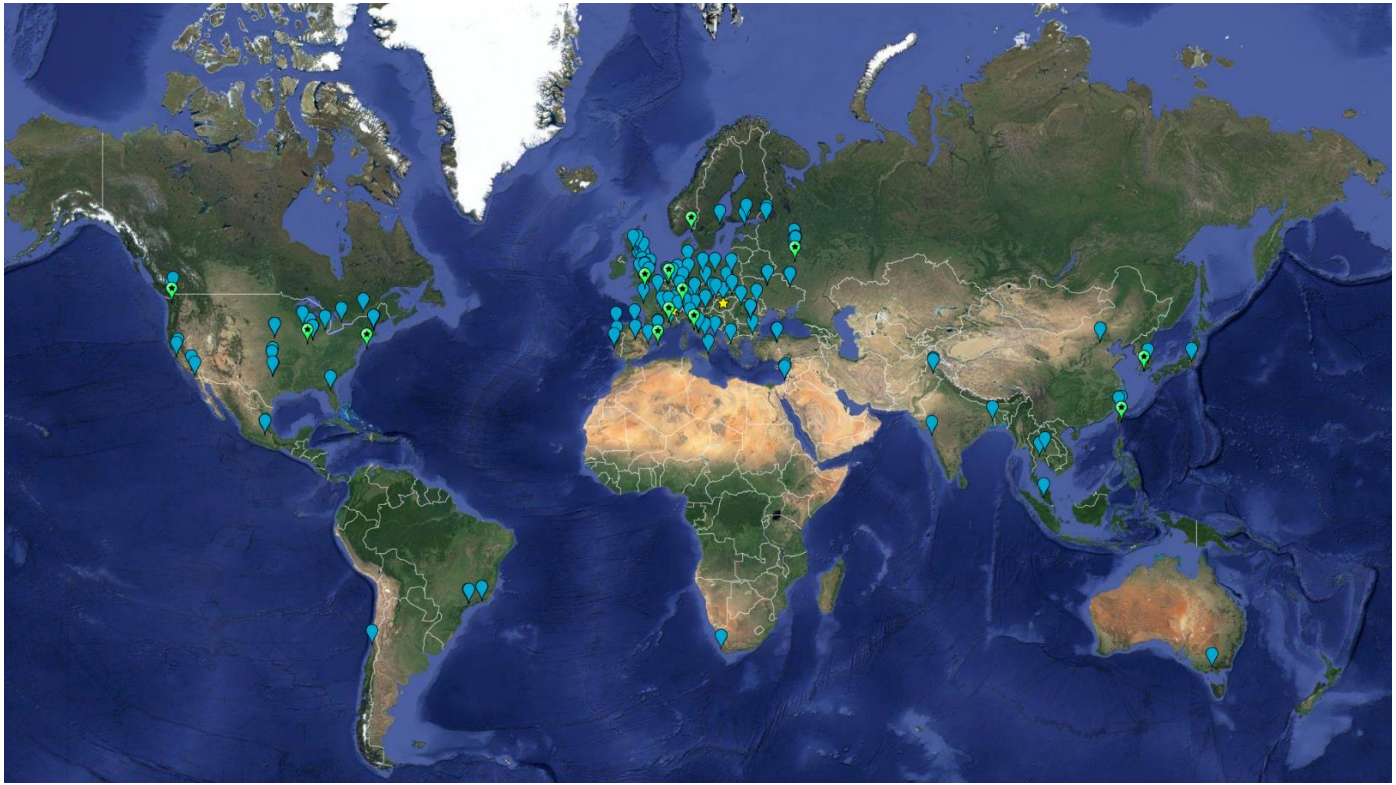
Sources: CERN https://wlcg.web.cern.ch/; HADDOCK, WeNMR, @Bonvinlab https://wenmr.science.uu.nl/; Virgo, Pisa, IT; SKAO: the SKA-Low observatory, Australia https://www.skatelescope.org/ - OpenMOLE simulation on EGI - https://cdn.egi.eu/app/uploads/2022/04/EGI_Use_Cases.pdf; agent-based modelling of ICAs: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

Processing at scale for data intensive science



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

Not in one place: the worldwide LHC 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

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

One of these nodes: the Dutch National e-Infrastructure

- Joint SURF & Nikhef collective service – part of EGI, WLCG and FuSE
- hosts WLCG, but also LOFAR radio telescope data, and ~100 other projects
- 59 PByte near-line storage (tape), 42.5 PByte on-line (disk), 27.6 k cores (cpu)

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/

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:
  - 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 helped
  - AMD EPYC Genoa (Zen 4) has 96 cores/die
  - Intel Sapphire Rapids …
  - 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

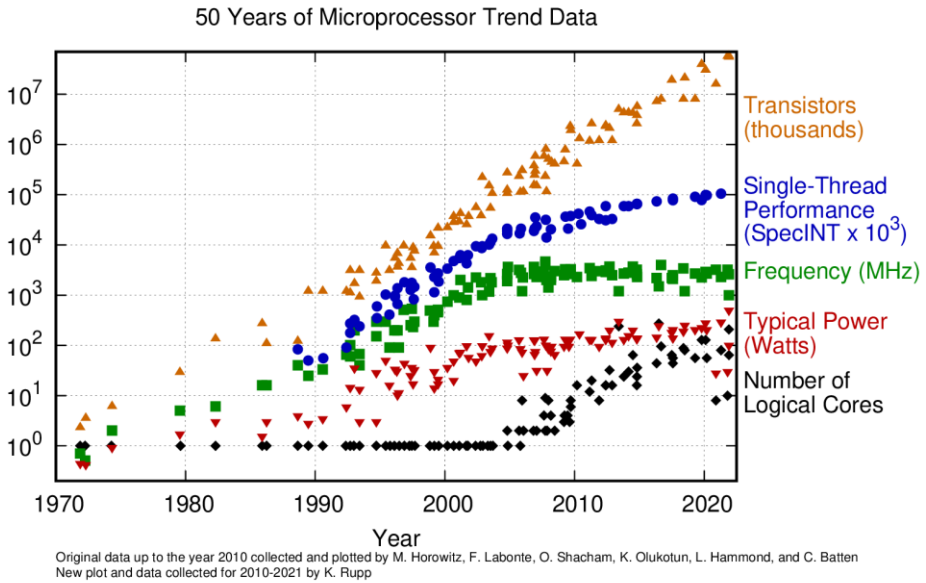
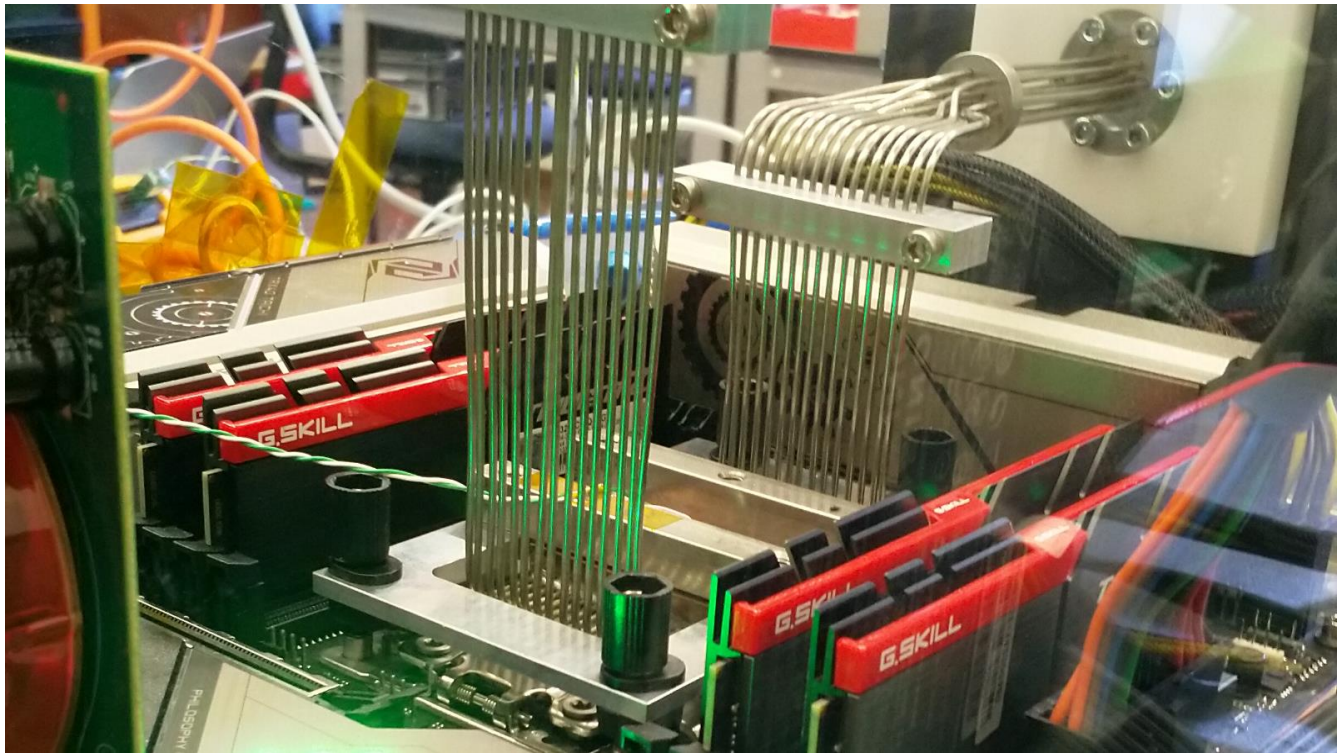


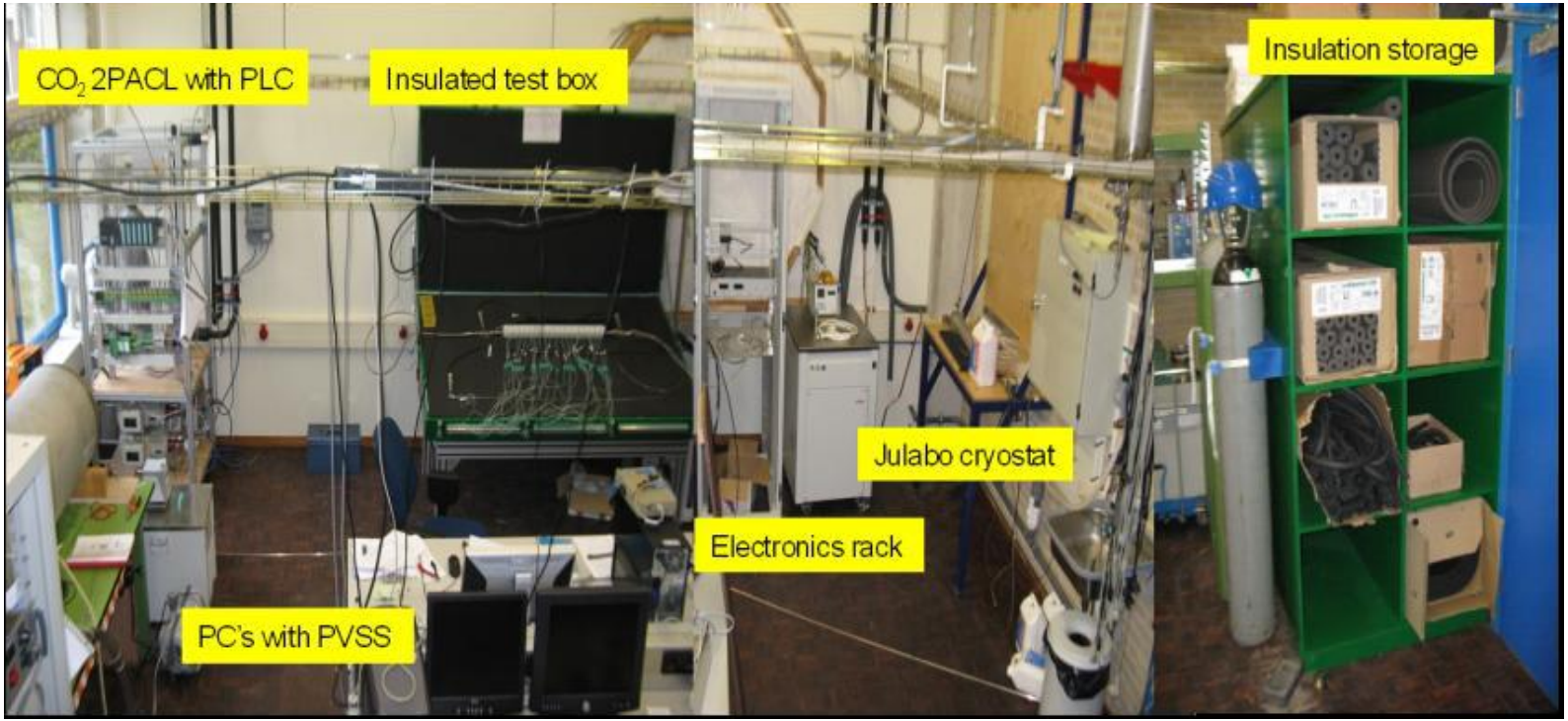
Image: K Rupp, https://github.com/karlrupp/microprocessor-trend-data

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 …

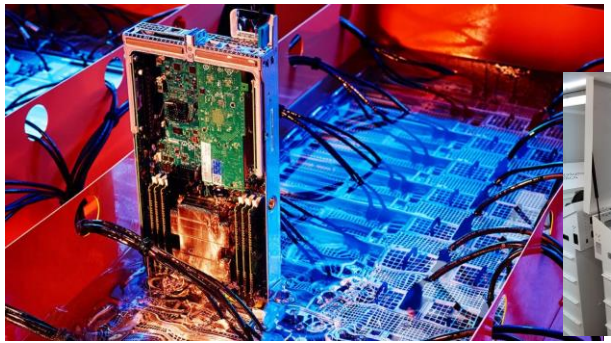


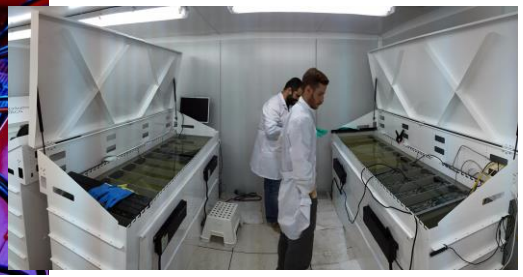
Nikhef 2PA LCO2 cooling setup. Image from Bart Verlaat, Auke-Pieter Colijn *CO2 Cooling Developments for HEP Detectors* https://doi.org/10.22323/1.095.0031

Getting the 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

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)

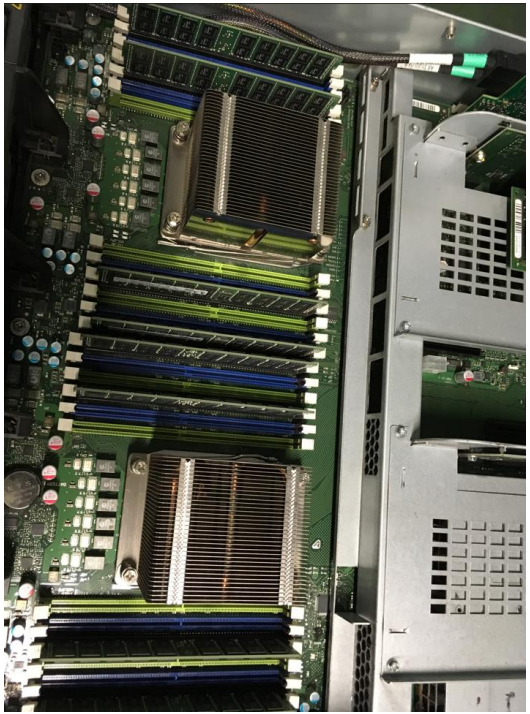


Image: dual-socket Fujitsu system at the Xenon experiment site, 2019. source: Tristan Suerink, Nikhef

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

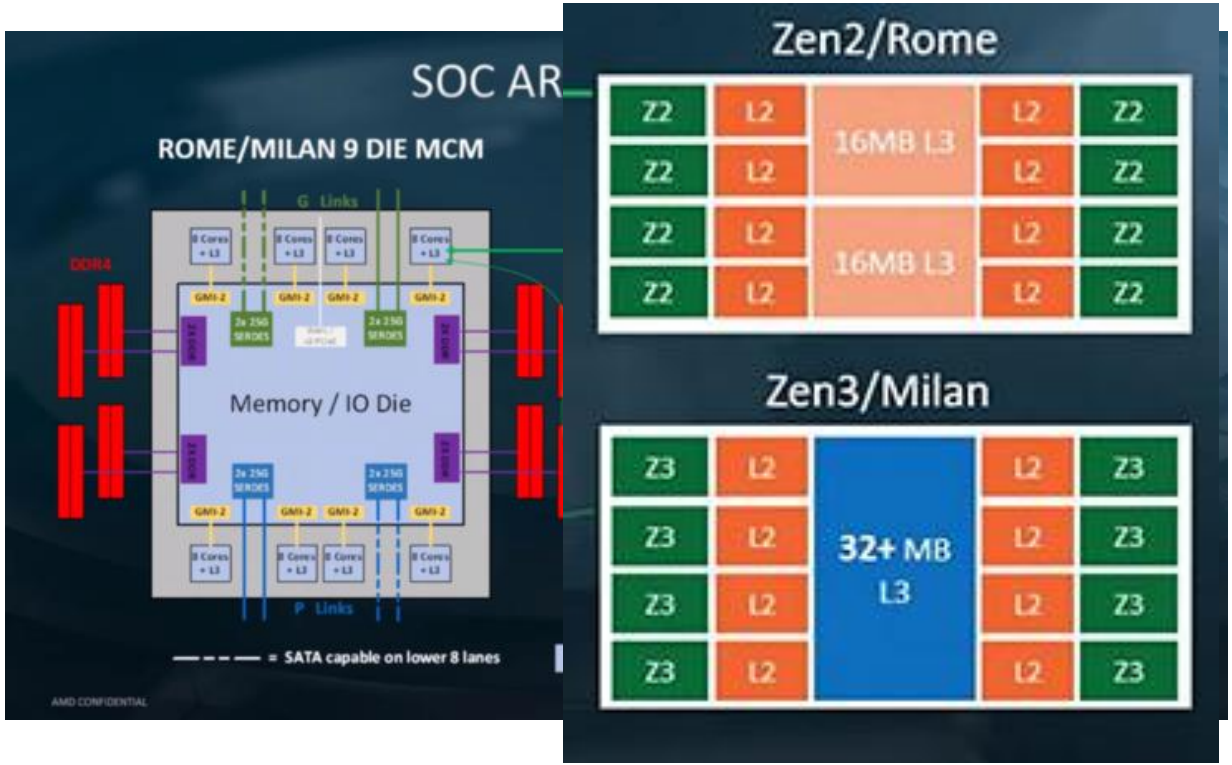
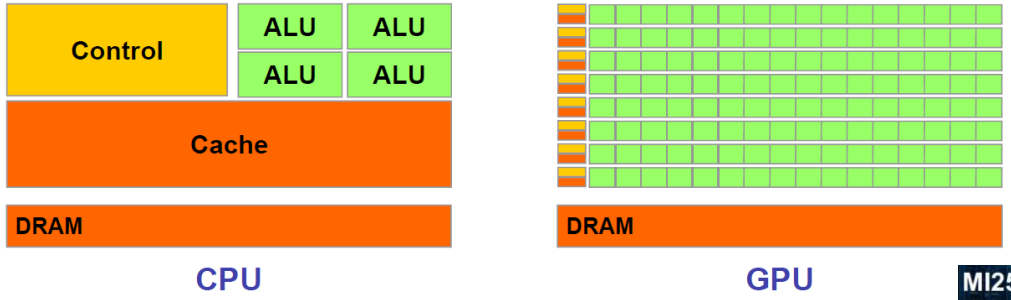


Image source: AMD, retrieved from https://m.hexus.net/tech/news/cpu/135479-amd-shares-details-zen-3-zen-4-architectures/

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!

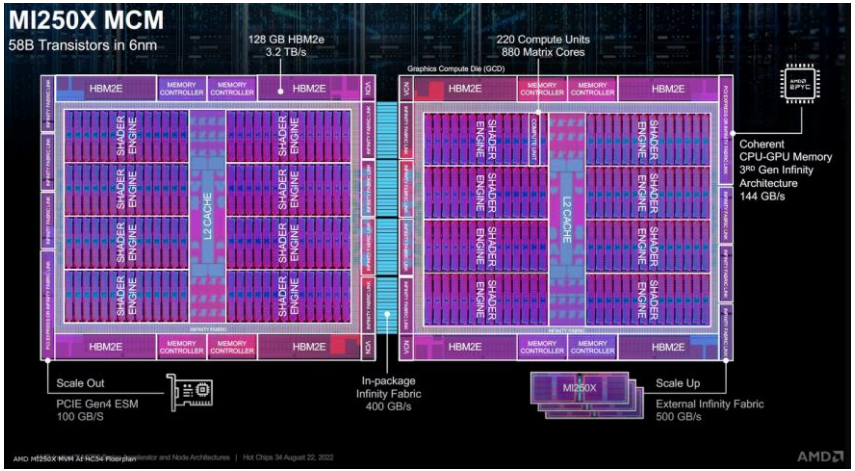
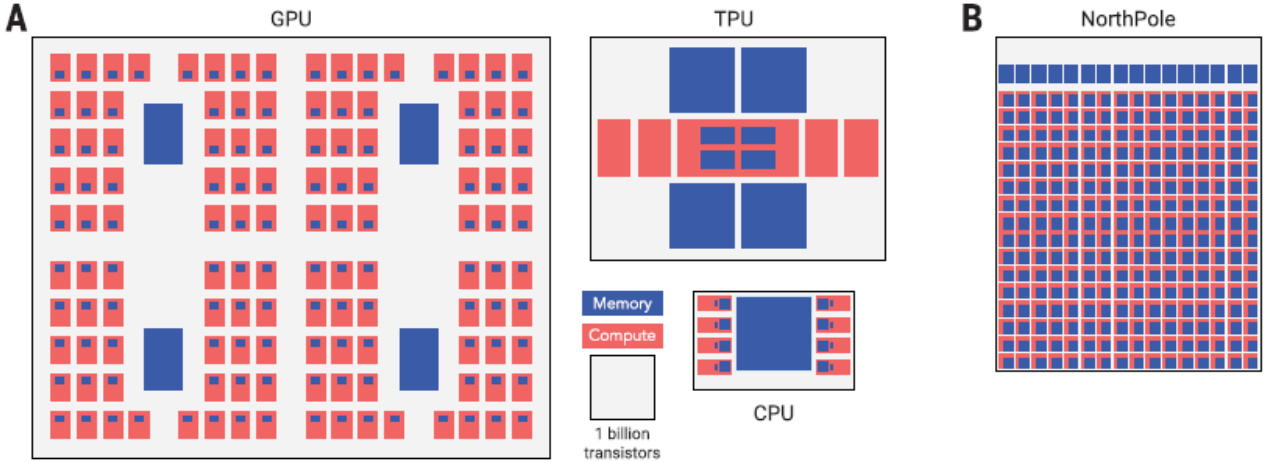


Image: 'Massively Parallel Computing with CUDA', Antonino Tumeo Politecnico di Milano, https://www.ogf.org/OGF25/materials/1605/CUDA_Programming.pdf
Floorplan image of die: AMD MI250 GPU, slide source: AMD

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/

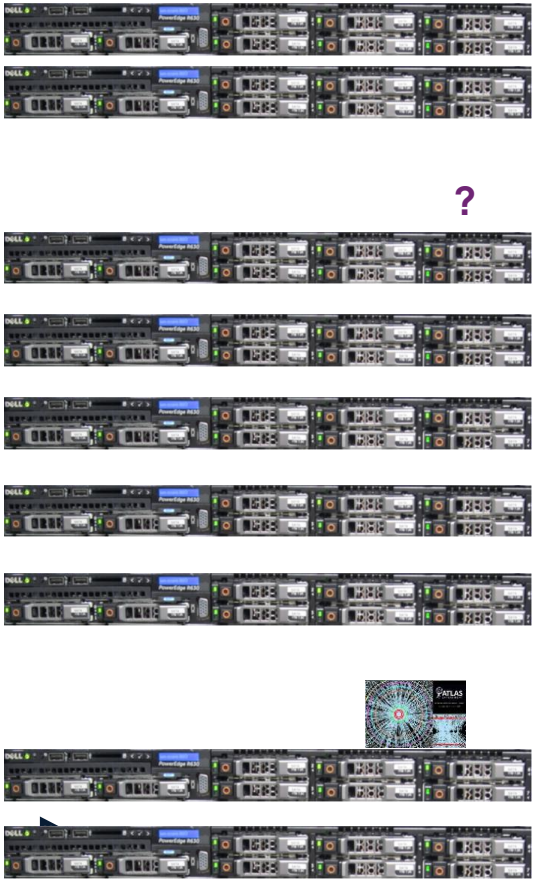
If large-scale IT does not quite fit … ahum …

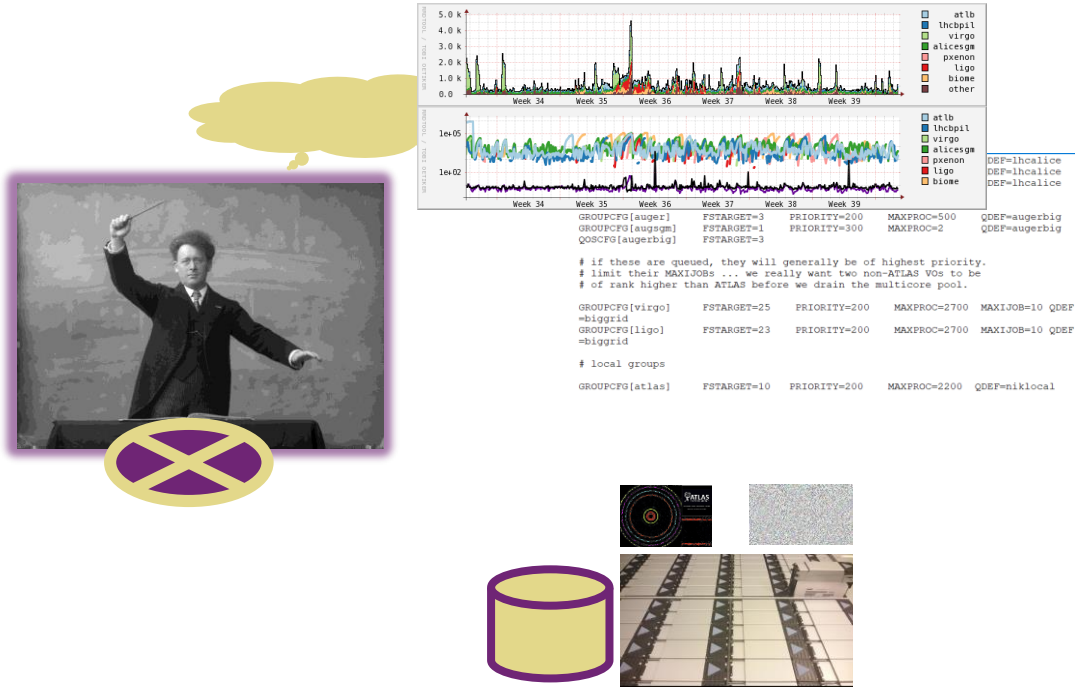
Image source: https://lambdalabs.com/products/blade

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

Scaling up – beyond one lone motherboard

Cluster computing and ‘conveniently parallel’ HTC



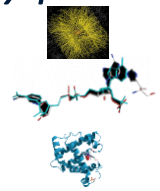


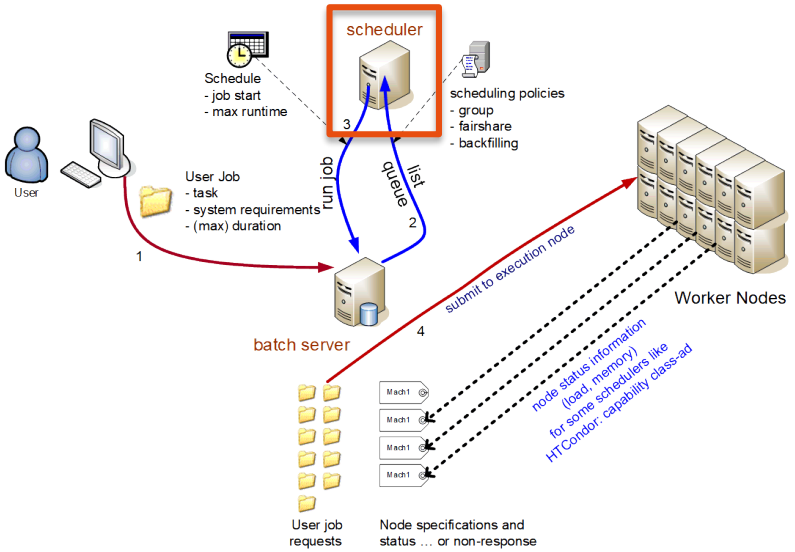
- ‘like milking cows’ (if you feed them lots of power first)
- parallel access to data comes at a cost of high IOPS

Batch system platform

Many things are *conveniently parallel*

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





challenge not in parallelism itself

- we have had HPC systems for ages

but

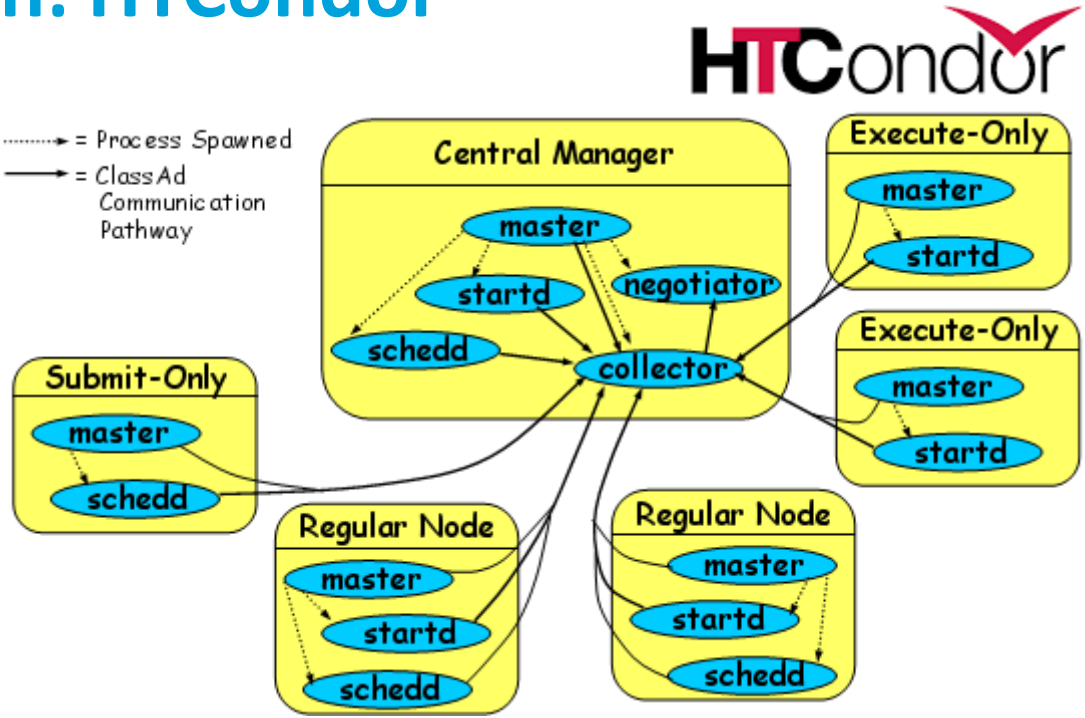
- large numbers of single-core jobs
- heterogeneous workloads
  sharing the same set of worker nodes
- computing with concurrent data access

```
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, UWMadison; https://research.cs.wisc.edu/htcondor/CondorWeek2008/condor_presentations/desmet_admin_tutorial/

Physical farms: selecting the ‘worker nodes’

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

- **balanced features** for node throughput: CPU, storage, memory bandwidth & latency, NIC & network speed
- **single-socket** multicore systems are fine, typical: 64-128 cores per system
- **network**: 2x25Gbps
  (+ ‘out of band’ management like IPMI)
- **memory**: 8 GiB/core
- **local disk**: 4TB NVME PCIe Gen4 x4

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

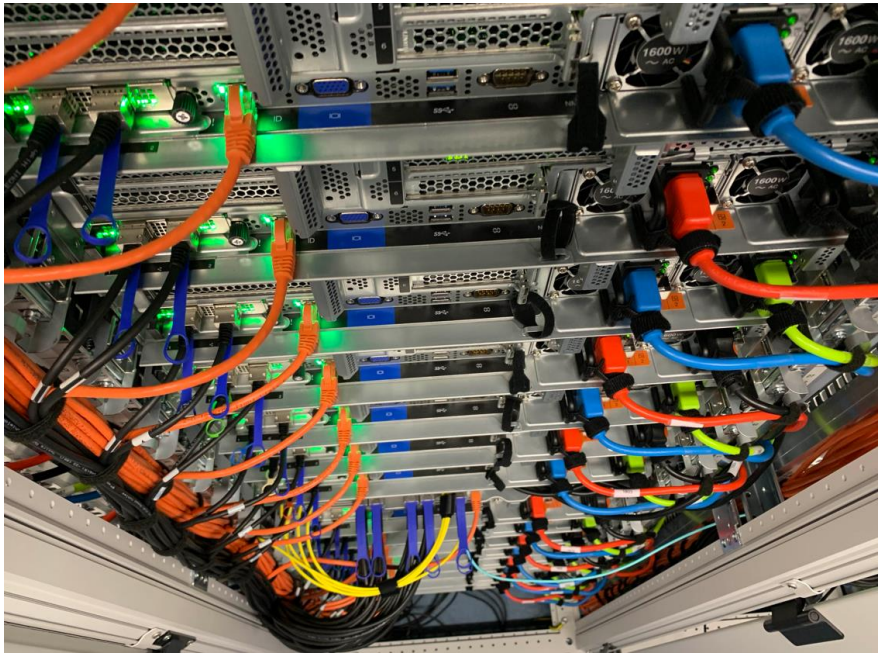
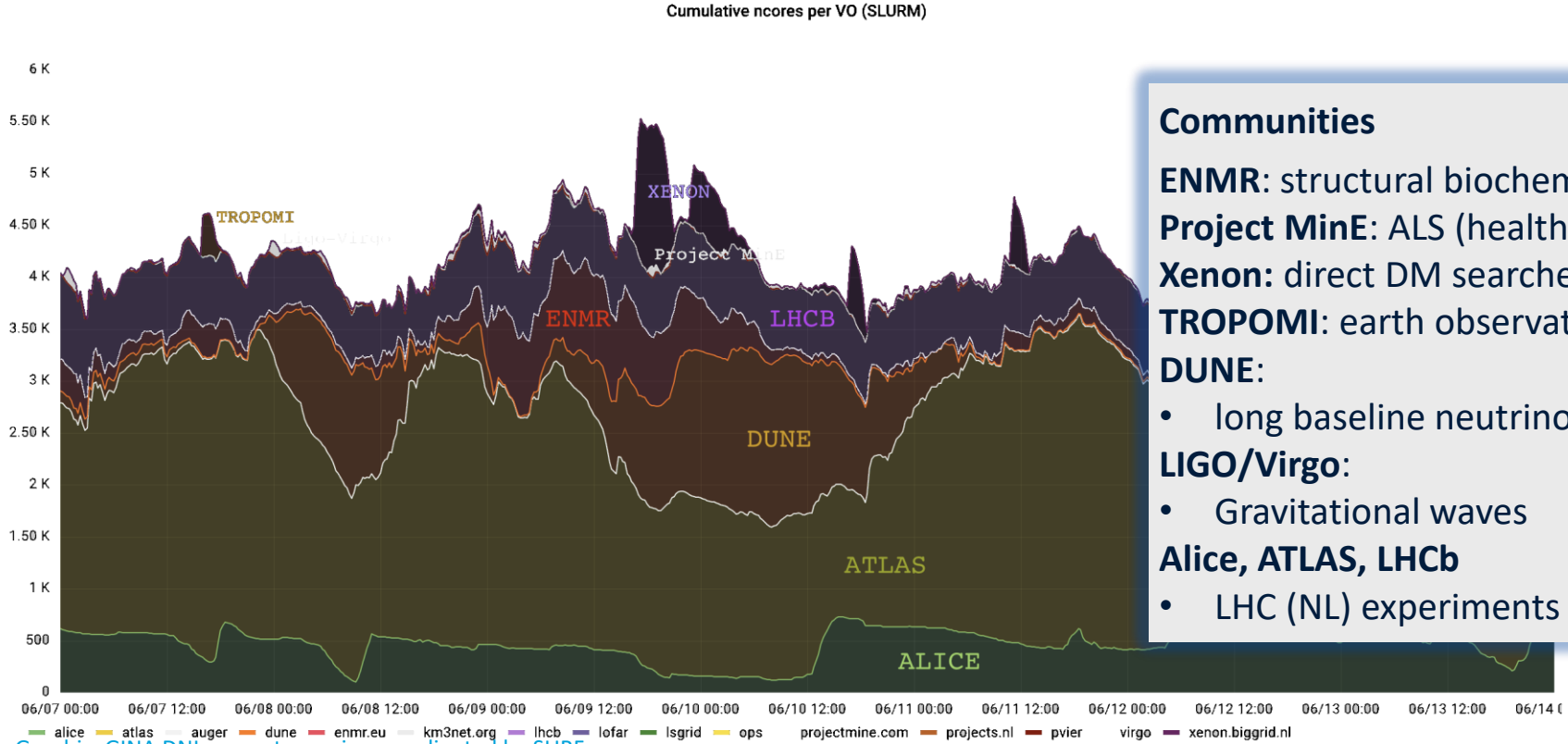


Image: Cluster ‘Lotenfeest’ at the Nikhef NDPF, acquired March 2020. Lenovo SR655 with AMD EPYC 7702P 64-Core single-socket

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

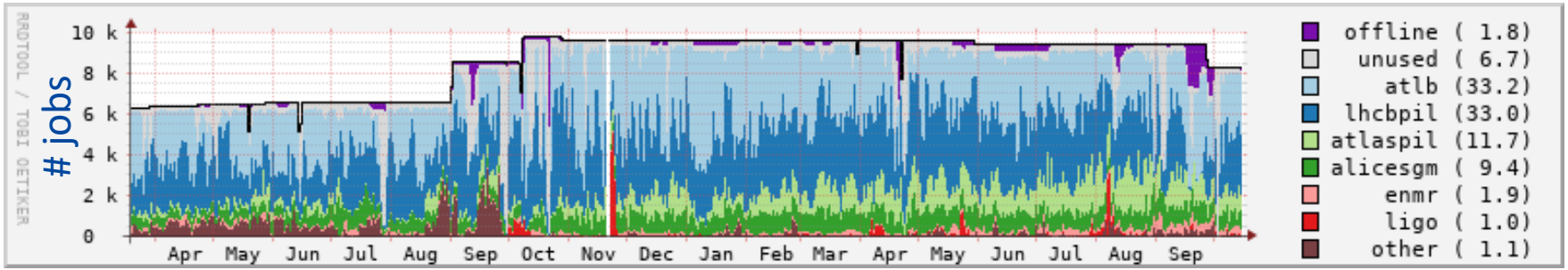
- LHC (NL) experiments

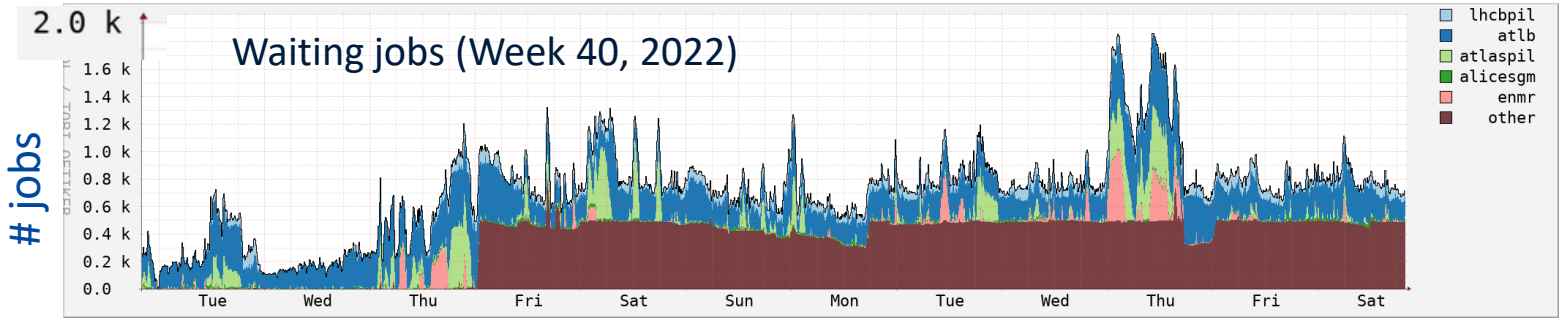
Graphic: GINA DNI compute service coordinated by SURF

NDPF ‘WLCG and Dutch National Infra’ cluster

Running jobs:

period: March 2021 .. October 2022





drainage event on Sept 27 are nodes being moved to the LIGO-VIRGO specific cluster; Source: NDPF Statistics overview, https://www.nikhef.nl/pdp/doc/stats/
‘other’ waiting jobs are almost all for the Auger experiment - GRISview images: Jeff Templon for NDPF and STBC

Estimated Response Time (and predicting it)

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

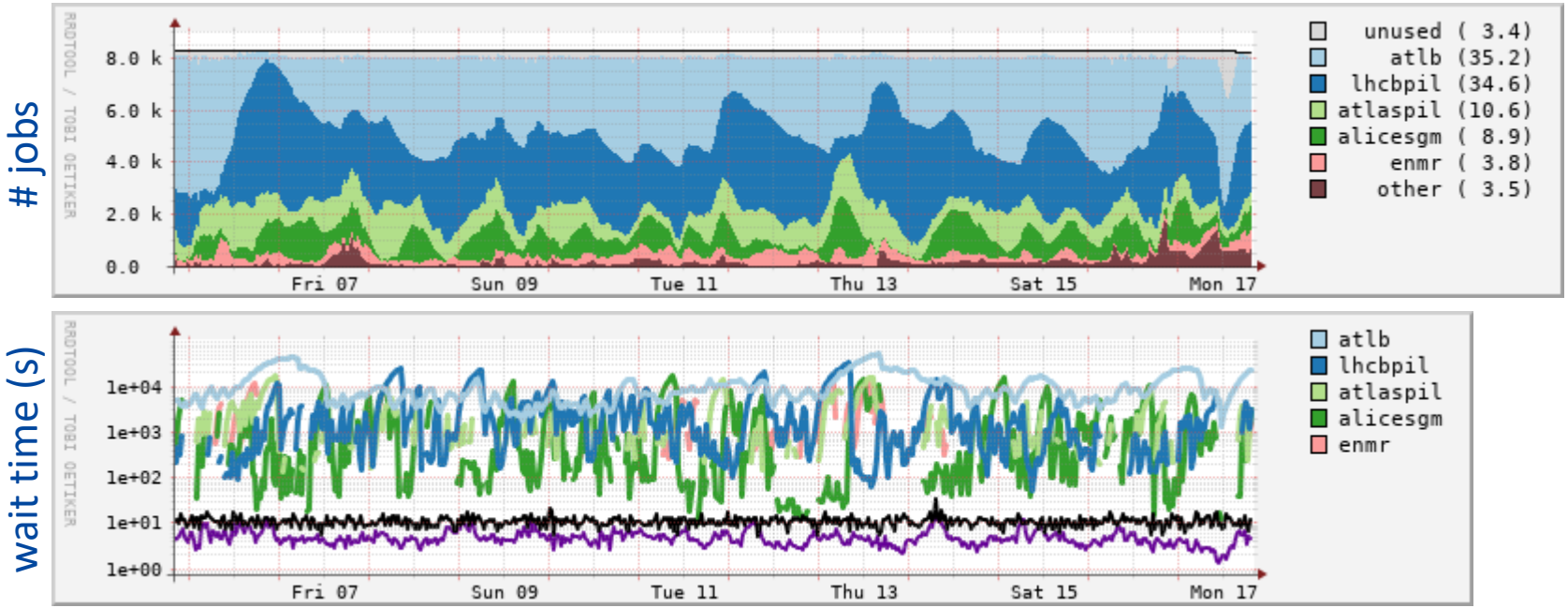


Image: Nikhef NDPF DNI “Grid” cluster. Period: October 6-17, 2022; top-5 communities; GRISview images: Jeff Templon

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*

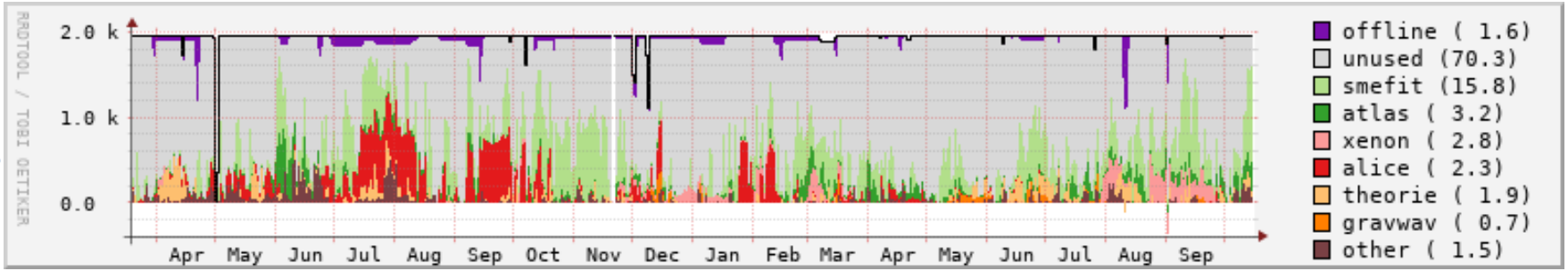
- ***response time*** is more important than efficiency
- fast turn-around/short waiting times by heterogeneous (‘competing’) user base
- data access must be parallelism-ready, but is ‘always’ local on-site
- local storage credentials and sharing with desktop and Jupyter environments

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

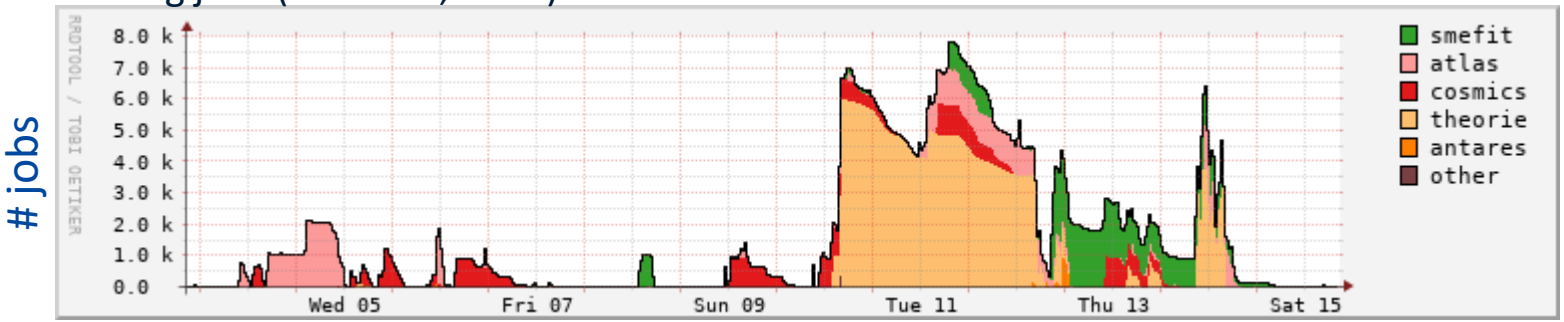
NDPF local analysis cluster ‘Stoomboot’

period: March 2021 .. October 2022

Running jobs:



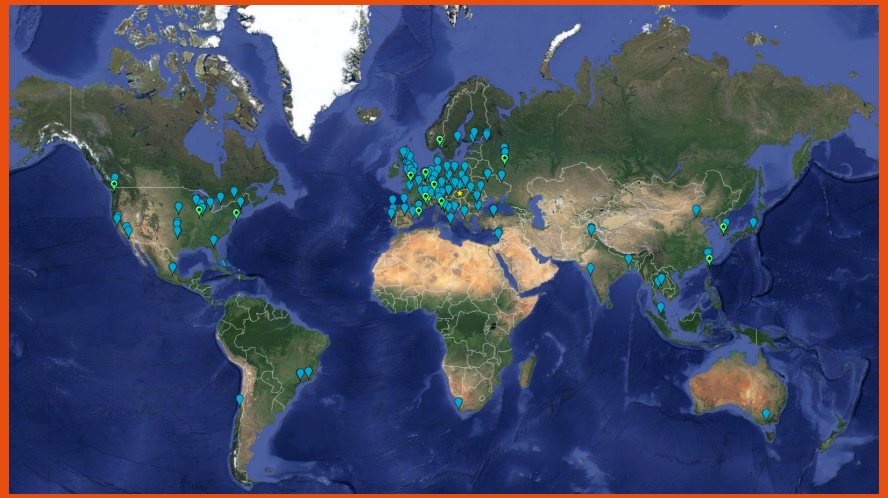
Waiting jobs (Week 40, 2022):



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

More of *more than one* …

The physical layer … and managing software-defined infrastructure



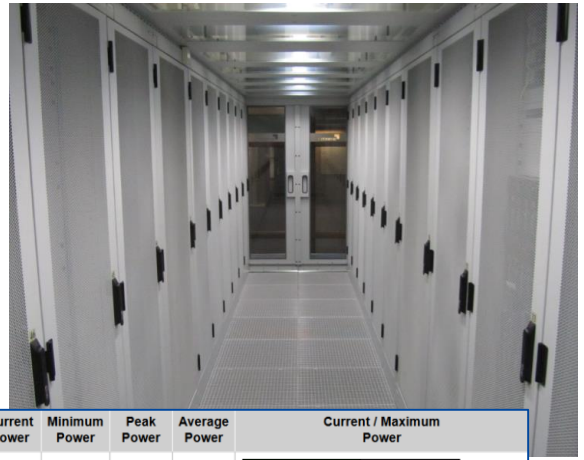
Large-scale IT: the worldwide LHC Computing Grid and beyond

Fancy an interactive console install?

Images: Nikhef Housing H234b NDPF science processing data centre

Where to put them: a brief look at data centres

- ‘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
  - Nikhef ~2 MW,
  - Meta Zeewolde (now cancelled) would have been 160 MW
- data centre efficiency metric: $PUE = \frac{E_{total}}{E_{IT\_equipment}}$



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

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

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, …

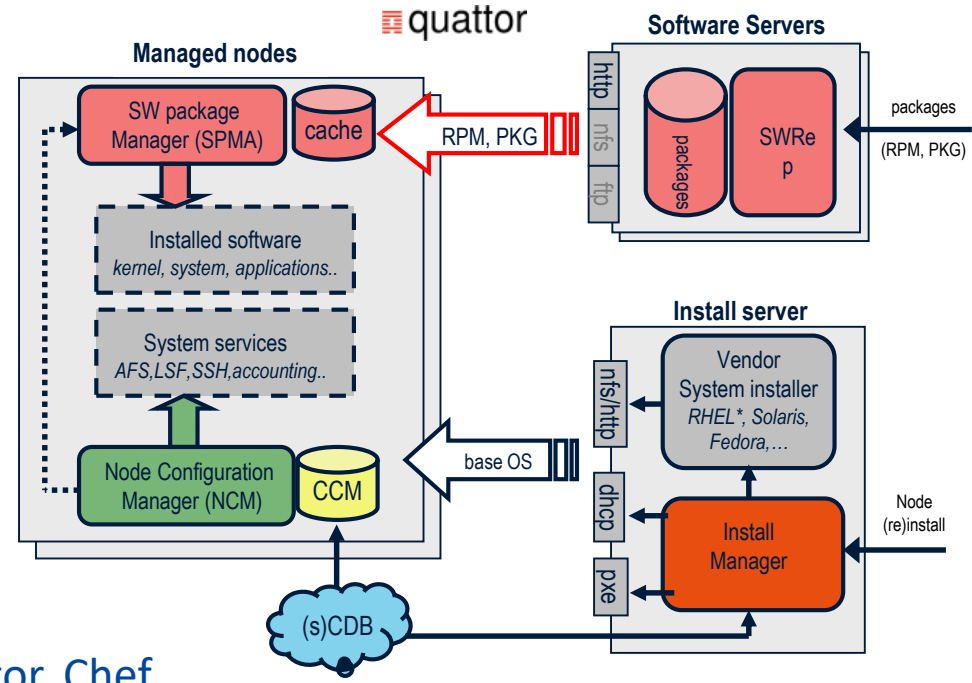


Illustration: German Cancio, CERN, quattor.org, used here as example; see also: ansible.com, saltproject.io, theforeman.org, cfengine.com, puppet.com, …

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 WeNMR portal)

driven from a configuration management DB

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*!

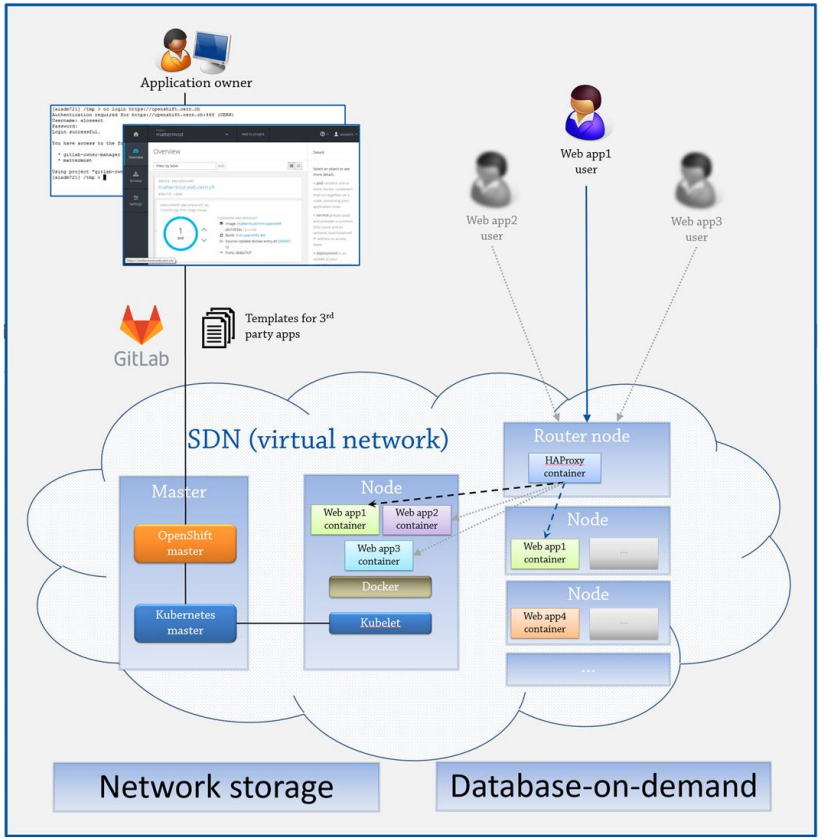
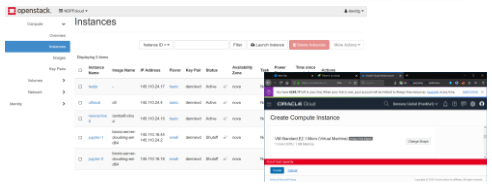
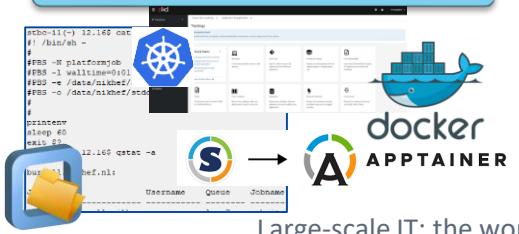


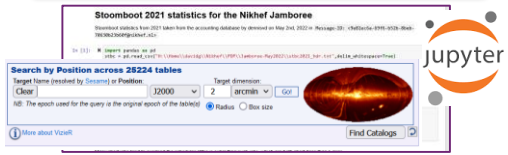
Image from CERN's OpenShift, A Lossent *et al* 2017 *J. Phys.: Conf. Ser.* **898** 082037 https://doi.org/10.1088/1742-6596/898/8/082037

Moving the management boundary

	Infrastructure-as-a-Service	Platform-as-a-Service	Software-as-a-Service
Guest	Application	Application	Application
	Data	Data	Data
	Runtime environment	Runtime environment	Runtime environment
	Middleware	Middleware	Middleware
Hyper visor	Operating system	Operating system	Operating system
	Virtualisation layer	Virtualisation layer	Virtualisation layer
Host	Physical server	Physical server	Physical server
	Storage devices	Storage devices	Storage devices
	Network	Network	Network





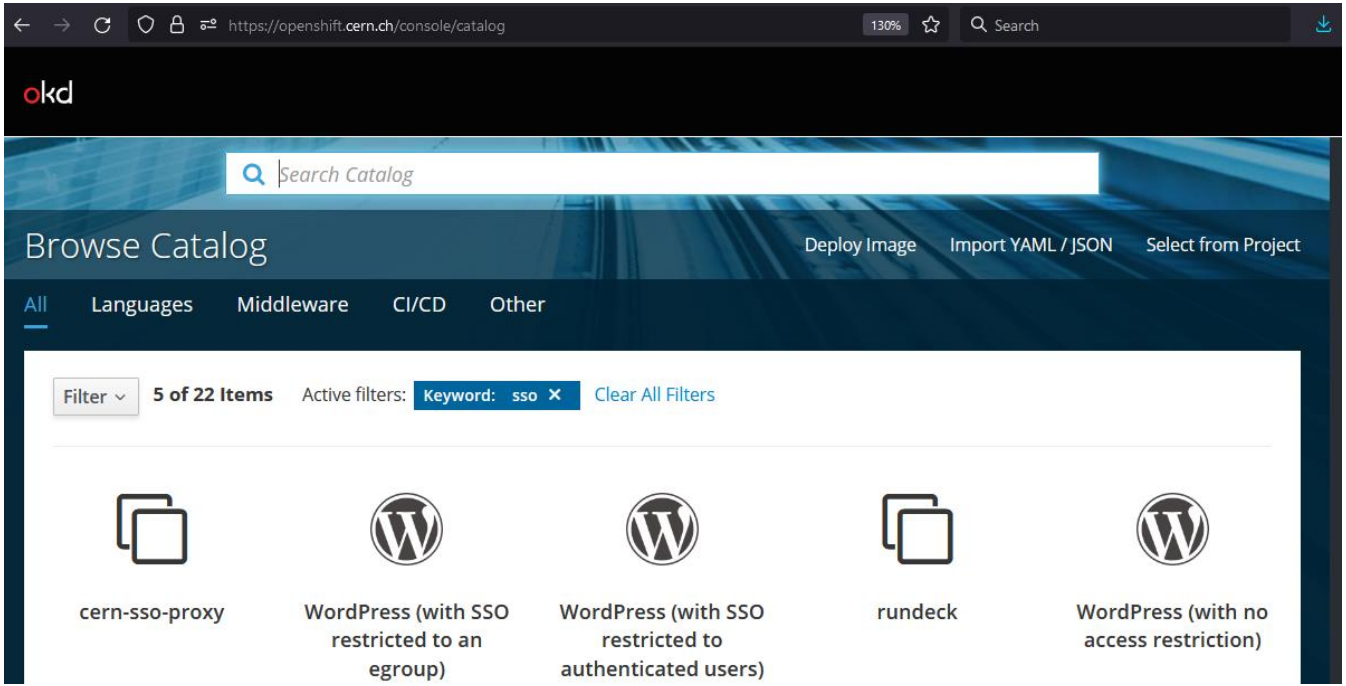


Astronomy catalogue: https://vizier.cds.unistra.fr/

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

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

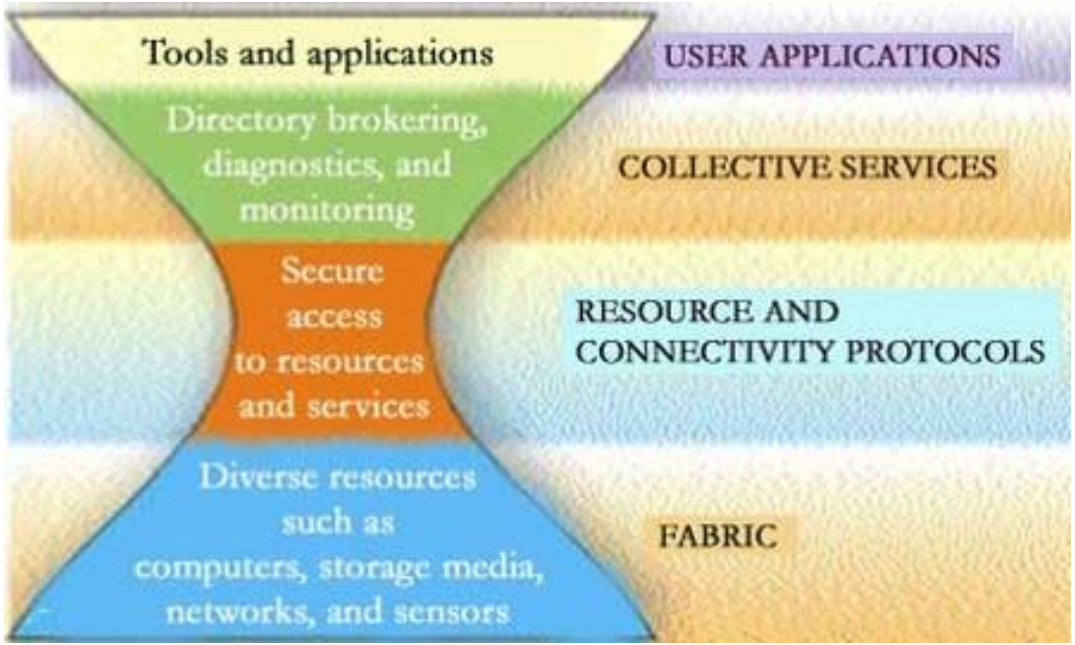
‘Cloudification’ eases systems management …



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’



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

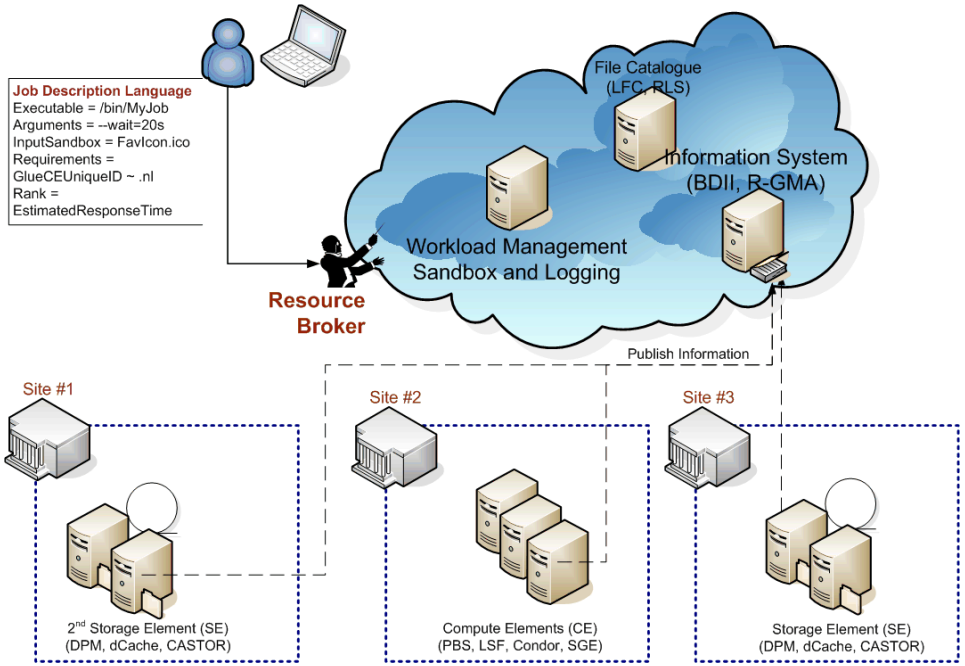
- a very simple stateless interface

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 …?



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

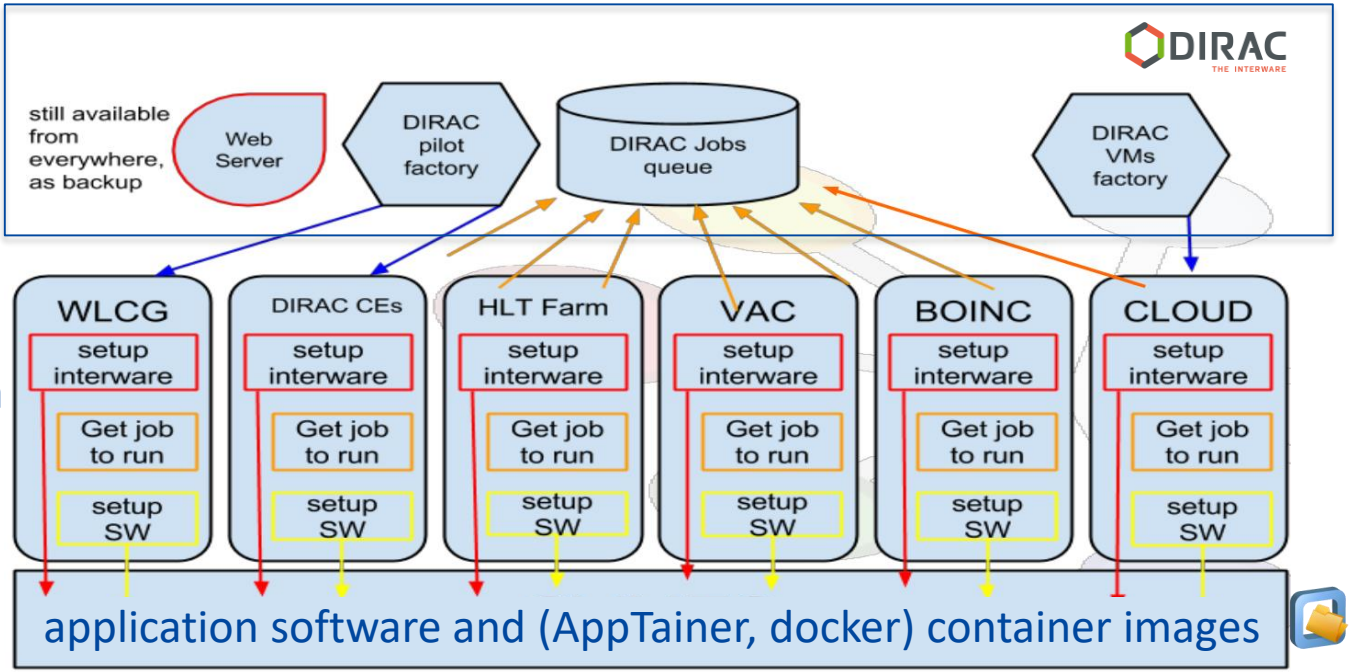
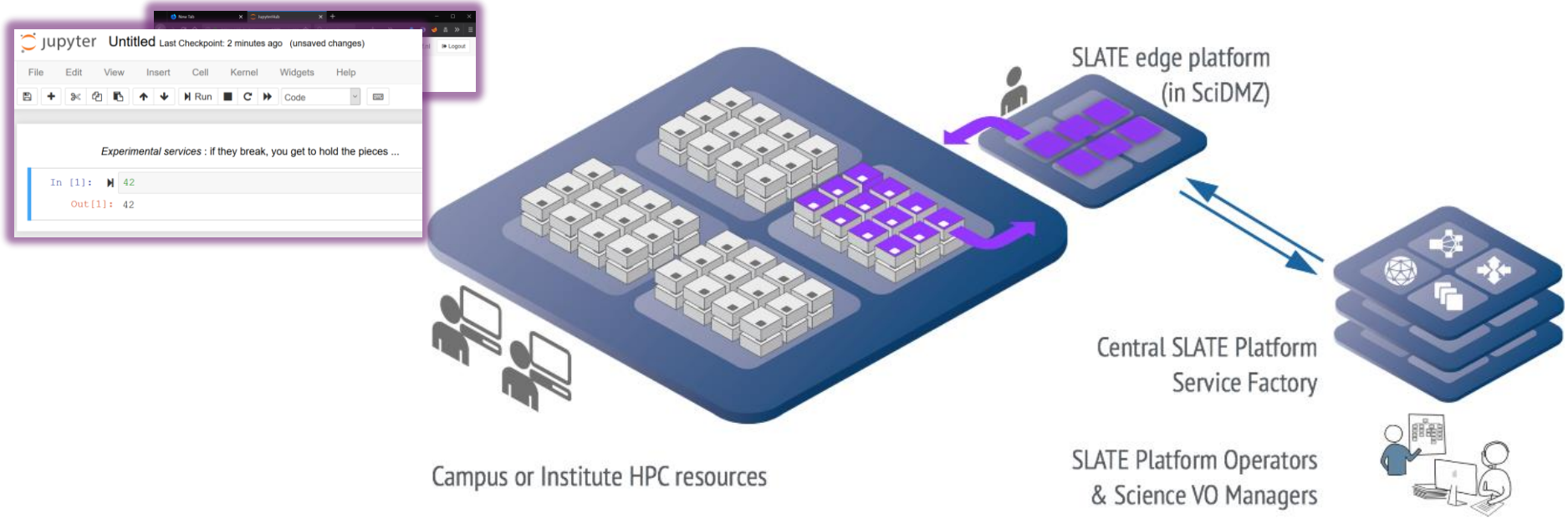


Image: DIRAC project, A. Tsaregorodtsev *et al.* CPPM Marseille, from https://dirac.readthedocs.io/ ; CVMFS (CERN VM File System) is a common software distribution platform using distributed signed data objects in a cached hierarchy using CDN techniques, see https://cernvm.cern.ch/fs/

An overlay network of containers

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

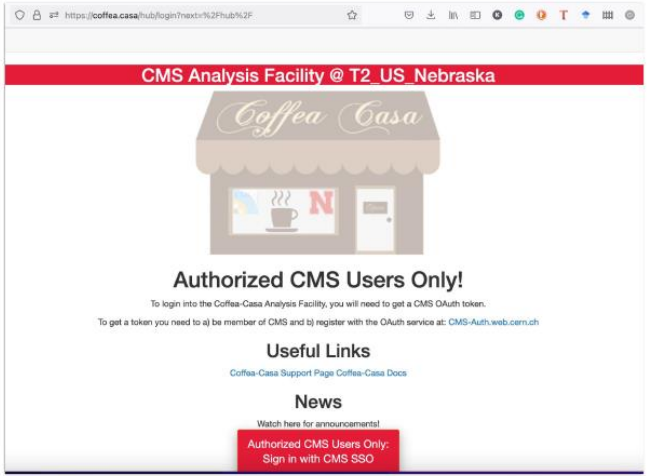


‘alien containers’ HPC integration - container computing, using curated application images

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

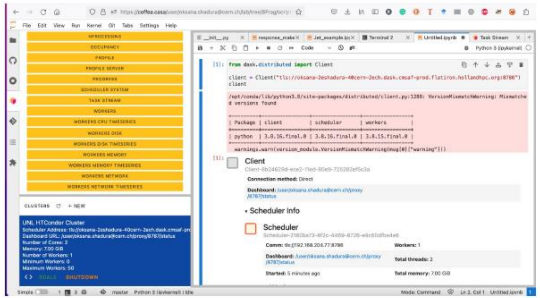


Powered by CMS IAM instance

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

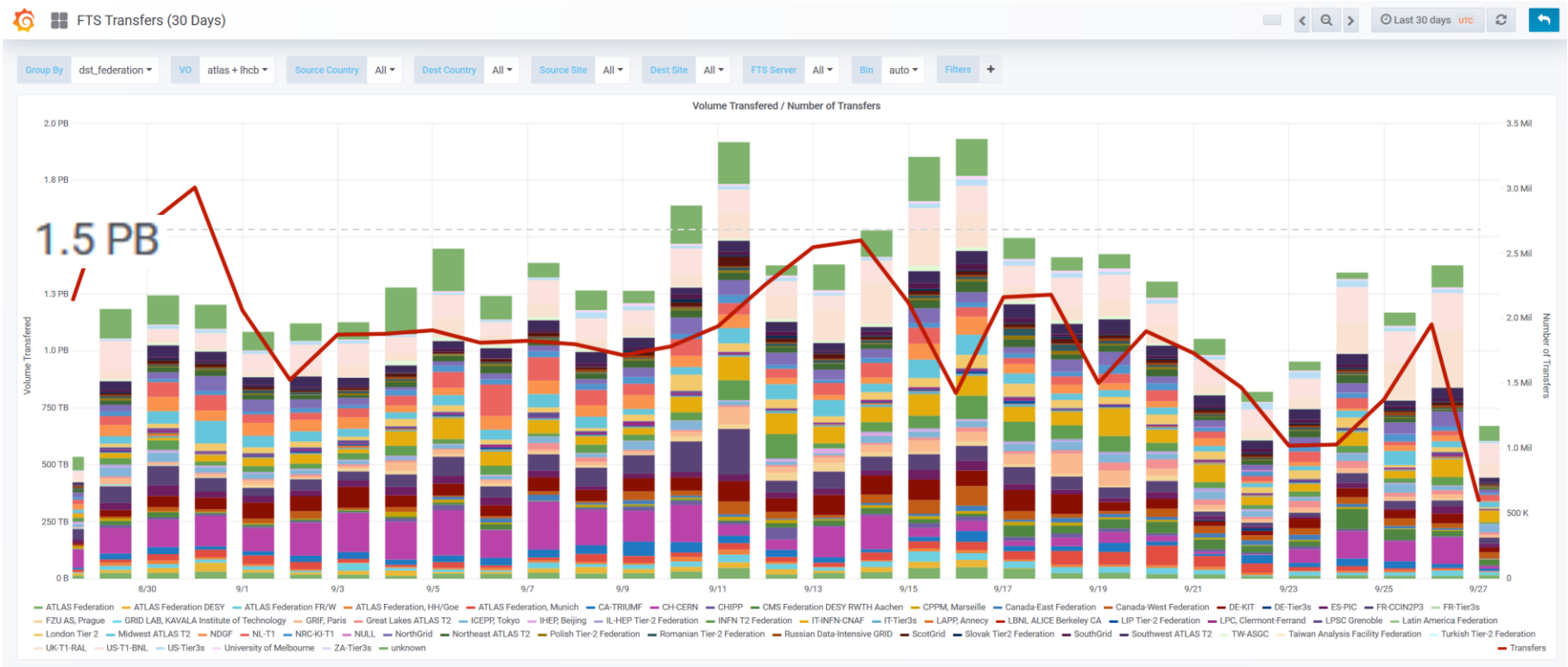


sk can be used inside Jupyter or you can simply nch it through Jupyter and connect directly from r laptop

14

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

High *throughput* computing is in the end about data



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 storage 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, (and still: CXFS), … but also NetApp, HDS, Dell-EMC, &c
Common open source: BeeGFS, gluster, dCache, CephFS, Lustre, …

And storage is usually *tiered* – fast local → online (spinning) disk → near-line (tape)

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)

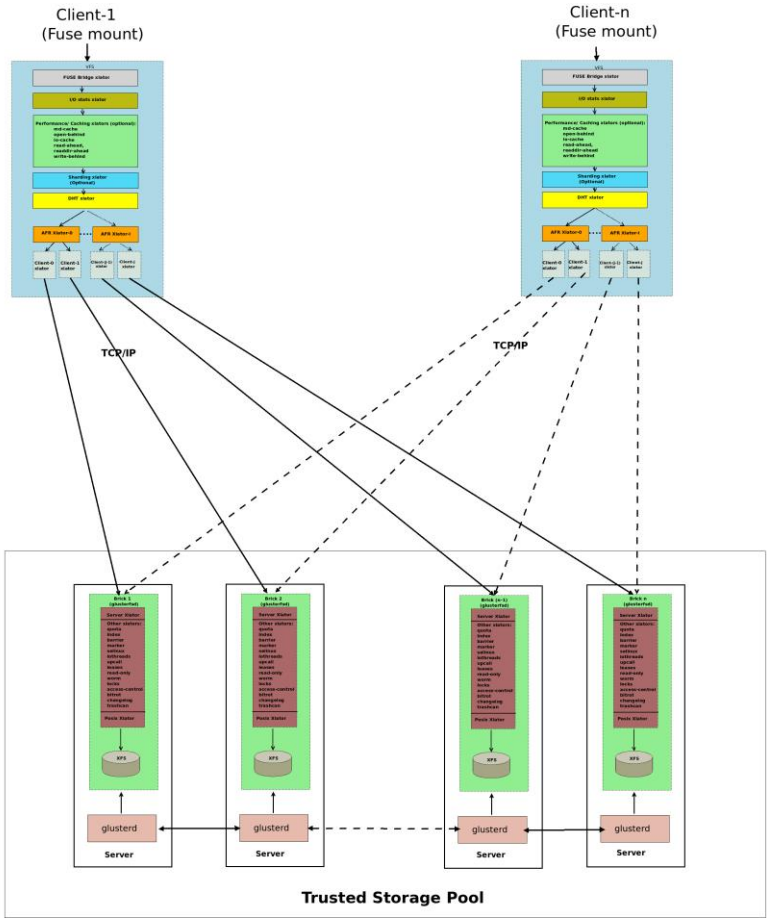
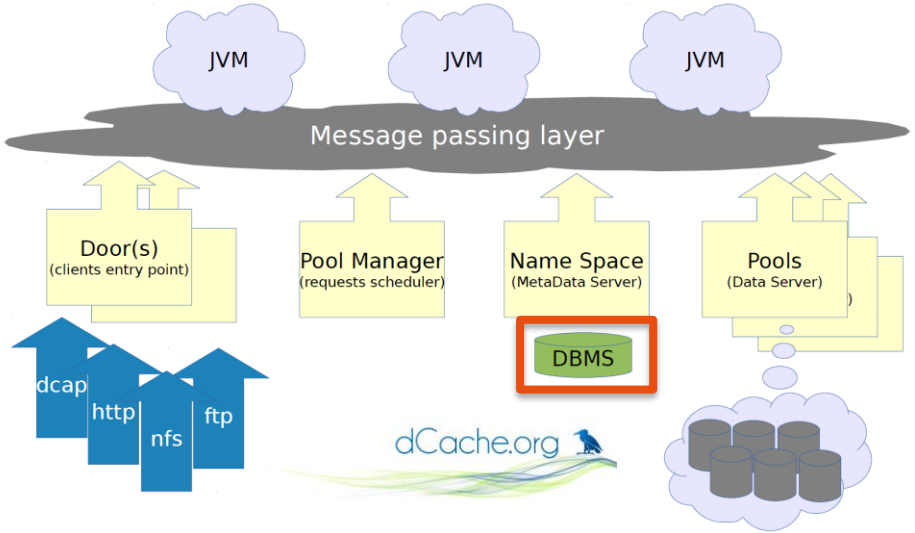
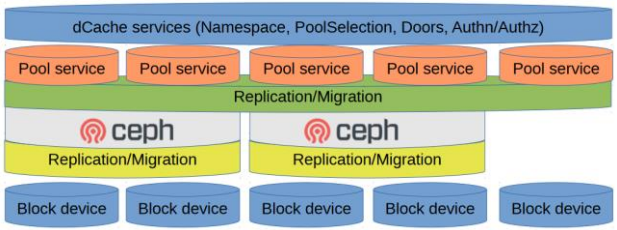


Image source Gluster community: https://docs.gluster.org/en/main/Quick-Start-Guide/Architecture/

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

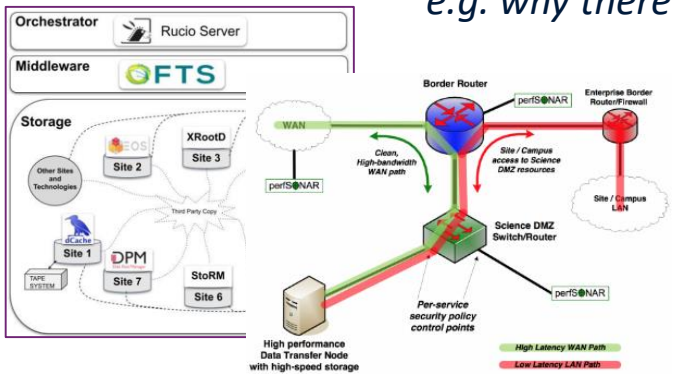
- can be widely (long latency) distributed
  - Nordic Data Grid Facility: Sweden is quite long (~16ms RTT), and Ljubljana to Umeå is ~30ms RTT (~ 2900km)
- redirect-then-access model limits interactions with any single node across a long-distance links
- at 'cost' of POSIX features like *atime* or concurrent 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

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



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

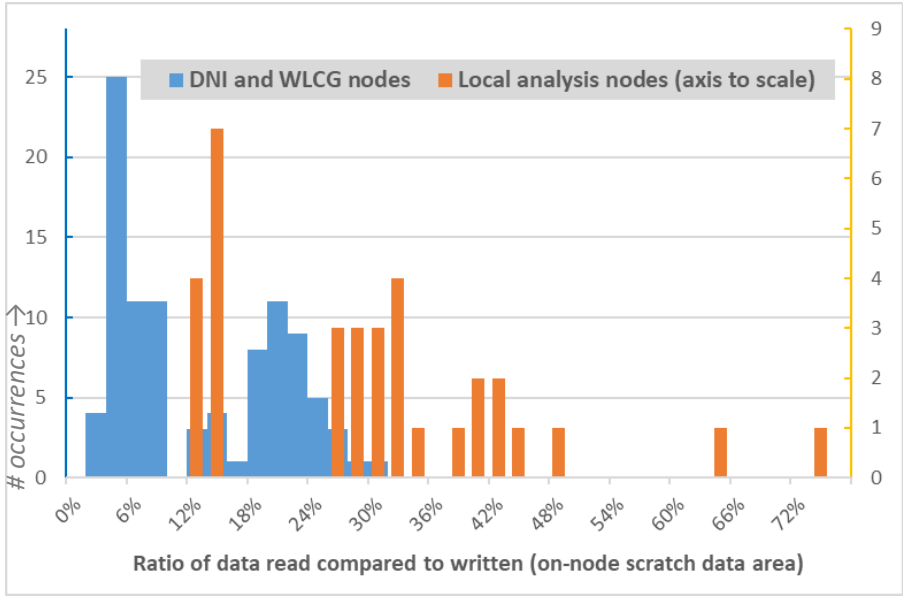
Especially with *WORN* storage: Write Once Read Never

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

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
- always from A-to-B?
  A conveyer belt will do much better!

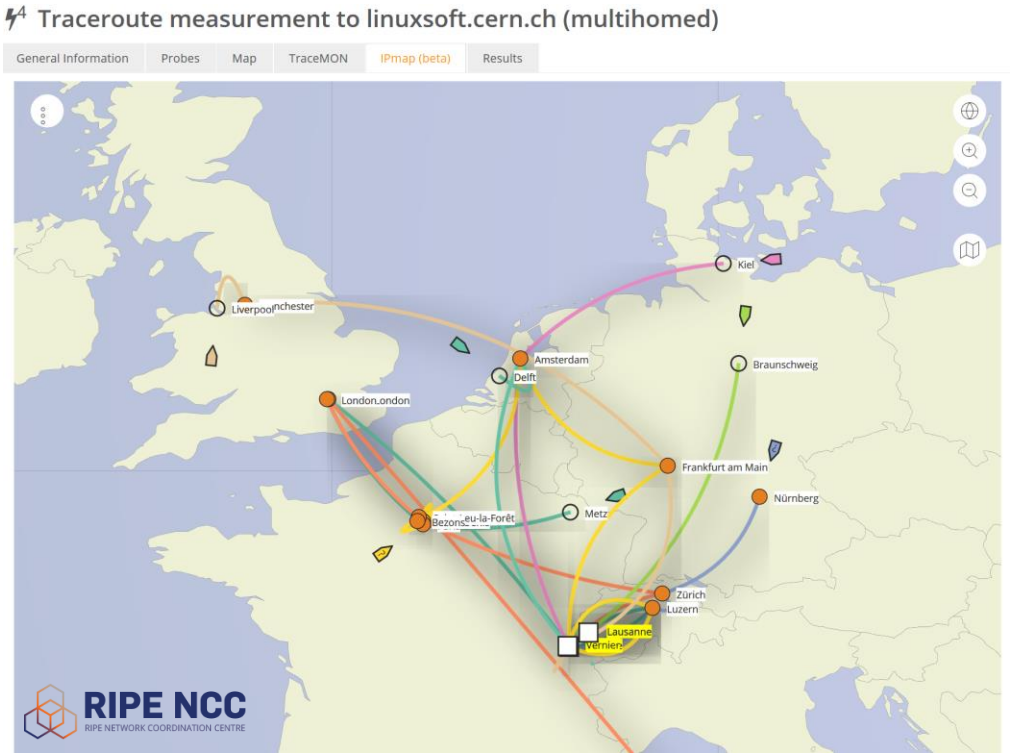
… although you still need
a hole to dump it in …

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



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*

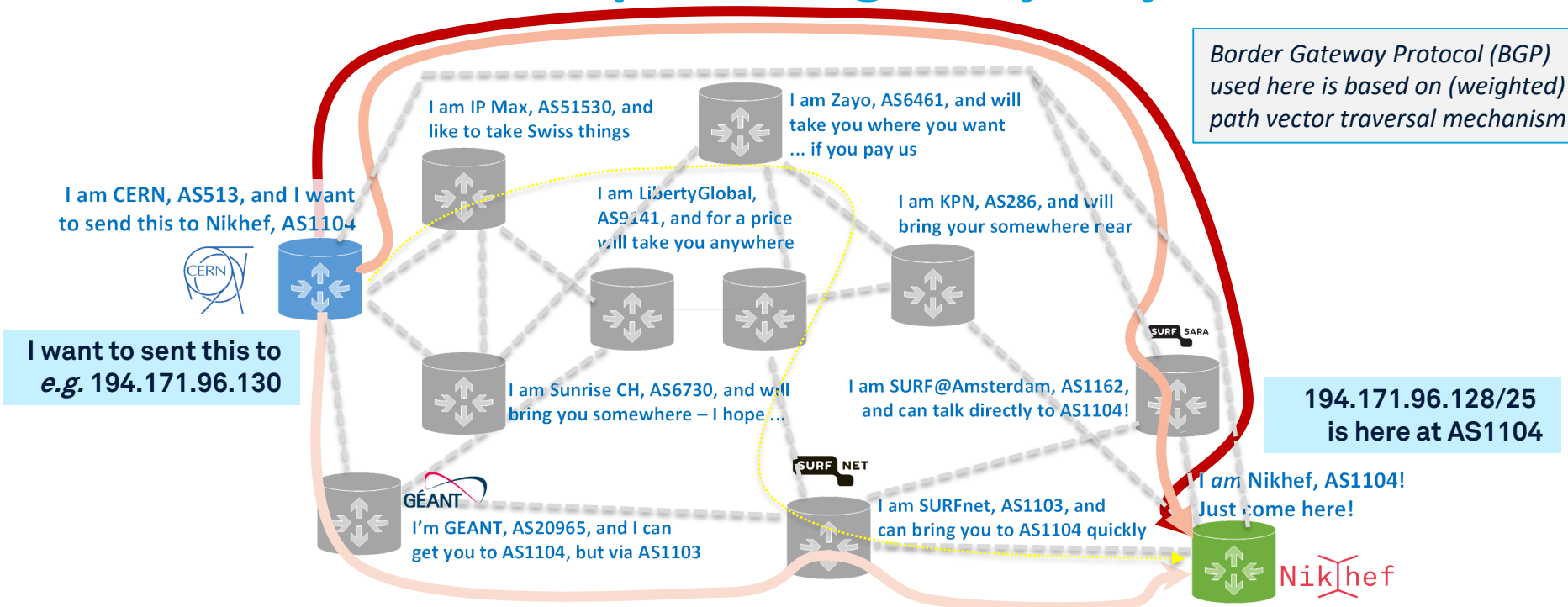
```
[root@kwark ~]# traceroute -6 -A -T gierput.nikhef.nl
traceroute to gierput.nikhef.nl (2a07:8500:120:e010::46), 30 hops max, 80 byte packets
 1  2a10-3781-17b6.connected.by.freedominter.net (2a10:3781:17b6:1:de39:6fff:fe6b:4558) [AS206238]  0.810 ms  1.052 ms  1.330 ms
 2  2a10:3780::234 (2a10:3780::234) [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?



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)

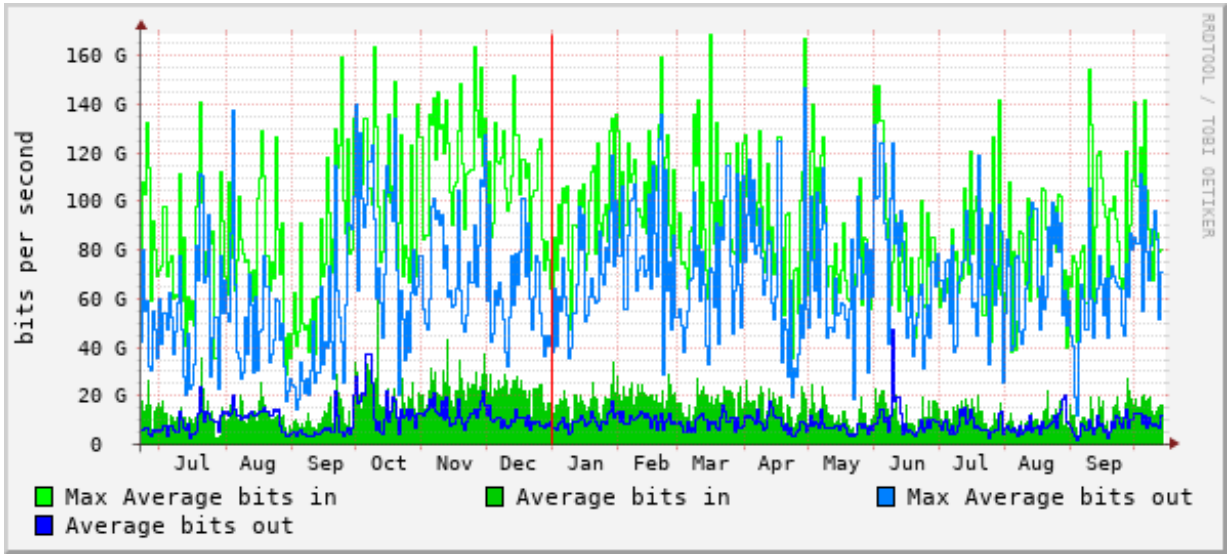
Announcing routes: the Border Gateway Protocol

```
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)

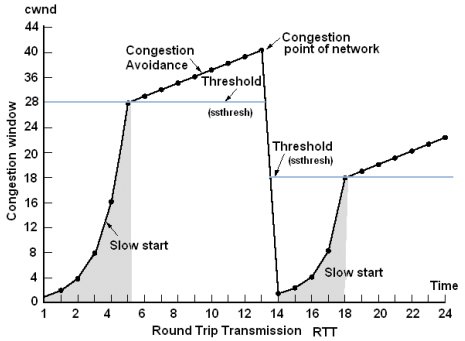
network limit (MSS 9000 byte, RTT: 150.0 ms, Loss: $2.304*10^{-11}$ ($2*10^{-09}$%)) : **100000.00 Mbit/sec.**

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

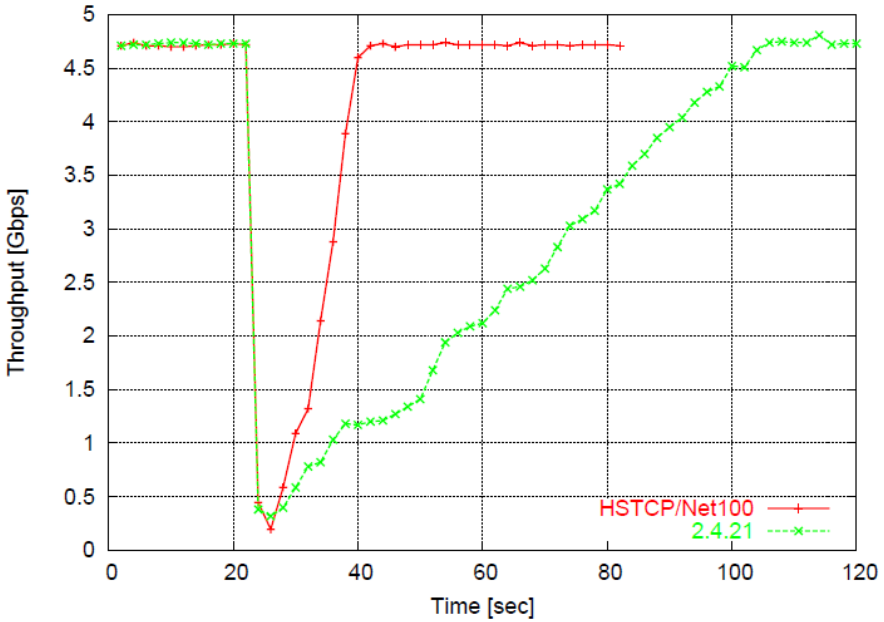


Figure 10: HSTCP versus stock TCP recovery time

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

LHCOPN

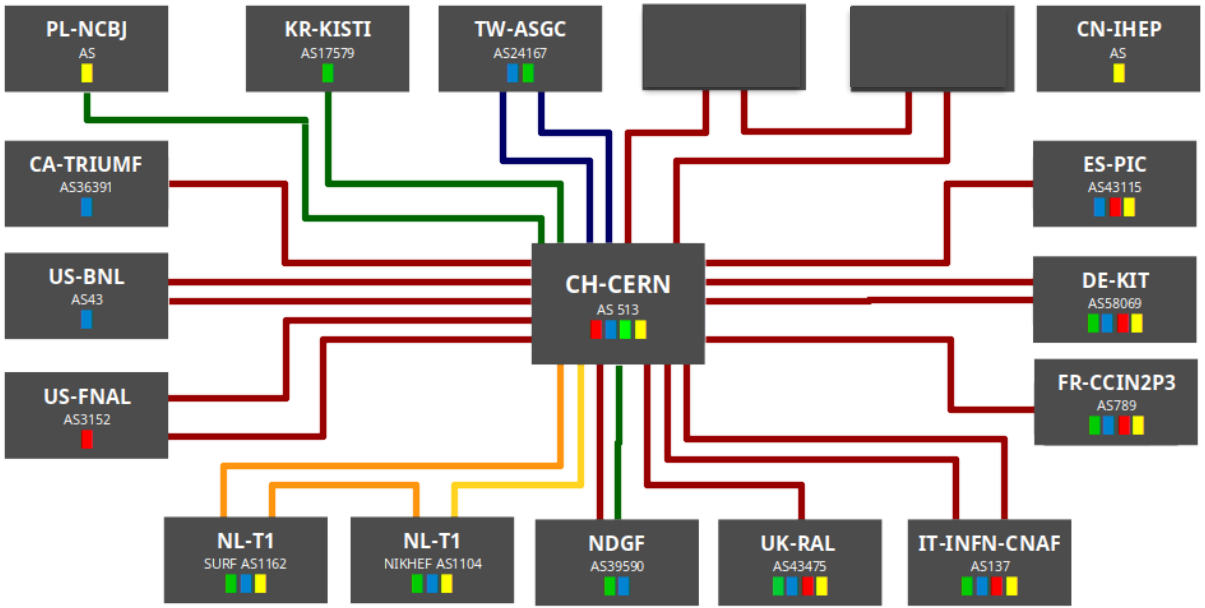
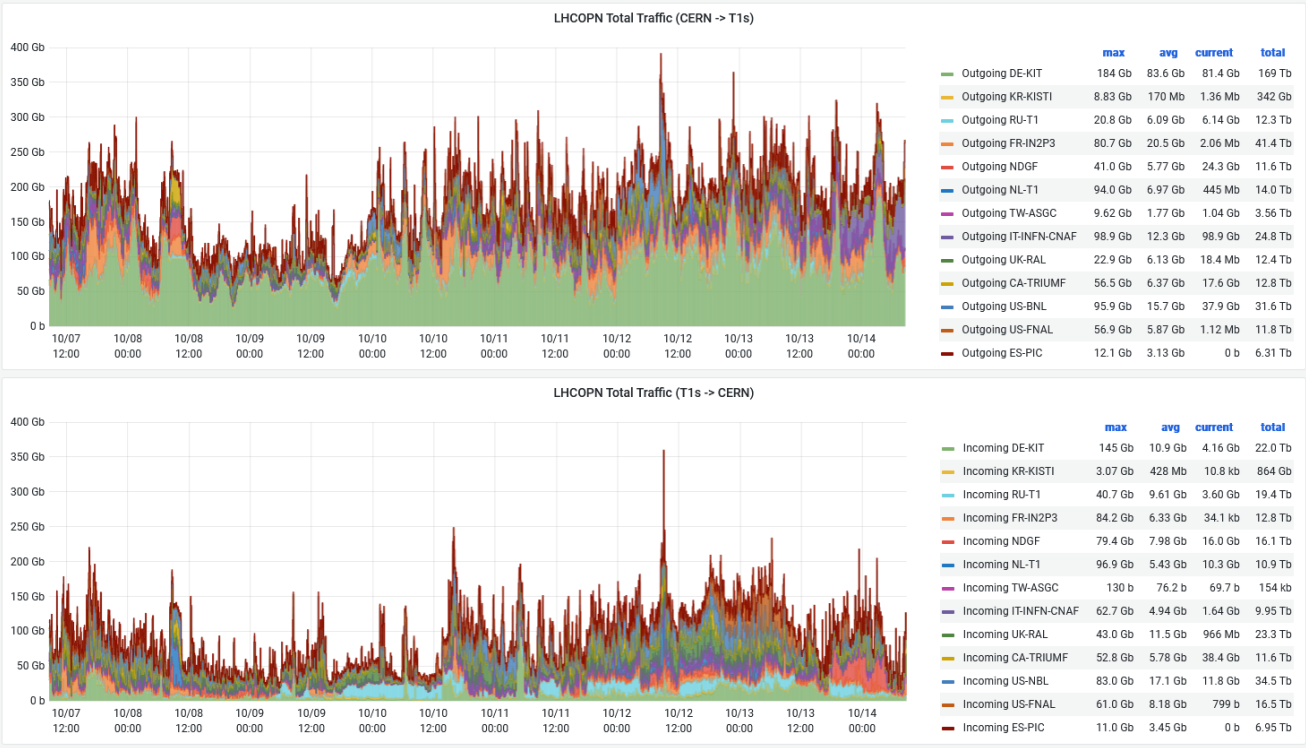


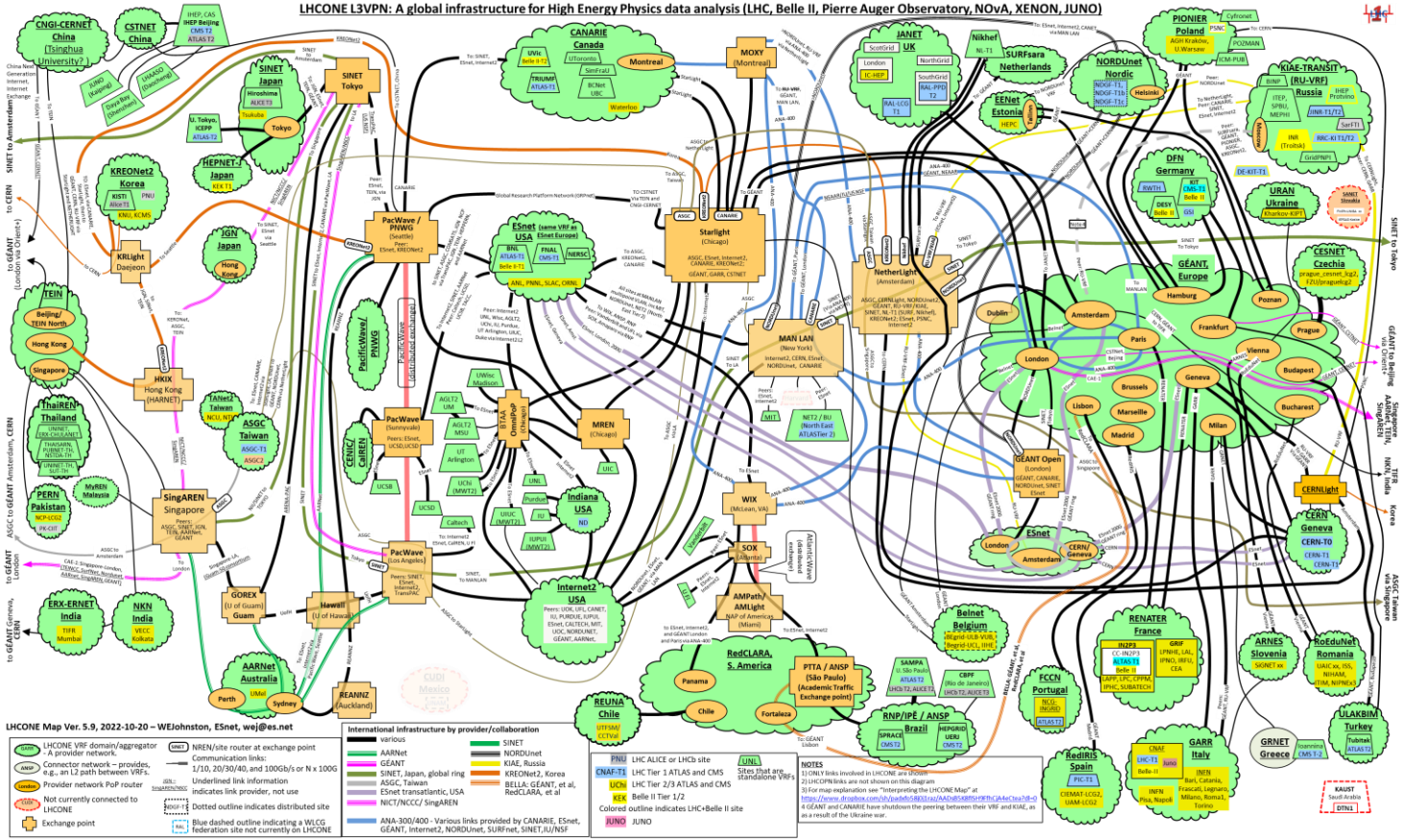
Image source: Edoardo Martelli, CERN, https://lhcopn.web.cern.ch/

LHCOPN – traffic levels for T0T1 data transfer



CERN OpenMonIT LHCOPN, period Oct 7 .. Oct 14 2022, from https://monit-grafana-open.cern.ch/d/HreVOyc7z/all-lhcopn-traffic

LHCone



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**

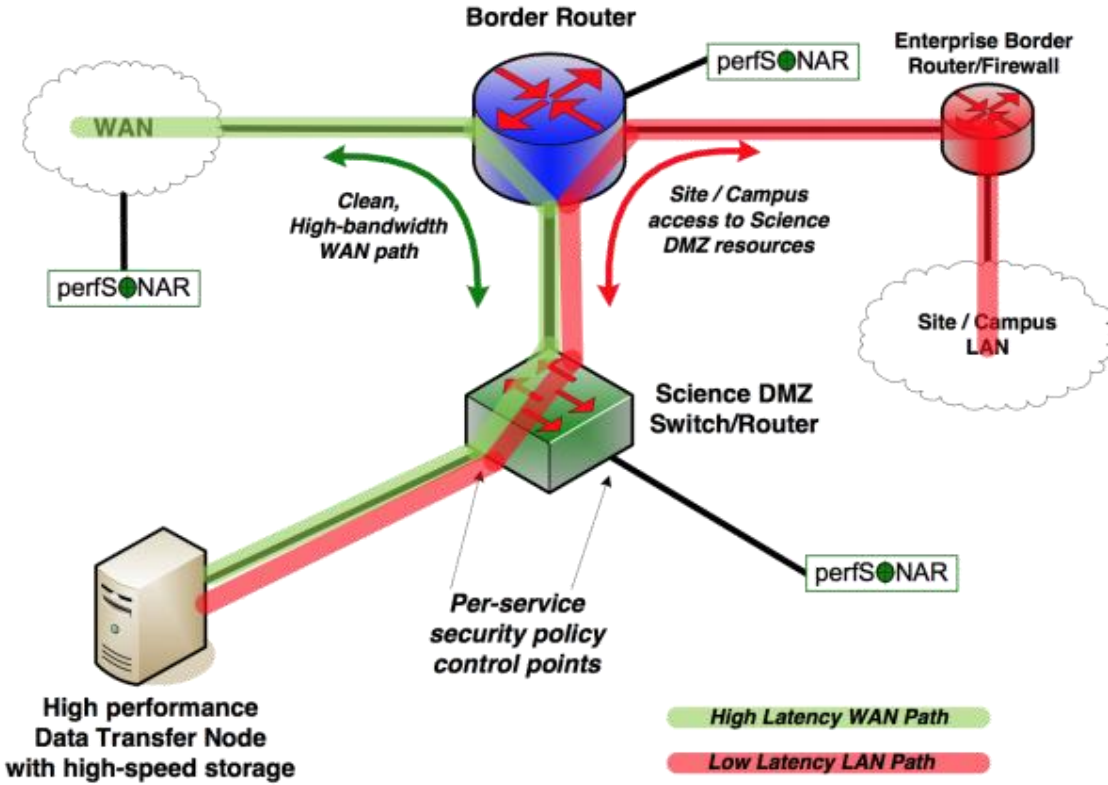
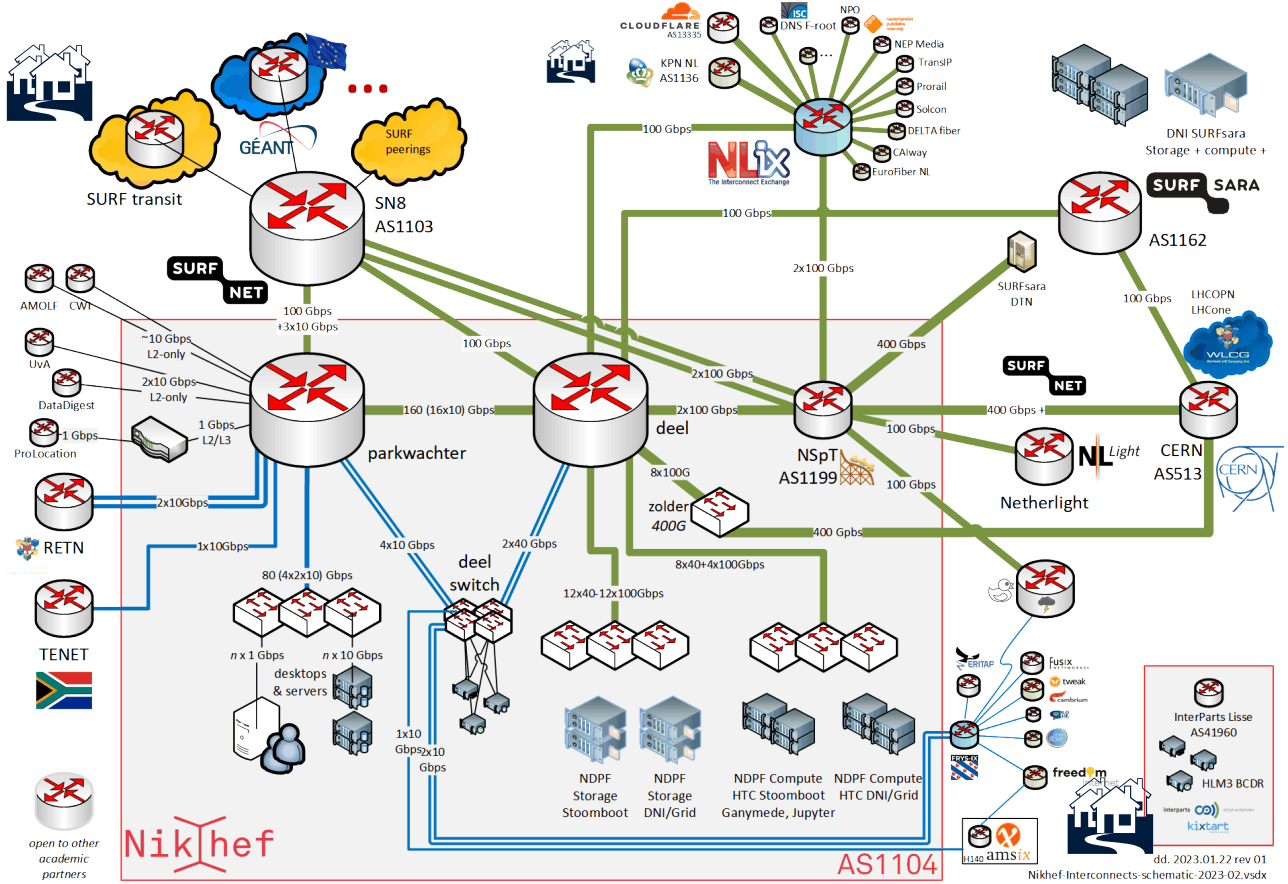


Image and ‘ScienceDMZ’ concept promulgated by ESnet (see fasterdata.es.net)

Just one random autonomous system: AS1104



AS1104
state as of 2022,
before the SR=1x and
400G upgrades

Exercising the network – sensor data and events

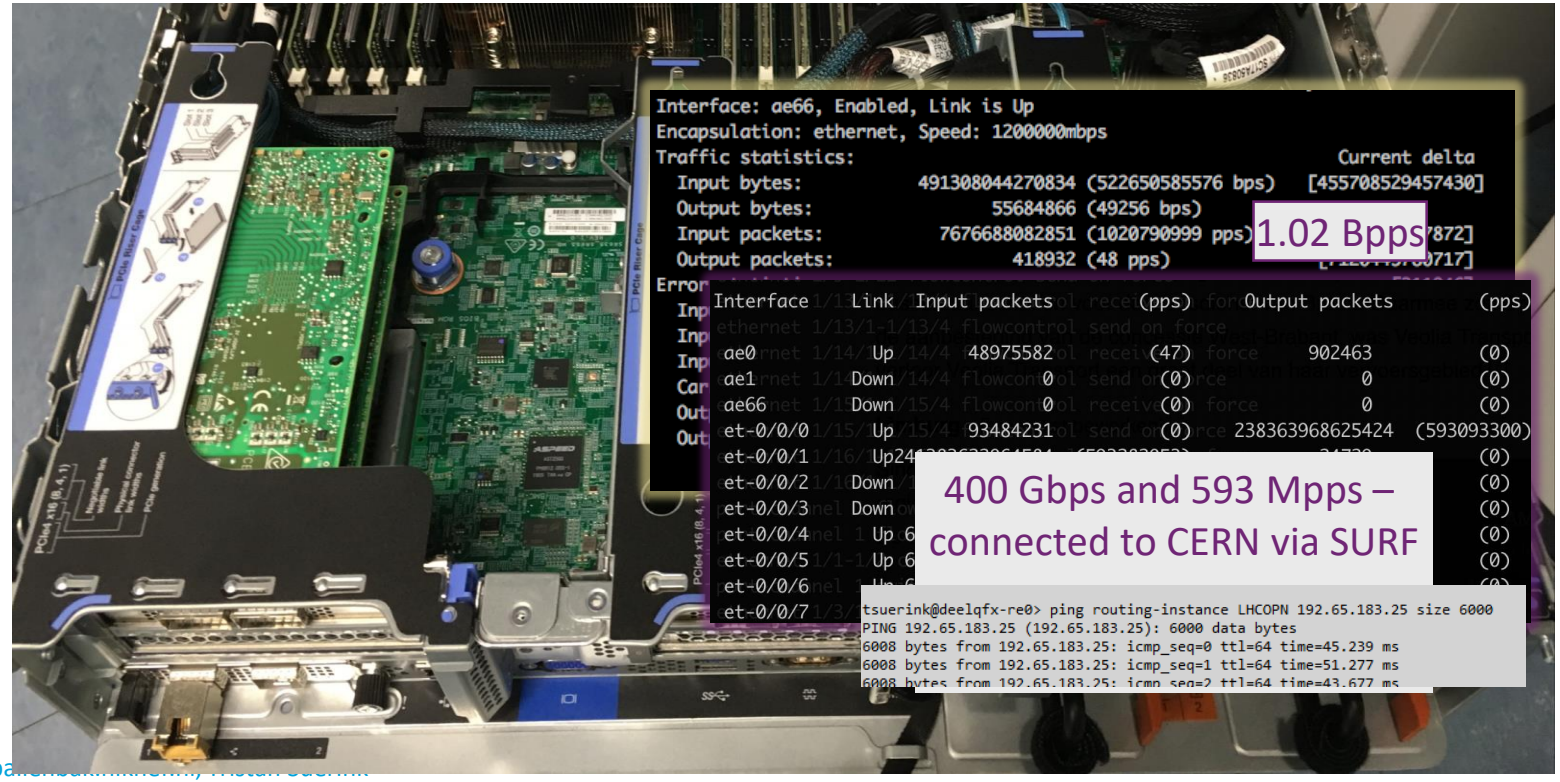
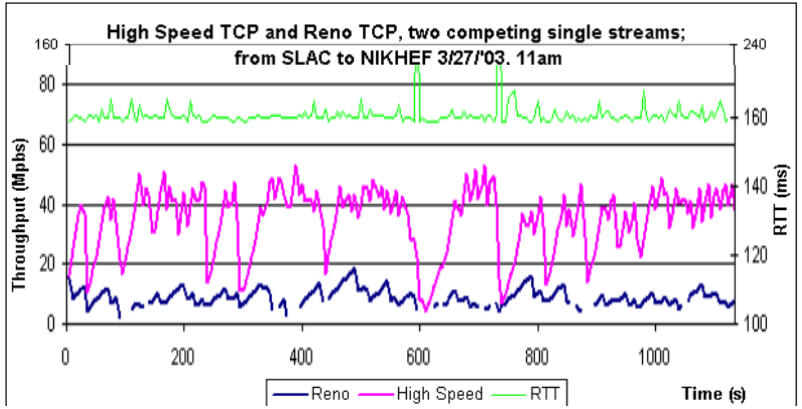


Image: ba

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



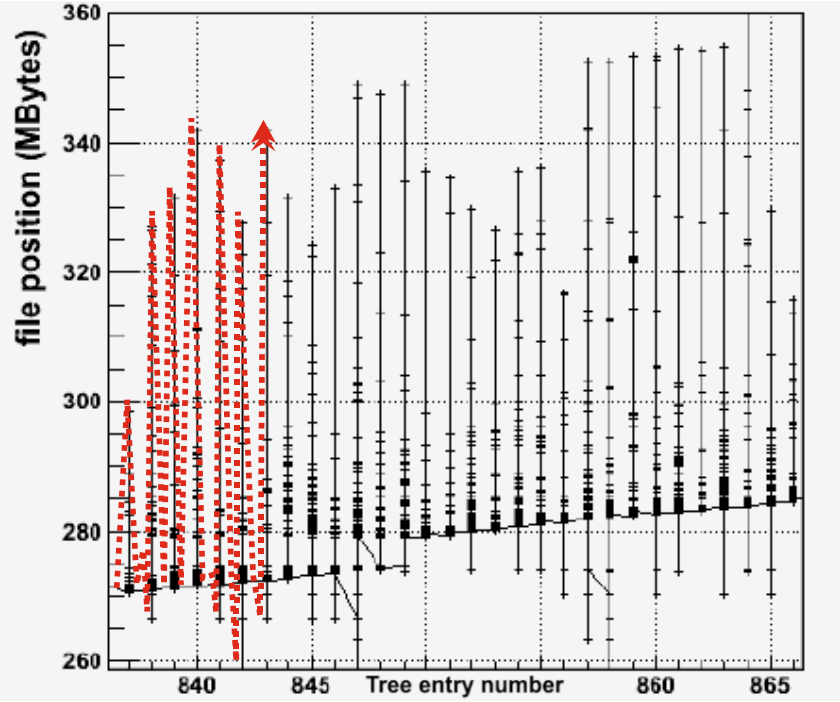
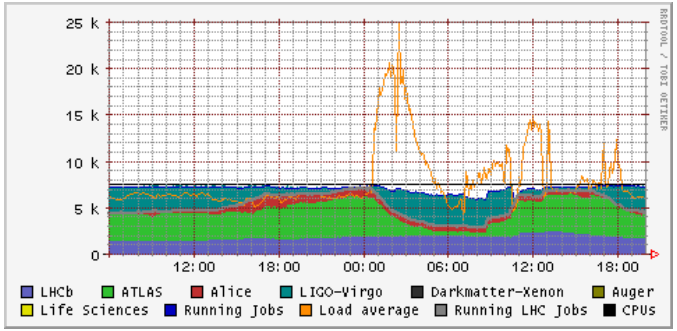


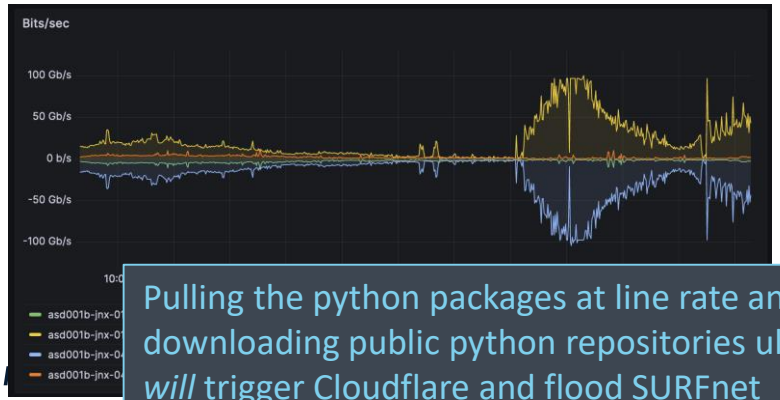
Image of TCP slow-start and packet loss impact (in Mpps): Antony Antony et al., Nikhef, for DataTAG, 2003(!)
Right: base graphic: Philippe Canal “Root I/O: the fast and the furious”, CHEP2010 Access pattern reflects Root versions < 5.28, before Ttree caching and ‘baskets’

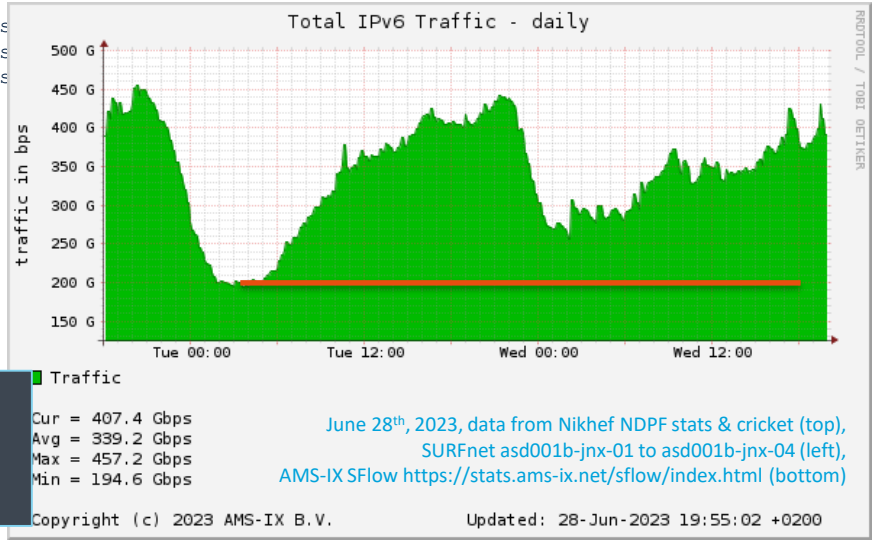
And sometimes traffic is triggered by researchers scaling up ‘accidentally’ from a laptop to a cluster without too much thought

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

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: 400 → 800G AMS-GVA ☺

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
ierotonde is een testomgeving waar SURF en Nikhef gaan experimenteren met nieuwe
ng beschikt over een internetsnelheid van 800 Gbit/s, wat meer dan 1000 keer sneller
gemiddeld huishouden in Nederland. De innovatierotonde stelt Nederlandse
e 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
ote hoeveelheden data snel te kunnen verwerken, is de verwachting dat 800Gbit/s
. 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

Op dez
Een goe
Examen
Wat geb
Vragen
Terug n

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

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

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

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.

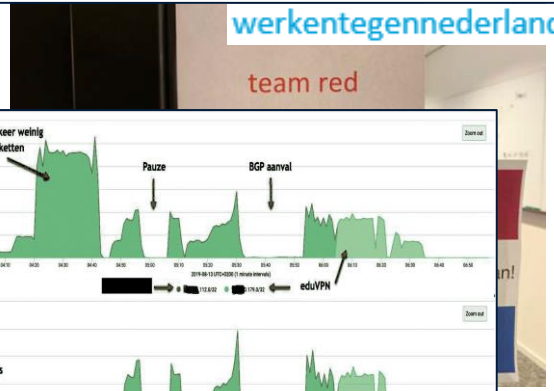


Image sources: belastingdienst.nl, rws.nl, nu.nl

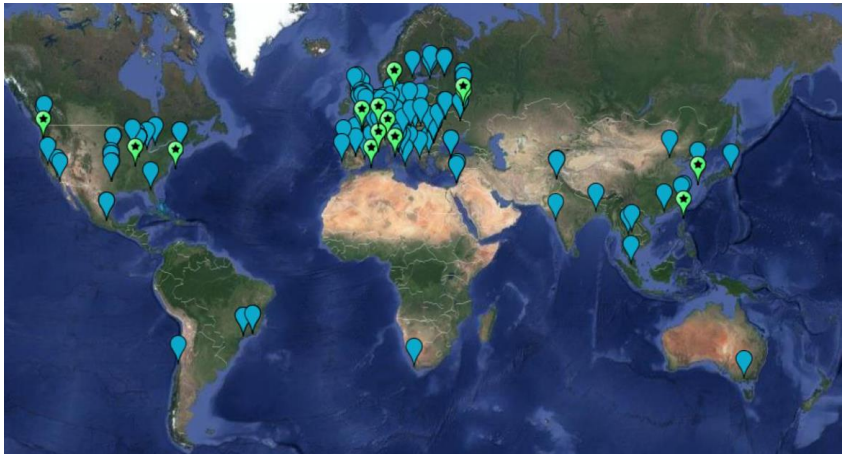
Access, Trust & Identity

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
- ~60 countries & regions
- ~20000 users

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 EDG and the HEP 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, vis

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 grou

To be completed by the User :

It is **MANDATORY** to provide the following inform
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:

Last	First	Middle
SWIETZER	JOHN	JAMES

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 – 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

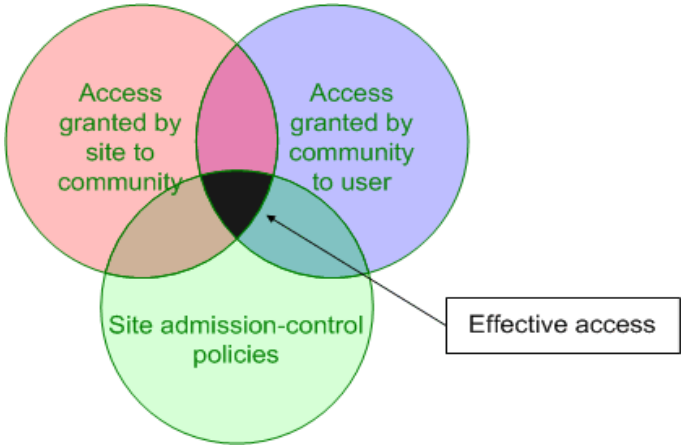
```
ldap_access_order = filter,authorized_service
ldap_access_filter = (|(memberOf=cn=gridSrvAdministrators,ou=DirectoryGroups,dc=farmnet,
dc=nikhef,dc=nl)(memberOf=cn=gridMWSecurityGroup,ou=DirectoryGroups,dc=farmnet,dc=nikhef
,dc=nl)(memberOf=cn=nDPFPrivilegedUsers,ou=DirectoryGroups,dc=farmnet,dc=nikhef,dc=nl))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

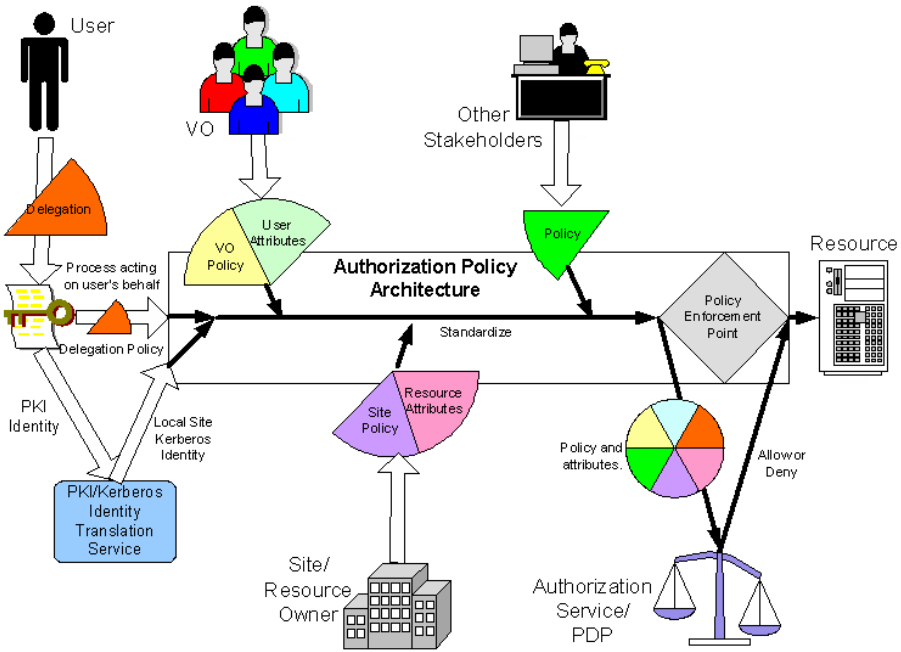
Policy example: Argus system, https://argus-documentation.readthedocs.io/en/stable/misc/examples.html; service-specific: sssd.conf ldap auth_provider

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)





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

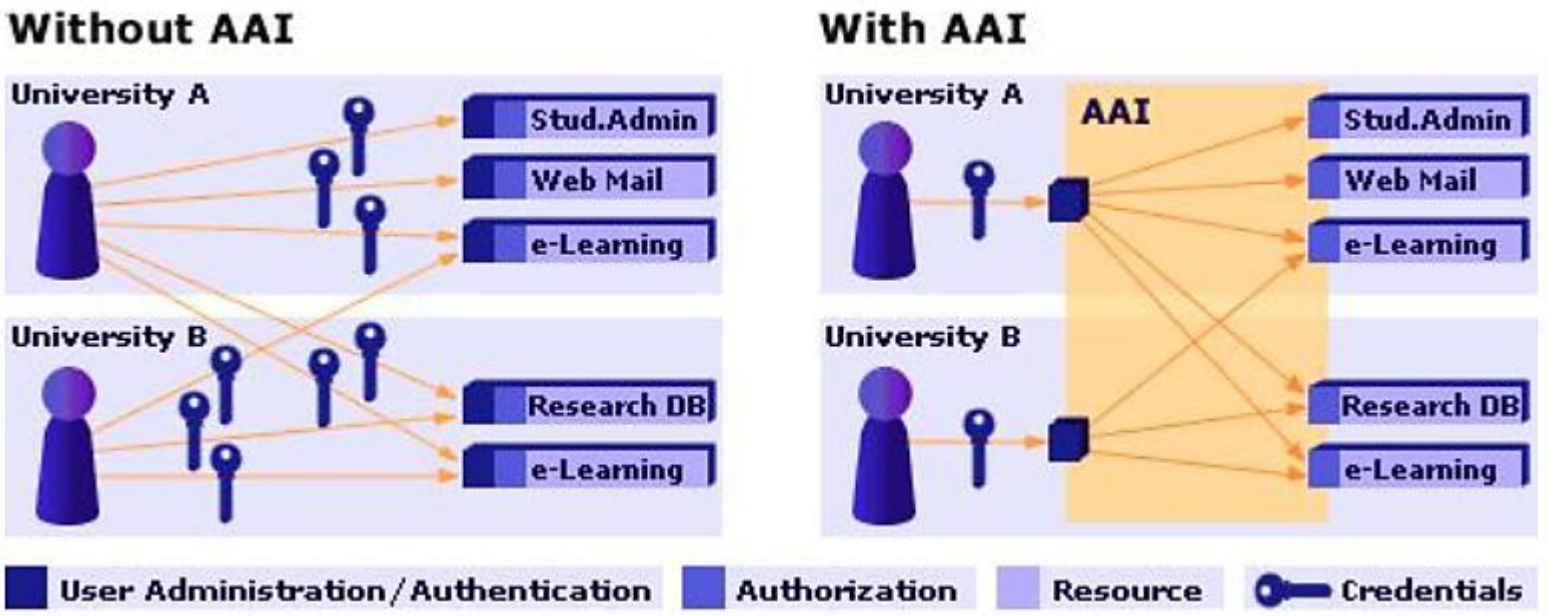
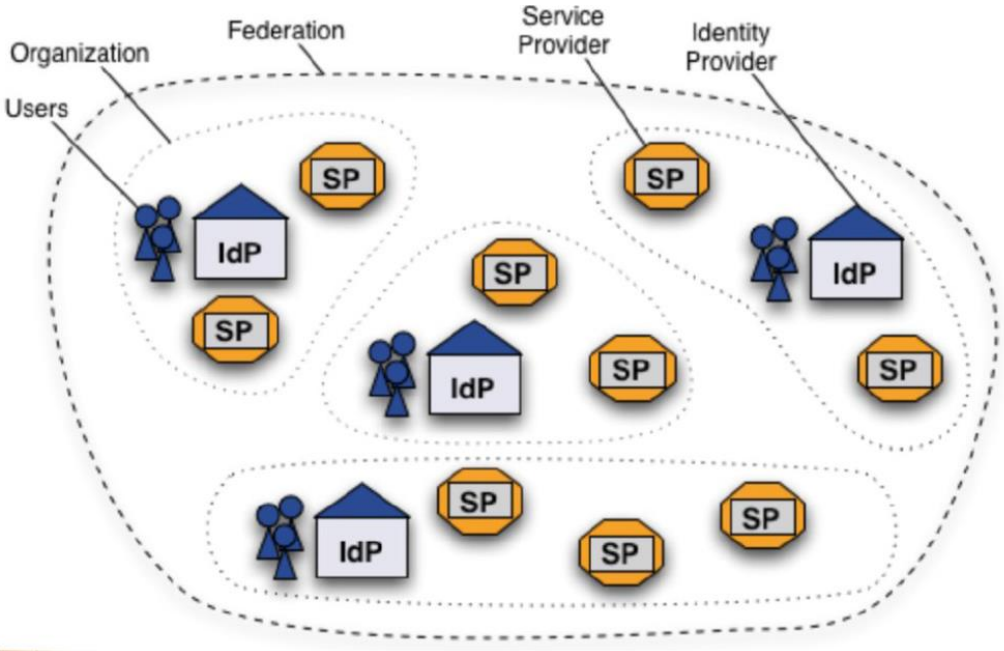


Image: AARC NA2 training module “Authentication and Authorisation 101” - https://aarc-community.org/training/aai-101/

Federation

portability of identity information across otherwise autonomous administrative domains

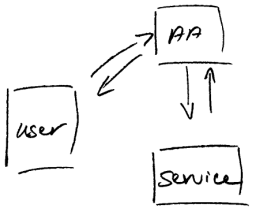


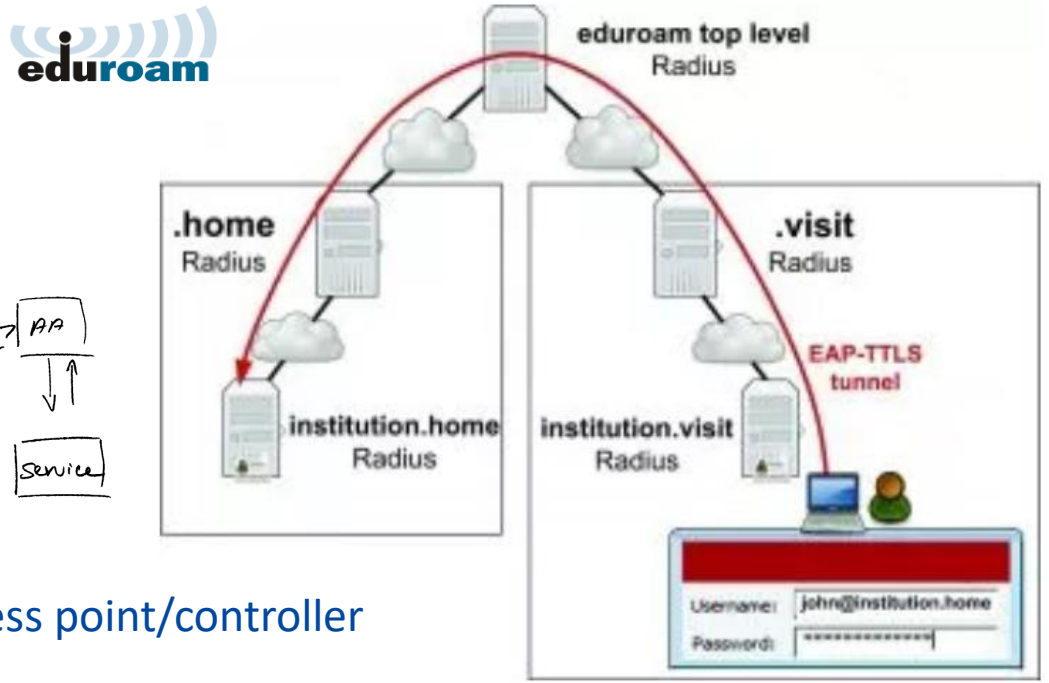
Shibboleth IdP image and SAML2 auth flow by SWITCH (CH) – see also https://refeds.org/ on federation structure and (assurance and security) guidelines

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

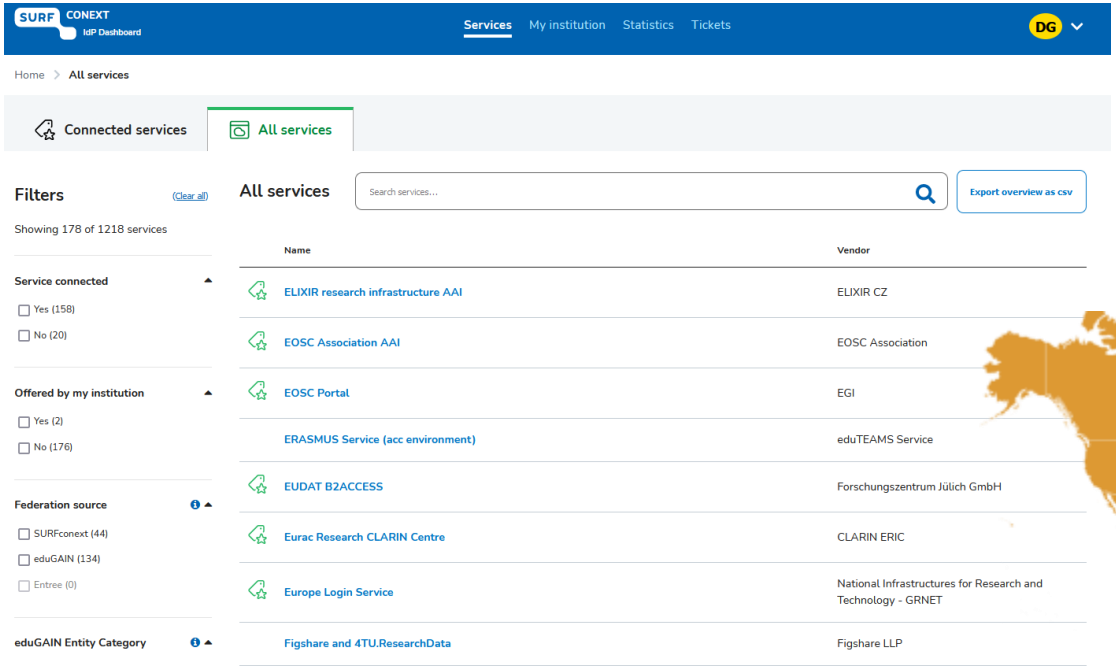


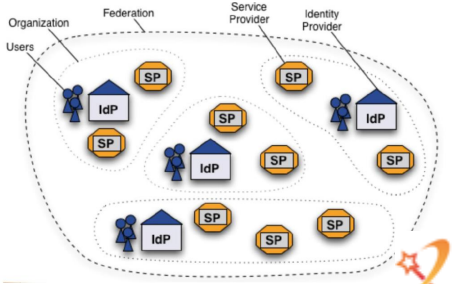


RADIUS server then instructs WiFi access point/controller

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

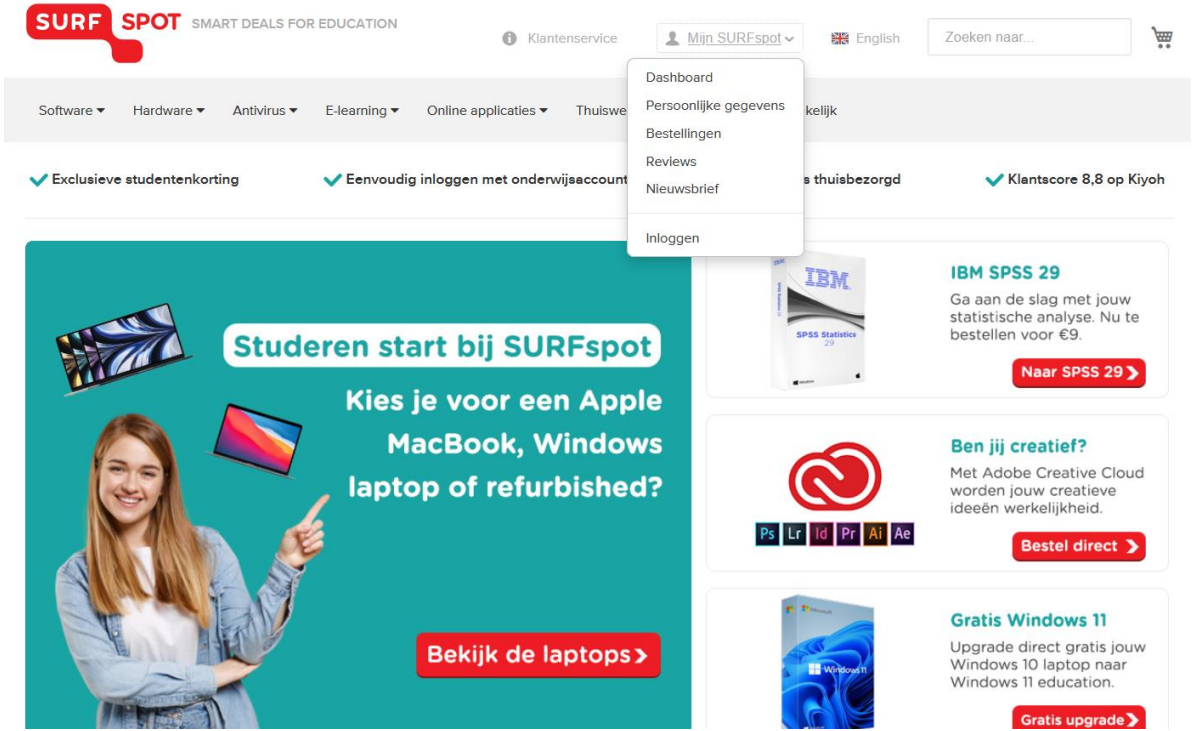




eduGAIN

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

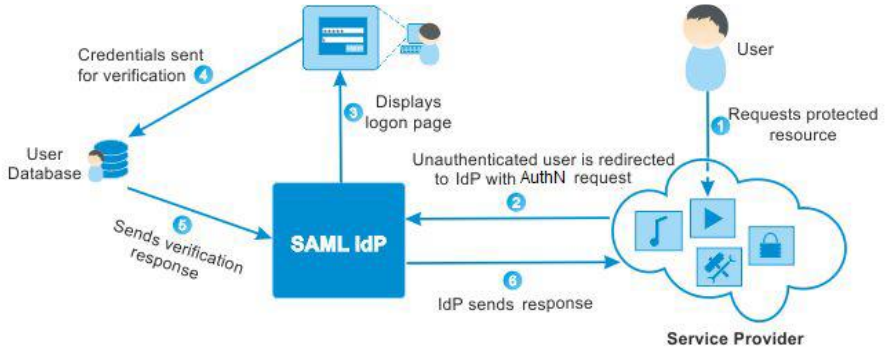
Your favourite federated service?



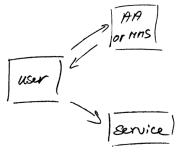
https://surfspot.nl/

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



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 message decoder: https://addons.mozilla.org/en-US/firefox/addon/saml-message-decoder-extension/ (Magnus Suther)

SAML WebSSO flow image: SWITCH, CH

Under the hood, this is 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"
     ...
```

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)

OIDC Fed - OpenID Connect Federation

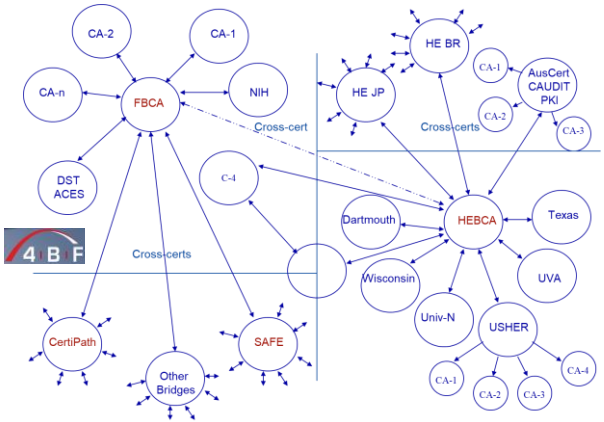
- federate end-points for OIDC Providers and Relying Parties (or OAuth2), with similar models

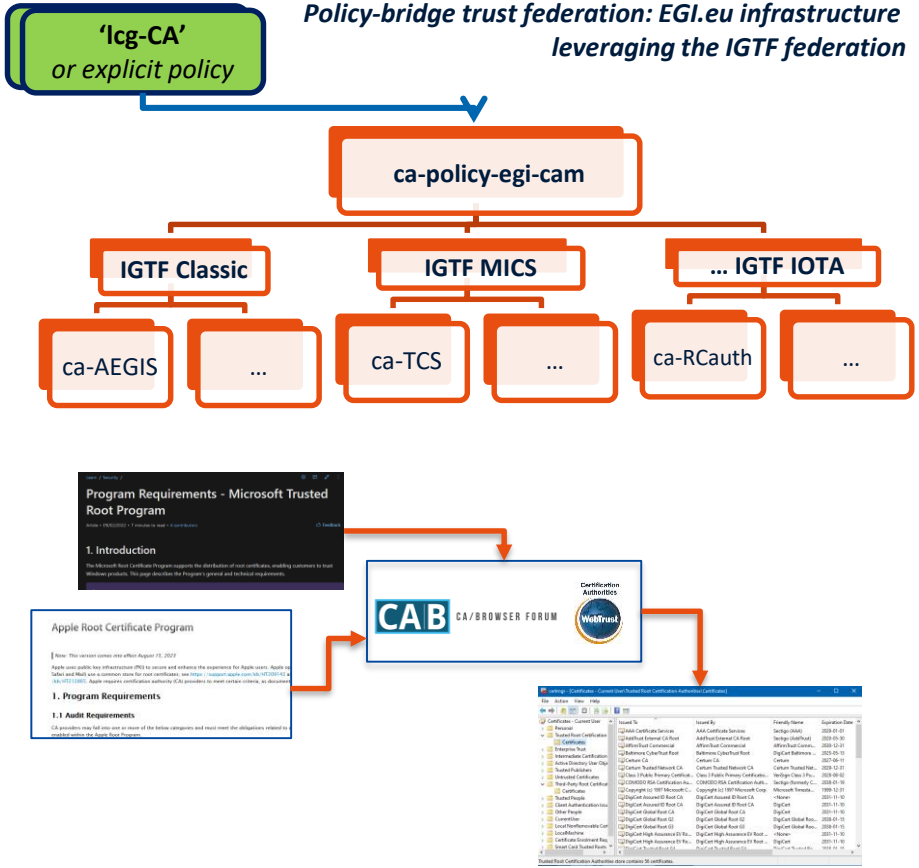
note federation based on 'ultimate trust' domains (e.g. cross-realm Kerberos) also exists …

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

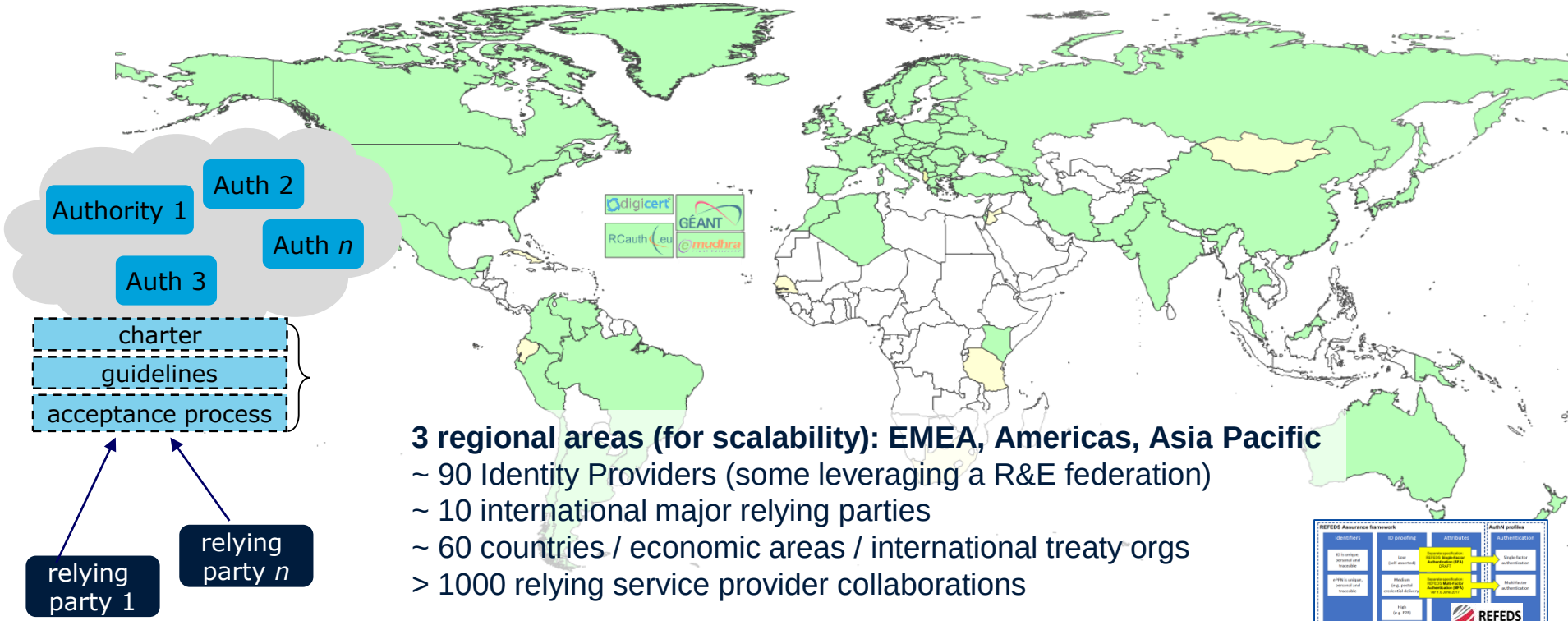
trust remains with the relying party
can be *bridged* by either cross-signing (left)
or by policy agreements (right)





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

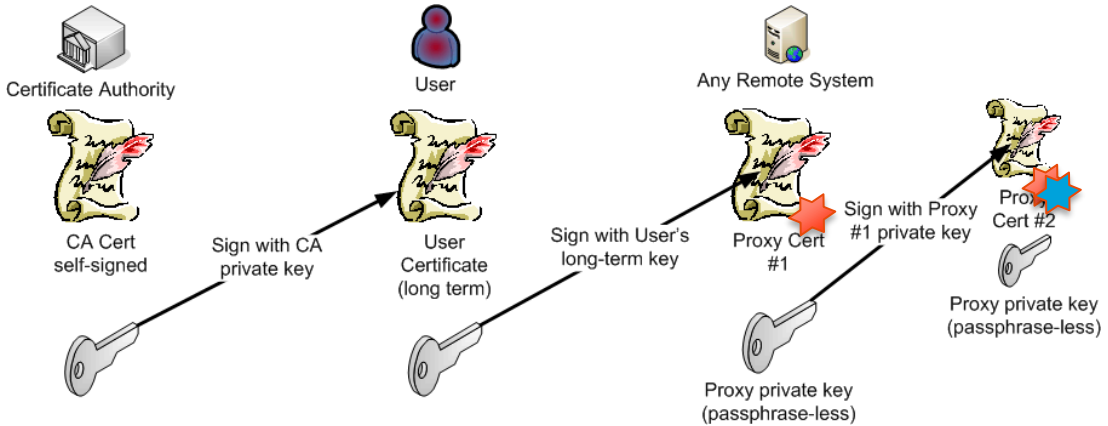
~ 60 countries / economic areas / international treaty orgs

> 1000 relying service provider collaborations

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

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



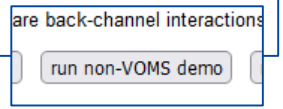
RFC3820 'proxy' certificates extend this concept to (constraint) identity delegation

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

Identity statement: an X.509 RFC5280 Certificate (textually)

```
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
```

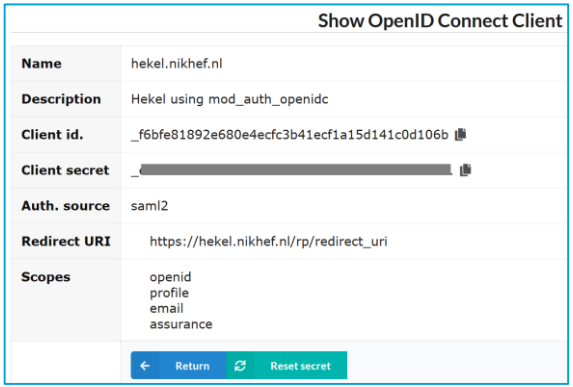
You should be able to get an ‘IGTF-DOGWOOD’ assurance certificate from RCauth.eu.
Go to https://rcdemo.nikhef.nl/ and select the ‘Basic demo’ and use ‘run non-VOMS’ to get and view your short-lived certificate

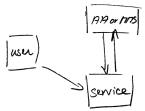


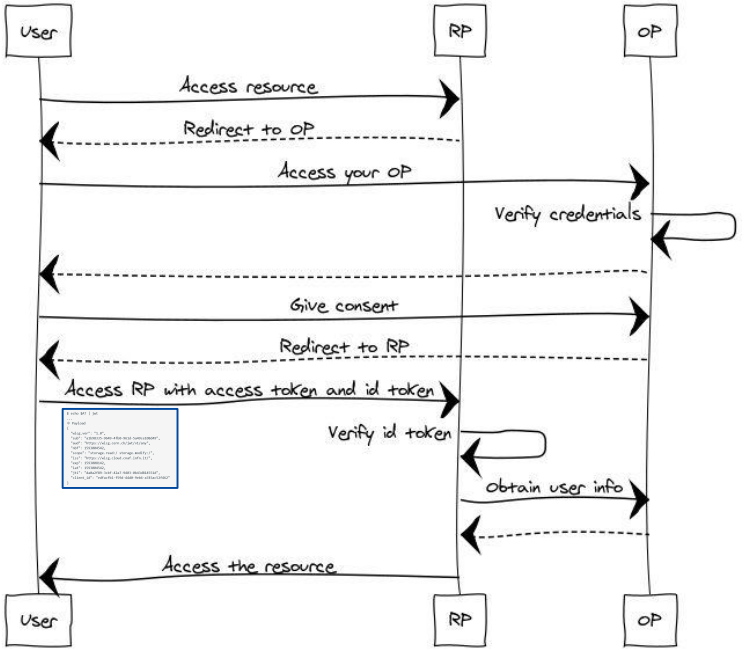
OpenID Connect and OAuth2

- 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’)





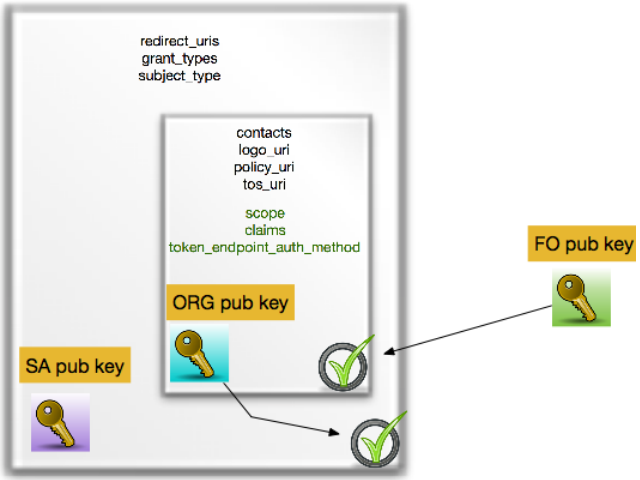


Shown is the ‘implicit flow’, other flows possible. Image source: AARC NA2 training on AAI 101
See https://openid.net/ for protocols and standardization work

OpenID Connect 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



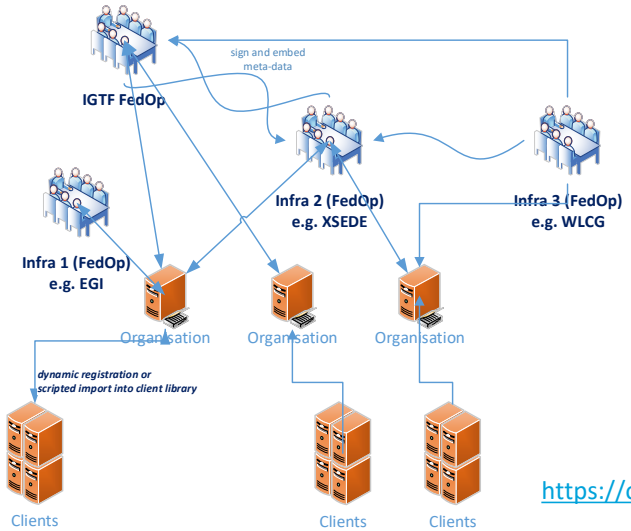


Image: Roland Hedberg, University of Umeå
OpenID Connect Fedrration:
https://openid.net/specs/openid-connect-federation-1_0.html

Federation: technology, interoperability, policy

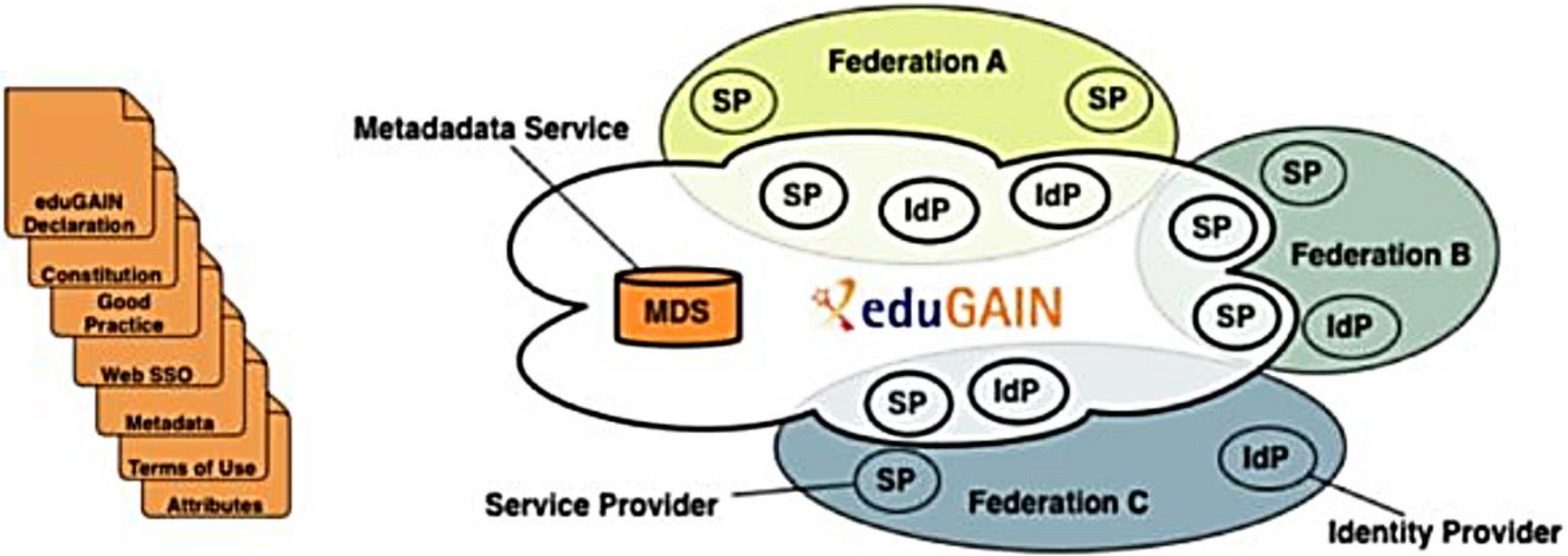
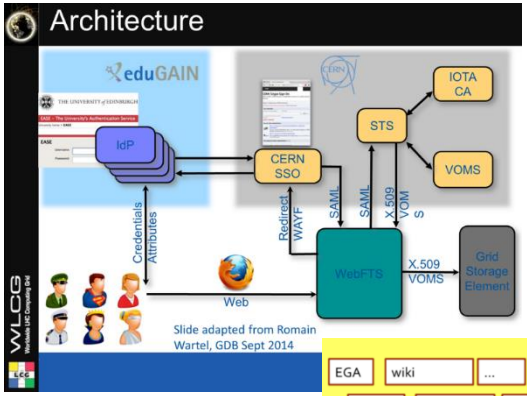


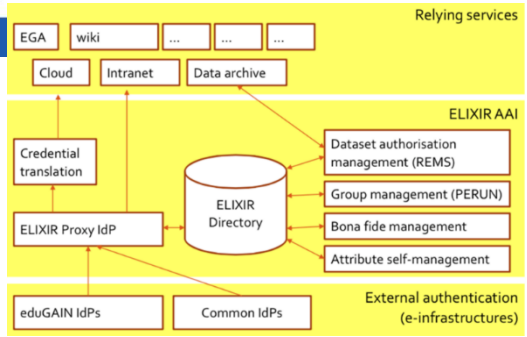
Image from SWITCH (CH) and edugain.org

Managing complexities of federation & identity

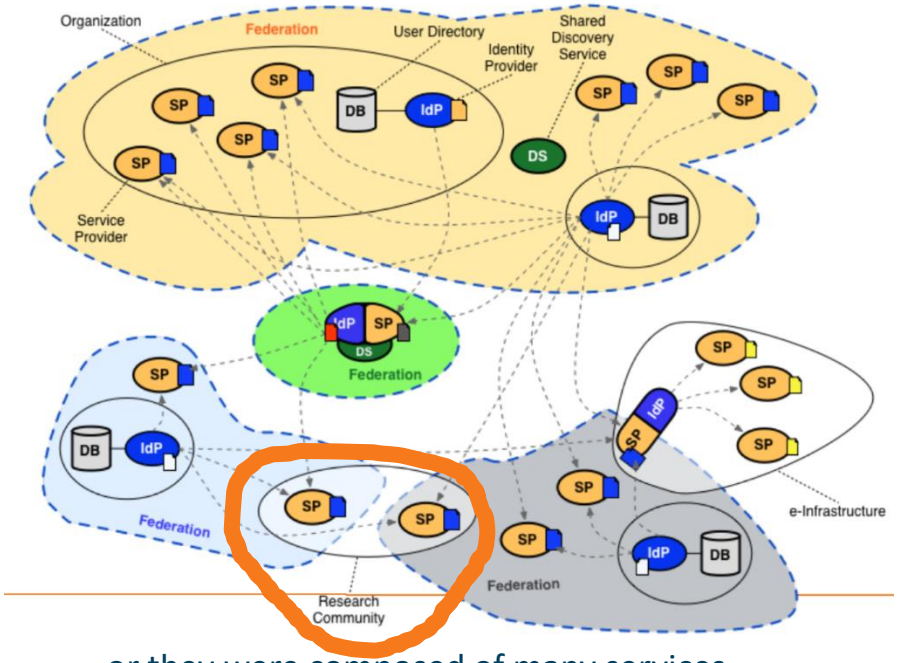


WebFTS prototype 'FIM4R' in wLCG Romain Wartel et al.

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

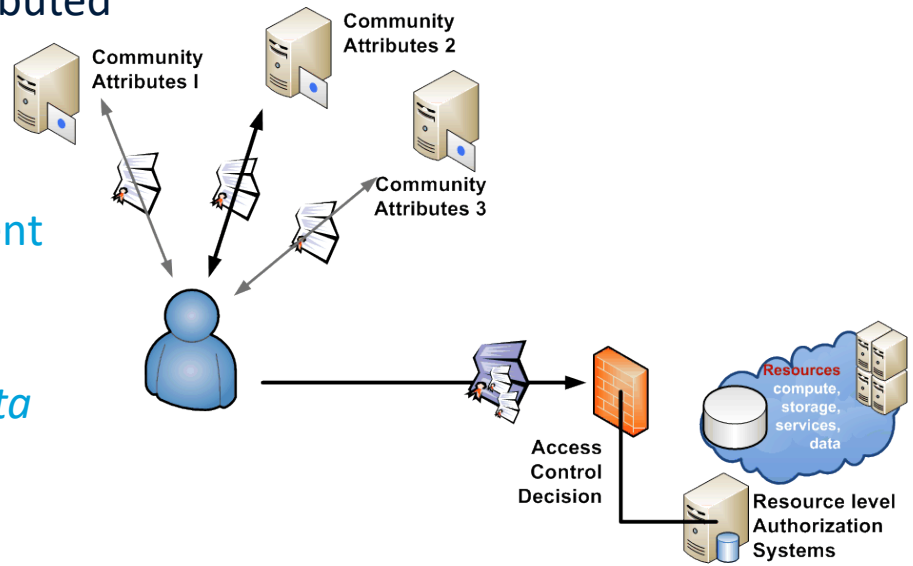
Multiple sources of authority: the community

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

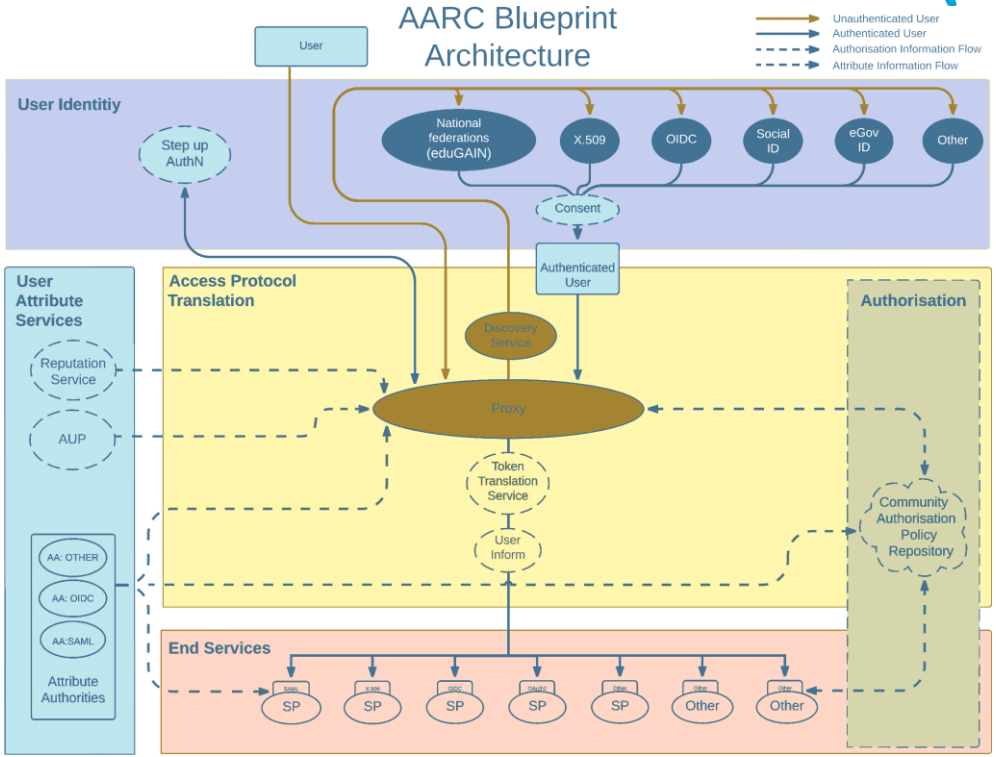
- *source of authority* for attributes is distributed

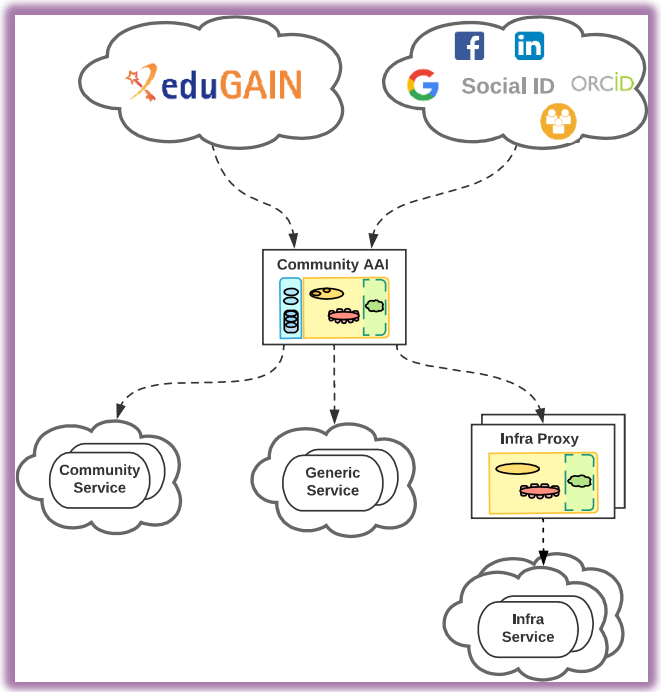
for example:
- community membership from an experiment
- affiliation status from home organisation

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



Most trust flows from the (research) community



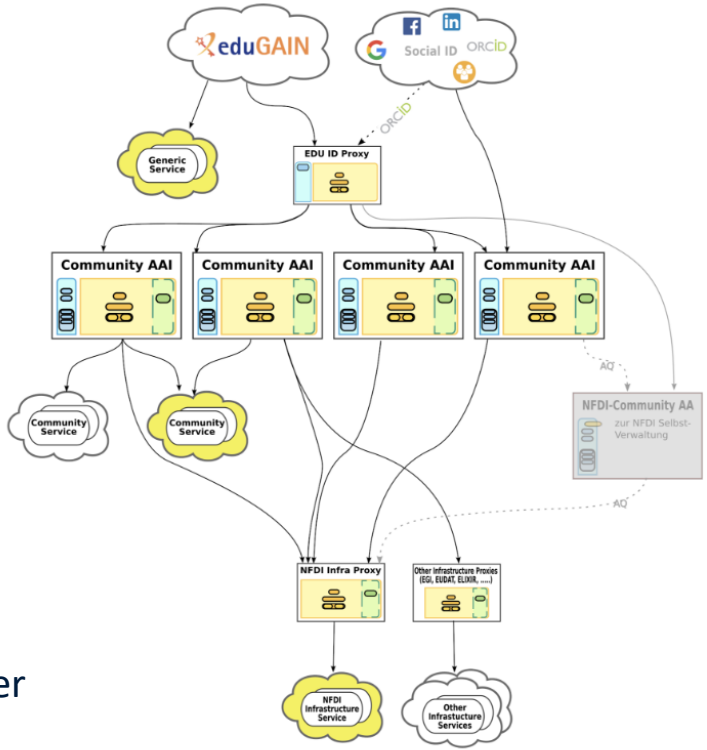


AARC Blueprint Architecture (2019) AARC-G045 https://aarc-community.org/guidelines/aarc-g045/; stacked proxies: EOSC AAI Architecture
EOSC Authentication and Authorization Infrastructure (AAI), ISBN 978-92-76-28113-9, http://doi.org/10.2777/8702

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 career



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

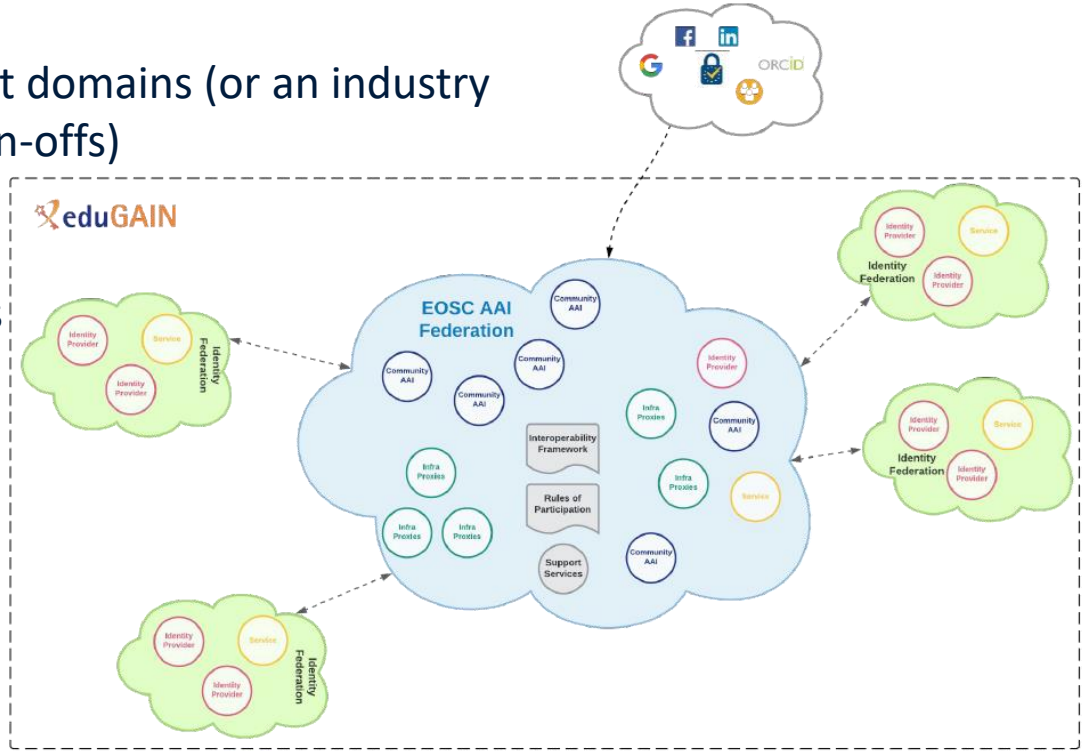
When many proxies from different groups come together

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

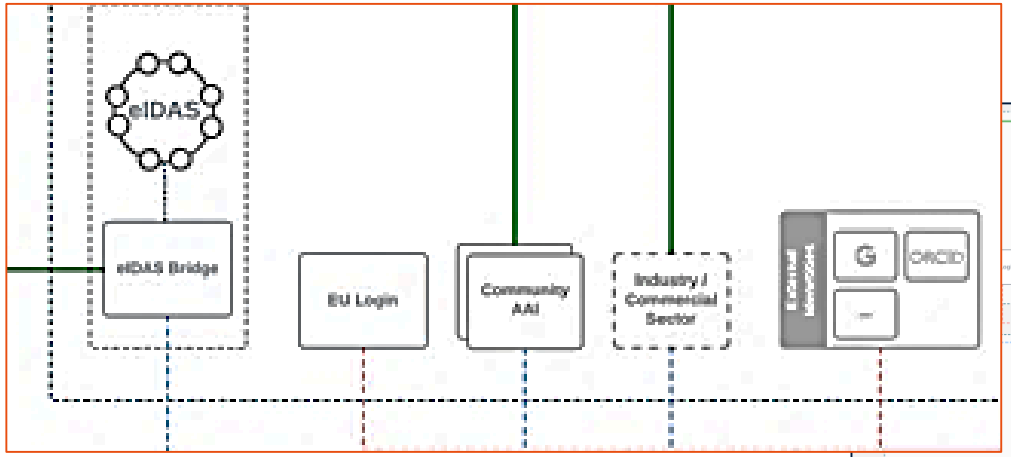
European Open Science Cloud (EOSC) AAI based on federations and proxies



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

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

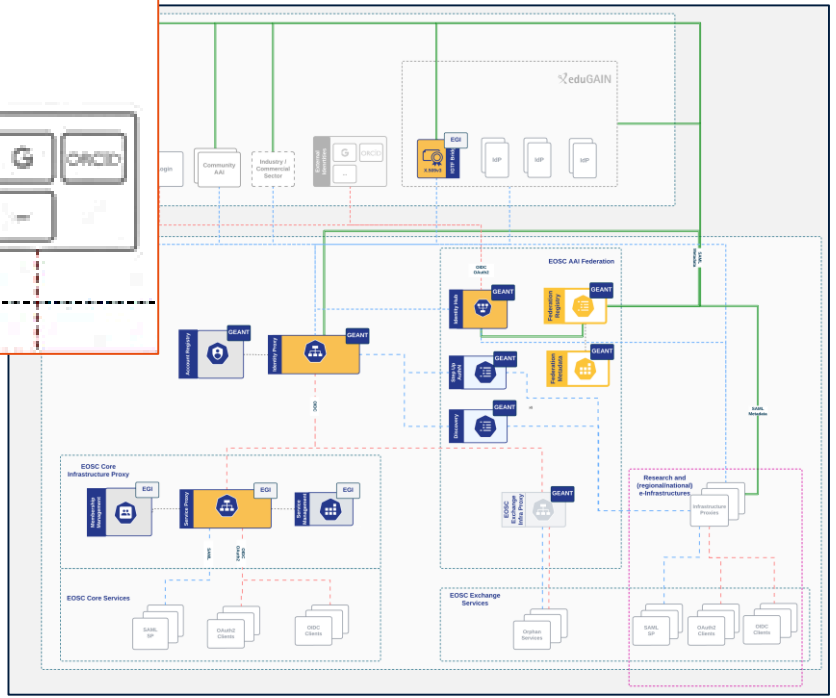
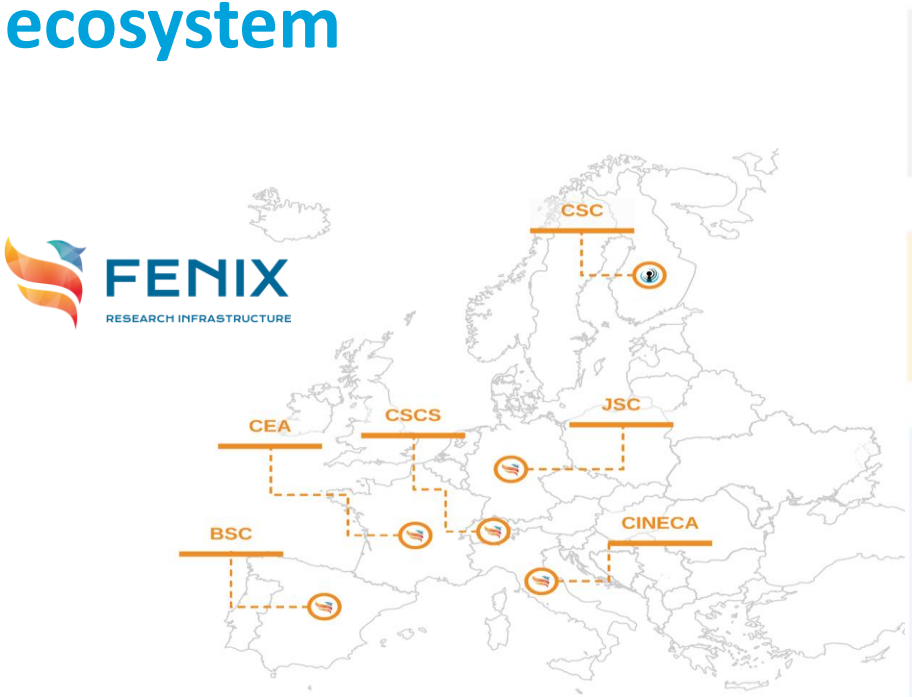
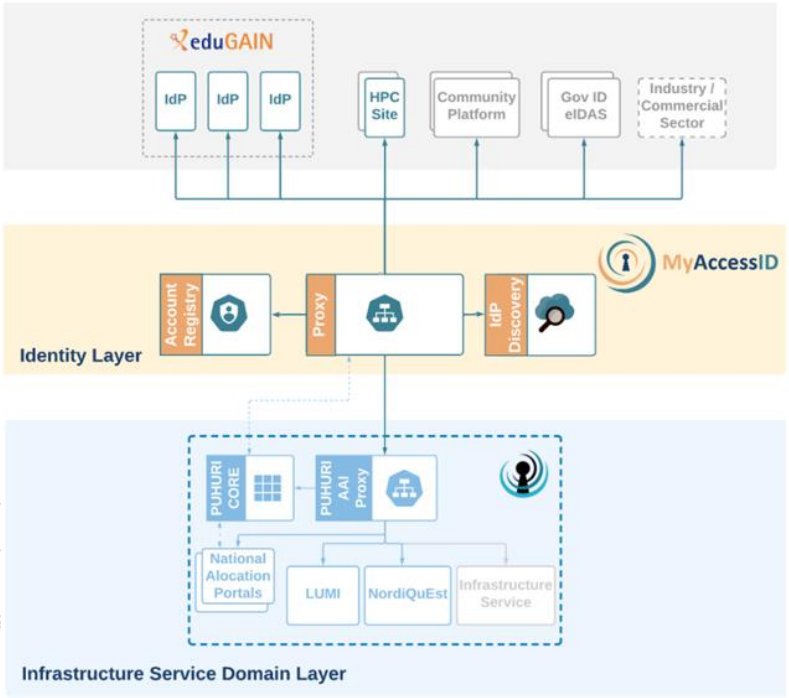


Image: EOSC AAI for the EOSC Core and Exchange Federation for the EOSC European Node by Christos Kanellopoulos, Nicolas Liampotis, David Groep (June 2023)

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



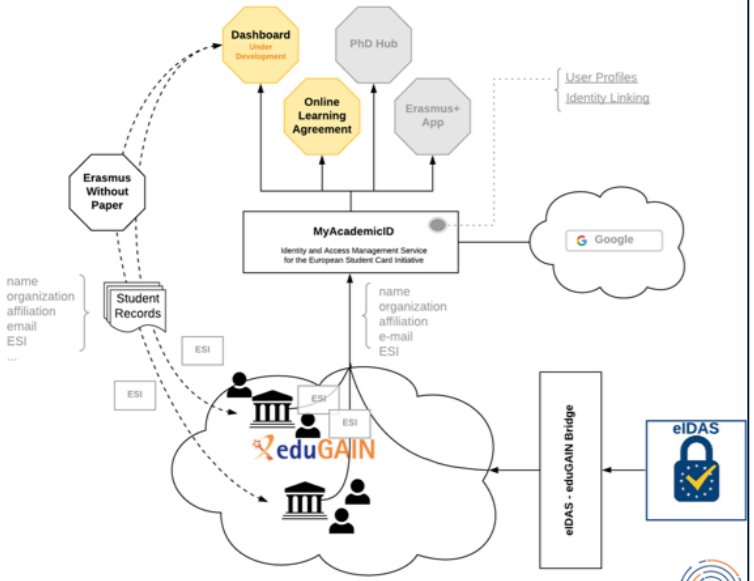


Fenix image via Christos Kanellopoulos, diagram via Anders Sjöström (NeIC, Puhuri) at the TNC23 workshop

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

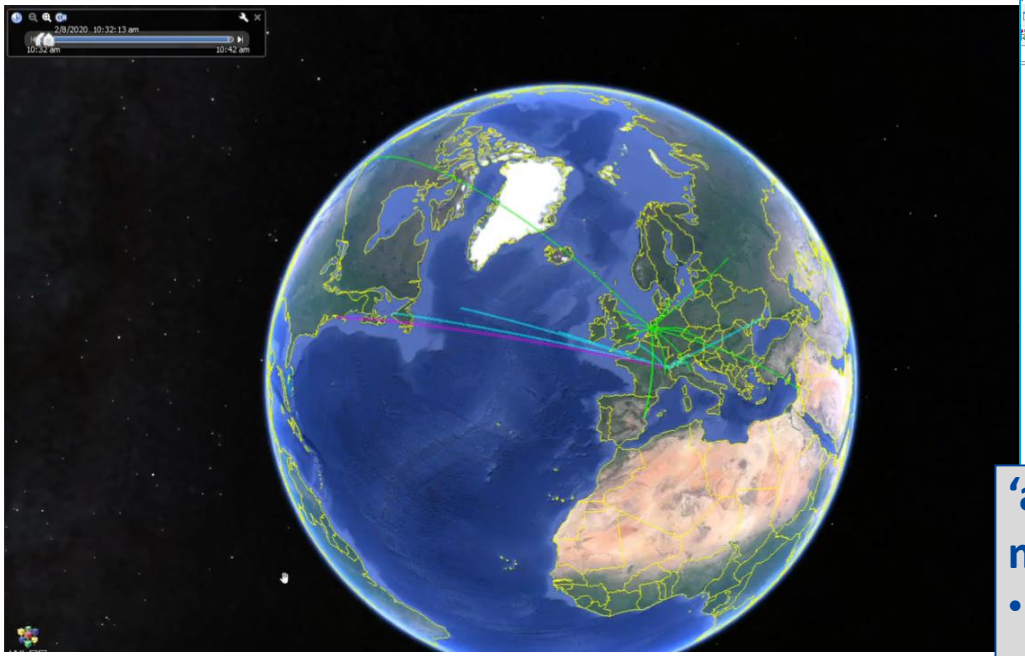


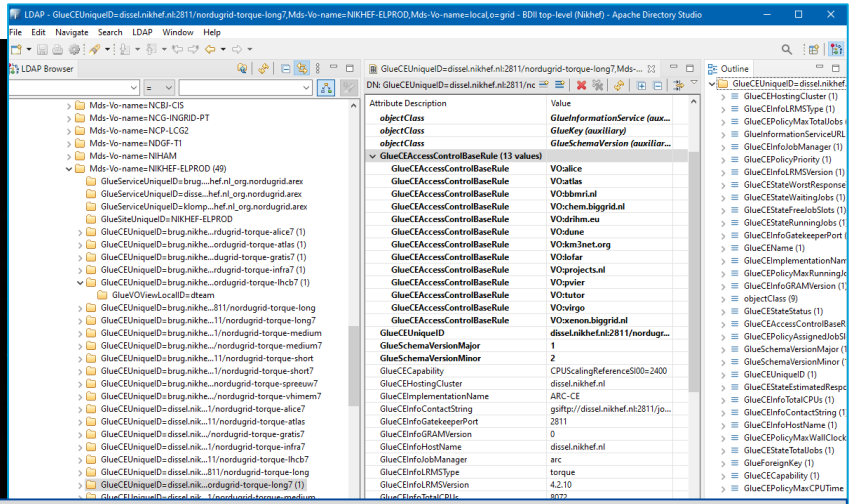
MyAcademicID

Putting it back together again

Common patterns in scalability

A global infrastructure of EGI, OSG and WLCG, …



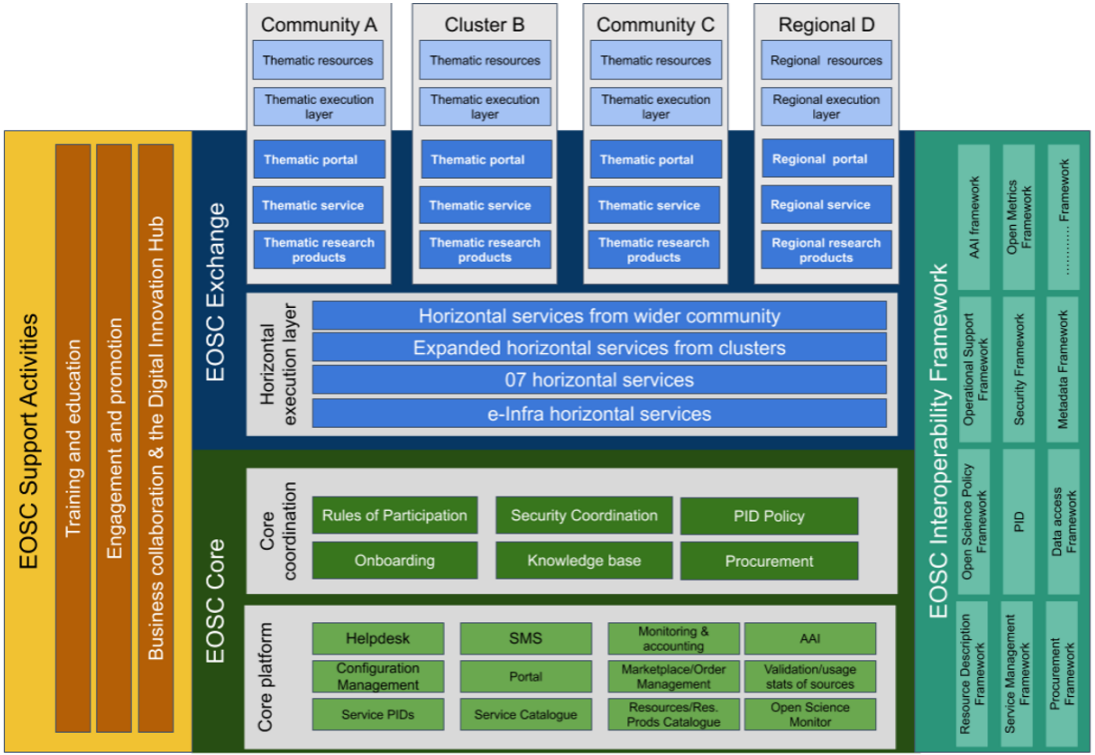


'an infrastructure with components matched to application need'

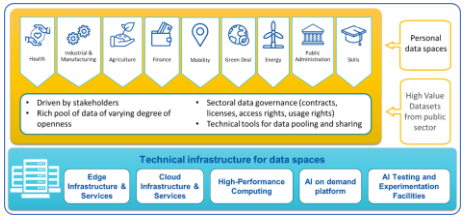
- systems architecture: compute (HTC clusters), networking, storage, and application structure
- in a balanced and {energy,cost}-efficient setup

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, …

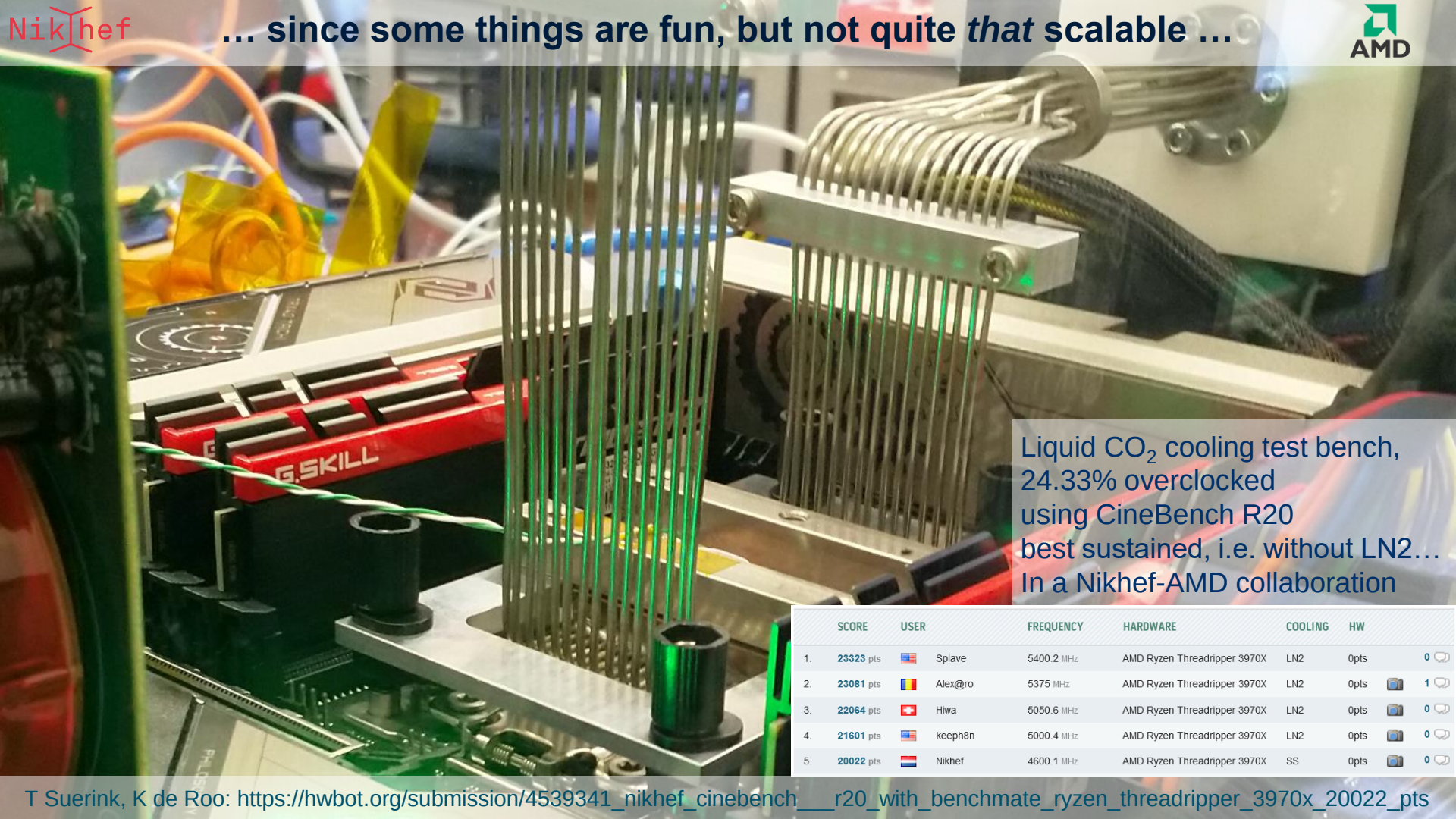


EOSC: https://eoscfuture.eu/wp-content/uploads/2022/04/EOSC-Core.pdf; data spaces image: https://digital-strategy.ec.europa.eu/en/library/building-data-economy-brochure

Did you discern 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
- Make central components passive and stateless (as possible) to allow scaling
  - 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

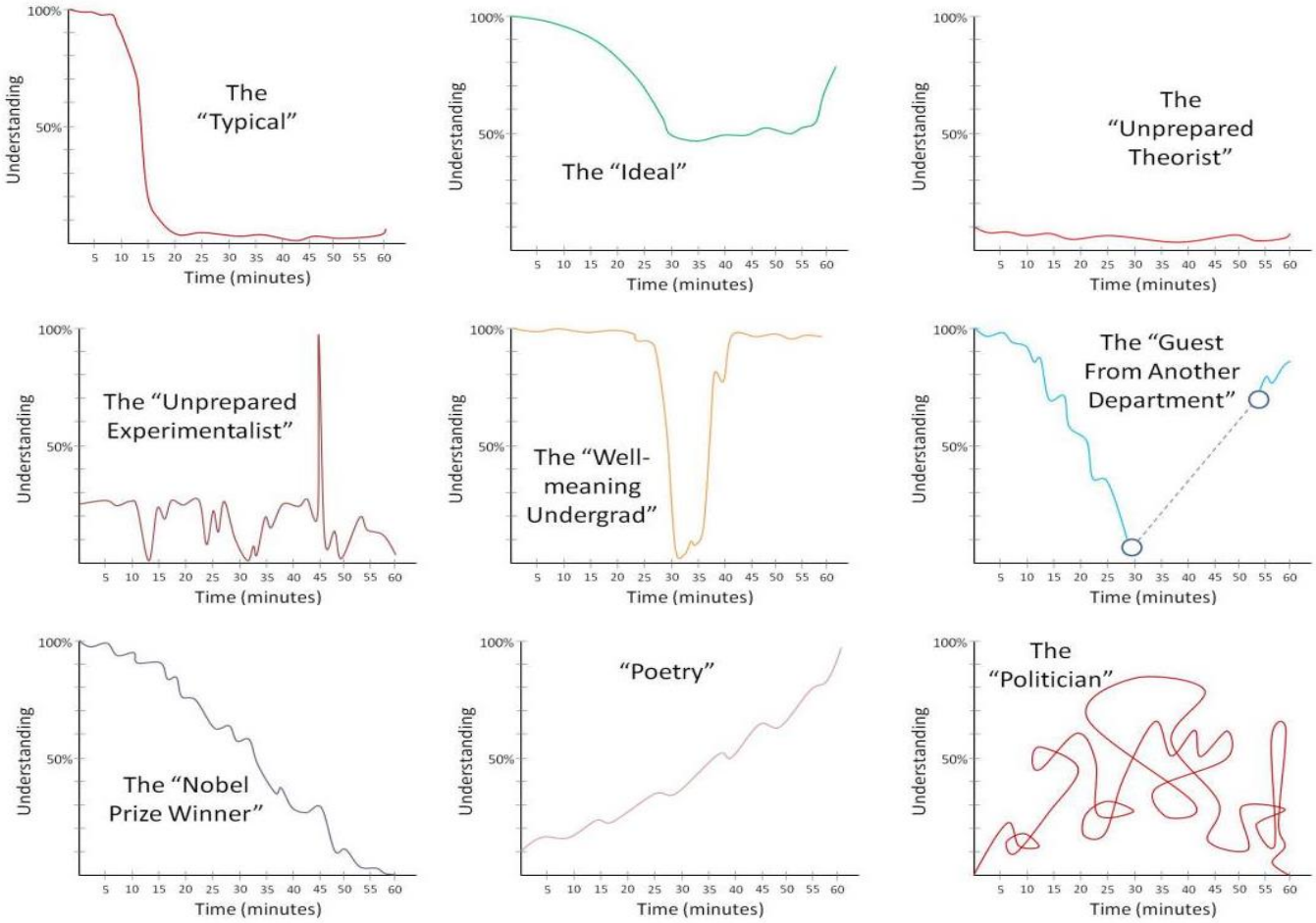
… since some things are fun, but not quite *that* scalable …



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

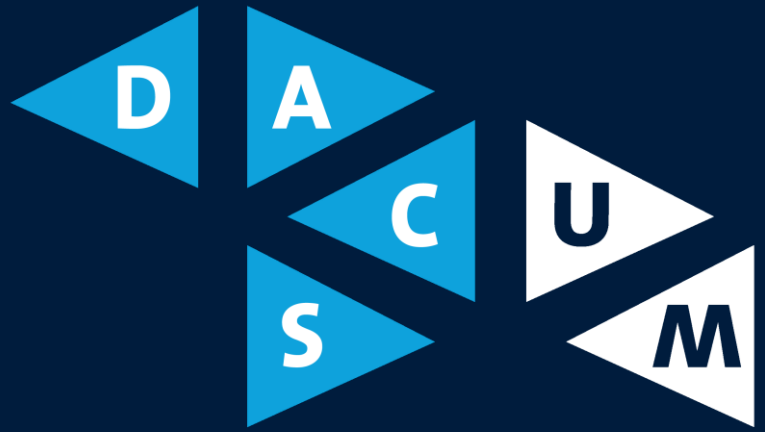


http://manyworldstheory.com/2013/10/03/the-9-kinds-of-physics-seminar/

More Q&A time!

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

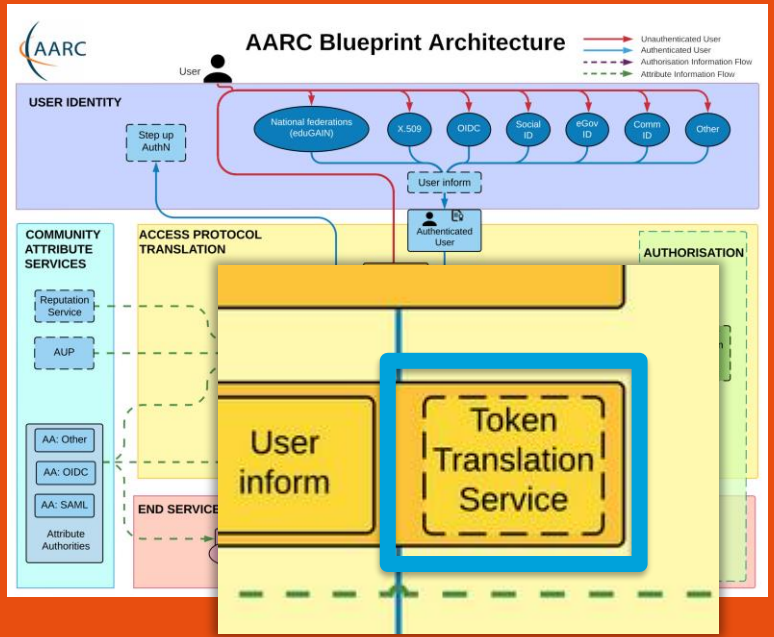
Nikhef



Maastricht University | Department of Advanced Computing Sciences

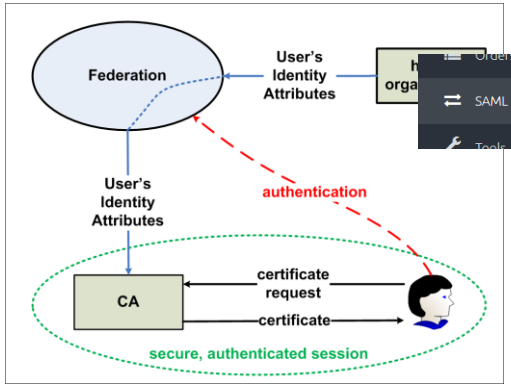
Distributed collaborative services
a more technical example with RCauth.eu

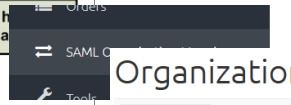
Credential translation in the AARC BPA
… building RCauth.eu
Leveraging federation and collaboration
for ubiquitous research credentials

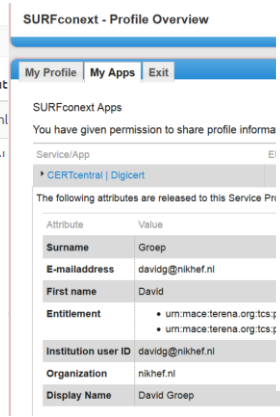


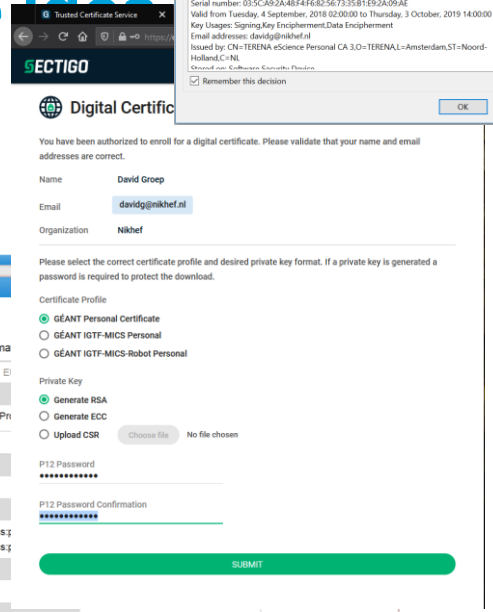
Bridges and Token Translation Services

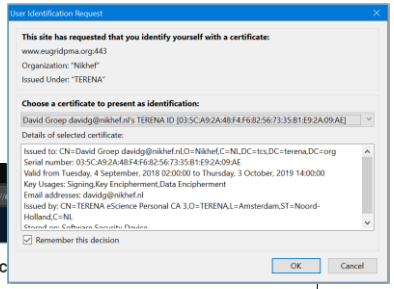
TCS - for users that manage to grasp the idea











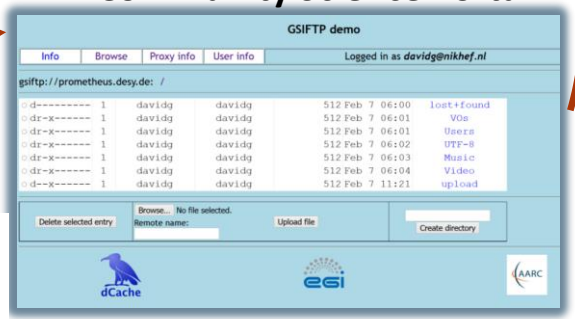
TCS is a SAML Service Provider (today by Sectigo)
to eduGAIN: where eligible authenticated users obtain client certificates for access to many research services
A globally recognized identity for all employees & students (they are automatically eligible!).

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

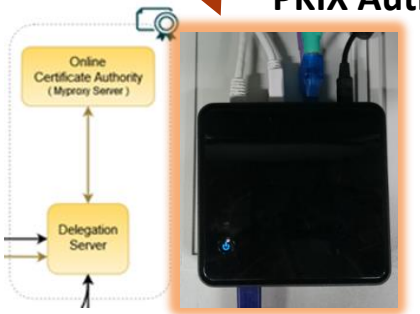
user facing | hidden back-end

Community Science Portal



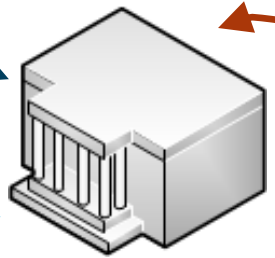
Infrastructure Master Portal Credential Store

IGTF accredited PKIX Authority

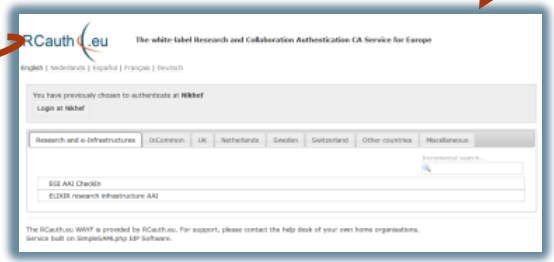


User Home Org
or Infrastructure IdP

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



REFEDS R&S
Sirtfi Trust



Policy Filtering WAYF to eduGAIN

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*

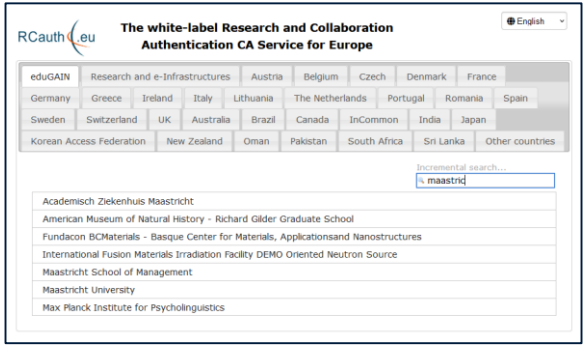
- **eduPersonPrincipalName** – scoped point-in-time unique identifier, which could be, but usually is not, privacy preserving: “davidg@nikhef.nl”, “P70081609@maastrichtuniversity.nl”
- **eduPersonTargetedID** – scoped transient non-reassigned identifier, like urn:geant:nikhef.nl:nikidm:idp:sso!*27c8d63ed42c84af2875e2984*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



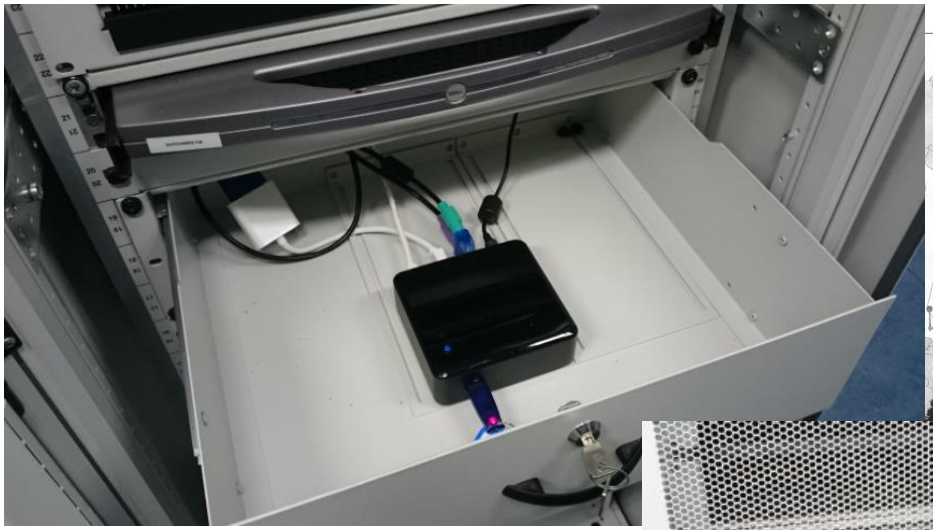
Parsed ID Token:

```
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
```

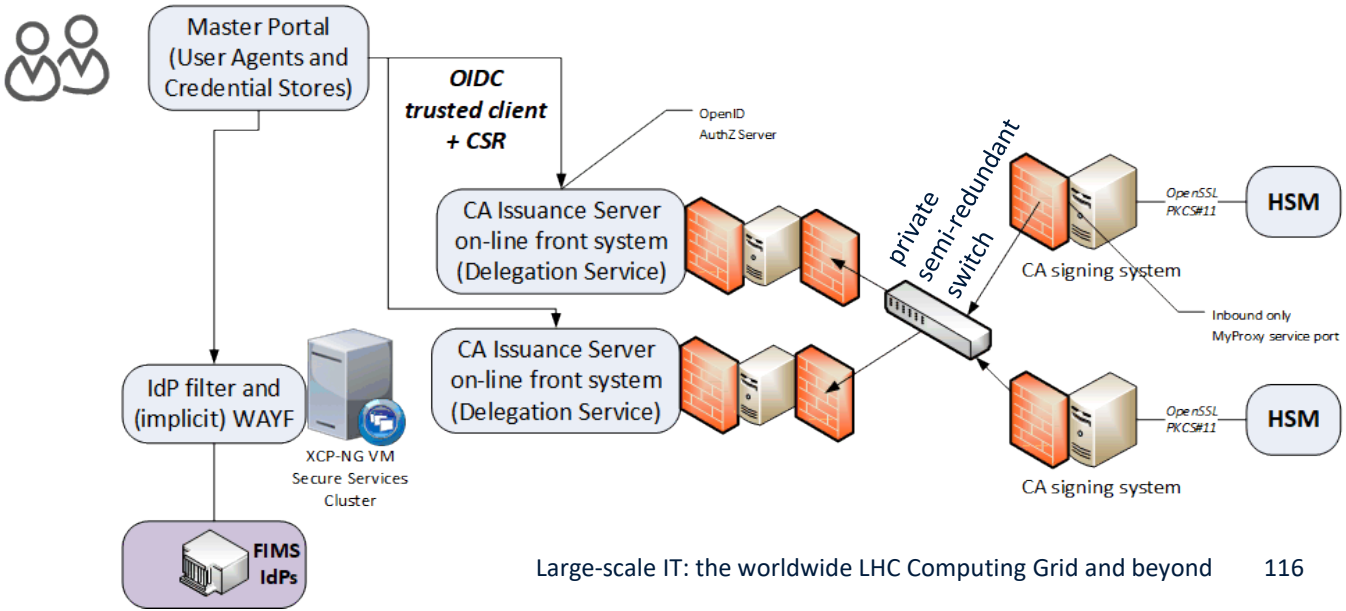
With a single, yet fully compliant, ‘Heath Robinson’ CA



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)

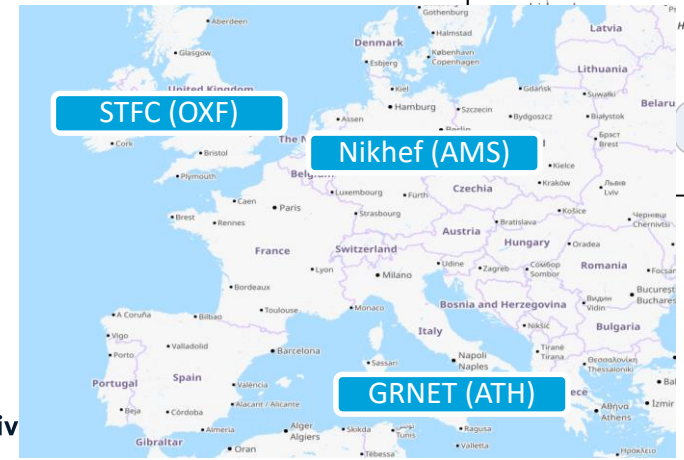
Eliminated SPOFs first using ‘local HA’

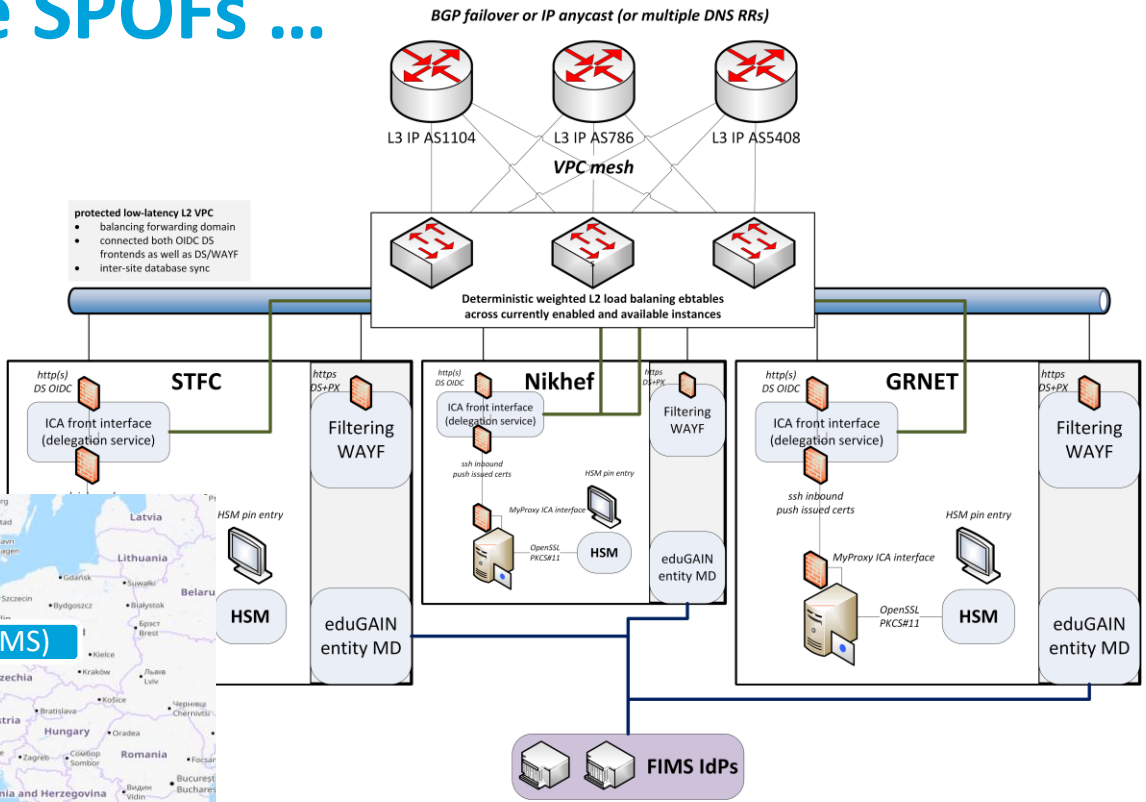


Since we do not like SPOFs …

Distributed High Availability setup

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



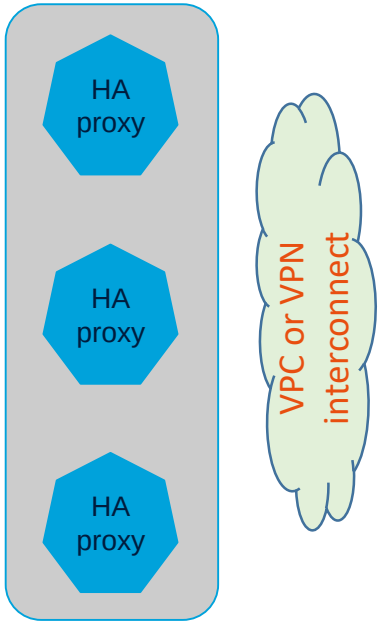


work supported by the EOSC Hub and EOSC Future Horizon Europe projects

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

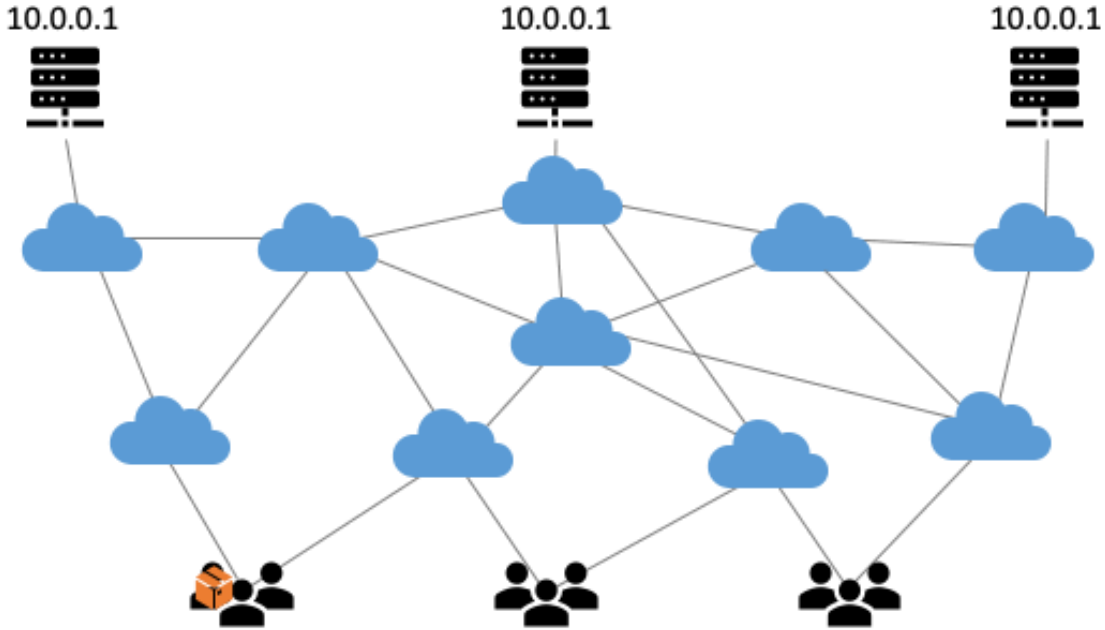
Straightforward proven solution is IP anycast

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



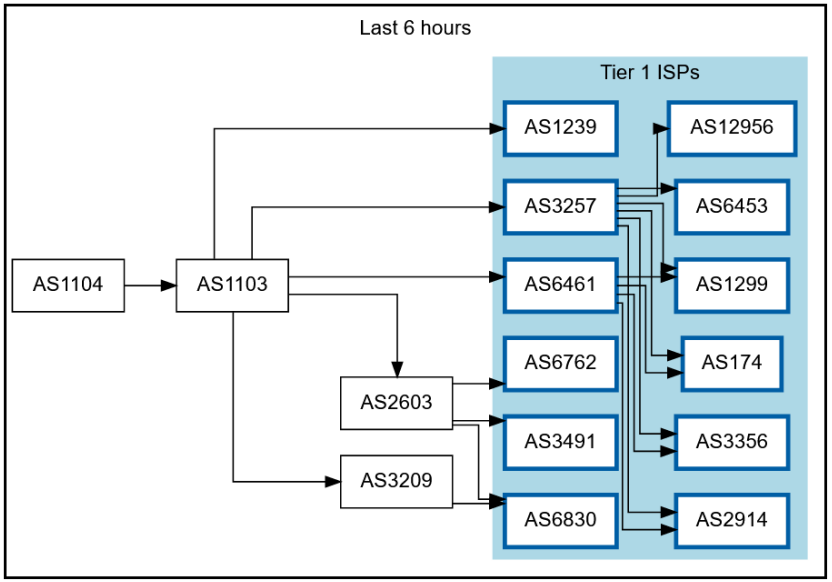
So we used

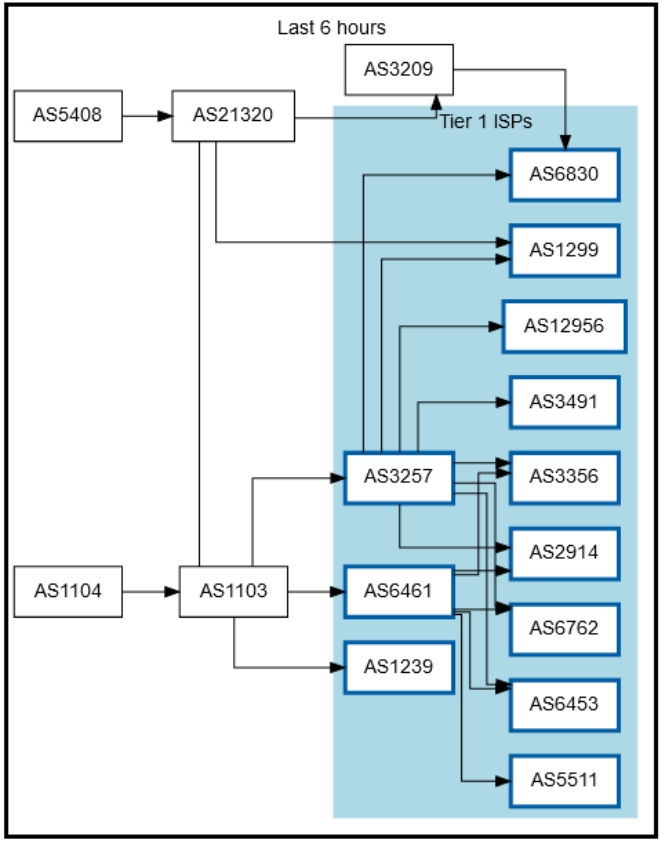
- 3 (for now: 2) sites
- one VM at each site exposing 2a07:8504:01a0::1
- smallest v6 subnet (/48)
- bird + a service probe
- each site's own ASN
- some IRR DB editing
- IPv4 is similar, with a /24

and some monitoring

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

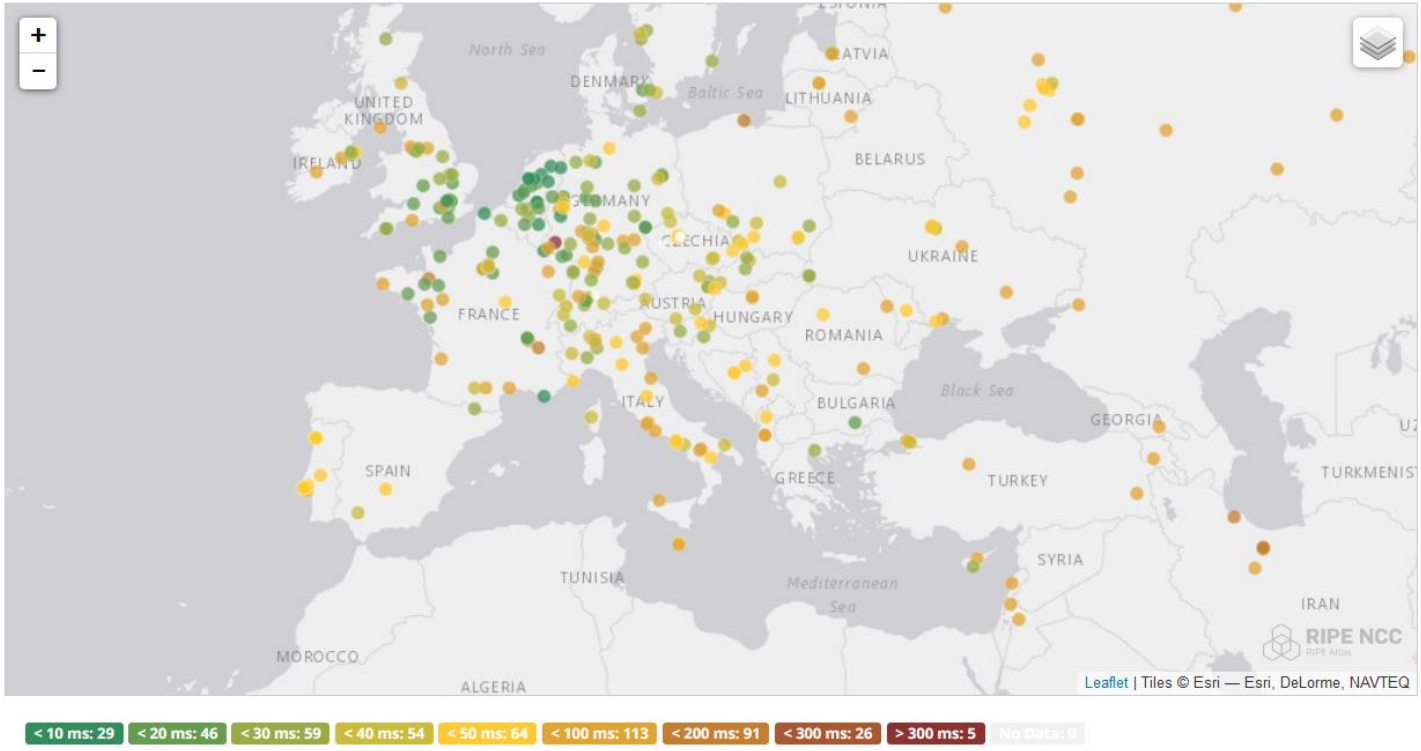
Getting 2a07:8504:1a0::/48 out there





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

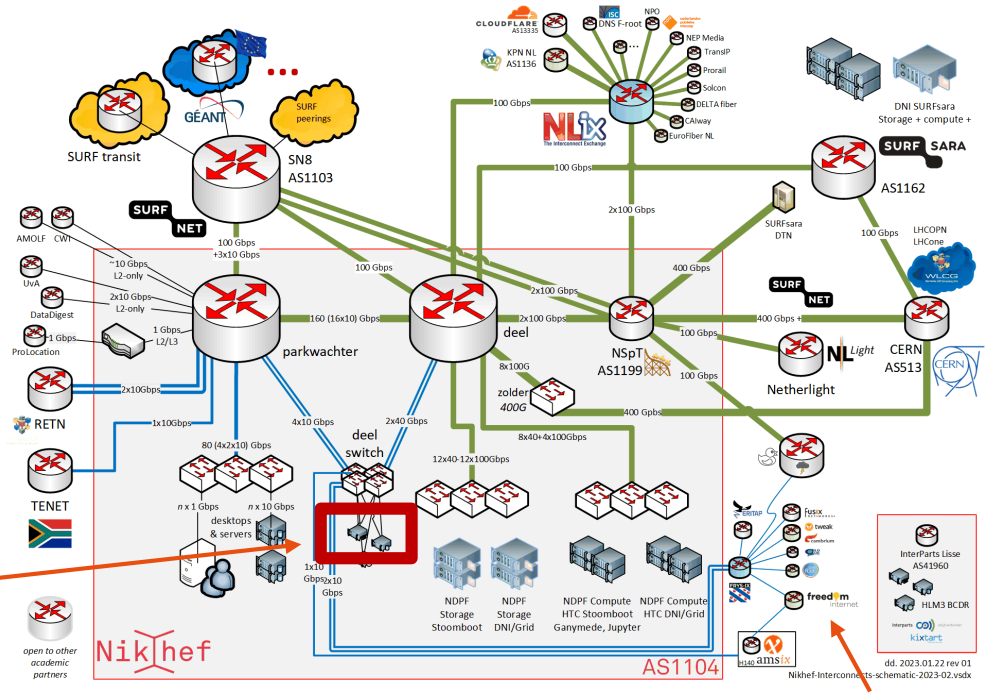
And you get reasonable load balancing in Europe for free



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

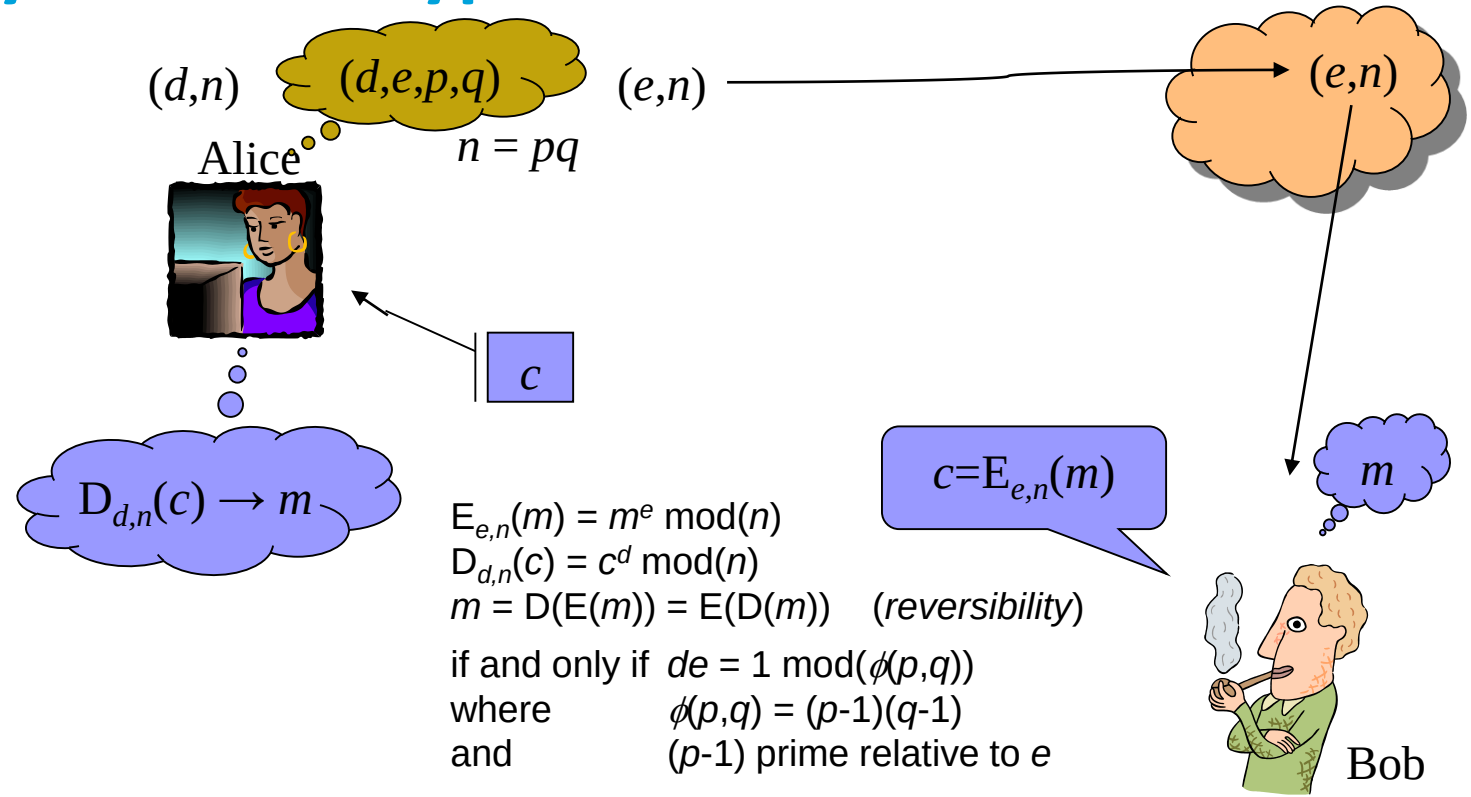
- pre-shared secrets, may be salted hashed on service side
- required: secure one-way hash function
- need a **protected channel** between identifiable end-points

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?



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

if and only if $de = 1 \bmod(\phi(p,q))$
where $\phi(p,q) = (p-1)(q-1)$
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)]

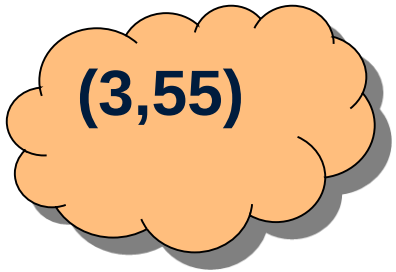
- Public Key: **(3,55)**
- Private Key: **(27,55)**

$E_{e,n}(m) = m^e \bmod(n)$
$D_{d,n}(c) = c^d \bmod(n)$
$m = D(E(m)) = E(D(m))$
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)**
- $c=E_{\mathbf{3;55}}(18)=18^3 \bmod(55) = 5832 \bmod(55) = \mathbf{2}$
- send message **"2"**



Decryption:

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

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

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

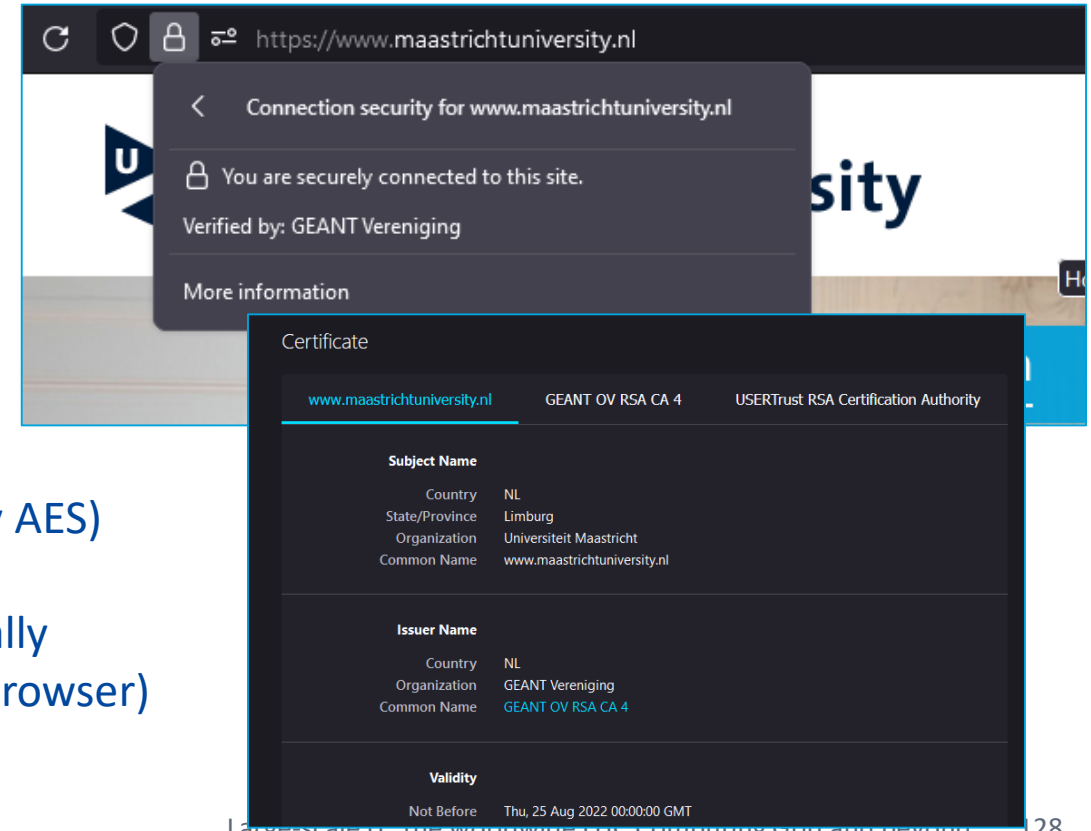
but also: the maximum plaintext is limited by the modulus length

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)



Other ancillary materials

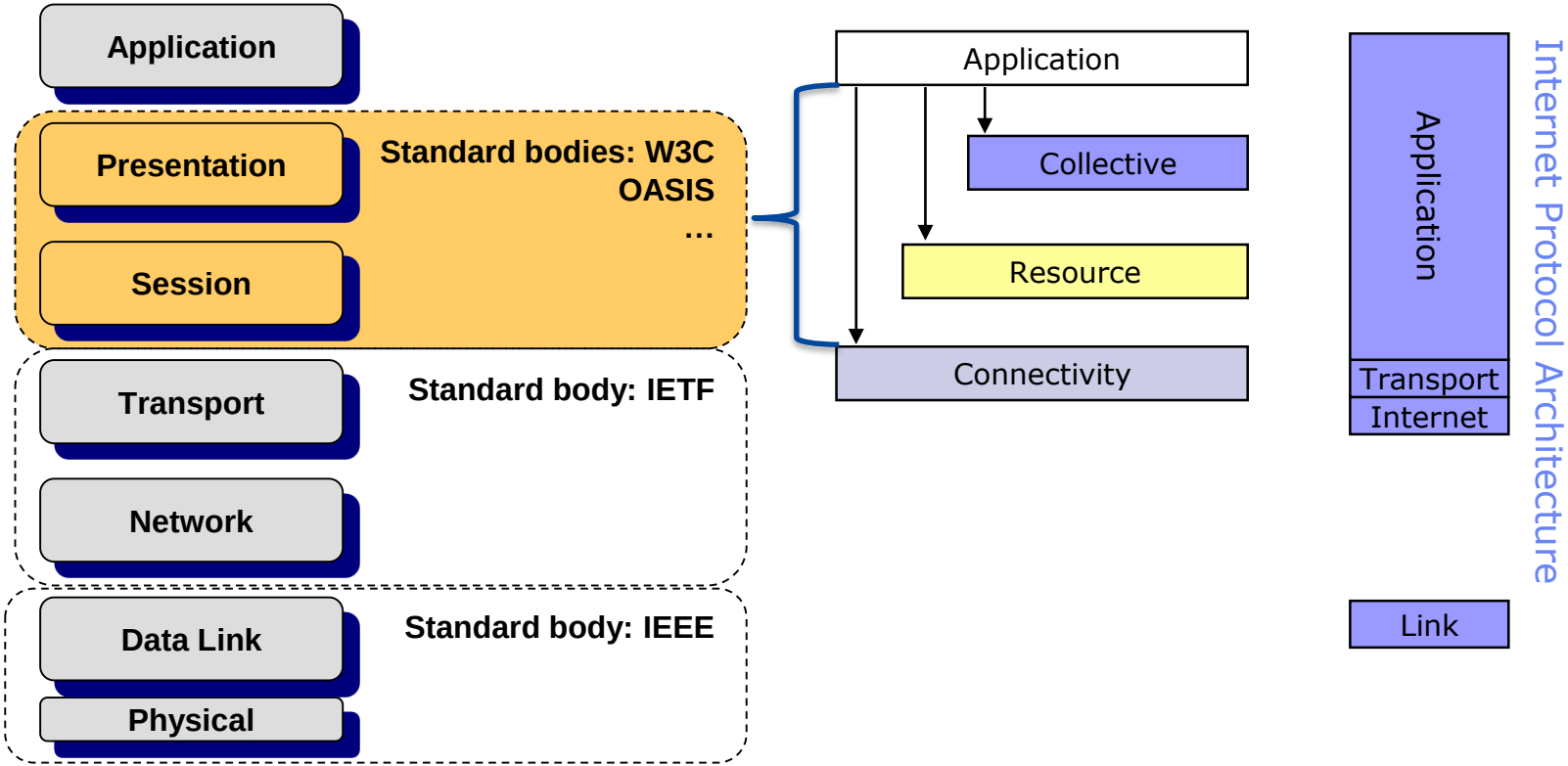
these generic slides do not form part of the module, but are just general background knowledge and example

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

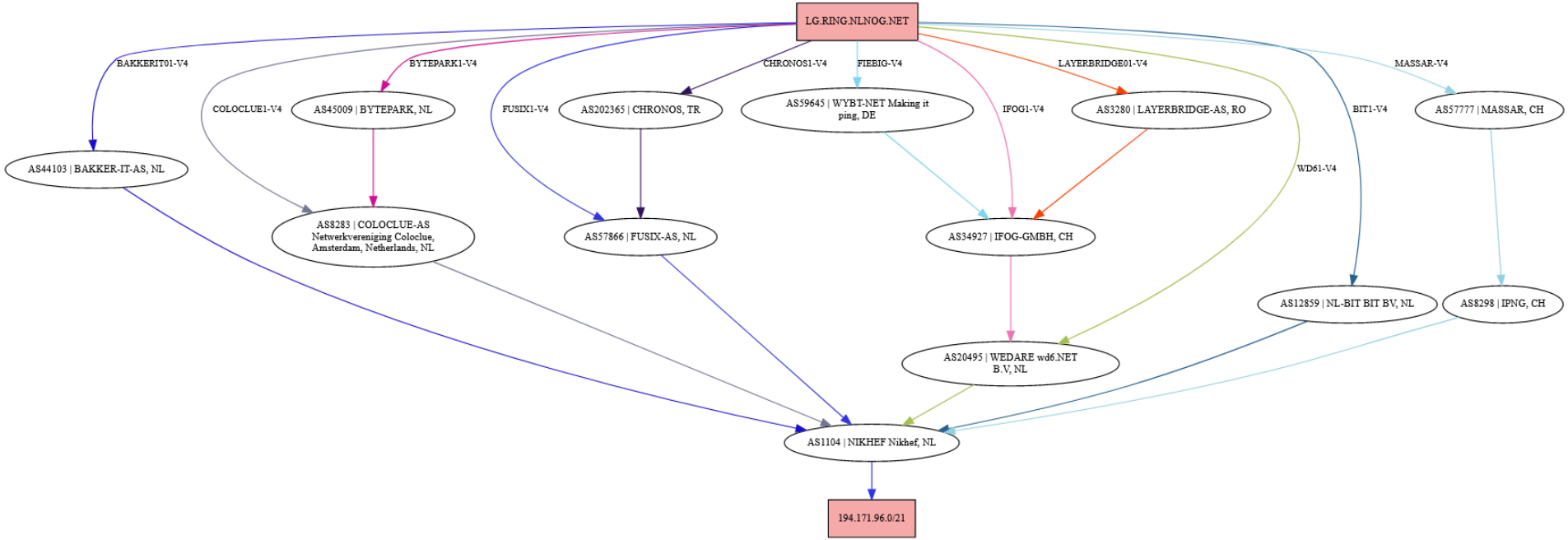
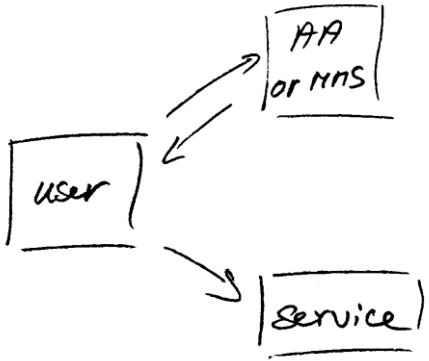
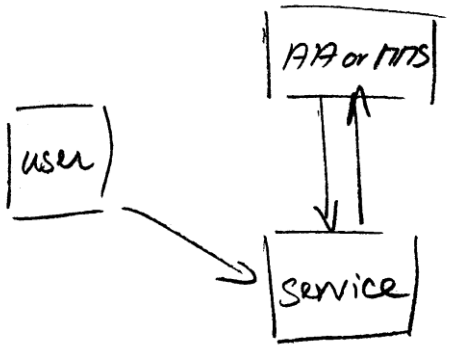


Image sources: NLNOG RING map https://lg.ring.nlnog.net/

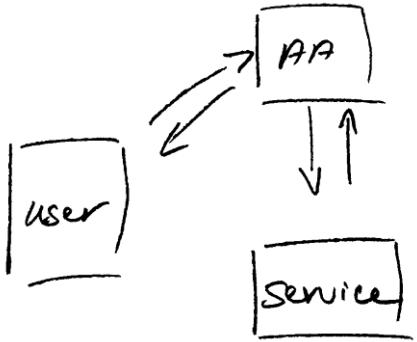
RFC2904 authorization models: three AuthZ flows



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

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.”