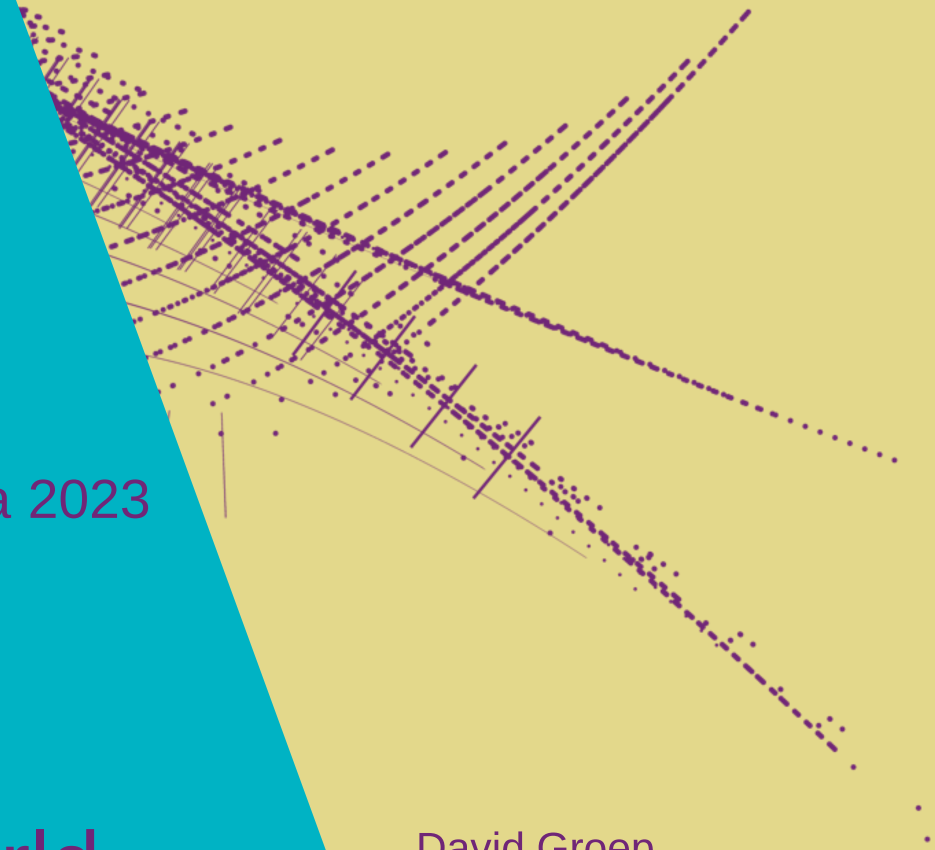
Nikhef

Maastricht University

Ignatius caput selectum deeltjesfysica 2023

Computing for (astro)particle physics at Nikhef and in the world



David Groep
March 2023

Data at the Large Hadron Collider at CERN

1964

Volume 13, Number 16 PHYSICAL REVIEW LETTERS 19 October 1964

BROKEN SYMMETRIES AND THE MASSES OF GAUGE BOSONS

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

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

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

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

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

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

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

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

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

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

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

into the form

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

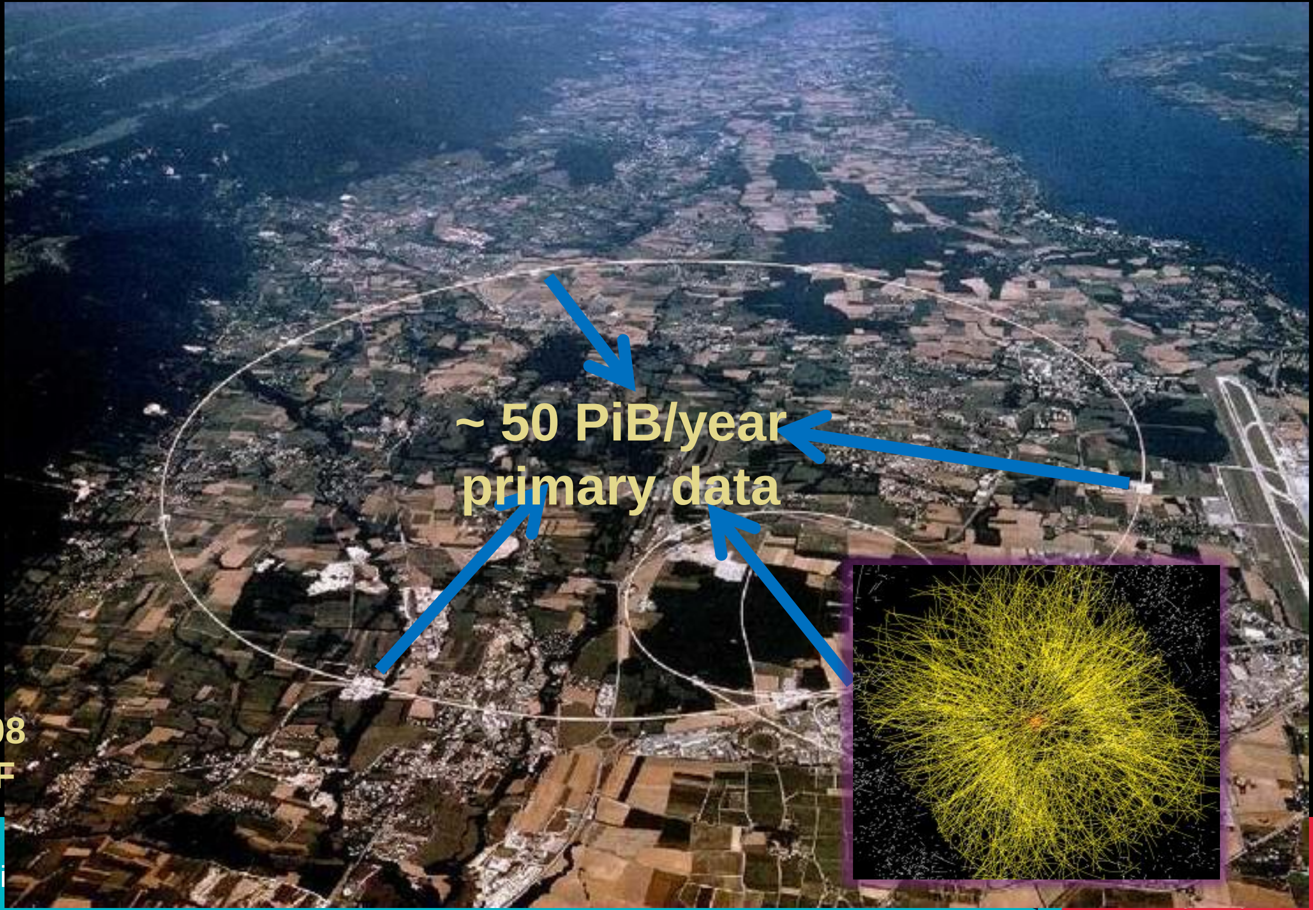
Equation (4) describes vector waves whose quanta have (bare) mass $e\varphi_0$. In the absence of the gauge field coupling ($e=0$) the situation is quite different: Equations (2a) and (2c) describe zero-mass scalar and vector bosons, respectively. In passing, we note that the right-hand side of (2c) 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.[5]

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

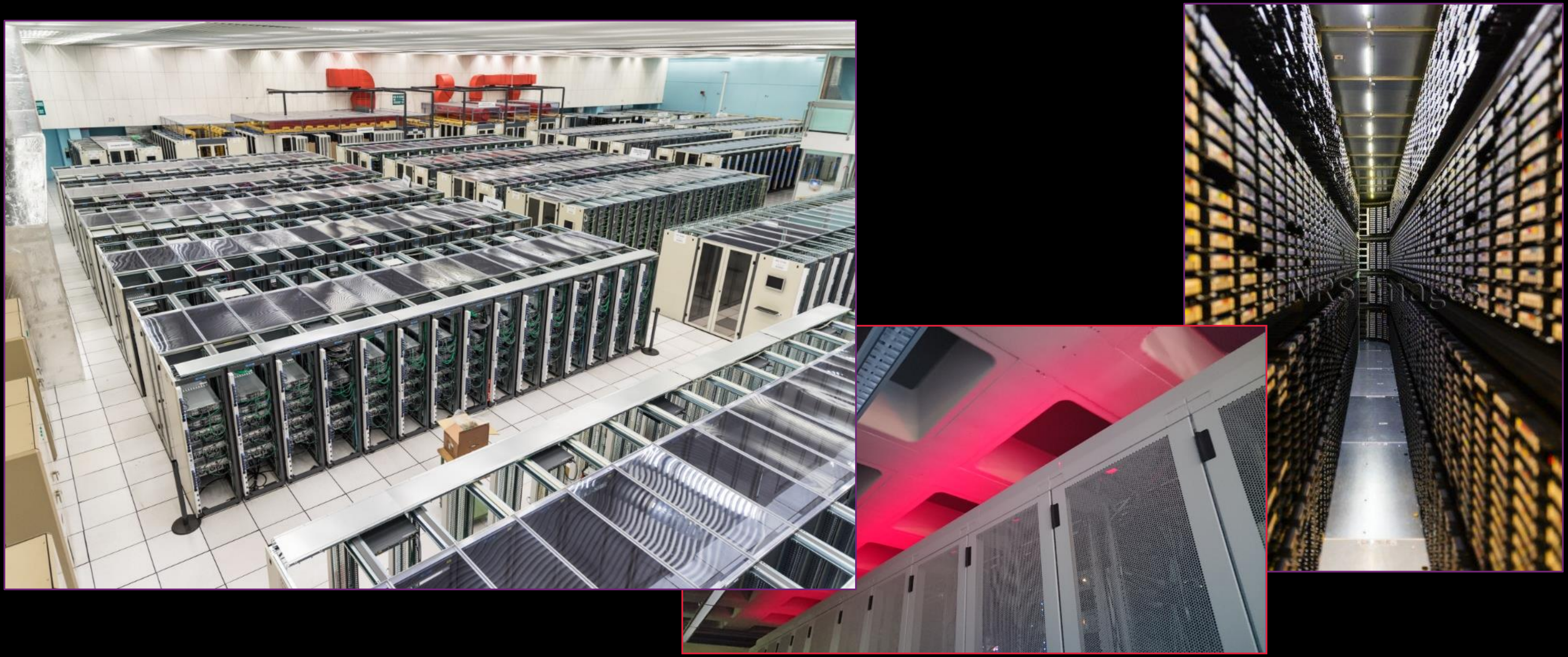
508

321

P. Higgs, Phys. Rev. Lett. 13, 508
16823 characters, 165kByte PDF

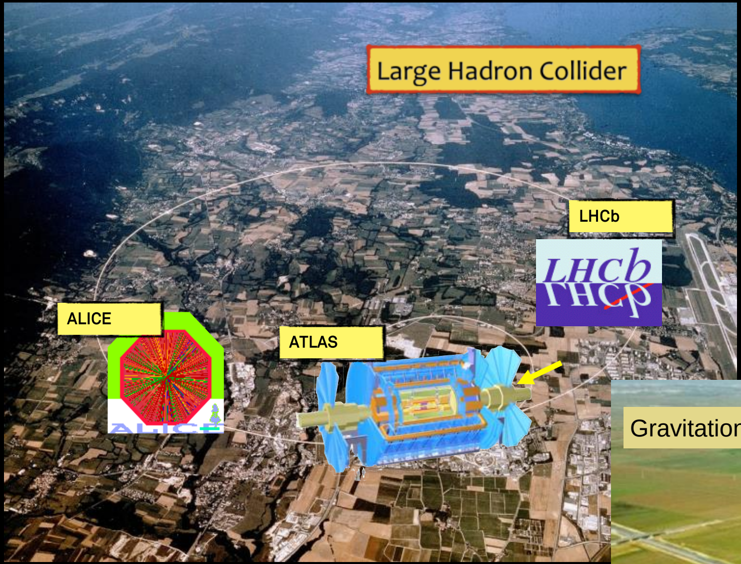


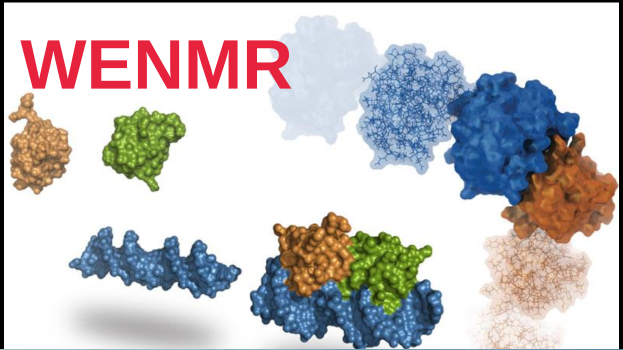
‘Big Science’ needs some computing …



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

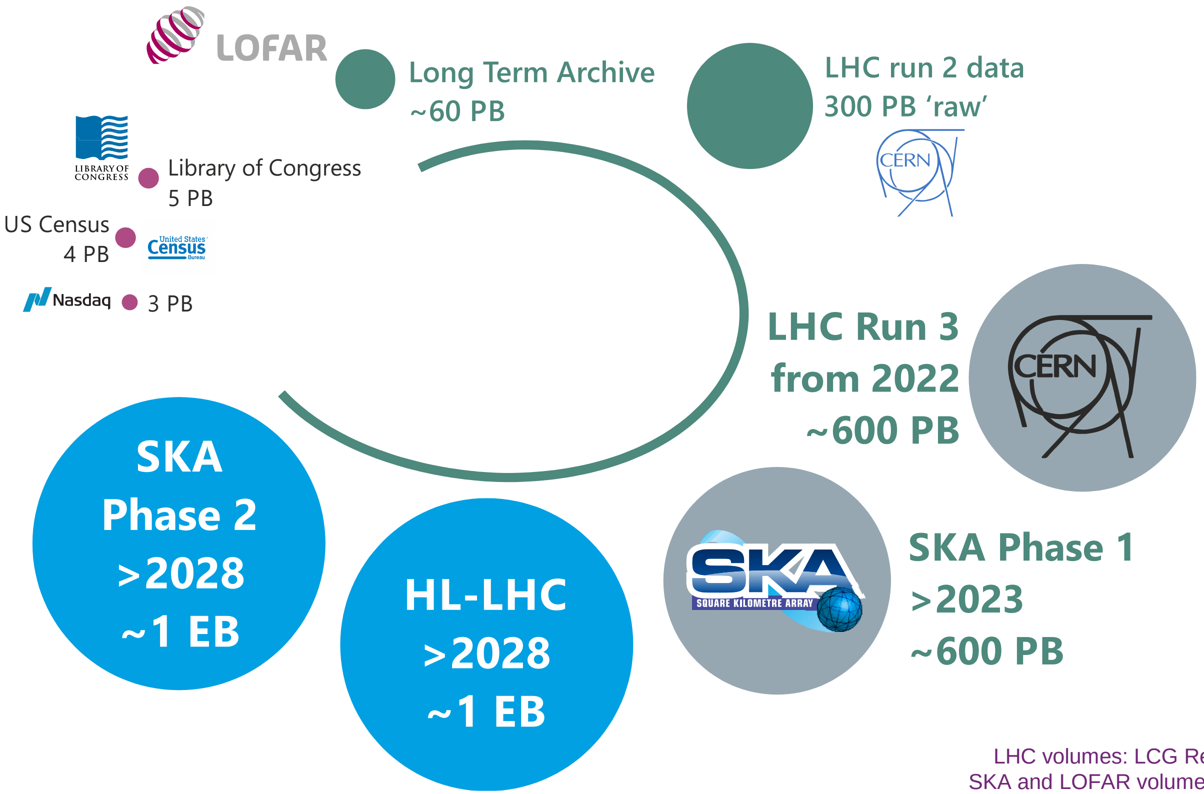
Larger scales for both facilities and computing





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/

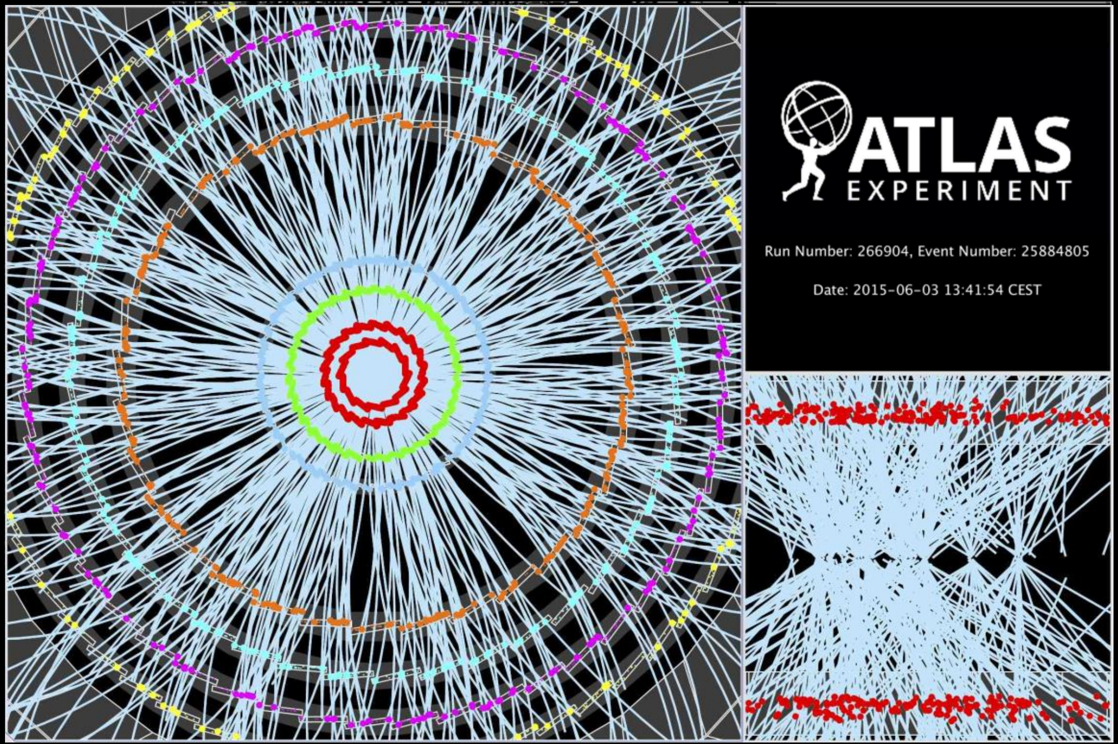
More data is coming!



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

Computing on lots of data – 40 Mevents/sec

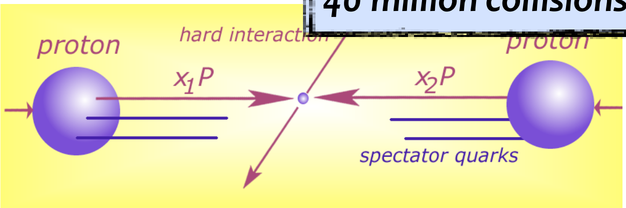
~ 10 seconds to compute a single event at ATLAS for ‘jets’ containing ~30 collisions

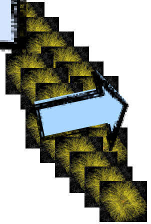


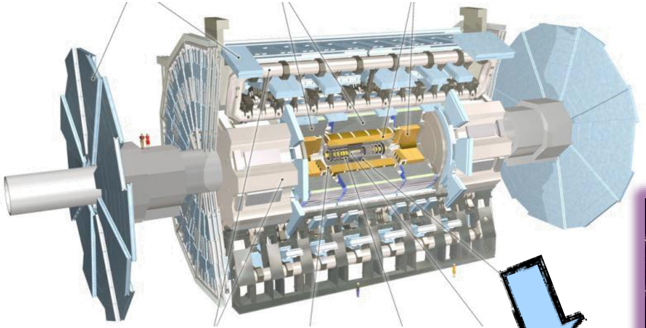
Display of a proton-proton collision event recorded by ATLAS on 3 June 2015, with the first LHC stable beams at a collision energy of 13 TeV; Event processing time: v19.0.1.1 as per Jovan Mitrevski and 2015 J. Phys.: Conf. Ser. 664 072034 (CHEP2015)

Detector to doctor workflow

40 million collisions / second







Trigger system selects 600 Hz ~ 1 GB/s data

Classify particles in collision and their physics properties:
- *electrons*
- *muons*
- *jets consisting of hadrons*
- *...*

Physics analysis by (PhD) students, in papers & analysis notes

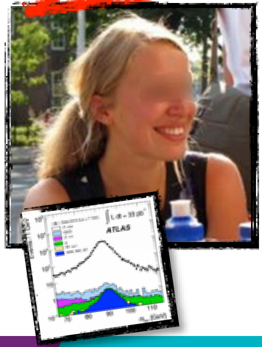


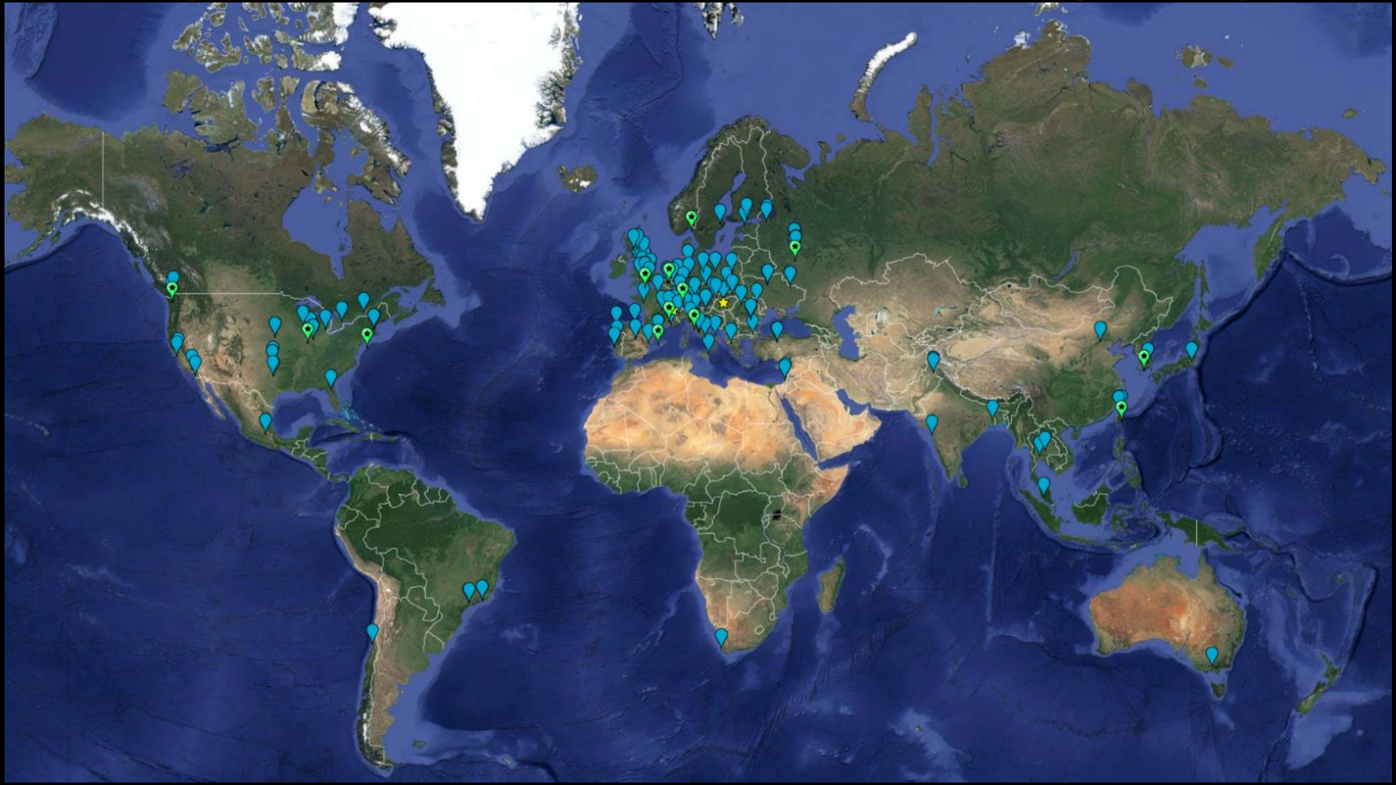
diagram adapted from Frank Linde; images: ATLAS collaboration, Nikhef. … and sorry for the GDPR-blur

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

Example: 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

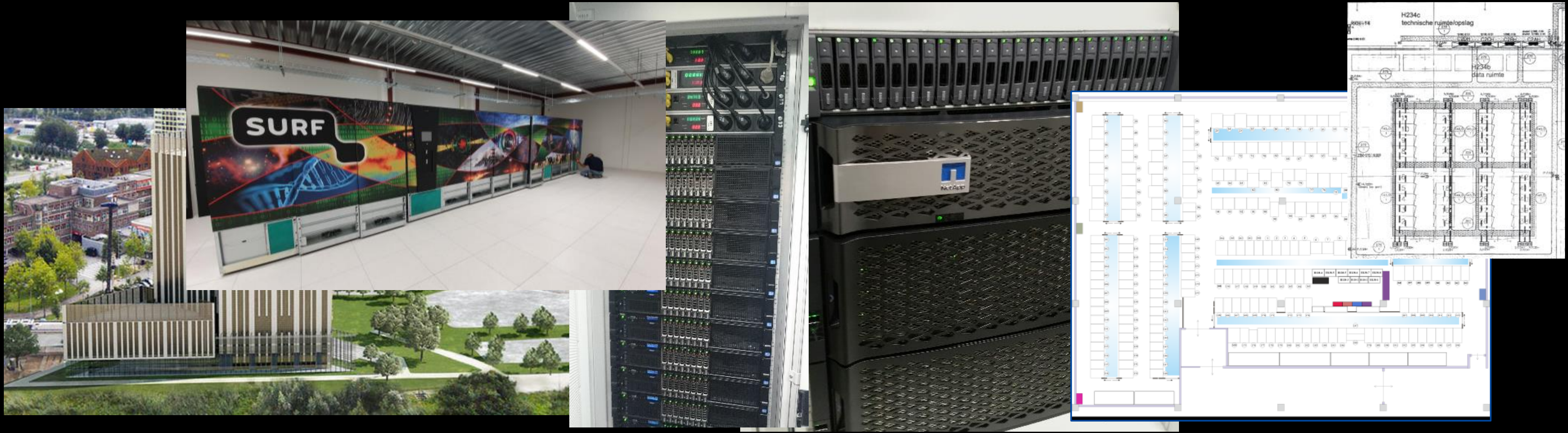
e-Infrastructures
EGI
PRACE-RI
EuroHPC
OpenScienceGrid
XSEDE (ACCESS)

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

Global distribution of computing and data placement

WLCG NL-T1 and the Dutch National 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/

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 on die
Intel Cascade Lake AP looked like a cludge
but now Sapphire Rapids appears better again

CPU design-level performance gains left

predictive execution
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

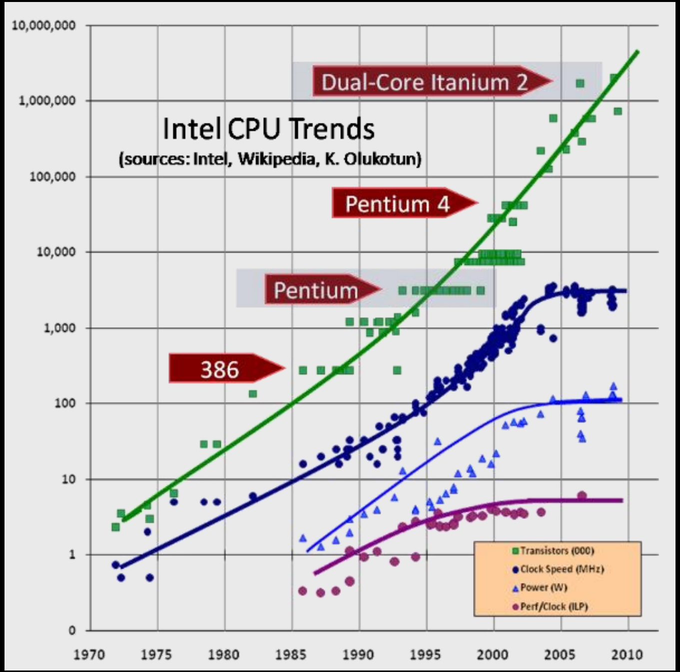
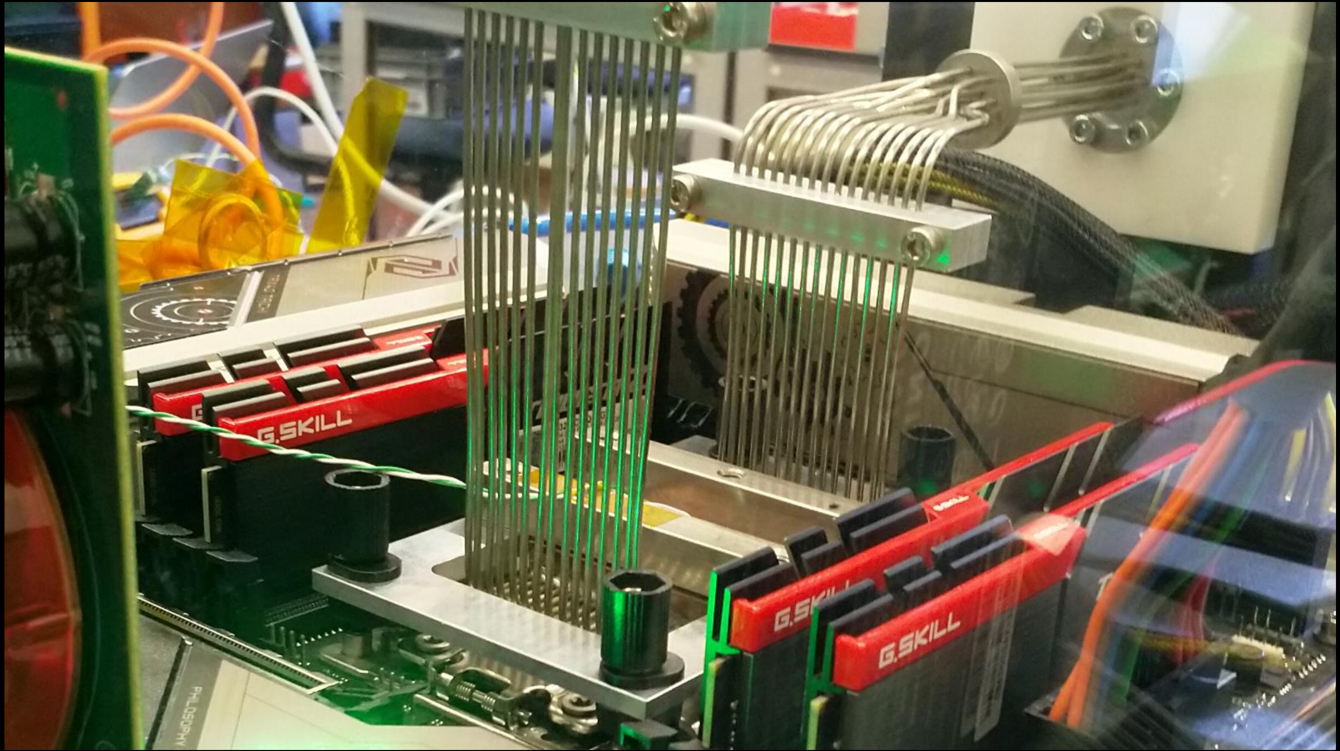


Image: Herb Sutter, *Dr.Dobbs Journal* 2004, updated 2009, see http://www.gotw.ca/publications/concurrency-ddj.htm

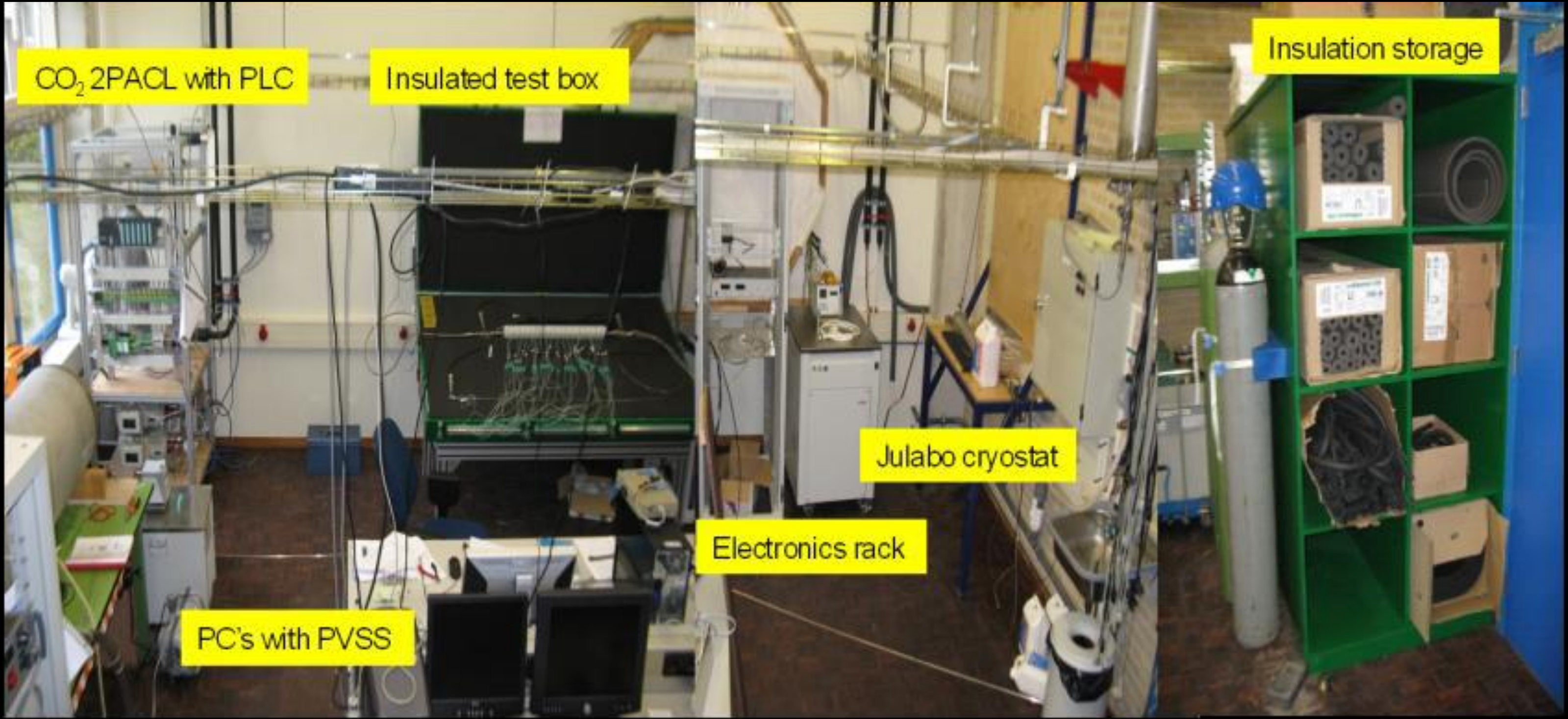
Nikhef

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



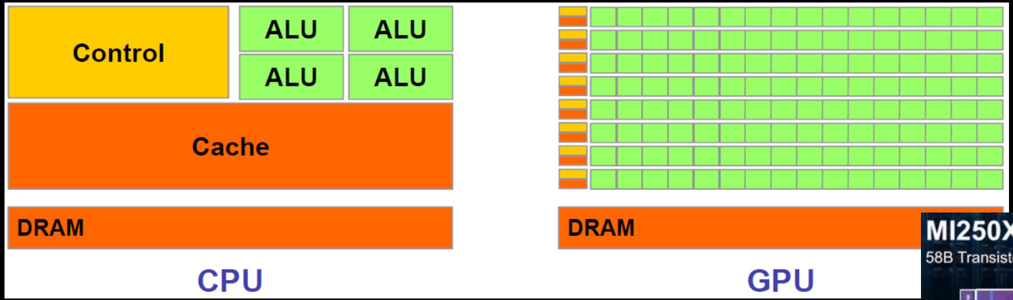
LCO2 cooling of an AMD Ryzen Threadripper 3970X [56.38 °C] at 4600.1MHz processor (~1.5x nominal speed) sustained, using the Nikhef LCO2 test bench system (https://hwbot.org/submission/4539341) - (Krista de Roo en Tristan Suerink)

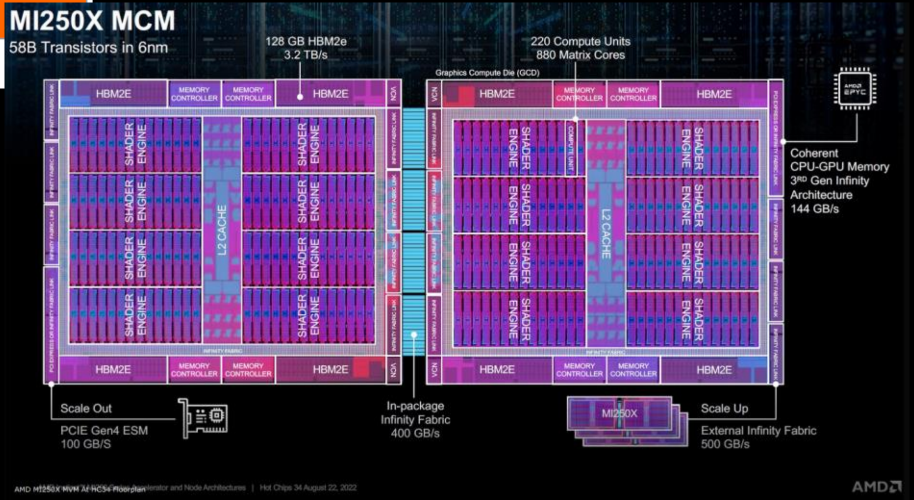
… since you then need this around it …



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

Accelerators – general purpose GPUs





- 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

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

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

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

Scaling up – beyond one lone motherboard

Physical farms: selecting the 'worker nodes'

For HTC applications
– like WLCG, SKA, WeNMR –
typically

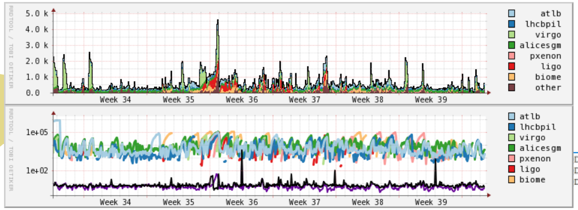
balanced features for node throughput
(CPU, storage, memory bandwidth, network)

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**

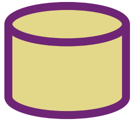
Image: Cluster 'Lotenfeest' at the Nikhef NDPF, acquired March 2020. Lenovo SR655 with AMD EPYC 7702P 64-Core single-socket

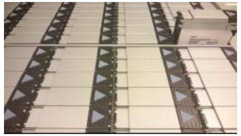
WLCG computing – conveniently parallel

?




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

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

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

# local groups

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





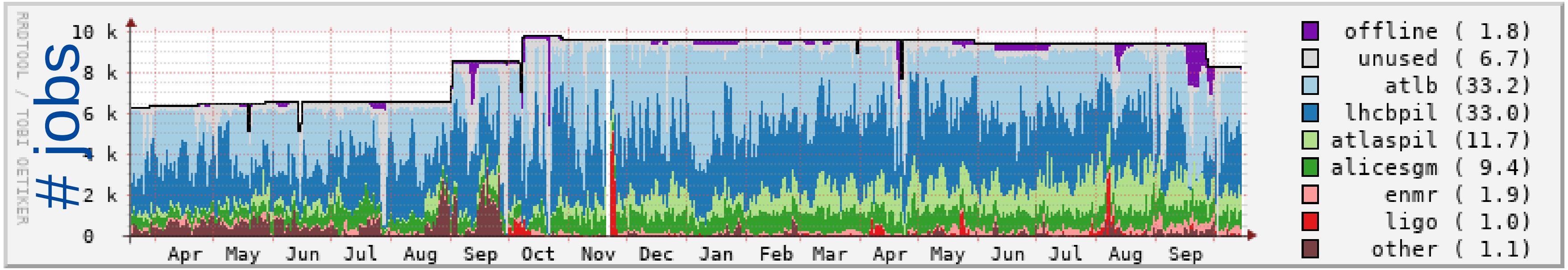
‘like milking cows’ (if you feed them lots of power first)

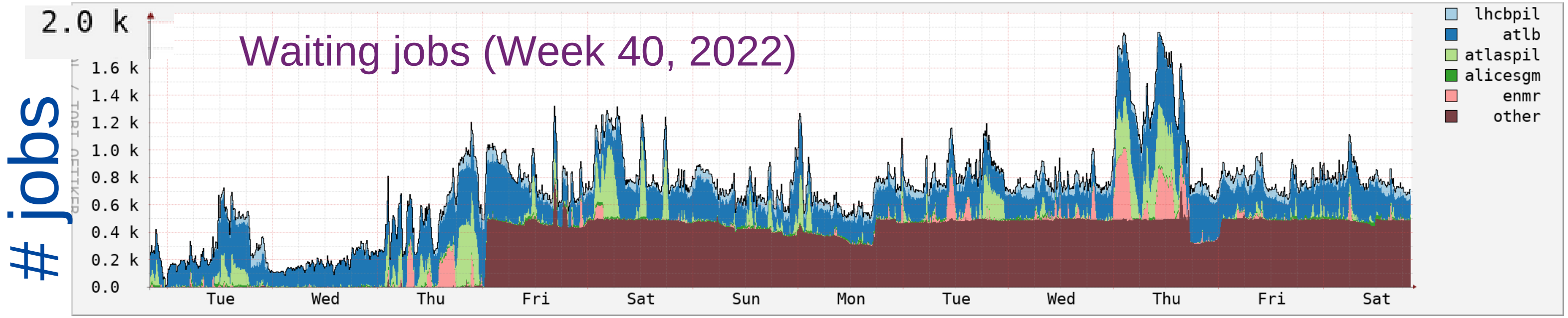
parallel access to data comes at a cost of high IOPS

NDPF ‘WLCG and Dutch National Infra’ cluster

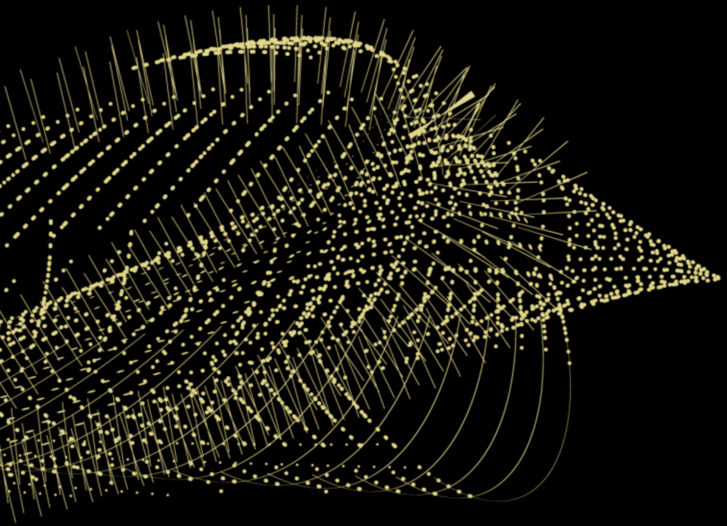
Running jobs:

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



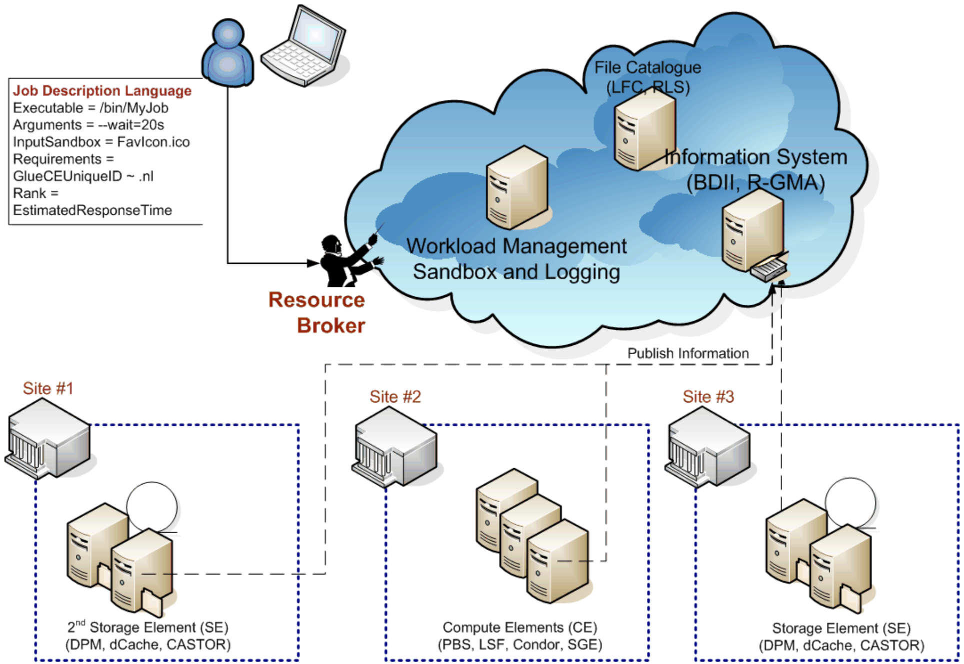
> 1 system
> 1 site
> 1 user group
> 1 organization
More than one …

More of more than one …

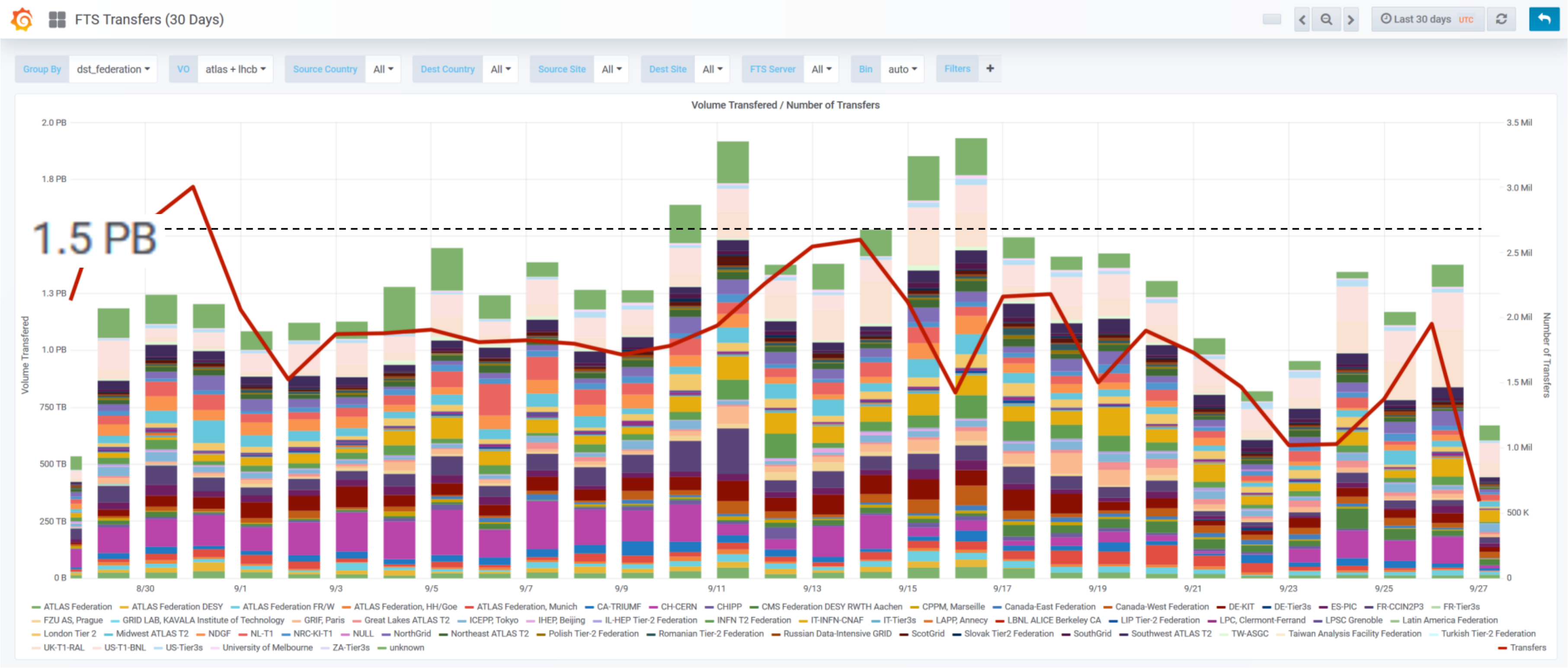
Fancy an interactive console install?

Images: Nikhef Housing H234b NDPF science processing data centre

Global computing and workload management



High throughput computing is also 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

Het vroege internet ...

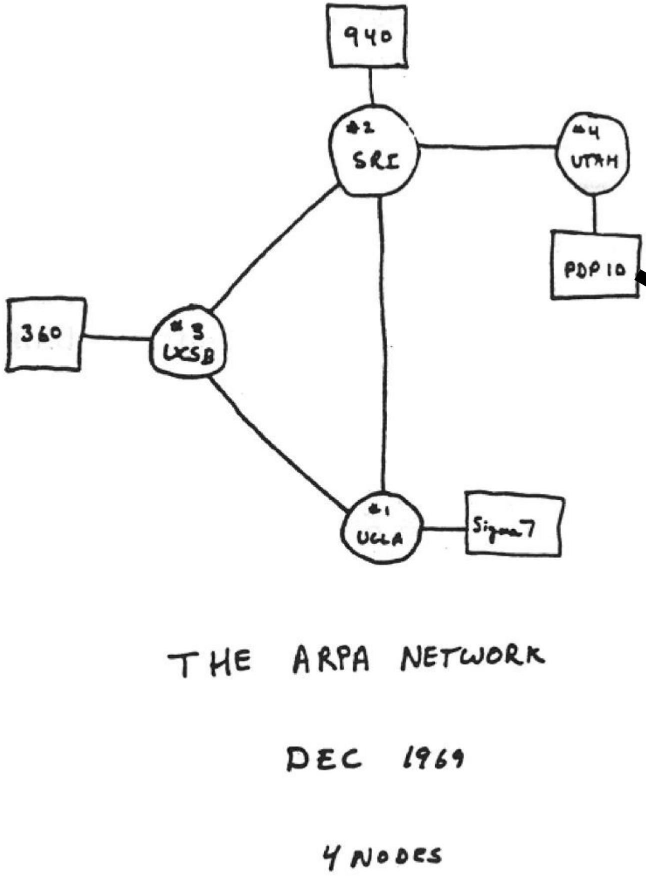
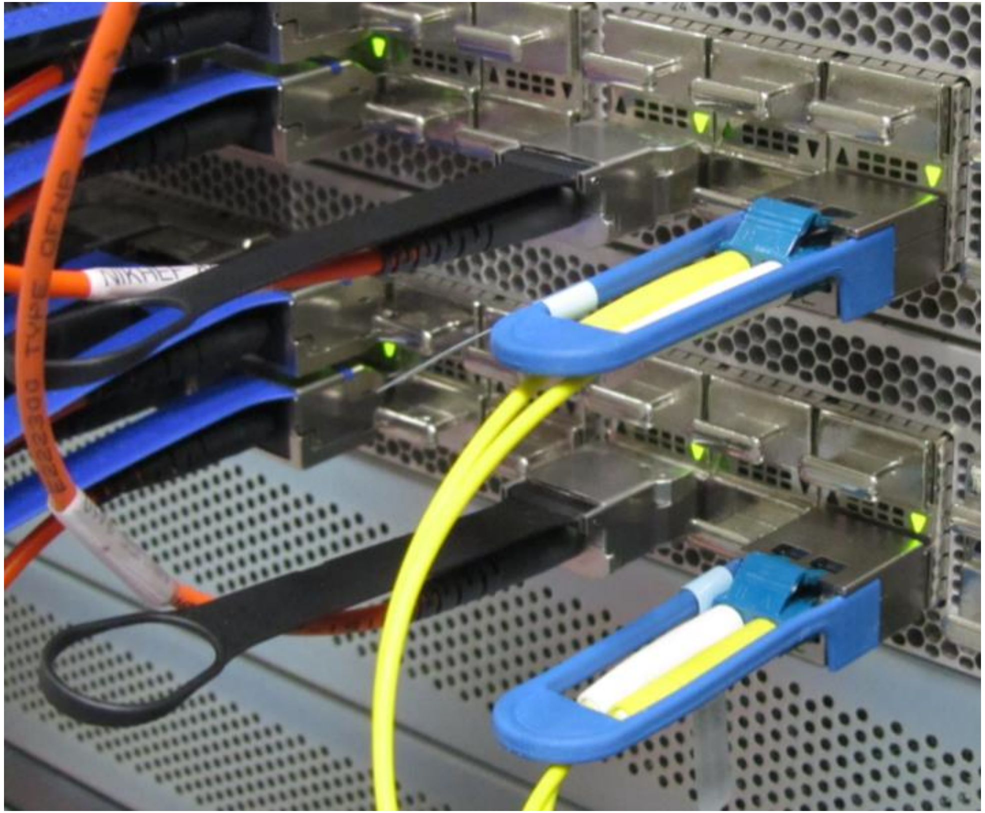


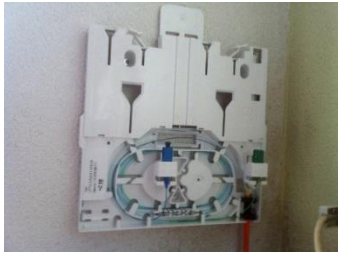
Image source: Alex McKenzie and "Casting the Net", page 56. See https://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography/atlas/arpanet2.gif ; acoustocoupler: Wikimedia

How does 100, or now 400 Gigabit per second look?

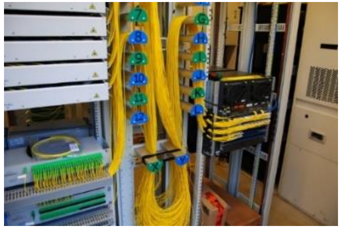


Nikhef
Data Processing
Facility
router 'deel'

Thuis 'FttH'
~1Gbps BX
single strand, SC



een KPN FttH
PoP in de wijk



vergelijk:
VDSL BR straatkast
voor als je nog
op xDSL koper zit

‘Elephant streams in a packet-switched internet’

Moving stuff around

wheelbarrows work fine in your garden
want to send it to different places?
Use waggons on a train, or ships,
going always from A-to-B anyway?

A conveyer belt will do much better!

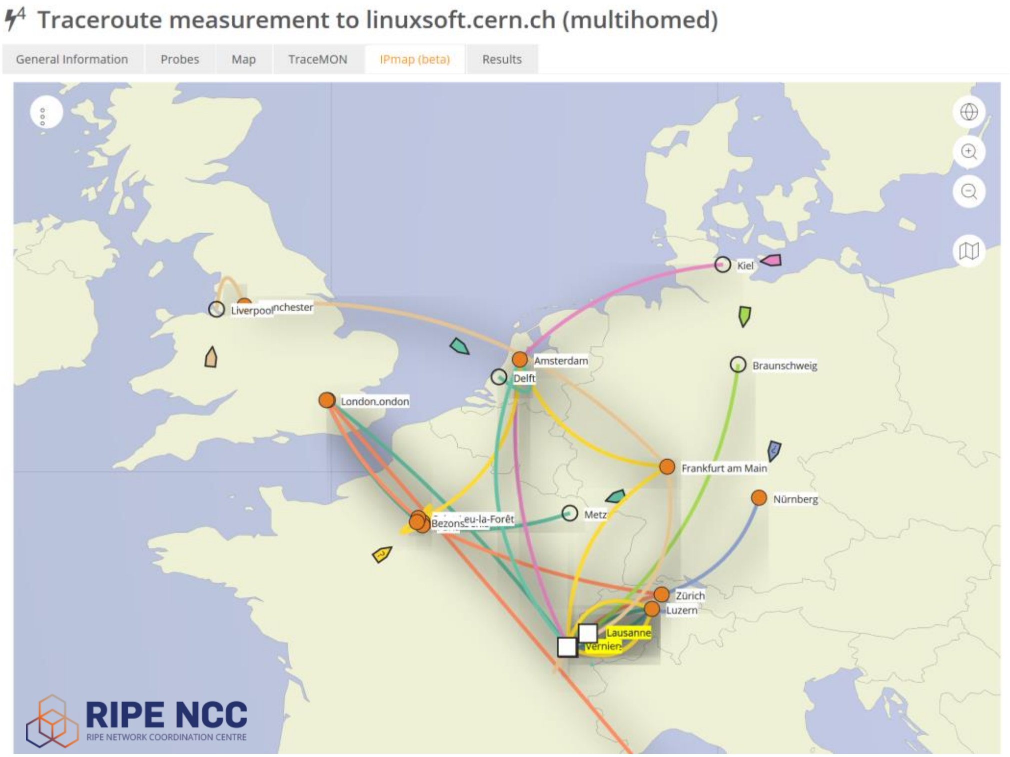
… although you still need
a hole to dump it in …

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
```

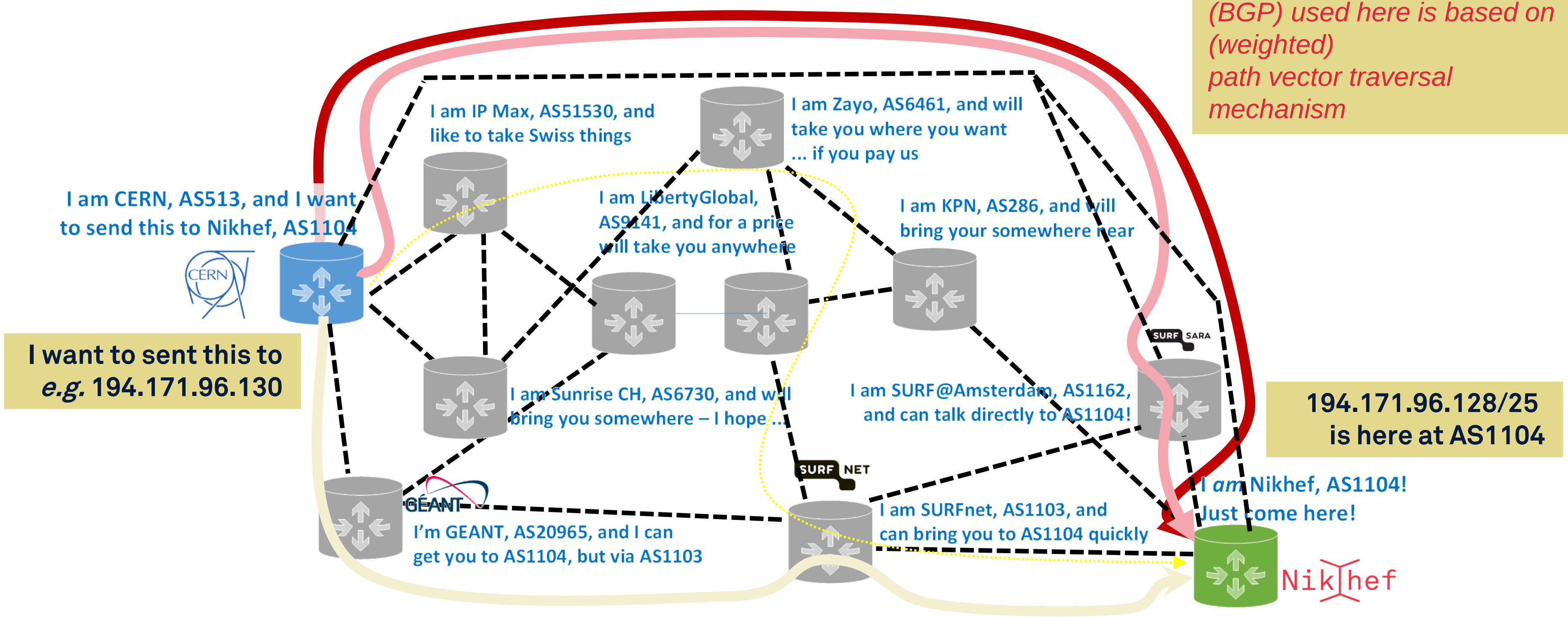
but from Interparts in Lisse, NH:

```
[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

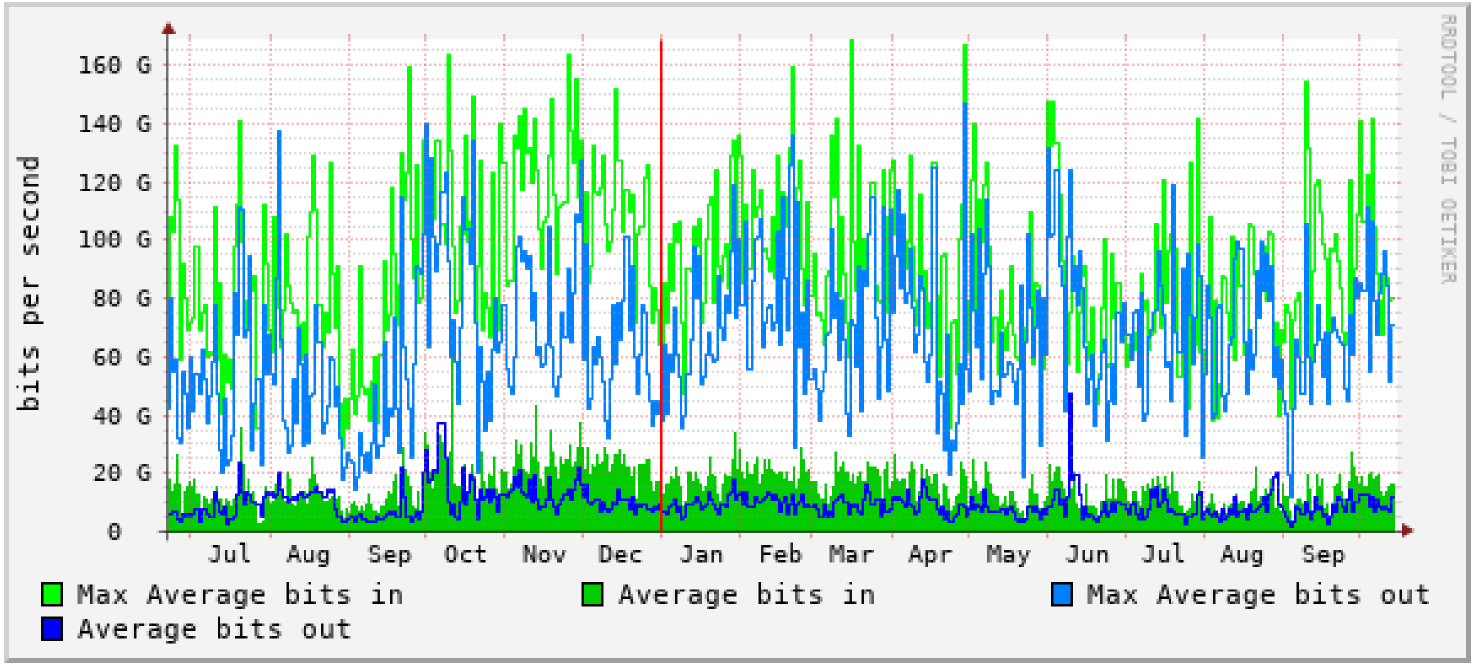
Where do internet packets go anyway?

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



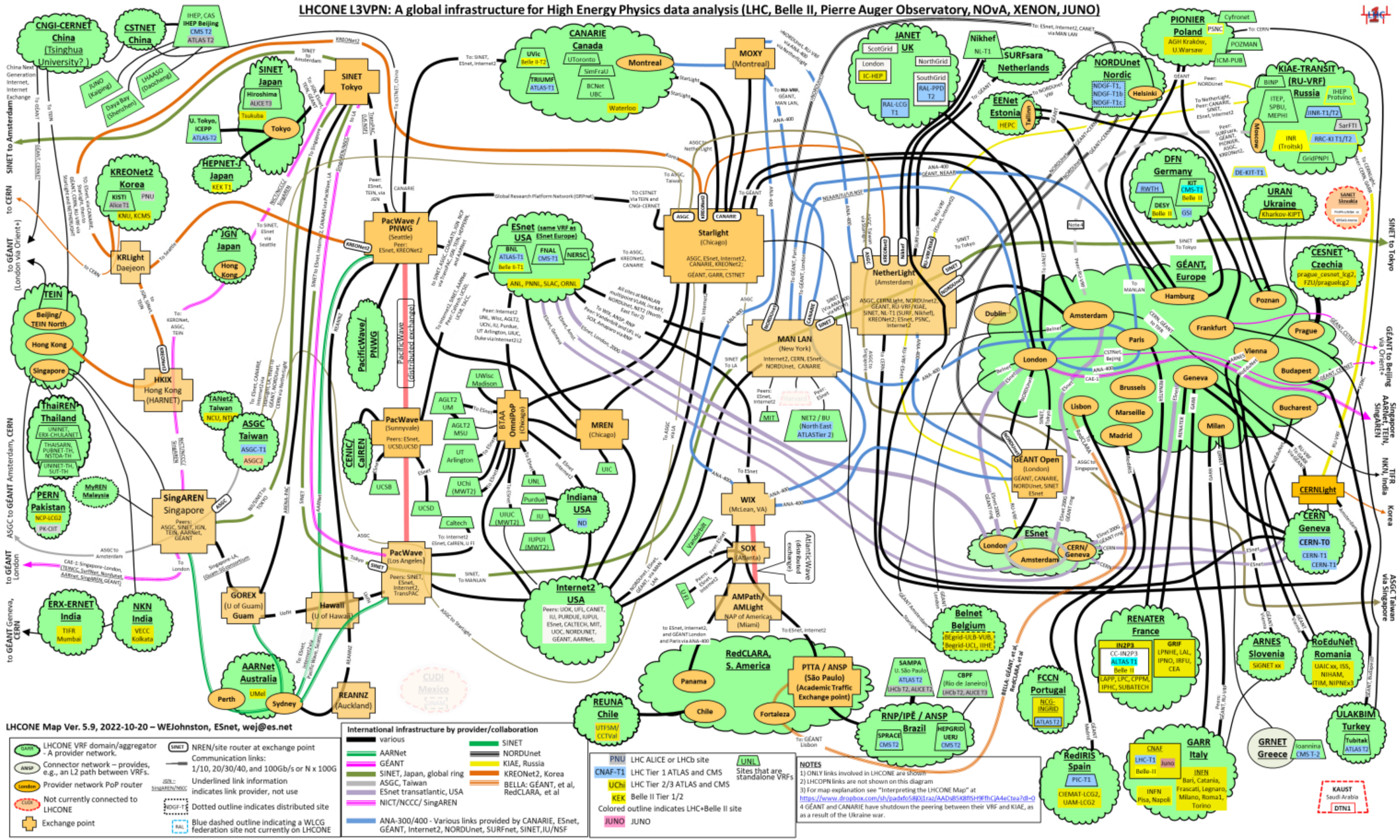
grey-dash lines for illustration only: may not correspond to actual peerings or transit agreements;
red lines: the three existing LHCOPN and R&E fall-back routes; yellow: public internet fall-back (least preferred option)

Typical data traffic to and from the processing cluster



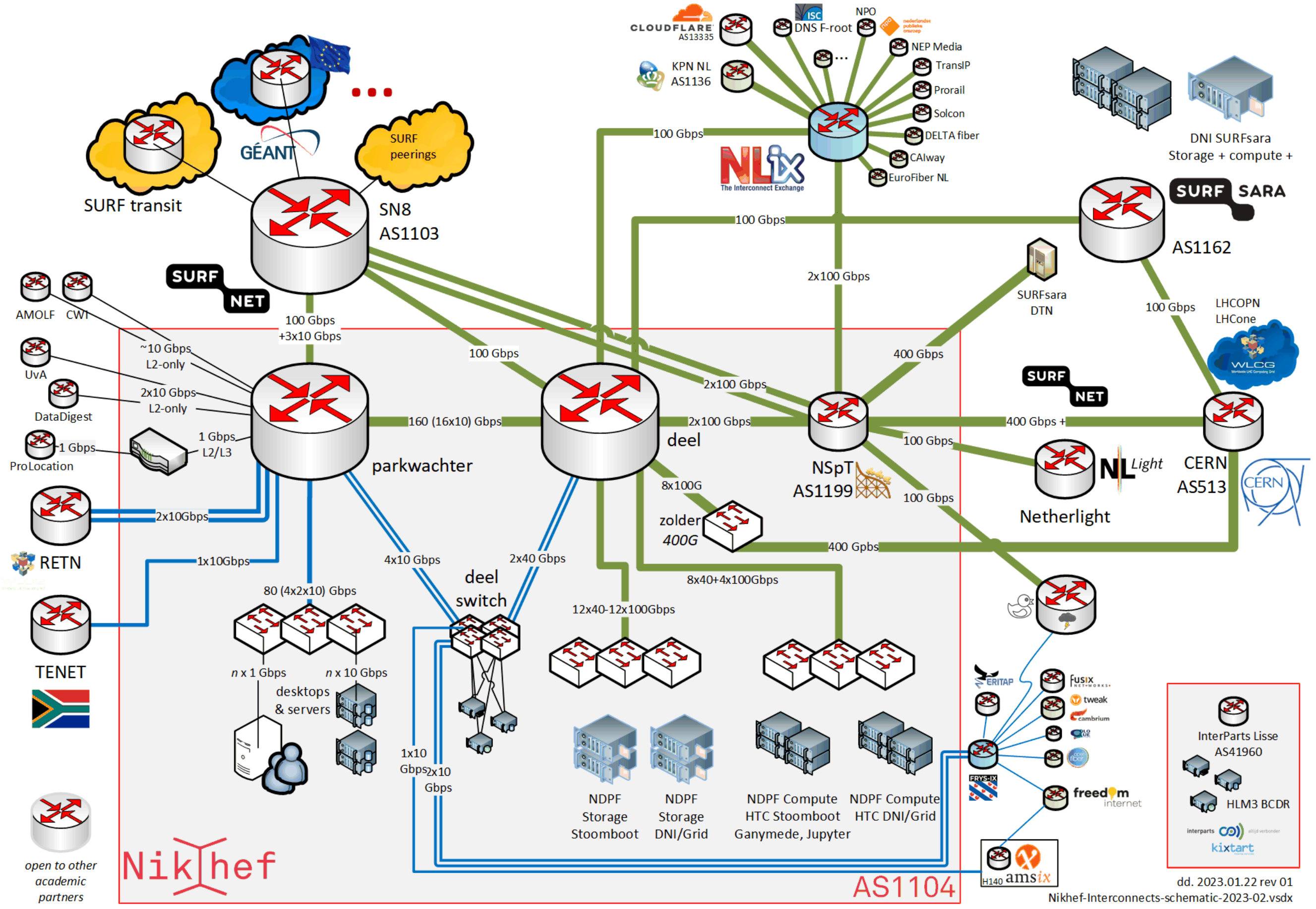
Source: Nikhef cricket graphs period June 2021 – October 2022 – aggregated (research) traffic to external peers from deelqfx – https://cricket.nikhef.nl/

LHCone



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

Just one random (smallish) autonomous system



AS1104

Exercising the network – sensor data and events

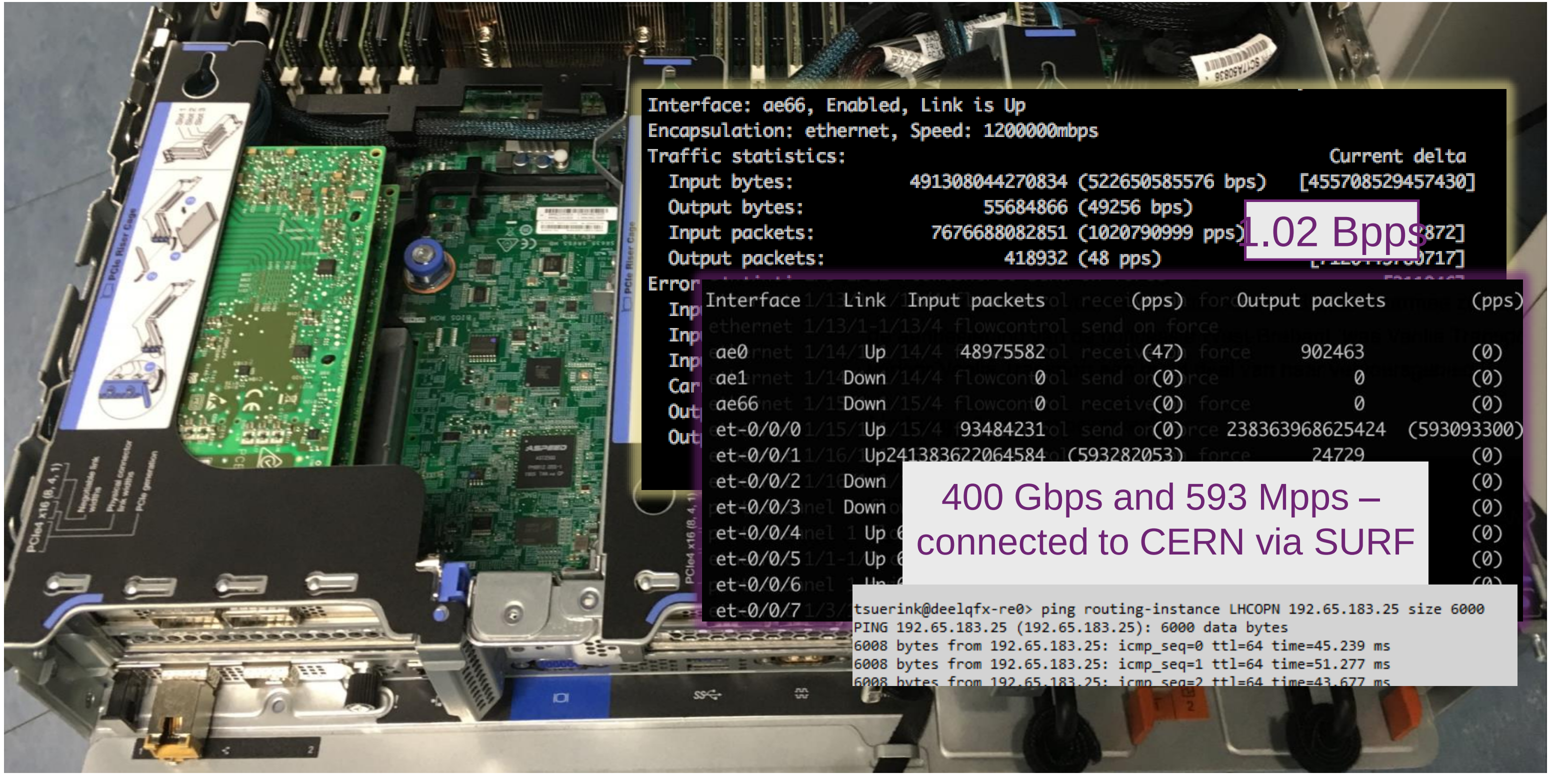


Image: ballenbak.nikhef.nl, Tristan Suerink

En … hoeveel gebruikt dat dan?

Eén server gebruikt zo'n 260W!

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

en het onderzoeksdatacentrum Nikhef (de 'glazen doos') kan 400kW aan – waar blijft dat dan?

De snelste CPU is voor ons niet altijd de beste (*sorry gamers!*). Want 5 jaar energie en beheer zijn even kostbaar als de server zelf!

WKO: Warmte Koude Opslag

21% van het vermogen is nodig om te koelen, maar:
we mogen 3500GJoule/jaar (~112 kWjaar, ~982 000 kWh)
aan studenten tegenover leveren om ze warm te houden !

Let’s go on tour!

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

Maastricht University

Nikhef