Een petabyte weinig? Spoorzoeken door data

Nikhef

Weekend van de Wetenschap
Amsterdam Science Park

2018-10-06

David Groep
davidg@nikhef.nl

Nikhef

Vind het grasveld met slagboom ...

Nikhef

in 40 plaatjes (en dat 1 000 000 keer)

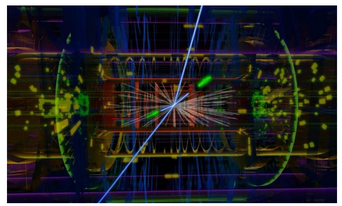
Botsingen ‘fotograferen’?

IMG_6592.JPG

JPG File

1.51 MB

digitale compact-camera, 5MP

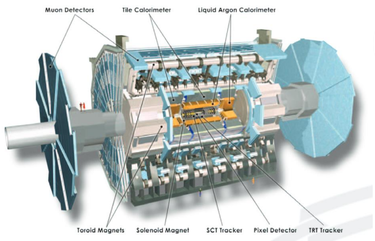


ATLAS_RAW_single_event.data

ROD File

1.60 MB

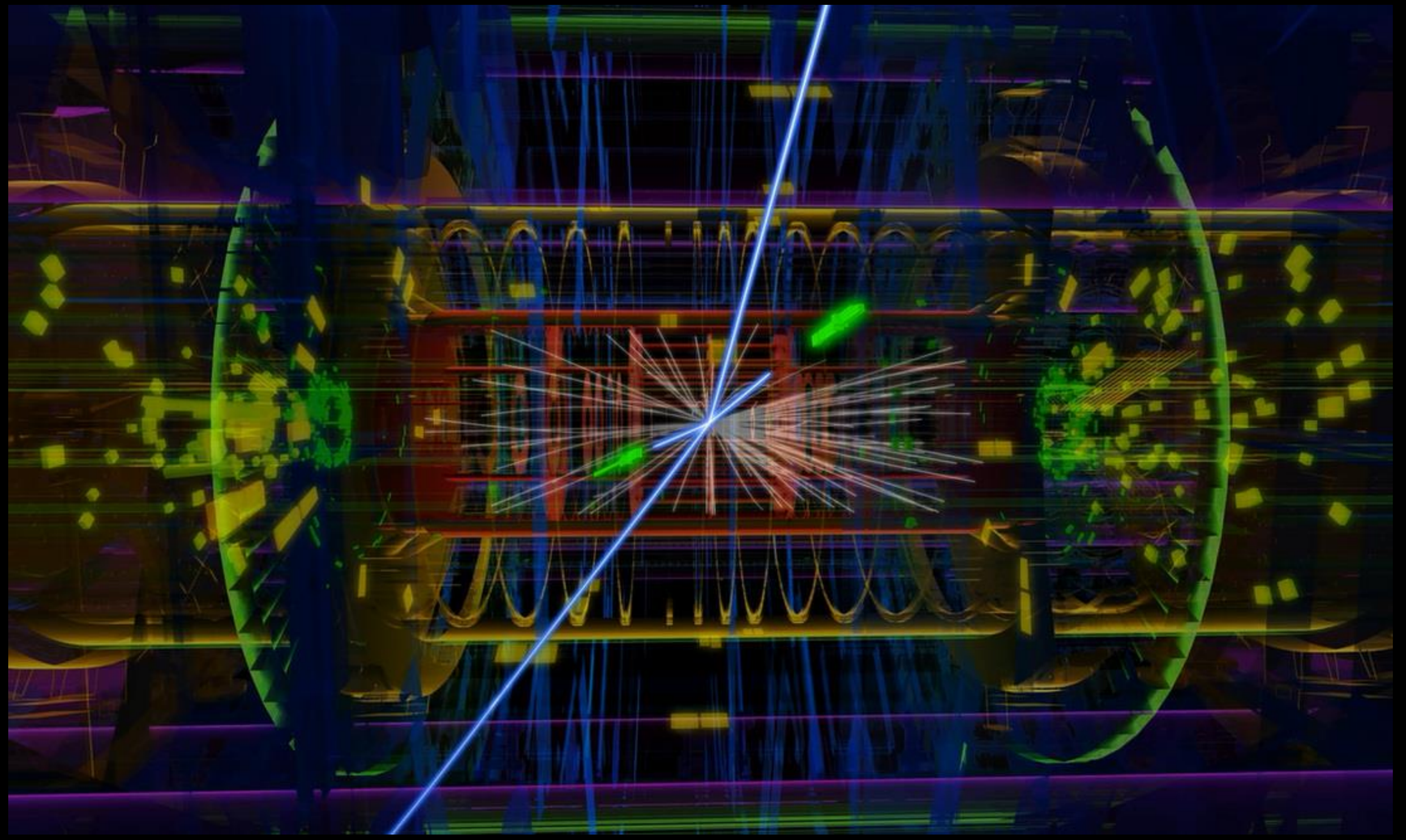
ATLAS detector, ~88 MP (maar niet alles licht elke keer op … gelukkig!)



aantal detectorelementen en raw data size: https://cds.cern.ch/record/1457044/files/ATLAS%20fact%20sheet.pdf

Van pixel naar sporen: veel rekenen!

*~ 10 seconden rekenen **per gebeurtenis** voor een ATLAS event met ‘jets’ met daarin ~30 botsingen*



Event processing time: v19.0.1.1 as per Jovan Mitrevski and 2015 J. Phys.: Conf. Ser. 664 072034 (CHEP2015)
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

Maar het is wel goed zoeken ...

Zachte botsingen	10^8
$W^{\pm} \rightarrow e^{\pm} \nu$	15
$Z^0 \rightarrow e^+ e^-$	1
Top-anti-top quarks	1
$bb \rightarrow \mu + X$	10^3
QCD jets, $p_T > 150$ *GeV*	10^2

Higgs deeltje: ~ 1 per dag

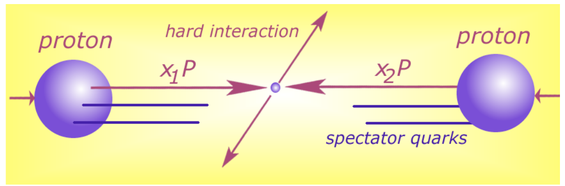
event rates from LHC Run 1

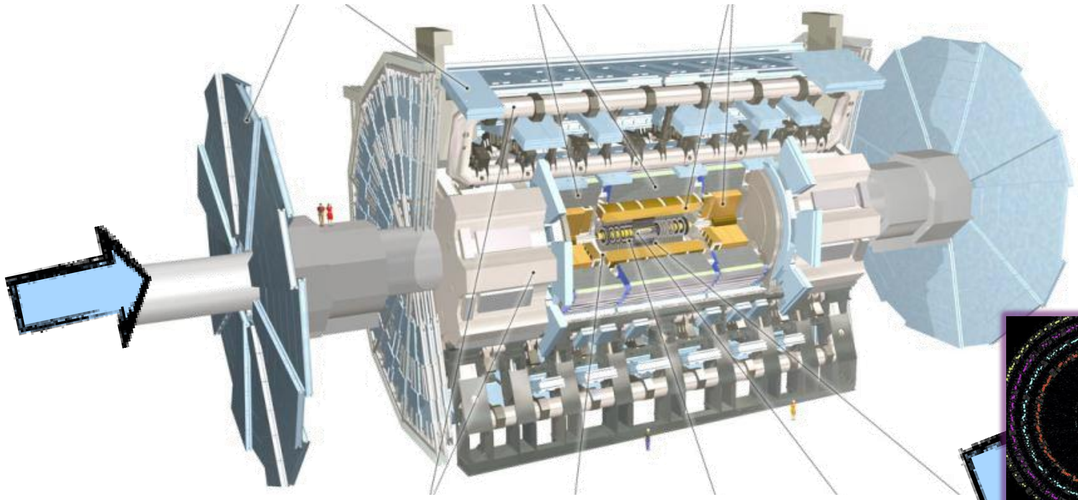
Selectie interessante gebeurtenissen uit 'achtergrond' van 1 op 10^{13} gebeurtenissen

- dit is equivalent met zoeken van
1 persoon op 1000 wereldpopulaties
- oftewel één speld in 20 miljoen hooibergen

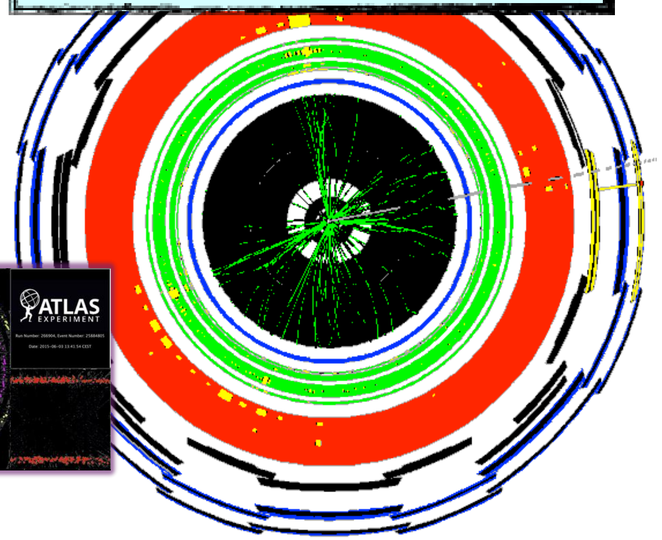
Hoe vind je ‘interessante’ events?

40 miljoen / seconde

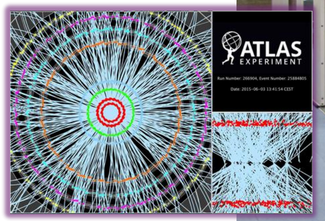




Trigger systeem selecteert 600 Hz ~ 1 GB/s data



Classificatie van deeltjes in de botsingen:
- *electronen*
- *muonen*
- *jets van hadronen*
- *...*



Analyse van botsingen door promovendi

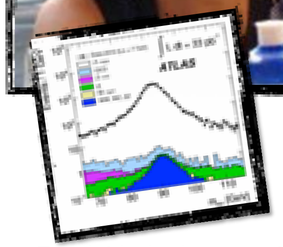
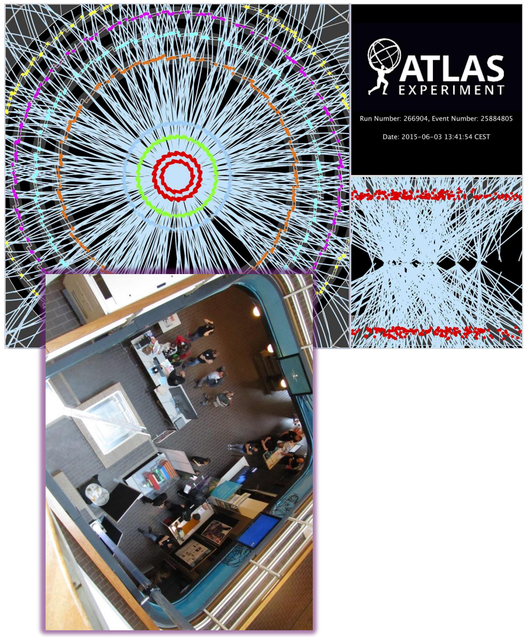


diagram: from Frank Linde; images: ATLAS collaboration, Nikhef. And sorry for the AVG-blur

Mega, Giga, Tera, … en Peta byte

1 Megabyte, MB = 1 000 000 bytes
[en 1 Mibibyte, MiB = 1 048 576 bytes]

x 1000



1 Gigabyte, GB = 1 000 MB
= 1 000 000 000 bytes

x 1000

1 Terabyte, TB = 1 000 GB
= 1 000 000 MB

x 1000

1 Petabyte, PB = 1 000 TB
ofwel 1 000 000 000 MB

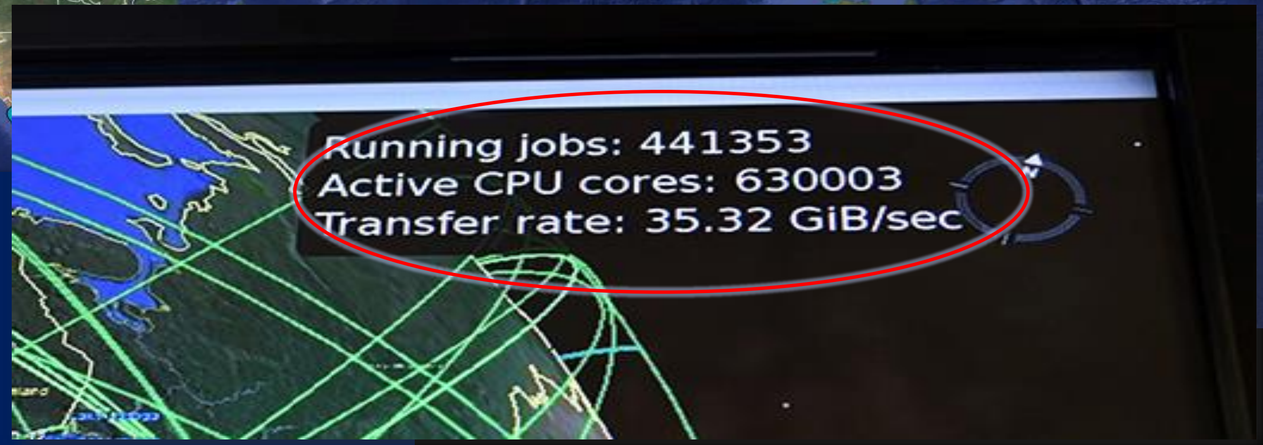
x 1000

Het spoor terug ...

Nikhef

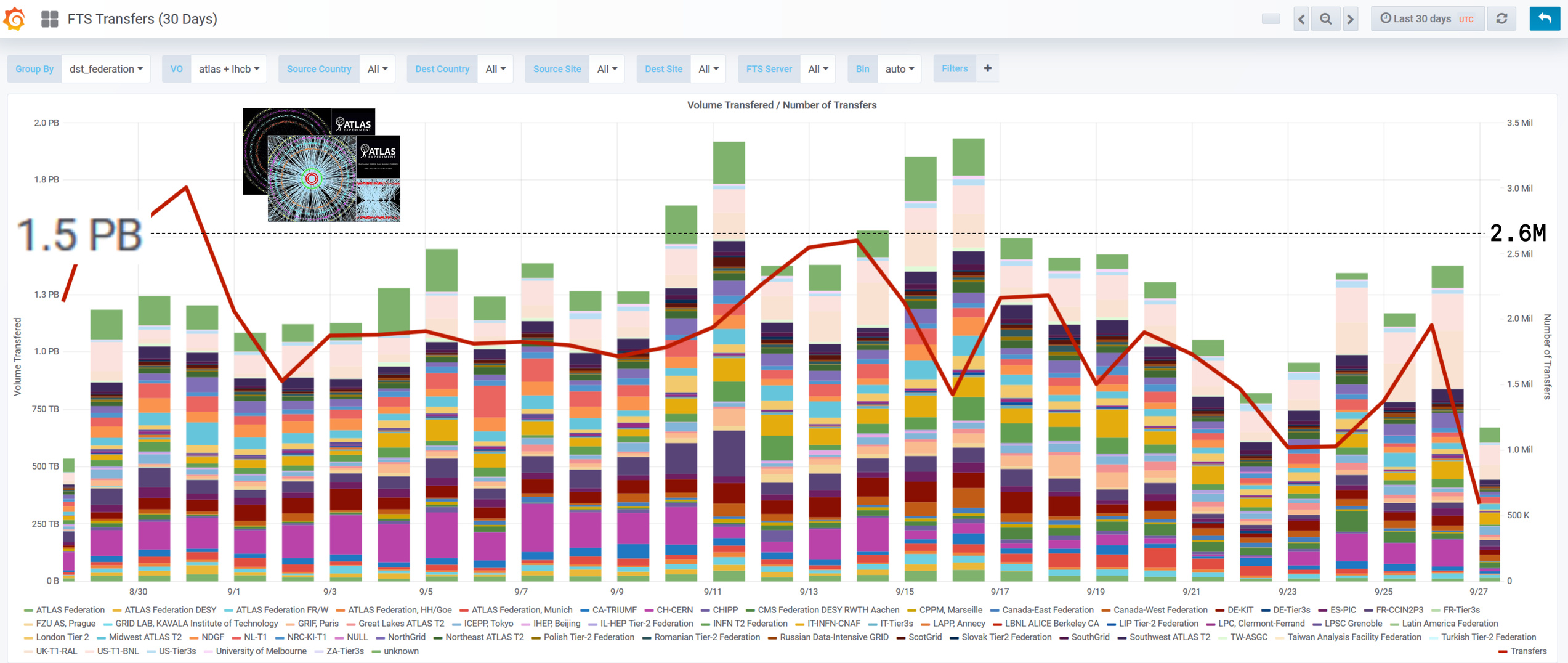
… van en naar de hele wereld

meer dan 170 instituten in 42 landen en economiën



- *CPU: ~ 350,000 modern rekenkernen*
- *Disk 310 PB*
- *Tape 390 PB*

Data verdelen over de wereld T0>T1s

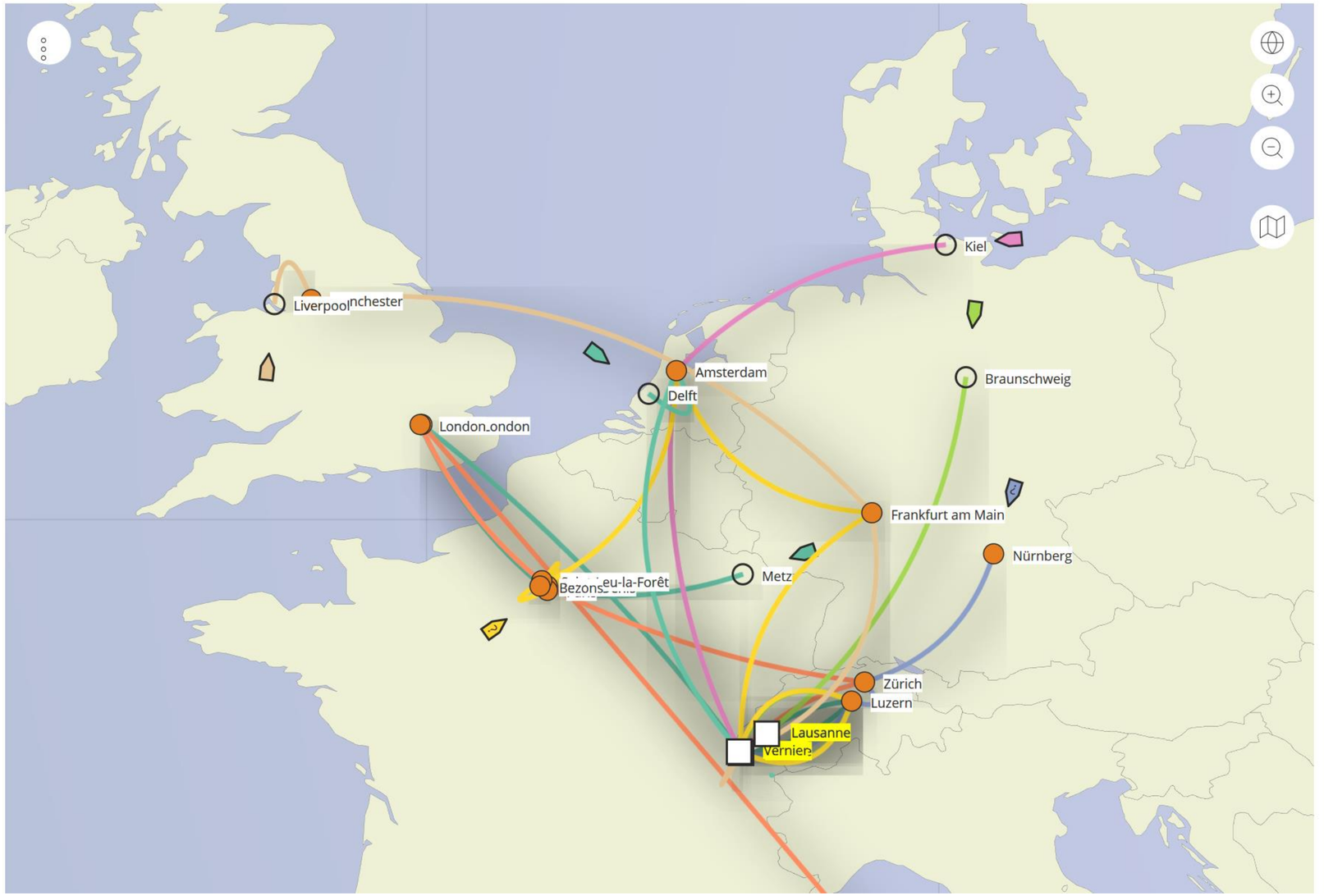


source: https://monit-grafana.cern.ch/d/000000420/fts-transfers-30-day on September 27, 2018

hoe gaan je thuis naar CERN?

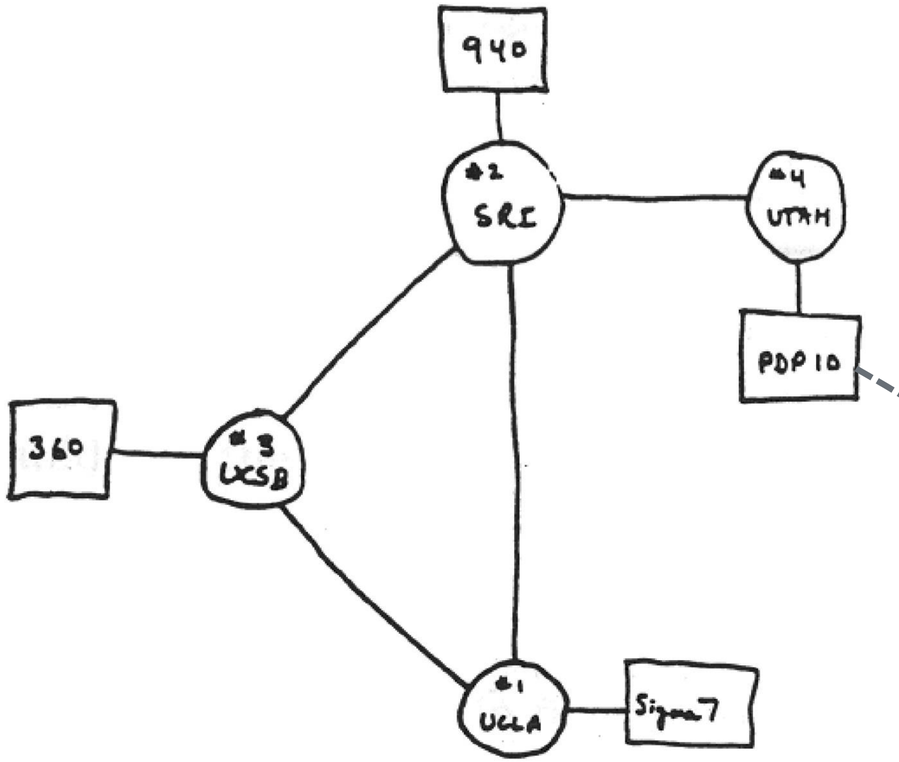
Traceroute measurement to linuxsoft.cern.ch (multihomed)

General Information | Probes | Map | TraceMON | IPmap (beta) | Results



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

Het vroege internet ...



THE ARPA NETWORK

DEC 1969

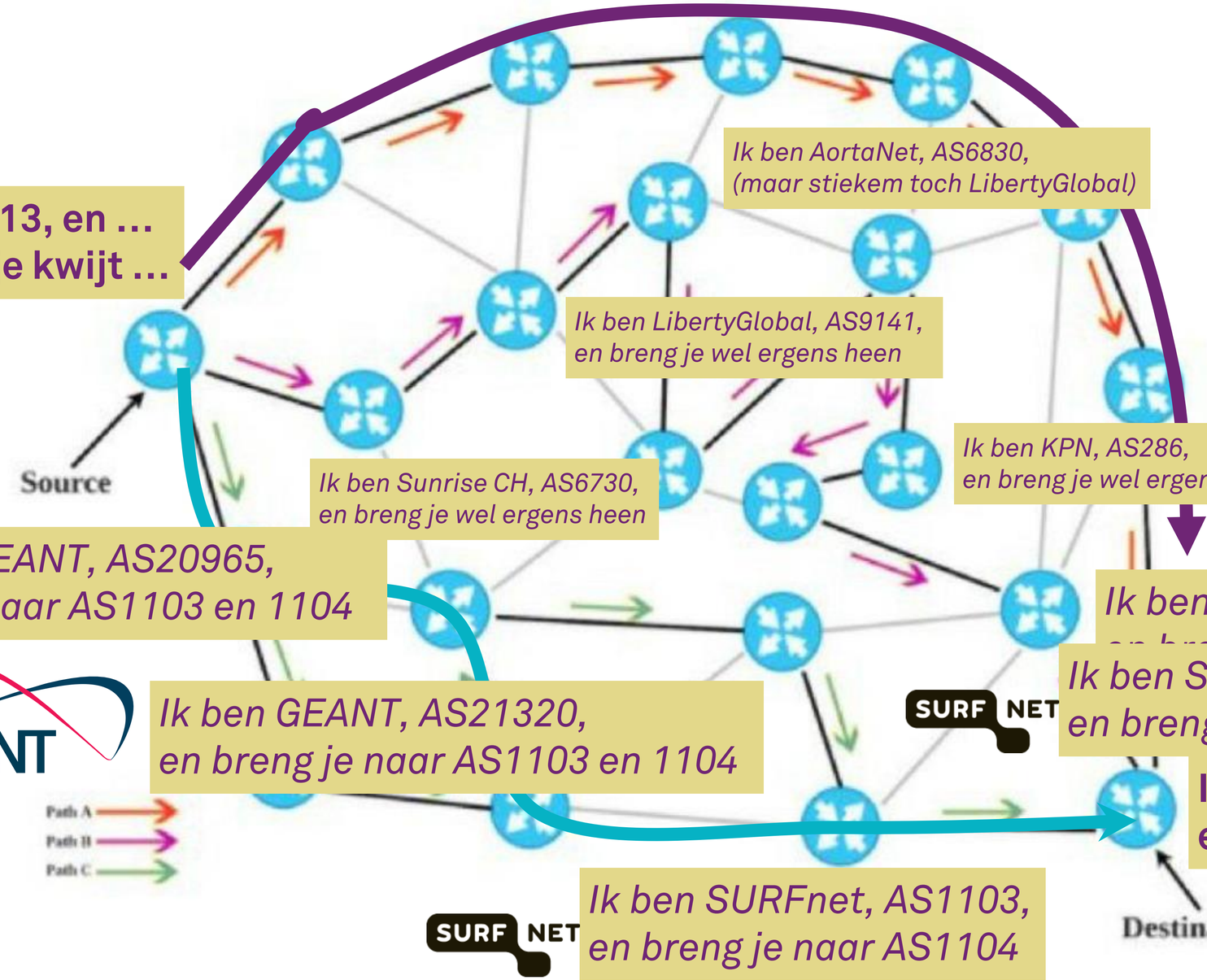
4 NODES

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

Sporen door het Internet

Ik ben CERN, AS513, en …
ik wil m'n pakketje kwijt …

188.184.38.9



Ik ben AortaNet, AS6830,
(maar stiekem toch LibertyGlobal)

Ik ben LibertyGlobal, AS9141,
en breng je wel ergens heen

Ik ben Sunrise CH, AS6730,
en breng je wel ergens heen

Ik ben KPN, AS286,
en breng je wel ergens heen

Ik ben ook GEANT, AS20965,
en breng je naar AS1103 en 1104

Ik ben GEANT, AS21320,
en breng je naar AS1103 en 1104

Ik ben SURFsara, AS1162,
en breng je direct naar AS1104!

Ik ben SURFnet, AS1103,
en breng je naar AS1104

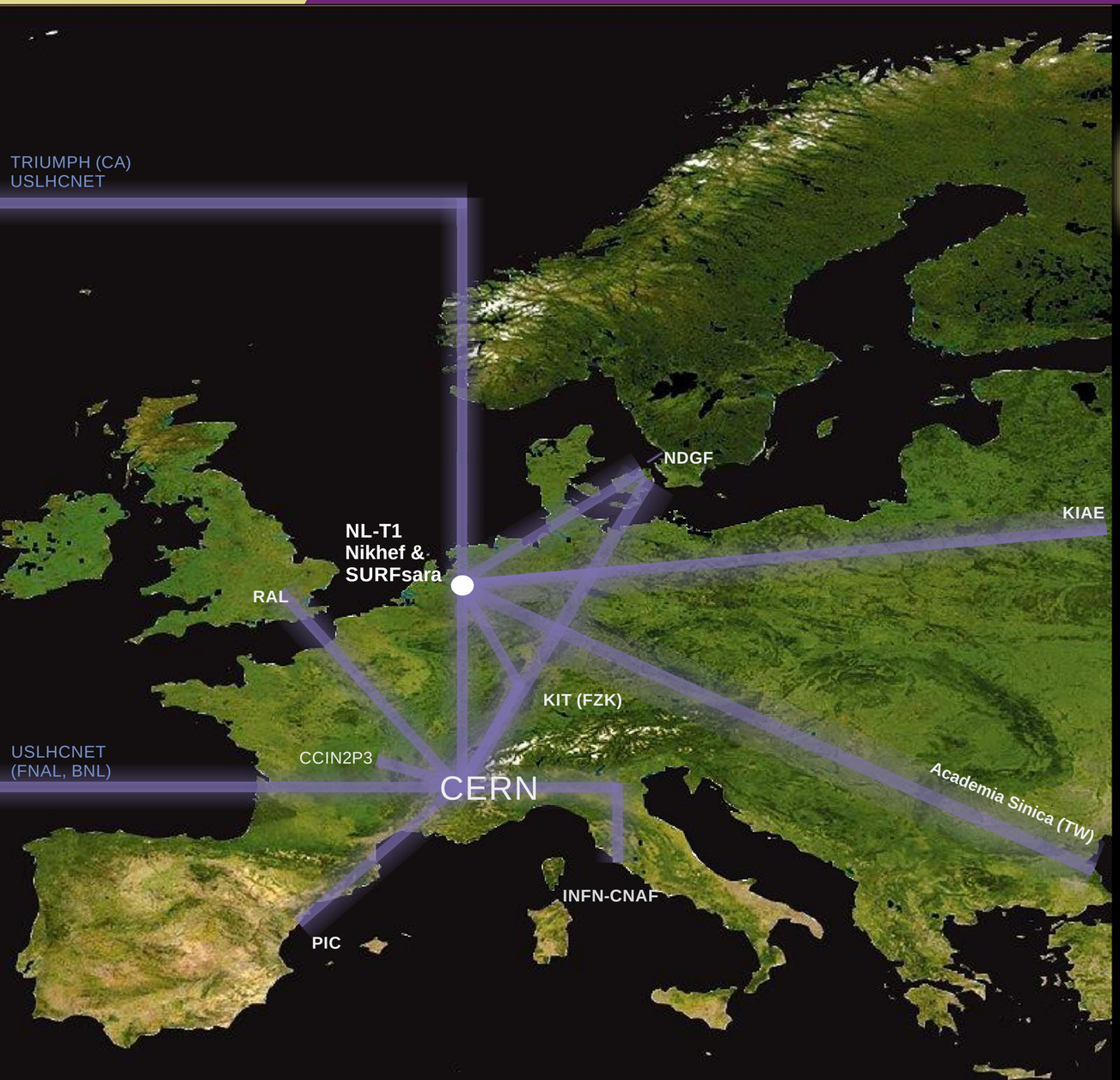
Ik ben SURFnet, AS1103,
en breng je naar AS1104

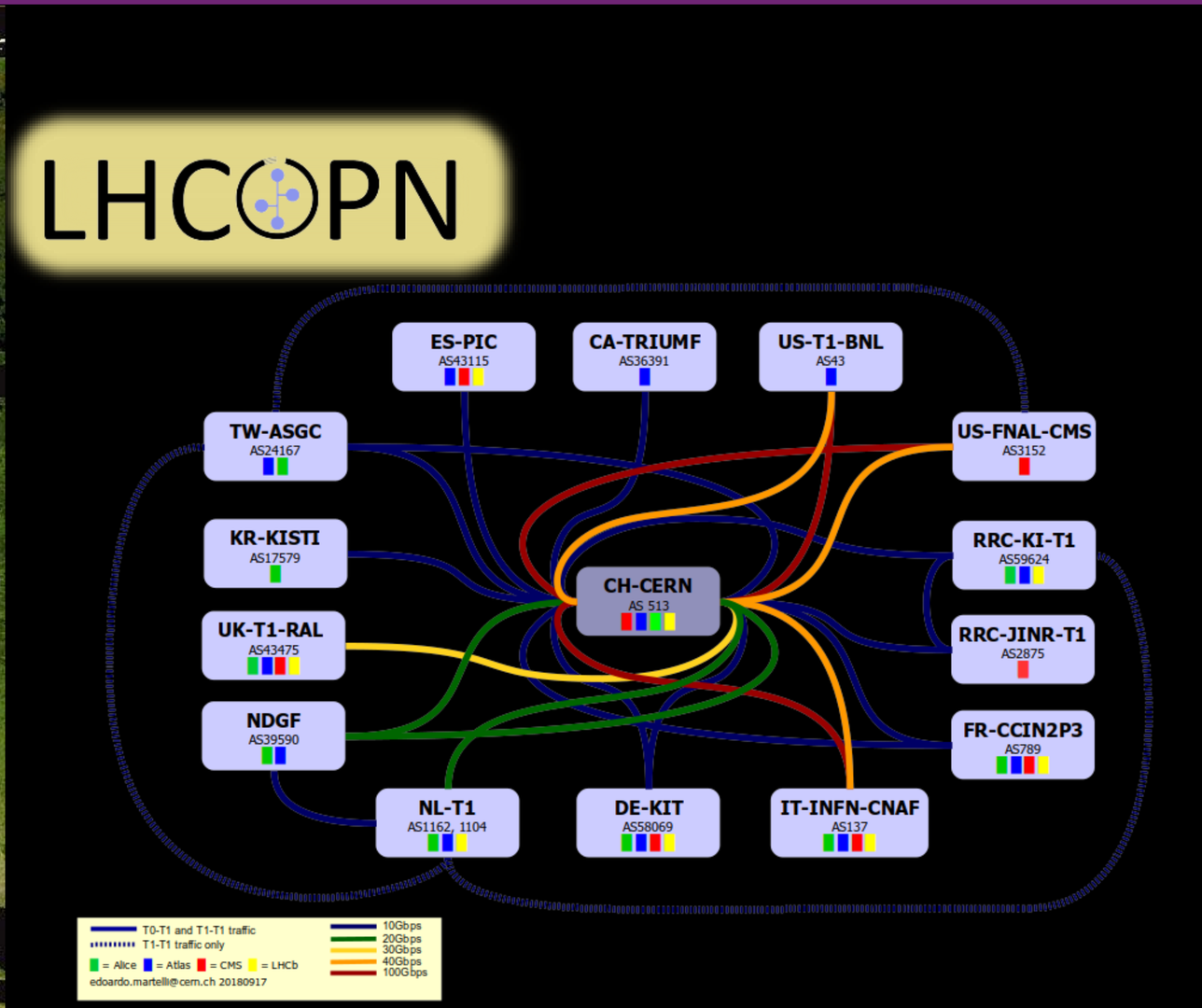
Ik ben Nikhef, AS1104
en ik heb schijfruimte voor je

194.171.96.128/25

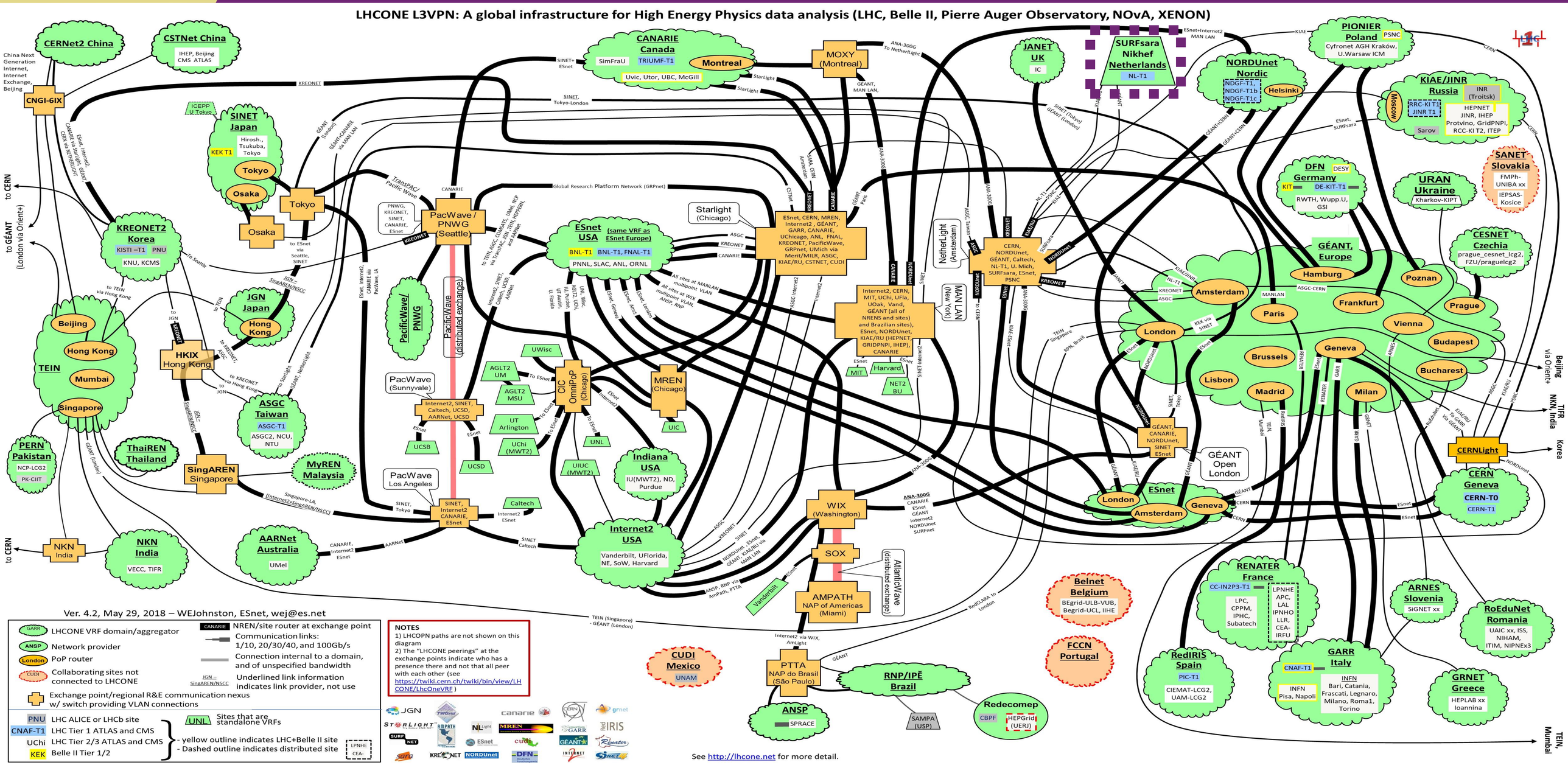
"segment routing"
image by David Penaloza, CISCO

LHC ‘Optisch Privaat Netwerk’

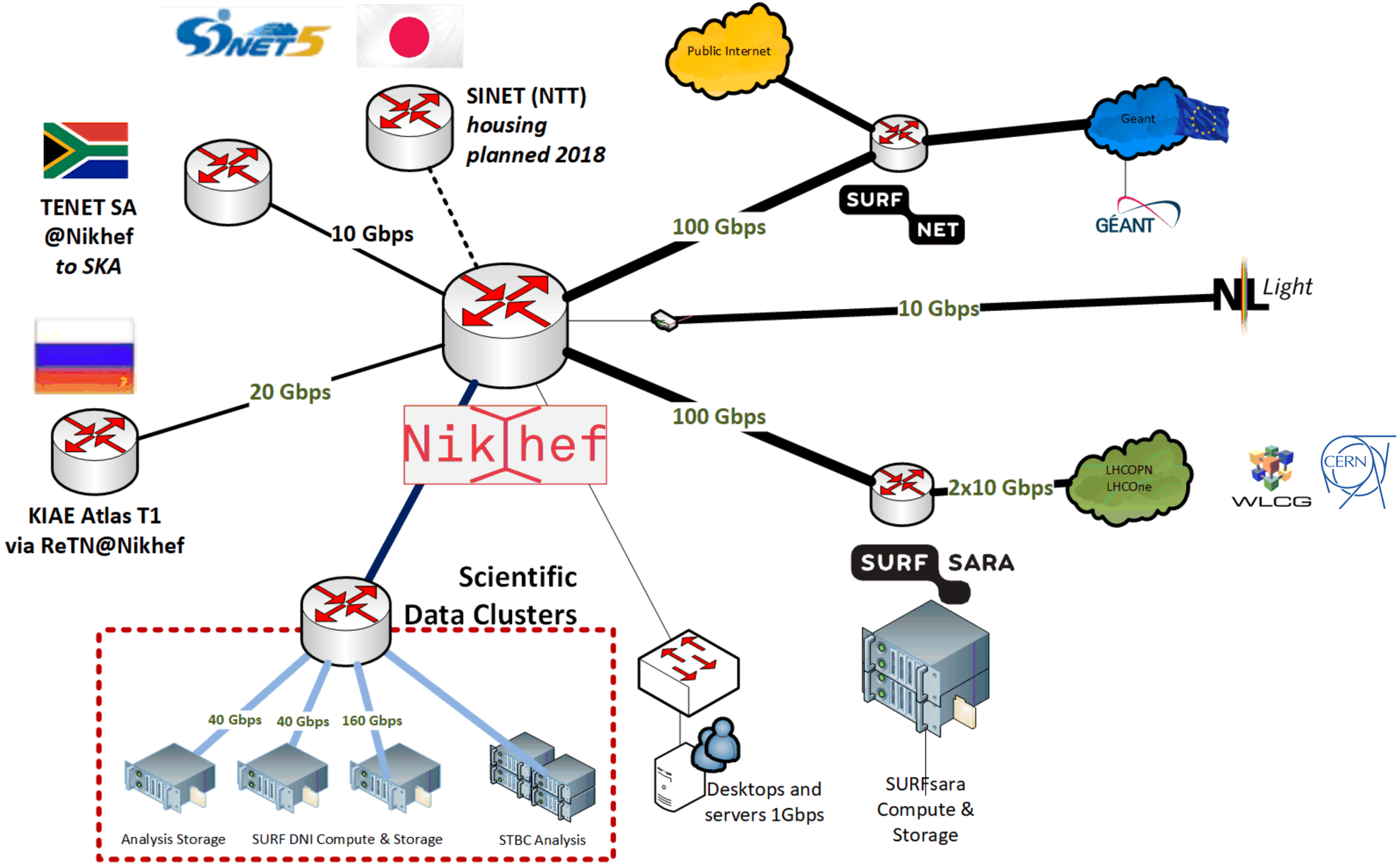




‘LHC Open Network Environment’

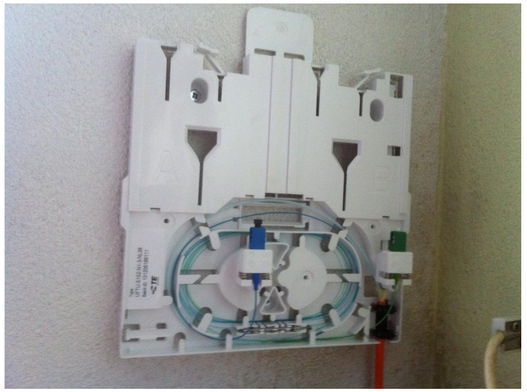


Een network infra voor ‘big data’

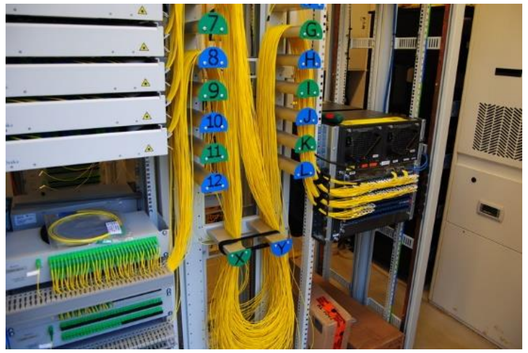


Hoe ziet 100Gbps eruit?

Thuis ‘FttH’
~1Gbps BX
single strand, SC



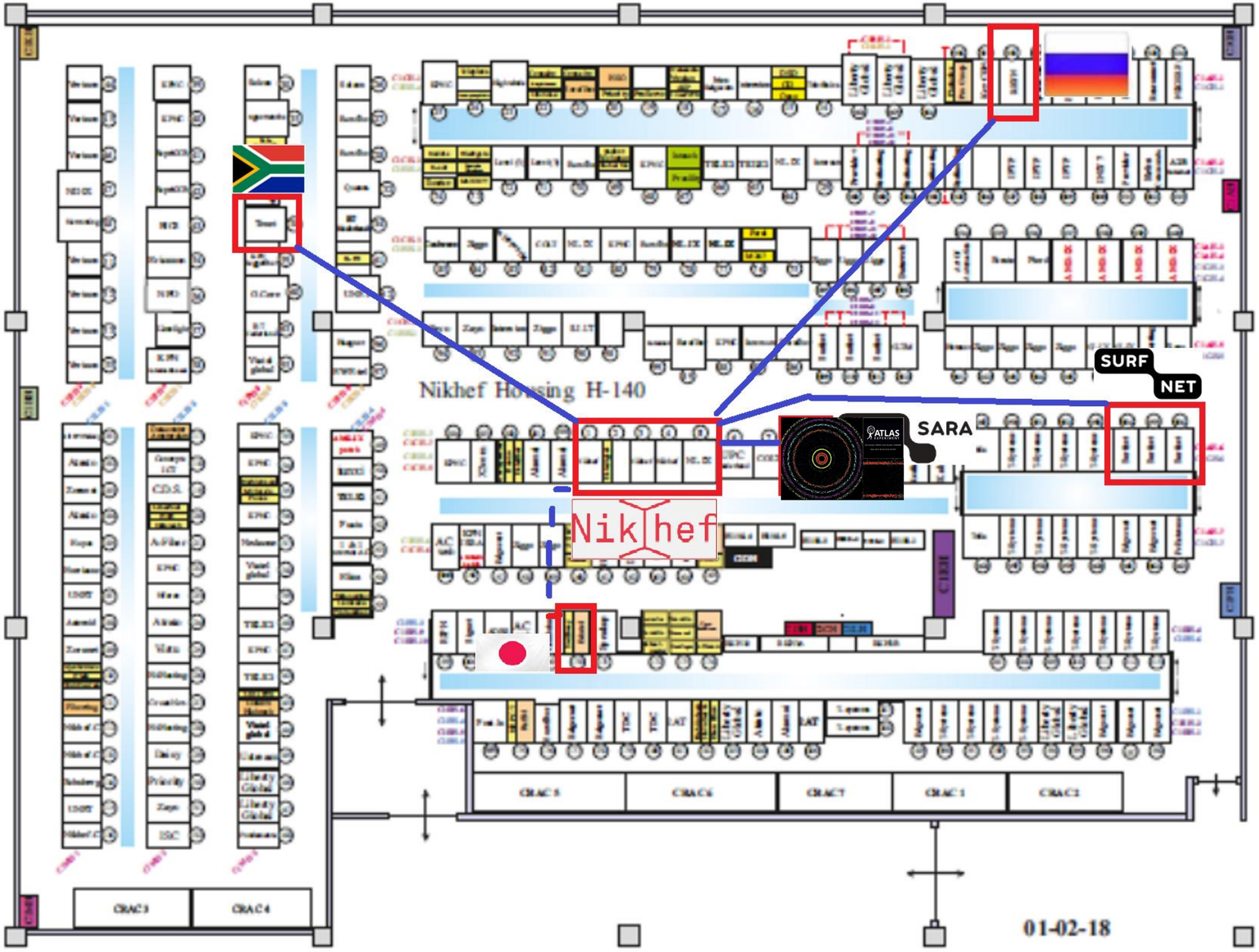
Nikhef
Data Processing
Facility
router ‘deel’

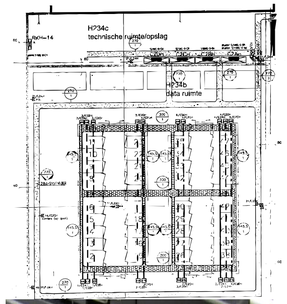


een KPN FttH
PoP in de wijk

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

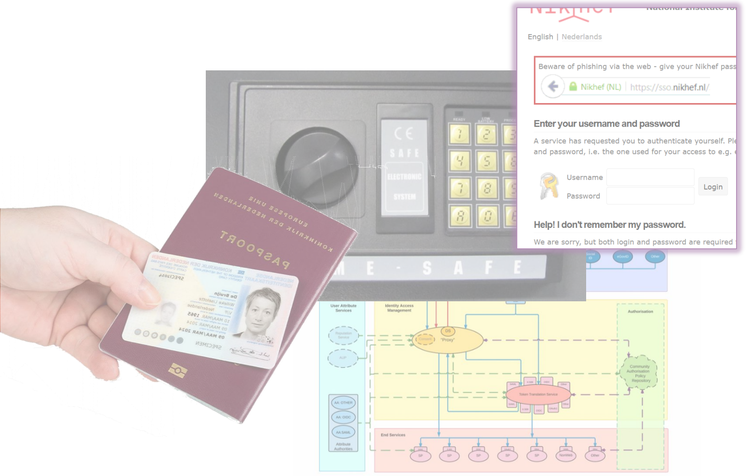
Het internet kun je best vastpakken





Onderzoeksinfrastructuur – de ‘cloud’

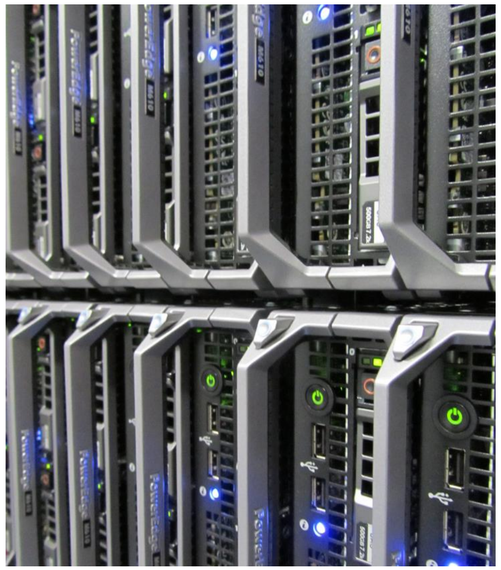
Samenwerking en beveiliging



we houden bij wie er mag rekenen, welk experiment wat gebruikt, en of onze rekenkracht goed gebruikt wordt …

… en we niet misbruikt worden om het internet aan te vallen!

Parallel rekenen



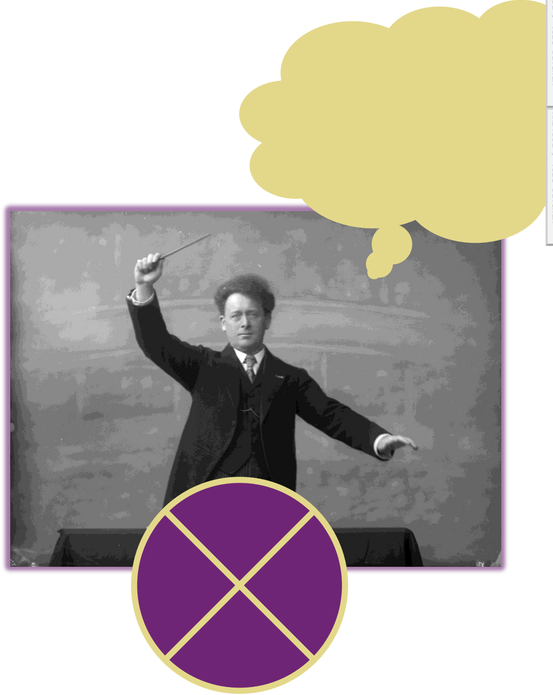
7000 rekenkernen (is ~300 servers) staan dag en nacht te rekenen aan onderzoeksdata

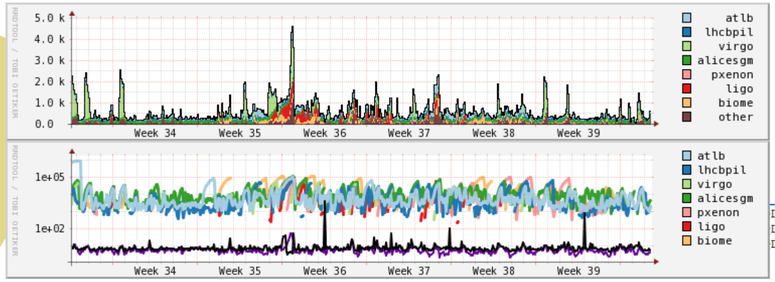
en parallelle data opslag

de ~5 petabyte gegevens bij ons moeten we ook in één dag kunnen lezen, anders staan die 7000 rekenkernen te wachten … die lezen ieder ~10MB per seconde!

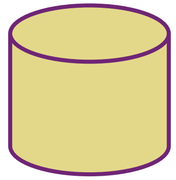
Parallel rekenen – ‘samen sterker’

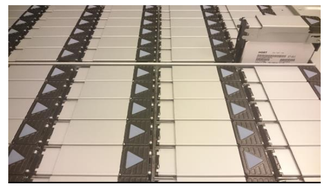
?






```
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
```



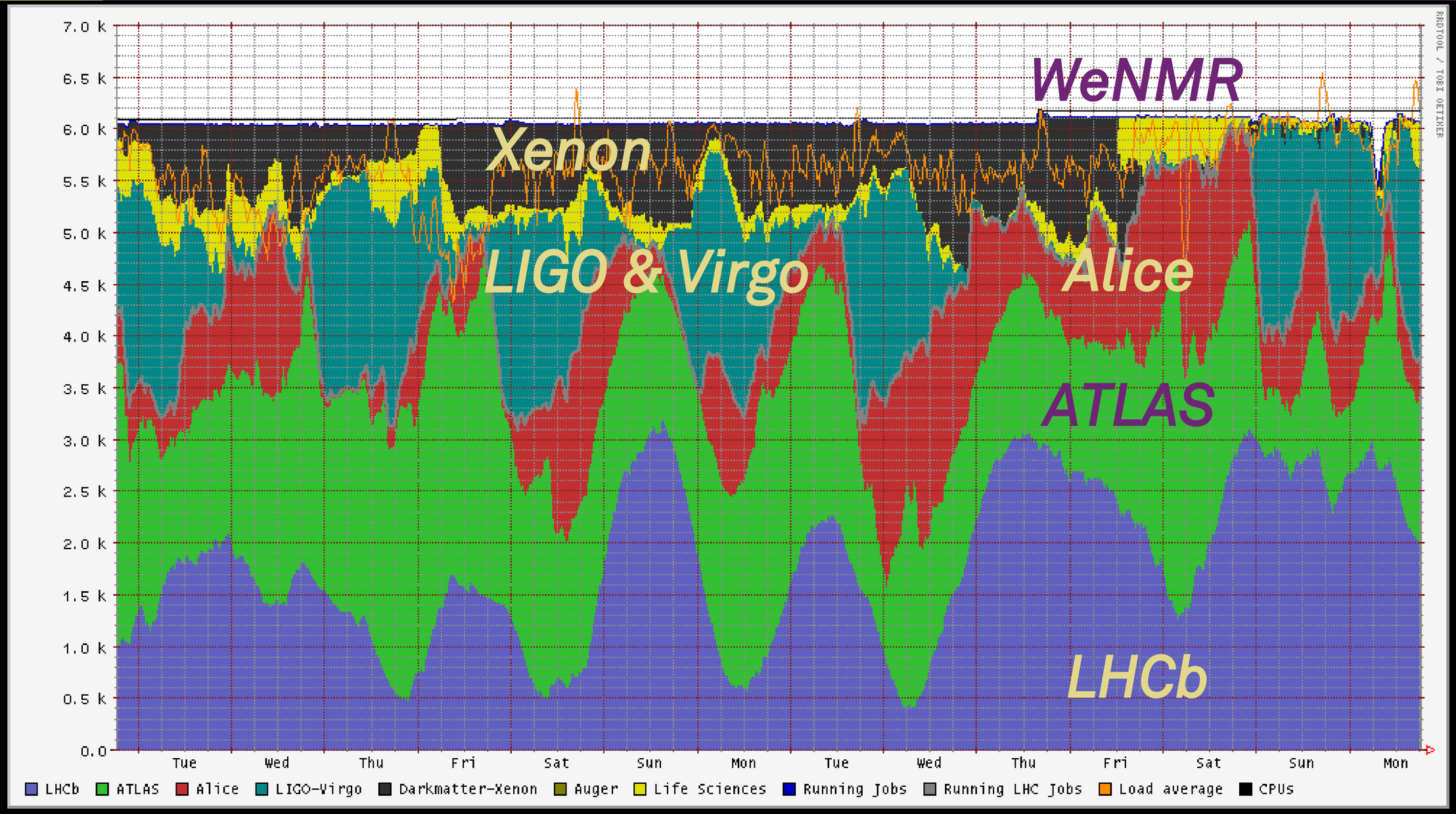


Lang leve … de wachtlijst

```
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
```

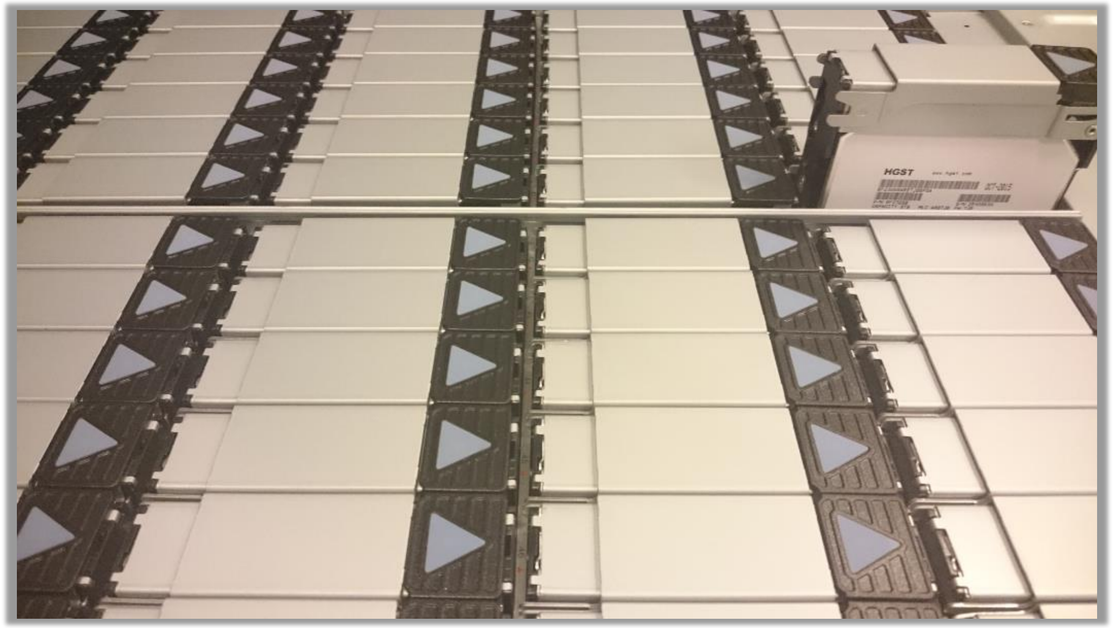
Nikhef

Zo hou je een rekencluster vol:



Verdeel en heers

DotHill (HGST): 480 TByte gross capacity/4RU



Kengetallen van opslag

- capaciteit (‘terabytes’)
- bandbreedte (‘megabyte per seconde’)
- aantal ‘opdrachten’ per seconds (“IOPS”)

Daarom kan onze data niet op een USB stick – en doet je ‘thuis NAS’ oplossing het ook niet

… hoe leuk ik mijn eigen opslagdoosje ook vind van slechts 15 Watt met 16Terabyte voor € 915 …

5 Petabyte lezen in 1 dag?
Dat is dus 61 GByte per seconde!
(en dus ~500 Gbps)

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 kW-jaar) aan de studenten leveren!


Nikhef
Intel Ethernet
Great Capability
Gbit LAN
Intel Ethernet
Great Capability
Gbit LAN
... want dit is wel leuk, maar niet echt sustainable ...
Collaboration of Intel™ and Nikhef PDP & MT (Krista de Roo) "CO_2 Inside"

Take a T-Byte!

Nikhef

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