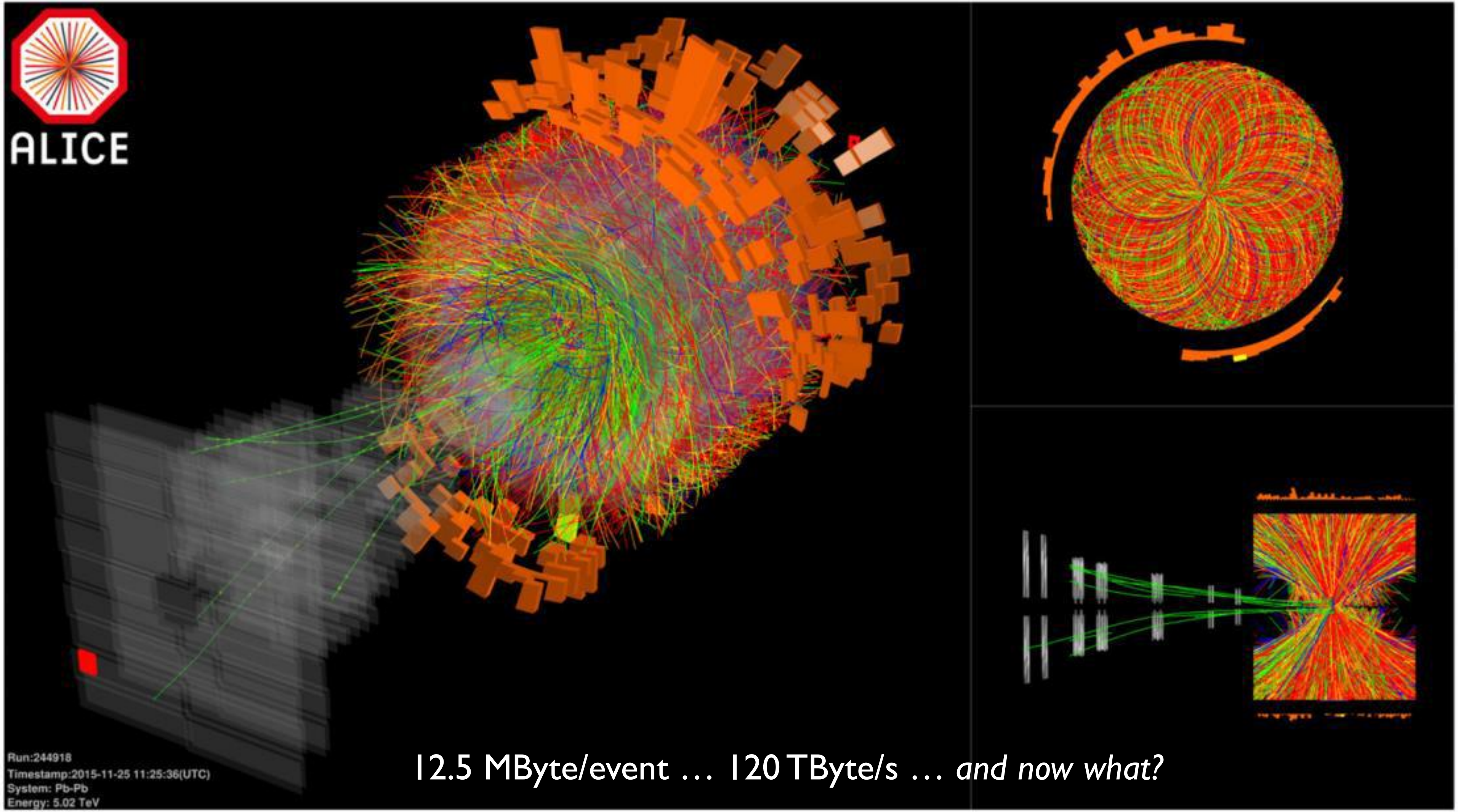
Accelerating Throughput – from the LHC to the World

David Groep

Nikhef

David Groep
Nikhef
PDP –
Advanced Computing for Research

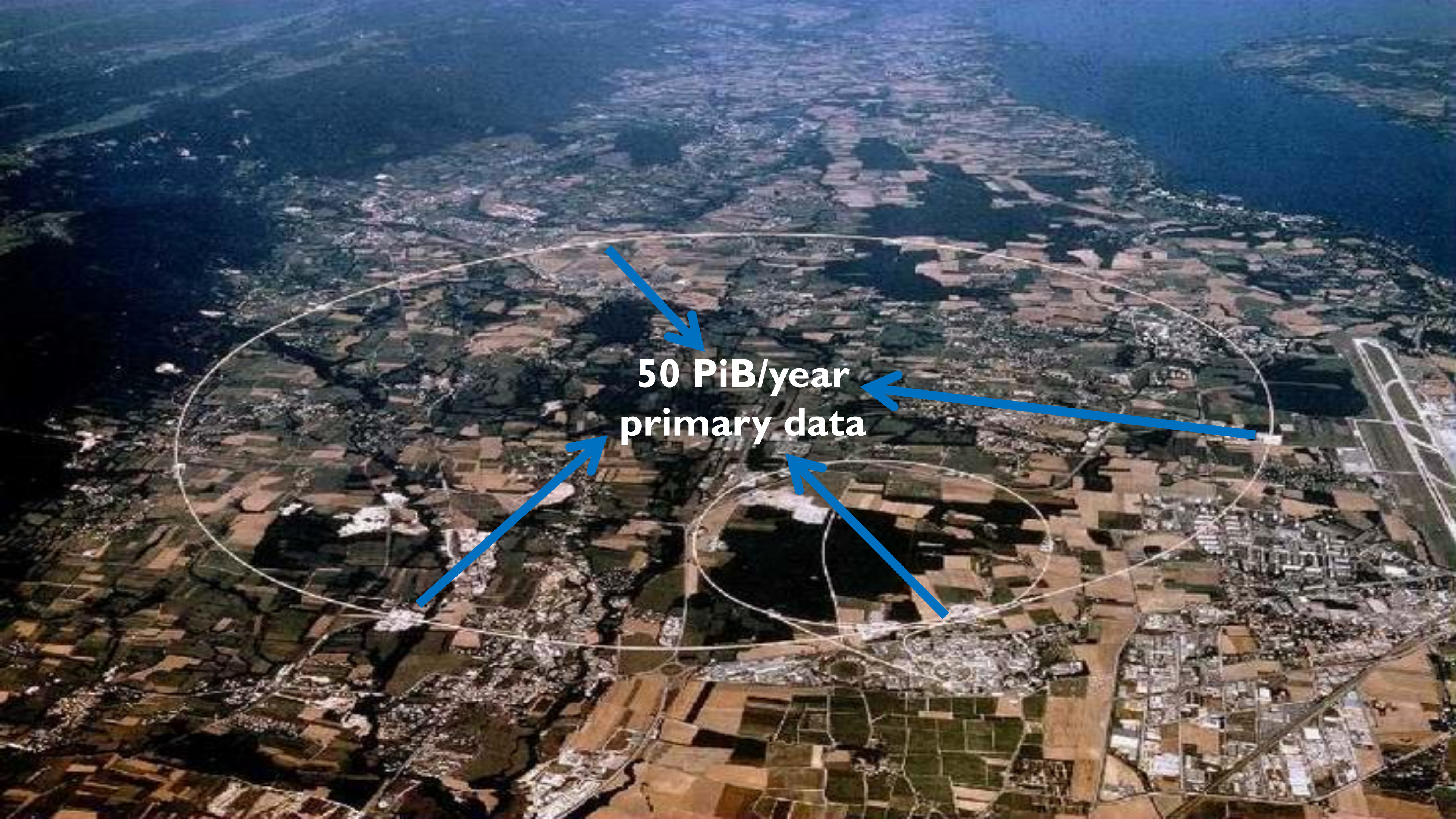


12.5 MByte/event … 120 TByte/s … *and now what?*

Kans Higgs deeltje:

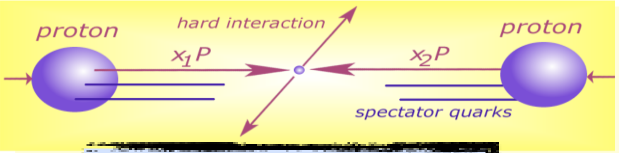
1 op de 1.000.000.000.000 bostingen

- Dit is equivalent met zoeken van 1 persoon op 1000 wereldpopulaties
- Oftewel één naald in 20 miljoen hooibergen

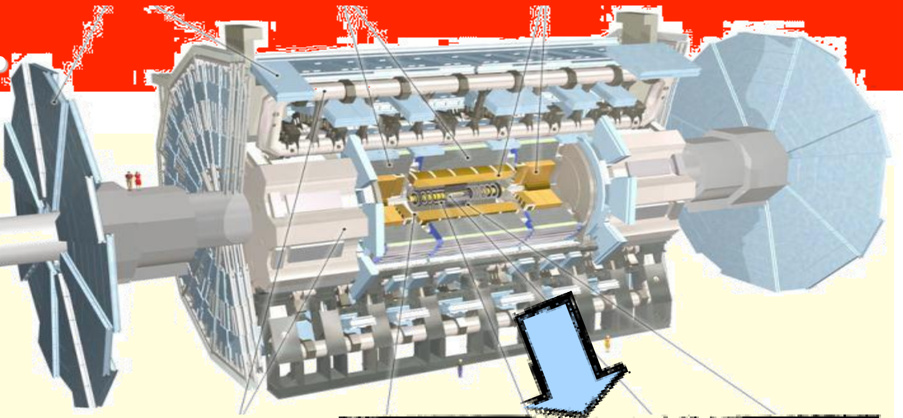


50 PiB/year primary data

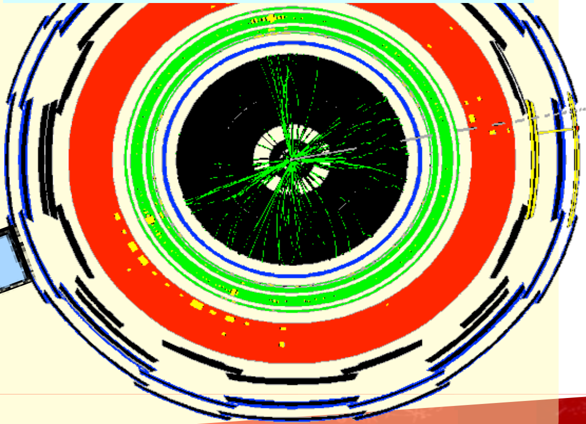
Detector to doctor …



40 miljoen / seconde



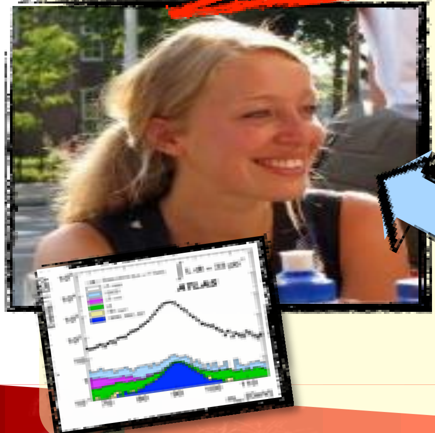
Trigger systeem selecteert 600 Hz
~ 1 GB/s data



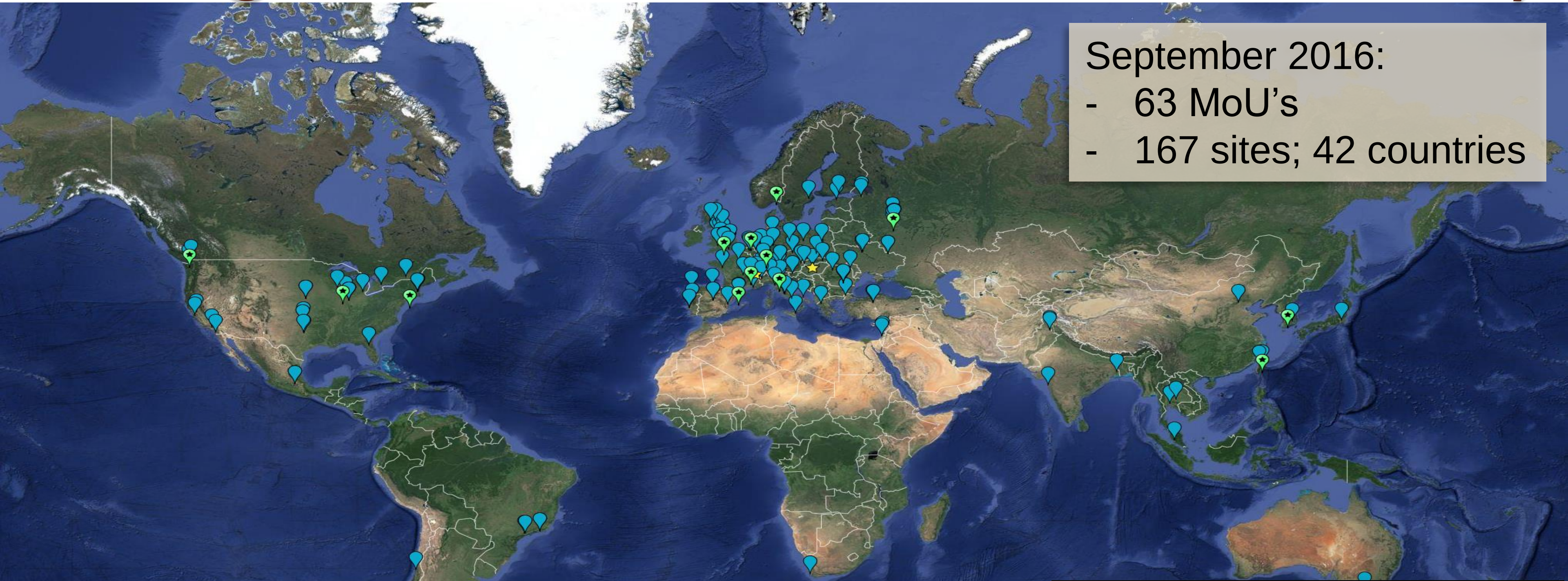
Analyse van botsingen door promovendi

and processing

Data distributie met GRID computers



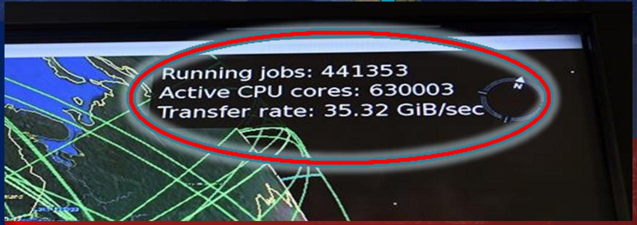
Building the Infrastructure … in a federated way



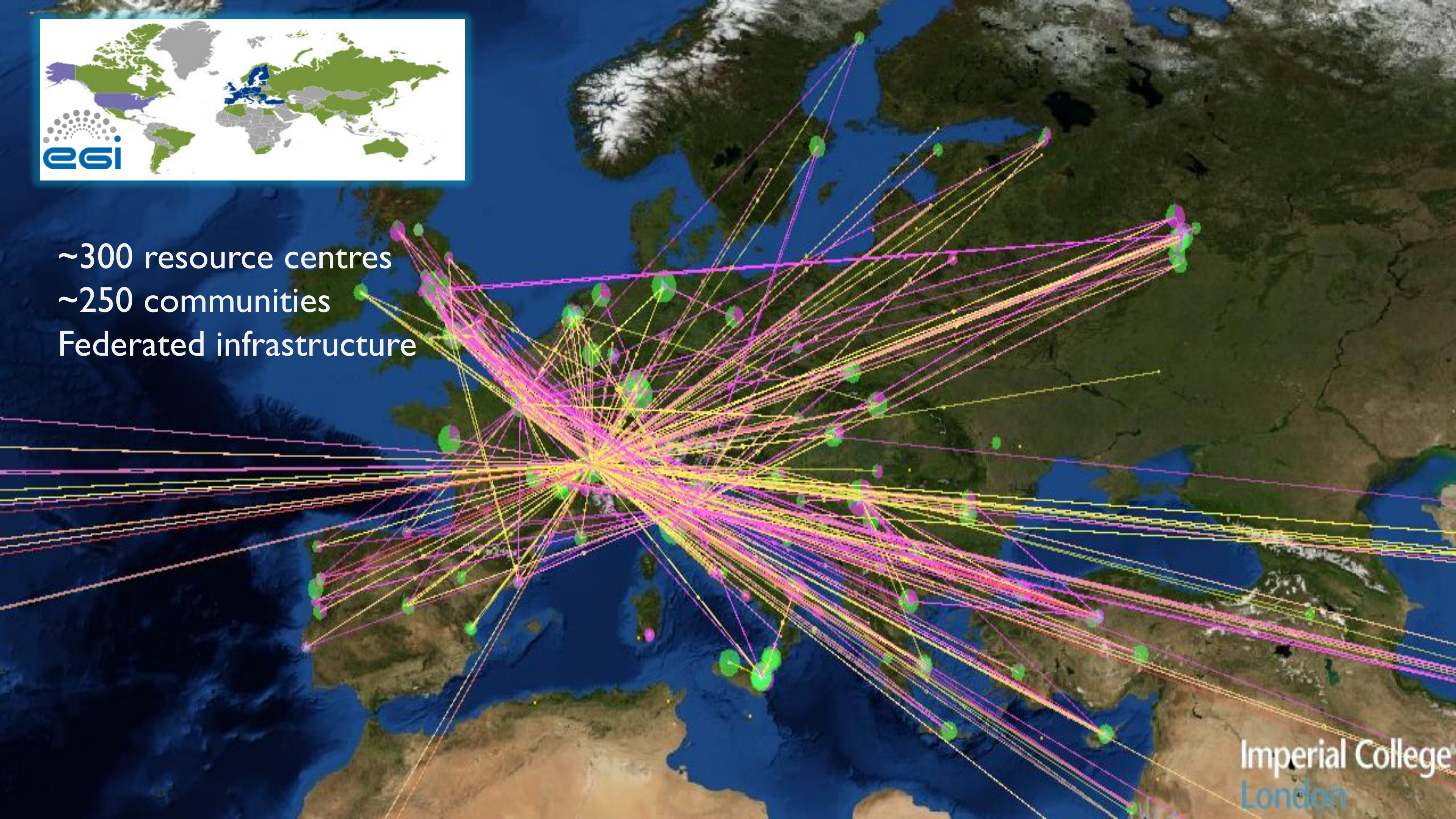
September 2016:

- 63 MoU's
- 167 sites; 42 countries

- ❑ CPU: 3.8 M HepSpec06
  - If today's fastest cores: ~ 350,000 cors
  - Actually many more (up to 5 yr old cores)
- ❑ Disk 310 PB
- ❑ Tape 390 PB



Running jobs: 441353
Active CPU cores: 630003
Transfer rate: 35.32 GiB/sec



~300 resource centres

~250 communities

Federated infrastructure

Global collaboration – in a secure way

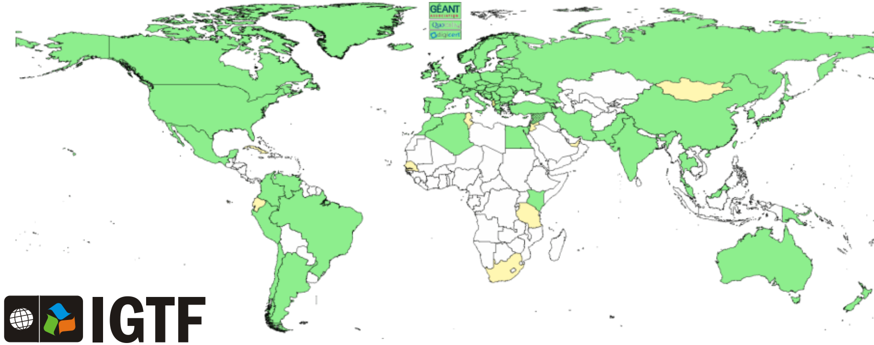
Collaboration is people as well as (or even more than) systems

A global identity federation for e-Infra and cyber research infrastructures

- Common baseline assurance (trust) requirements
- Persistent and globally unique

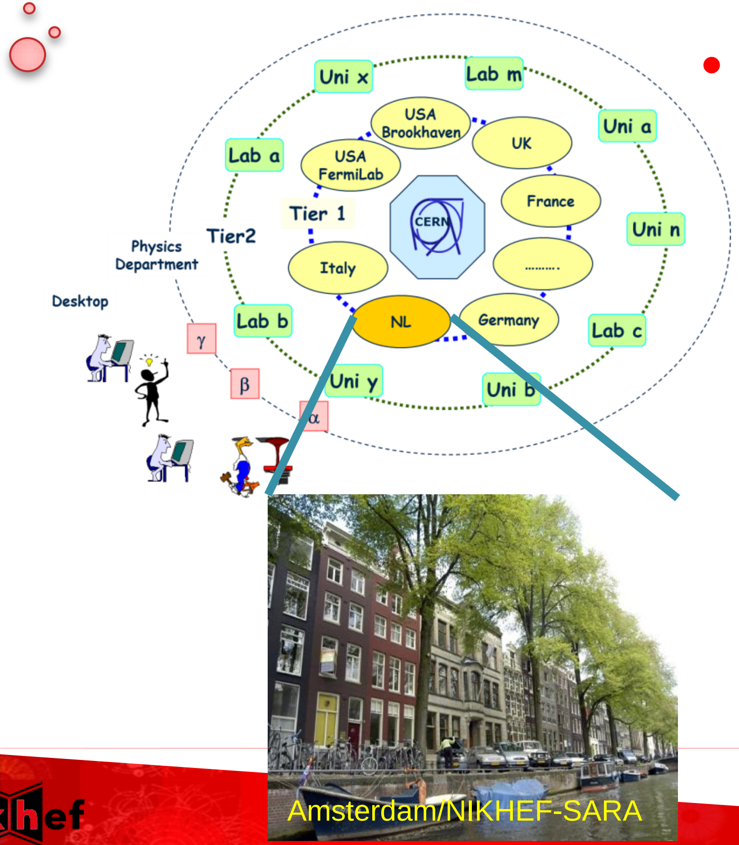
needs a global scope – so we built the Interoperable Global Trust Federation

- over 80 member Authorities
- Including your GÉANT Trusted Certificate Service

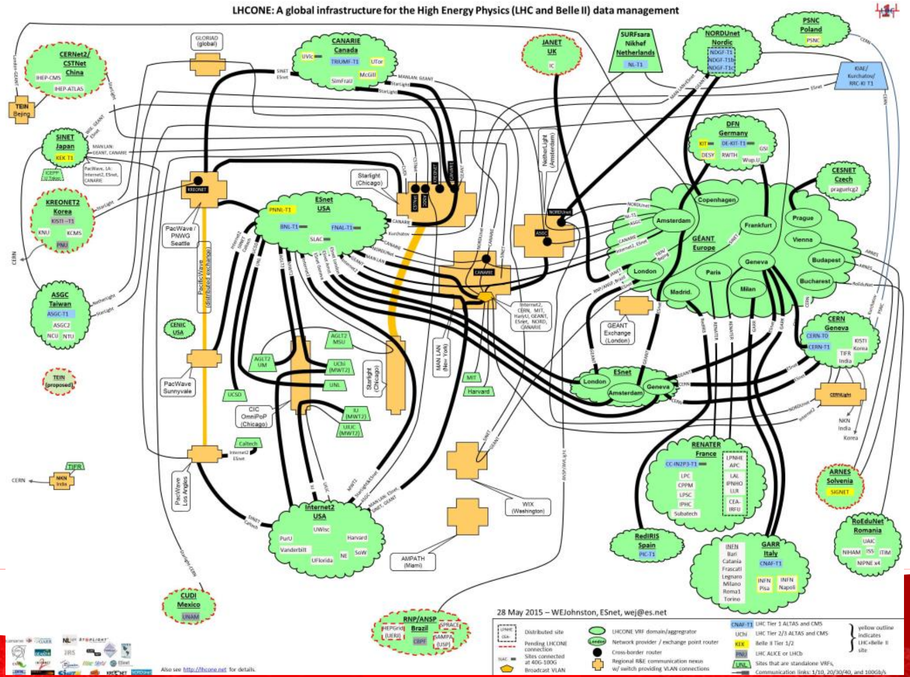


IGTF
Interoperable Global Trust Federation
AP | EU | TAG

Building the infrastructure for the LHC data

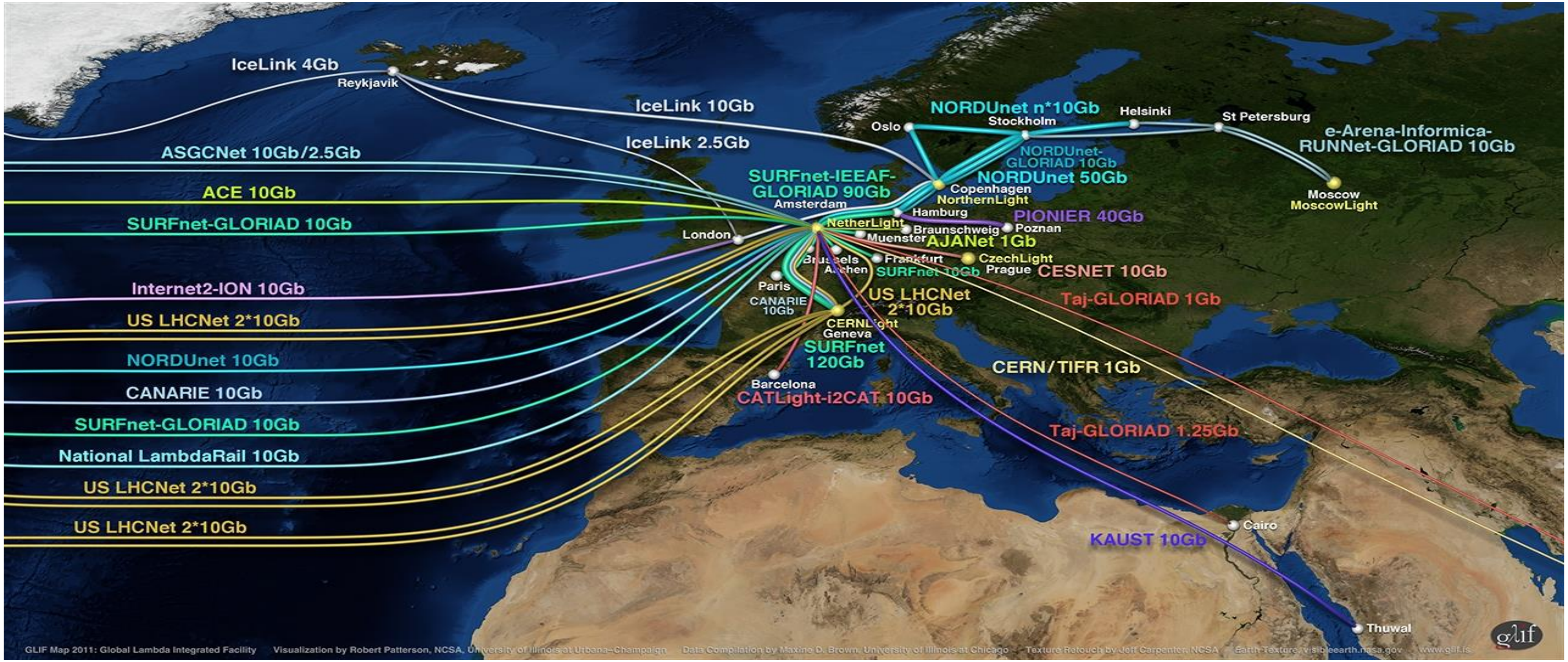


- From hierarchical data distribution to a full mesh and dynamic data placement


LHCONE: A global infrastructure for the High Energy Physics (LHC and Belle II) data management

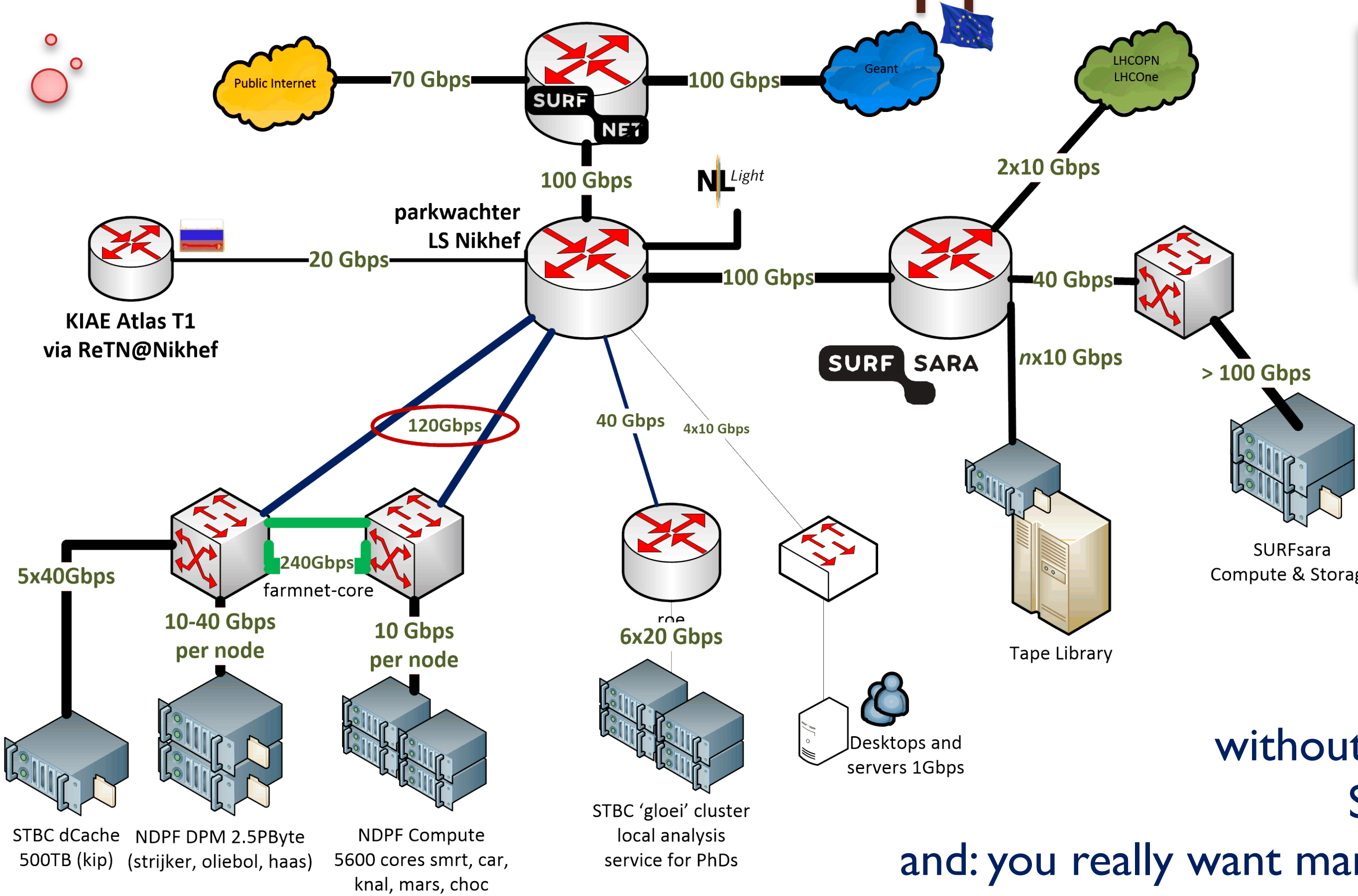
LHCOne graphic: lhcone.net

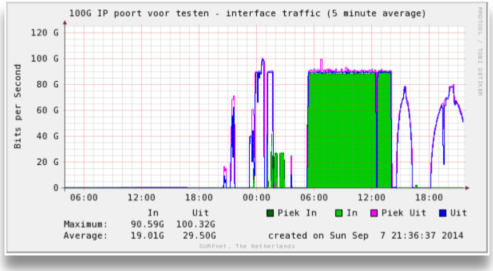
Connecting Science through Lambdas



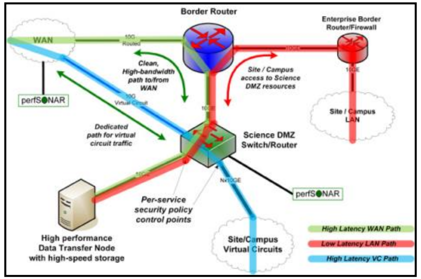
GLIF Map 2011: Global Lambda Integrated Facility   Visualization by Robert Patterson, NCSA, University of Illinois at Urbana-Champaign   Data Compilation by Maxine D. Brown, University of Illinois at Chicago   Texture Retouch by Jeff Carpenter, NCSA   Earth Texture: visibleearth.nasa.gov   www.glif.is

Network built around application data flow





```
            In        Uit          ■ Piek In  ■ In  ■ Piek Uit  ■ Uit
Maximum:    90.59G    100.32G
Average:    19.01G    29.50G       created on Sun Sep  7 21:36:37 2014
SURFnet, The Netherlands
```

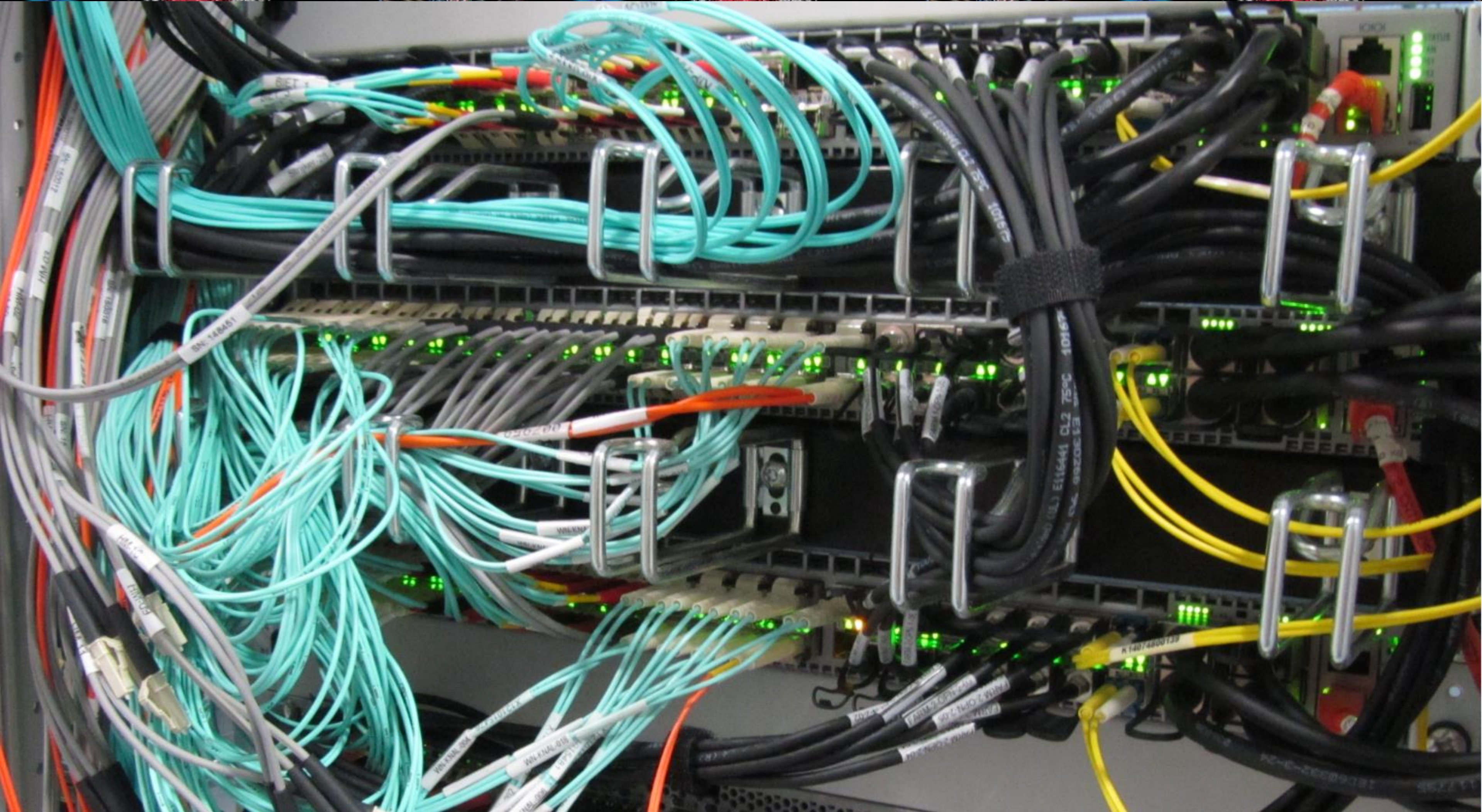


***Need** to work together!*

without our SURFsara peering,

SURFnet gets flooded ☺

and: you really want many of your own peerings



SN: 148451

Statistics

Dutch National e-Infrastructure coordinated by

"BiG Grid" HTC and storage platform services

- 3 core operational sites: SURFsara, Nikhef, RUG-CIT
- 25+ PiB tape, 10+ PiB disk, 12000+ CPU cores

@Nikhef

~ 5500 cores and 3.5 PiB

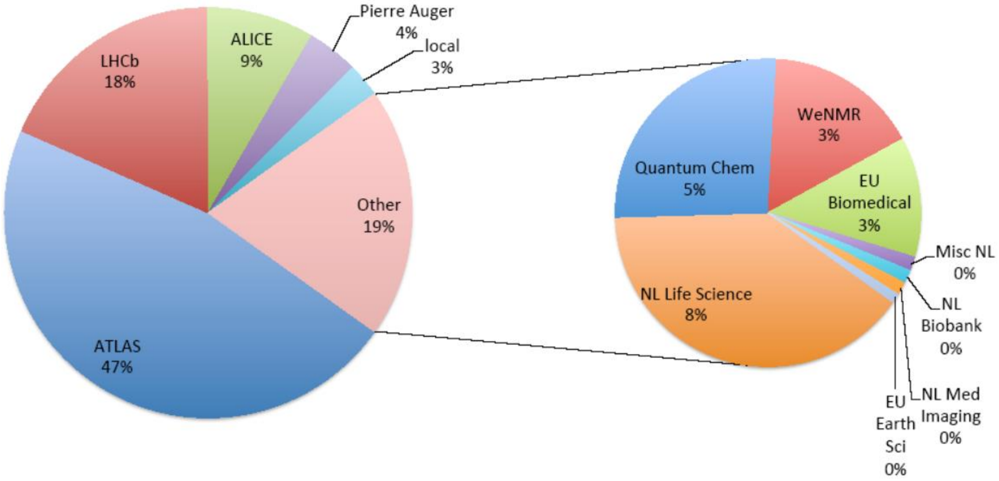
focus on large/many-core systems

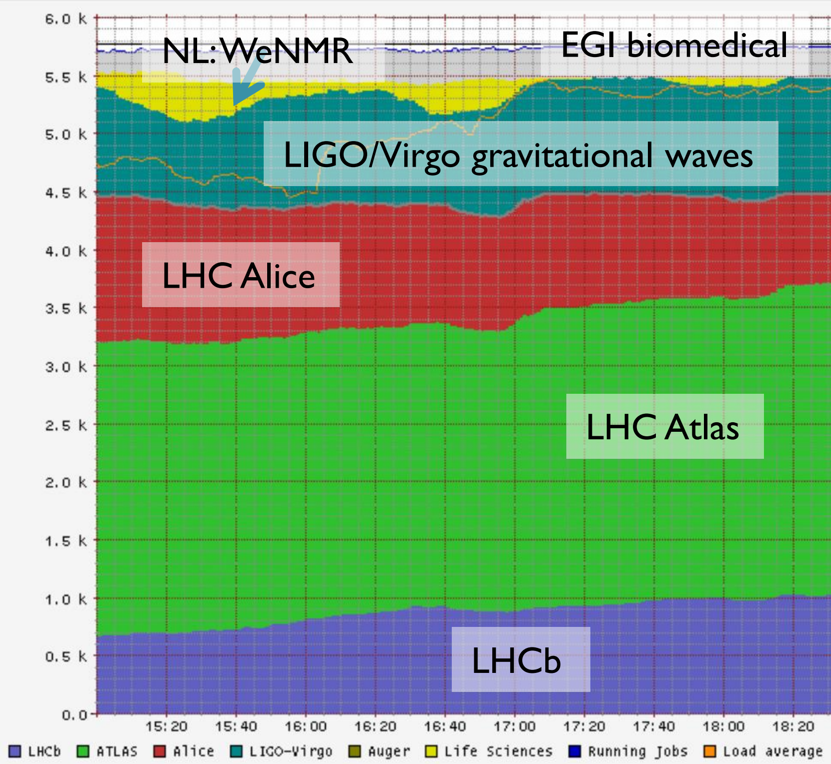
> 45 install flavours (service types)

and a bunch of one-off systems

Shared infrastructure, efficient infrastructure!

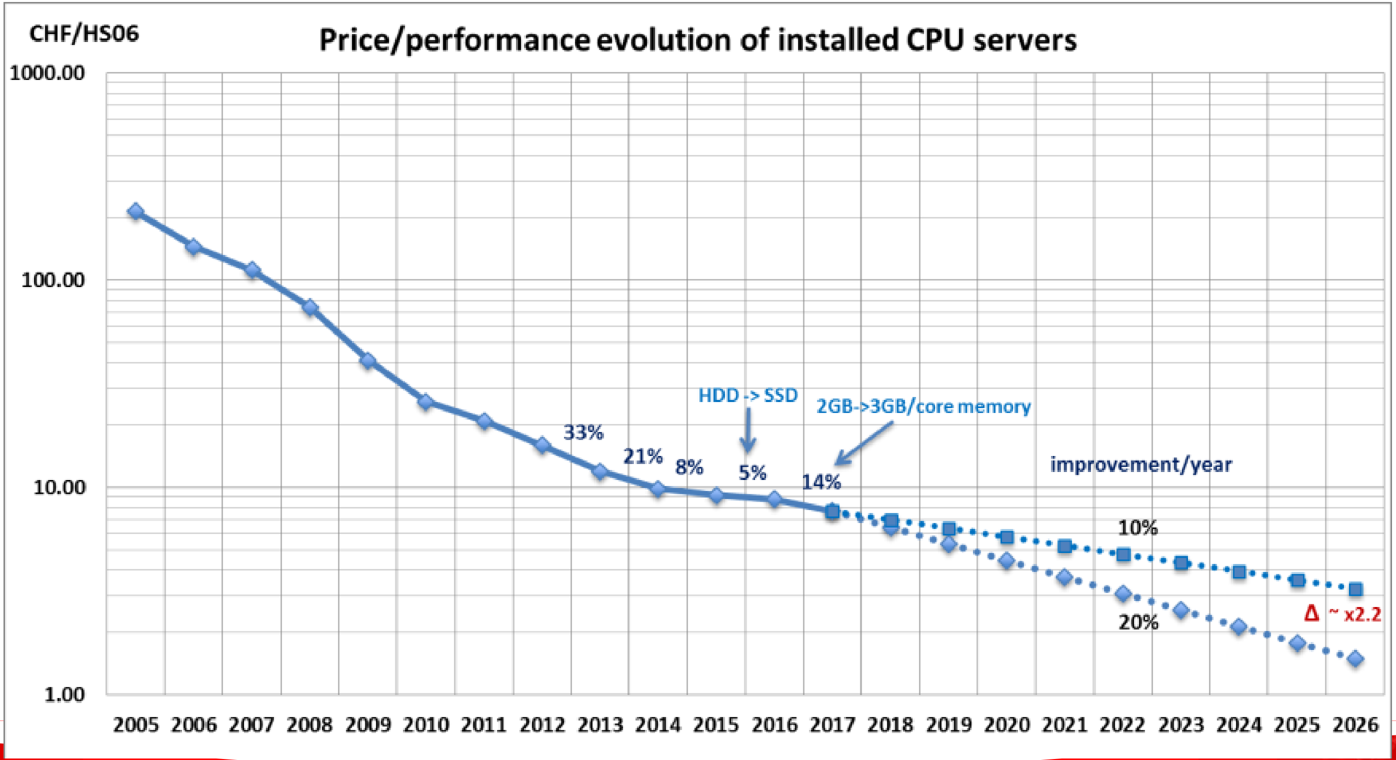
- >98% utilisation, >90% efficiency





Right:: NIKHEF-ELPROD facility, Friday, Dec 9th, 2016
Left: annual usage distribution 2013-2014

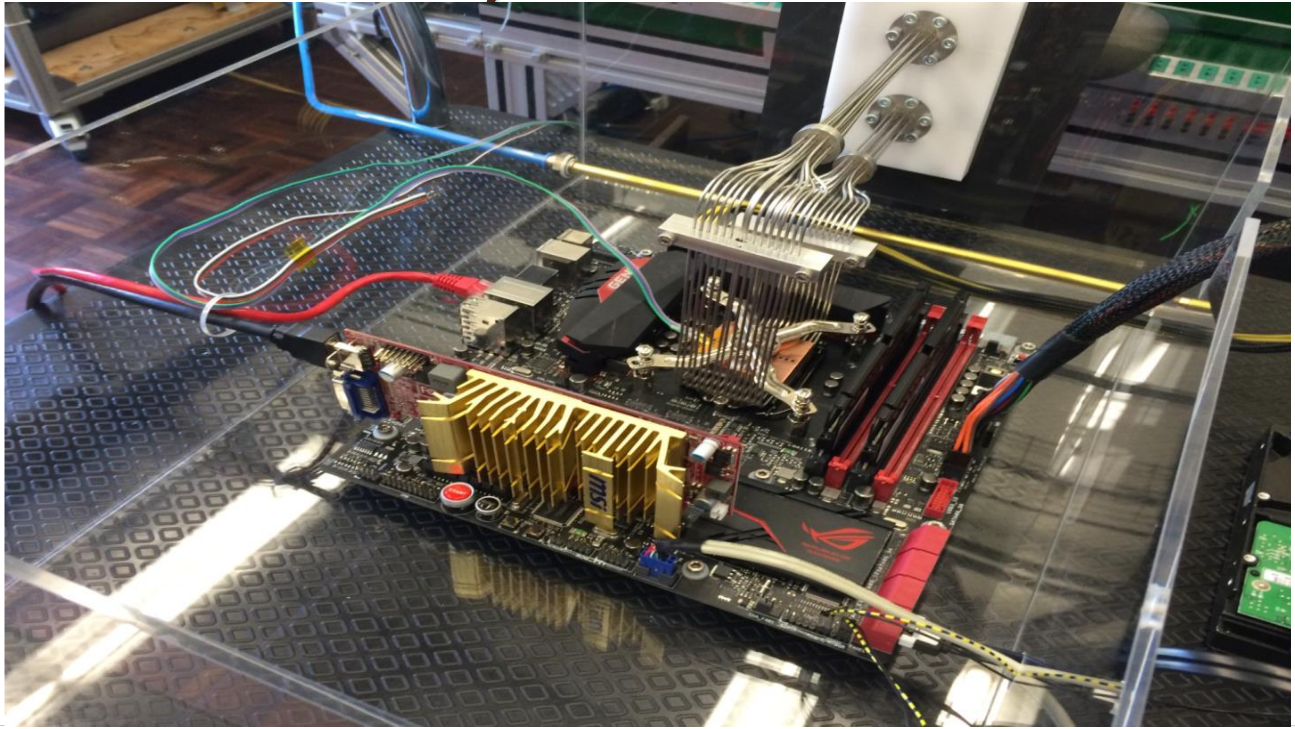
Waiting will not help you any more …



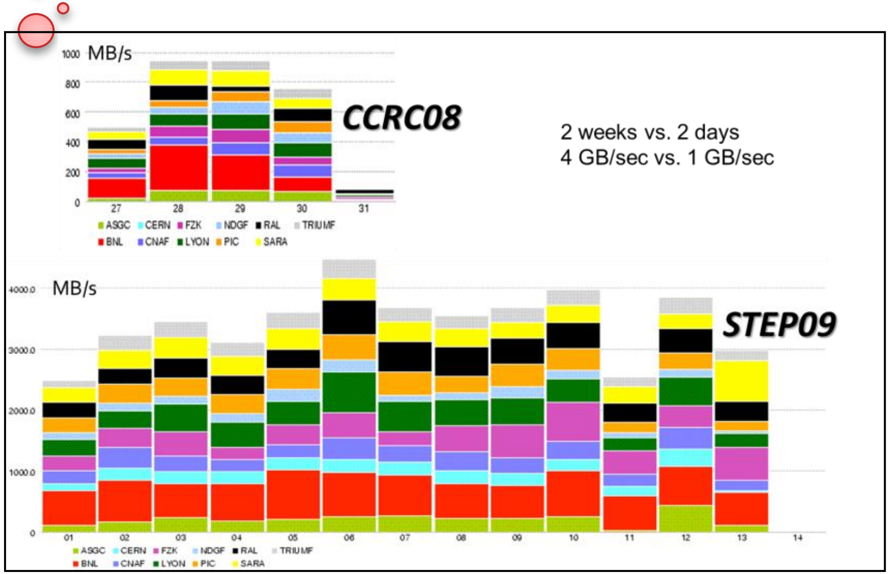
Helge Meinhard, Bernd Panzer-Steindel, Technology Evolution, https://indico.cern.ch/event/555063/contributions/2285842/

For (informed) fun & testing – some random one-off systems …

For (informed) fun & testing – some random one-off systems …

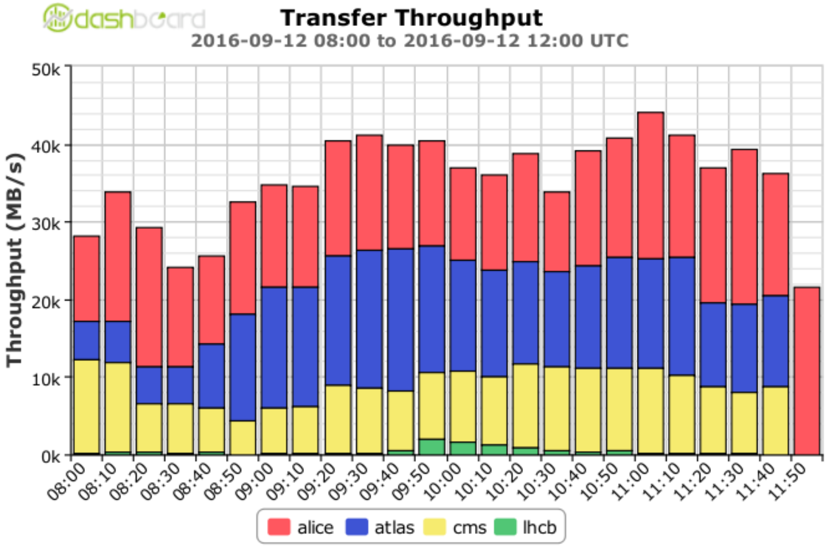


From SC04, CCRC08, STEP09, .. to today …



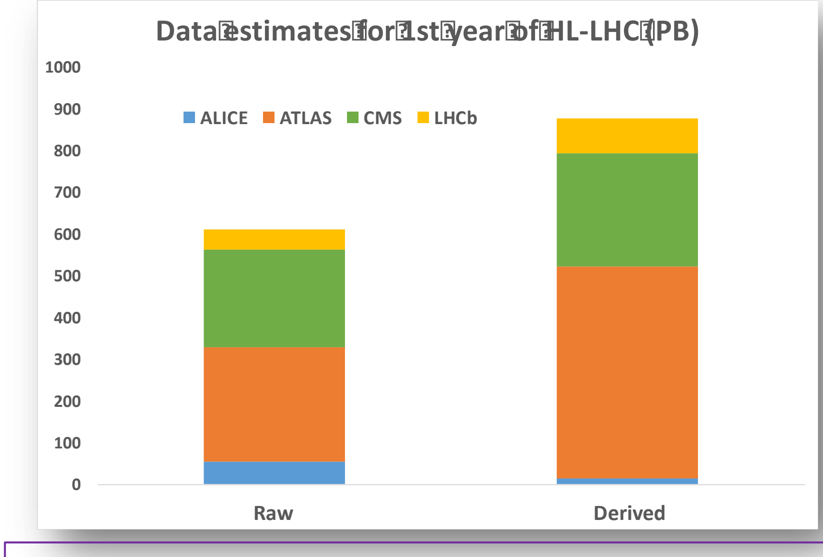
2 weeks vs. 2 days
4 GB/sec vs. 1 GB/sec

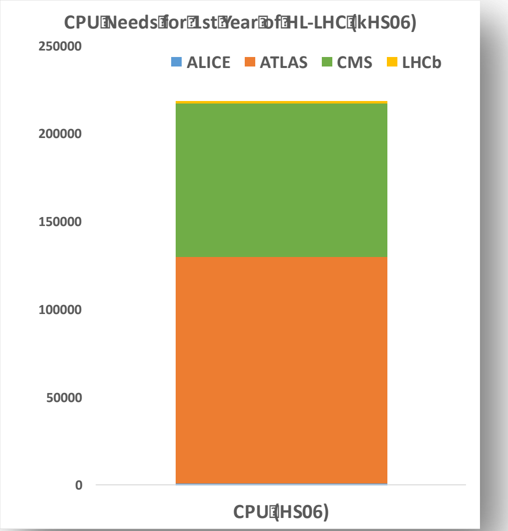
Global transfer rates increased to > 40 GB/s
Acquisition: 10 PB/mo (~x2 for physics data)



STEP09 : Jamie Shiers, CERN IT/GS; Throughput 2016: WLCG Workshop 2016, Ian Bird, CERN

… and tomorrow ?!





Data:

- Raw 2016: 50 PB → 2027: 600 PB
- Derived (1 copy): 2016: 80 PB → 2027: 900 PB

CPU:

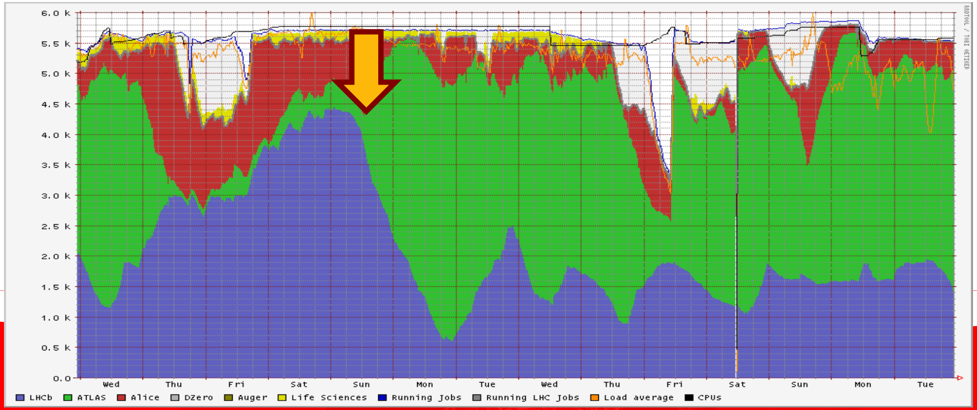
- x60 from 2016

Technology at ~20%/year will bring x6-10 in 10-11 years

Interconnecting compute & storage

‘data shall not be a bottleneck’

- 5500 cores process together
  ~ 16 GByte/s of data sustained
  or ~ 10 GByte/jobslot/hr
- are ‘bursty’ when many tasks start together
- and *in parallel* we have to serve the world



Infrastructure for research: balancing network, CPU, and disk

- CPU and disk both expensive, yet idling CPUs are ‘even costlier’
- architecture and performance matching averts any single bottleneck
- but requires knowledge of application (data flow) behaviour
  data pre-placement (local access), mesh data federation (WAN access)

This is why e.g. your USB drive does not cut it
– and neither does your ‘home NAS box’
… *however much I like my home system using just 15 Watt idle and offering 16TB for just € 915 …*

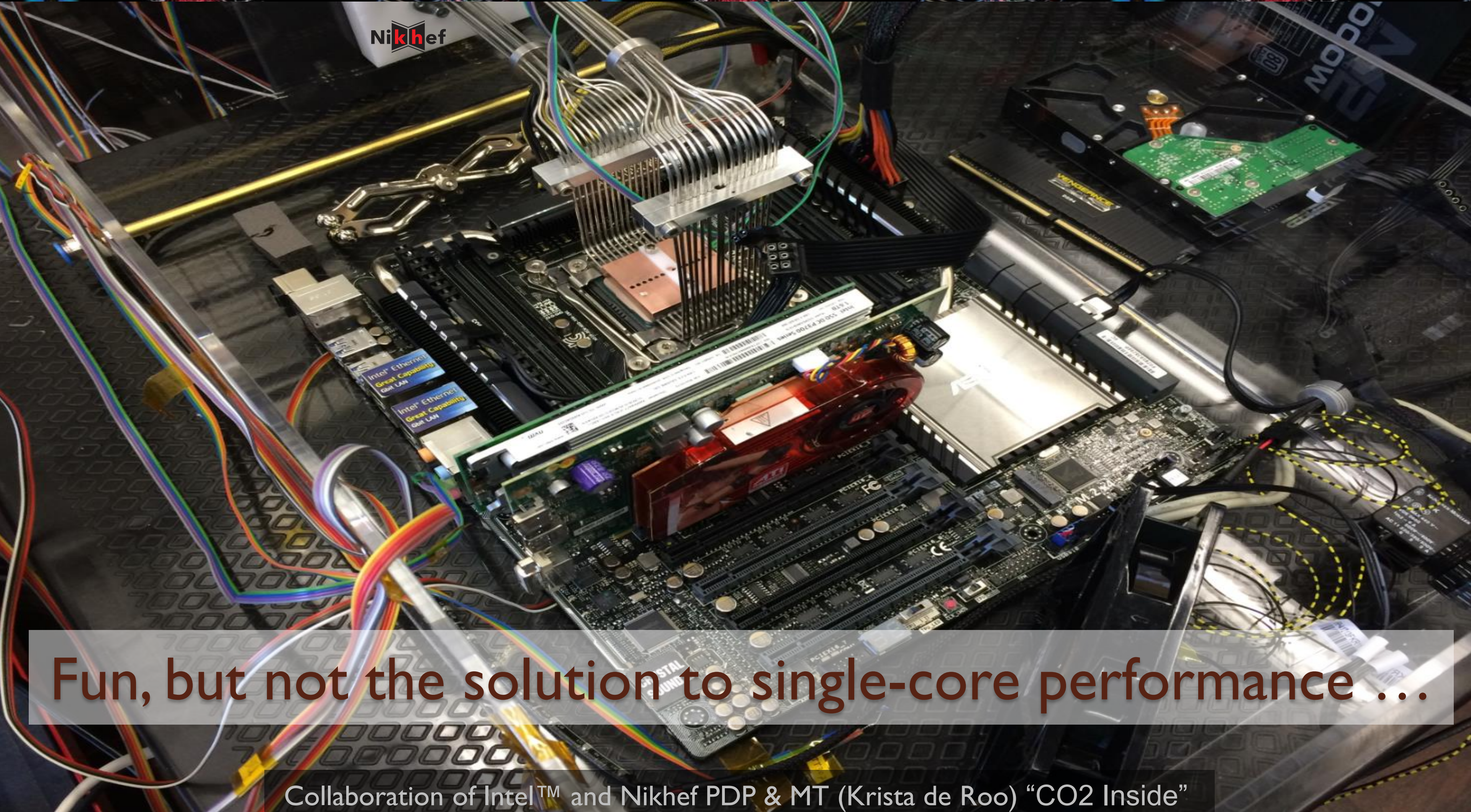
Getting more bytes through?

- Power 8: more PCI lanes & higher clock should give more throughput – *if all the bits fit together*
- Only way to find out is … by trying it!
  joint experiment with Nikhef and SURFsara on comparing IO throughput between x86 & P8

HGST: 480 TByte gross capacity/4RU

yet more is needed

- RAID card are now a performance bottleneck
- JBOD changes CPU-disk ratio
- closer integration of networking to get >100Gbps



Fun, but not the solution to single-core performance …

Collaboration of Intel™ and Nikhef PDP & MT (Krista de Roo) “CO2 Inside”