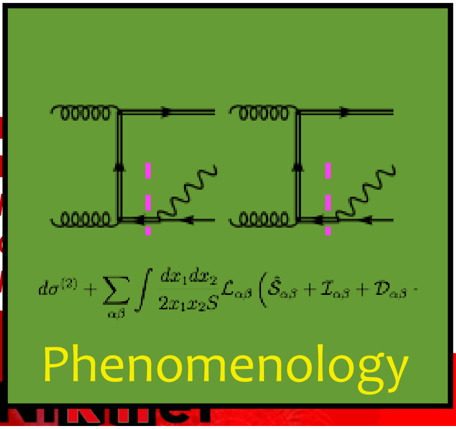
Nikhef – a Journey in Physics and Data Processing

David Groep
Nikhef
PDP - Advanced Computing for Research

Nationaal instituut voor subatomaire fysica

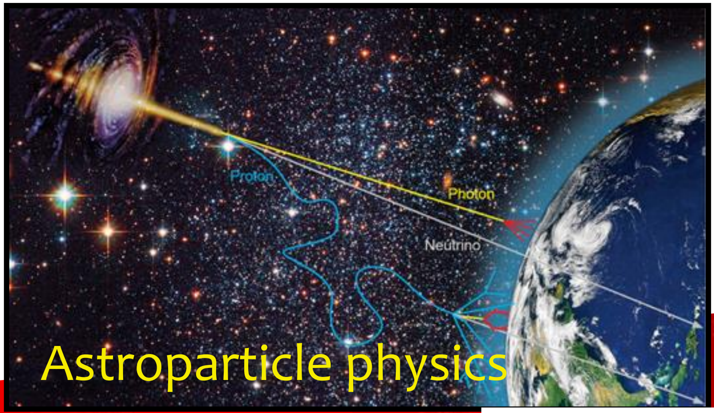
Verleggen van de grenzen van onze kennis

- **Accelerator-based particle physics**
  Experiments studying interactions in particle collision processes at particle accelerators, in particular at CERN;
- **Astroparticle physics**
  Experiments studying interactions of particles and radiation emanating from the Universe.


$d\sigma^{(2)} + \sum_{\alpha\beta} \int \frac{dx_1 dx_2}{2x_1x_2S} \mathcal{L}_{\alpha\beta} \left(\hat{S}_{\alpha\beta} + \mathcal{I}_{\alpha\beta} + \mathcal{D}_{\alpha\beta} -\right.$

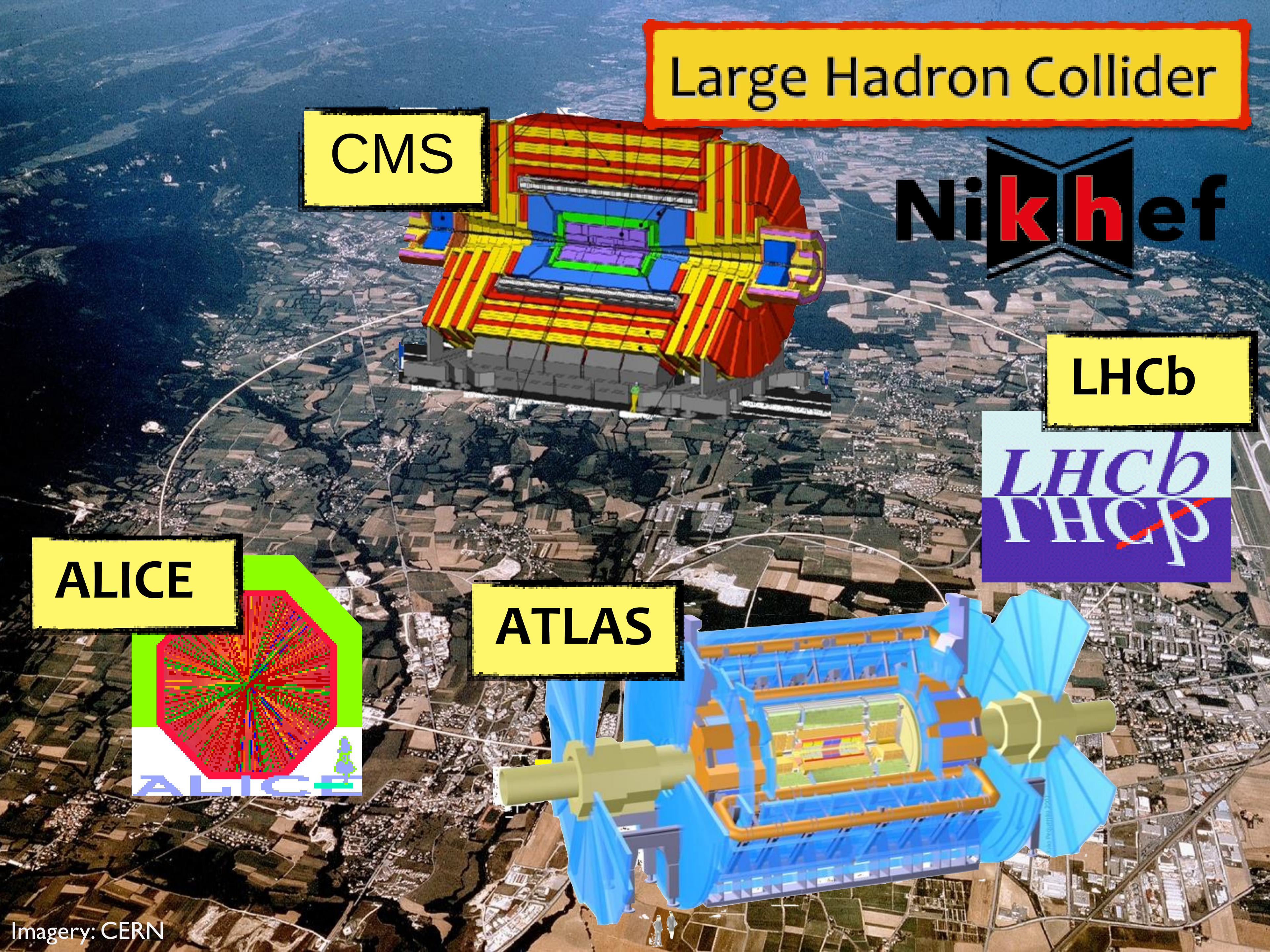
Phenomenology

Collider physics



Astroparticle physics

Large Hadron Collider



CMS

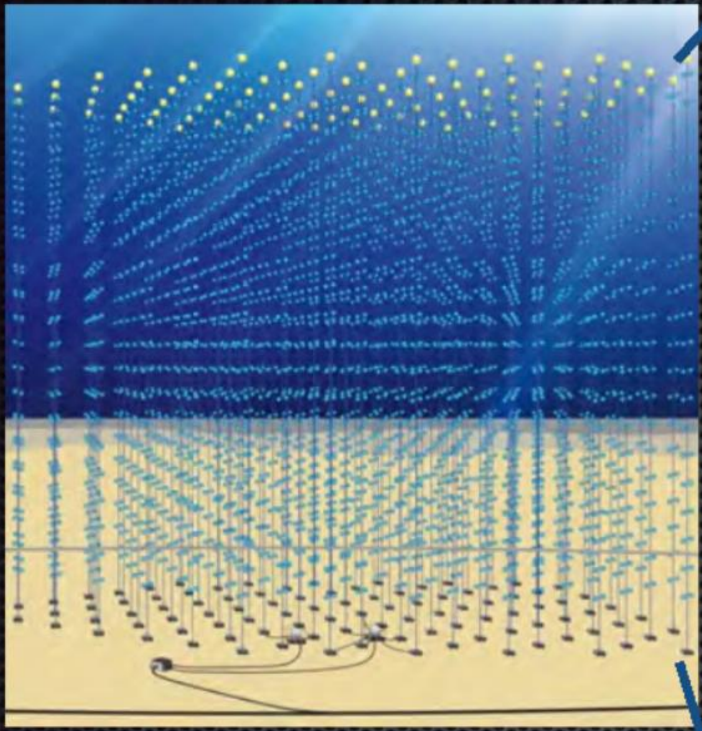
LHCb

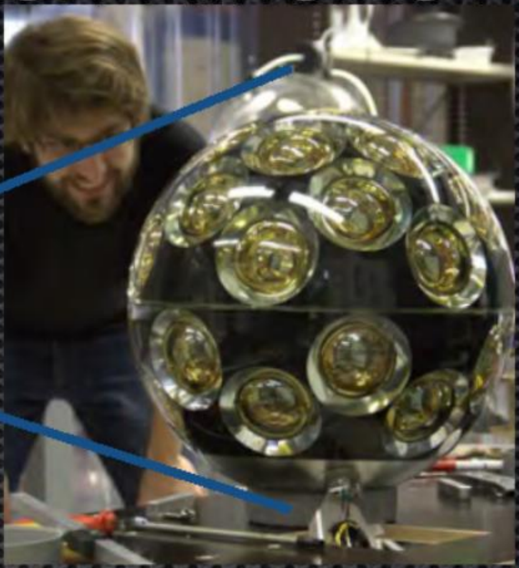
ALICE

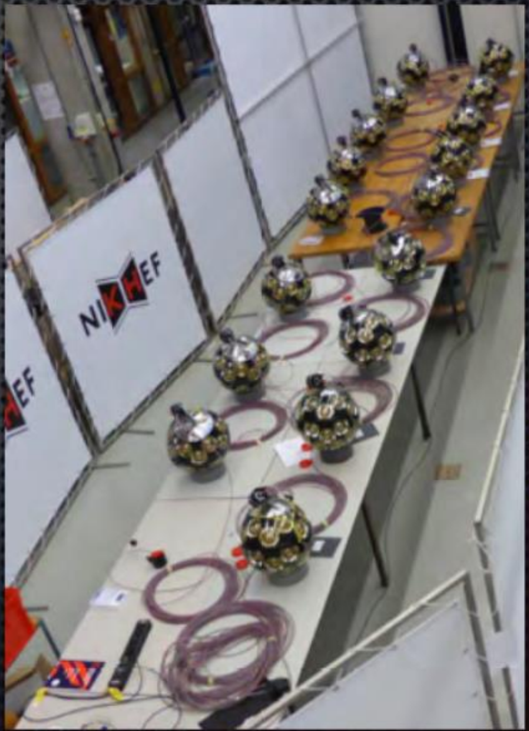
ATLAS

Imagery: CERN

Nikhefs neutrino-detector: KM3NeT

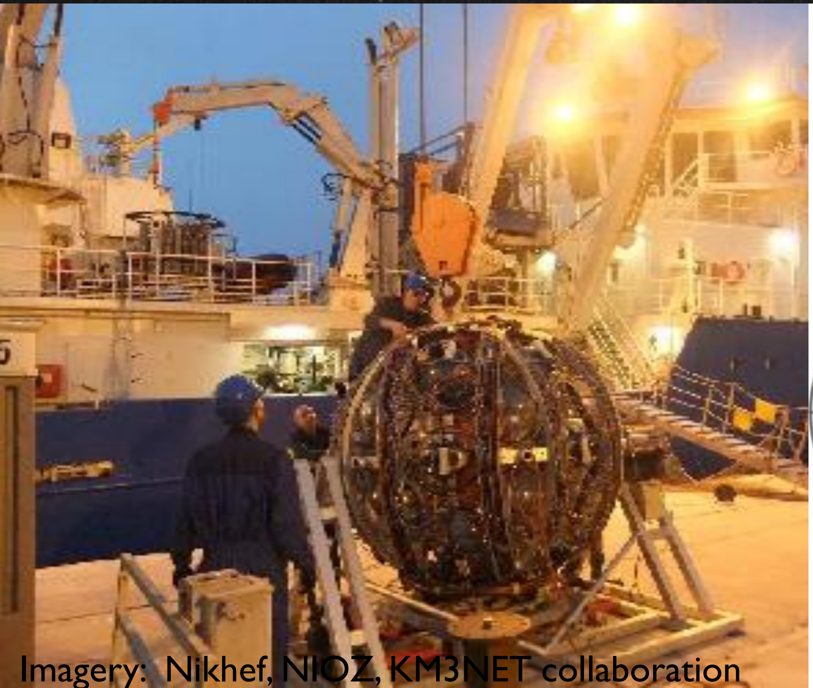




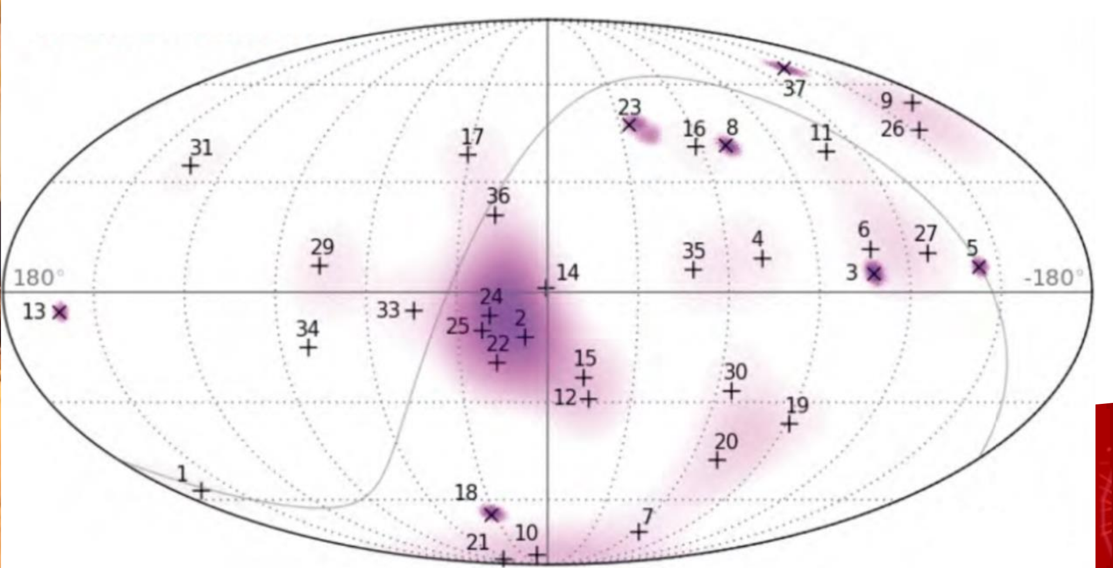


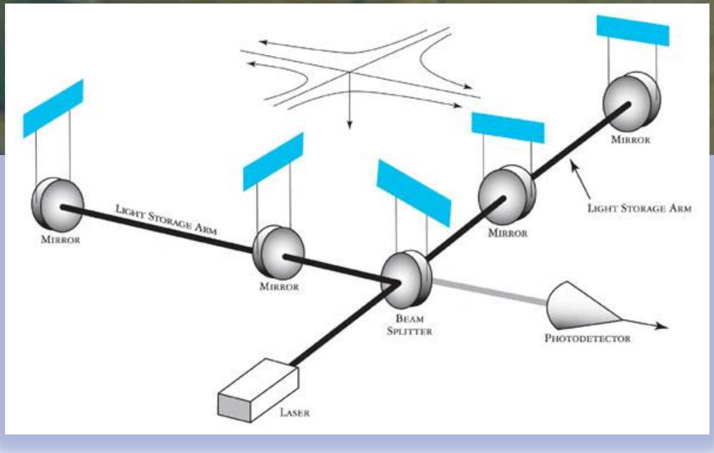
De Melkweg

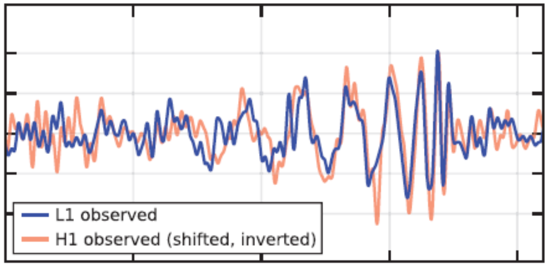
Little white structures prevent the HV bases and cables to touch each other



Imagery: Nikhef, NIOZ, KM3NET collaboration



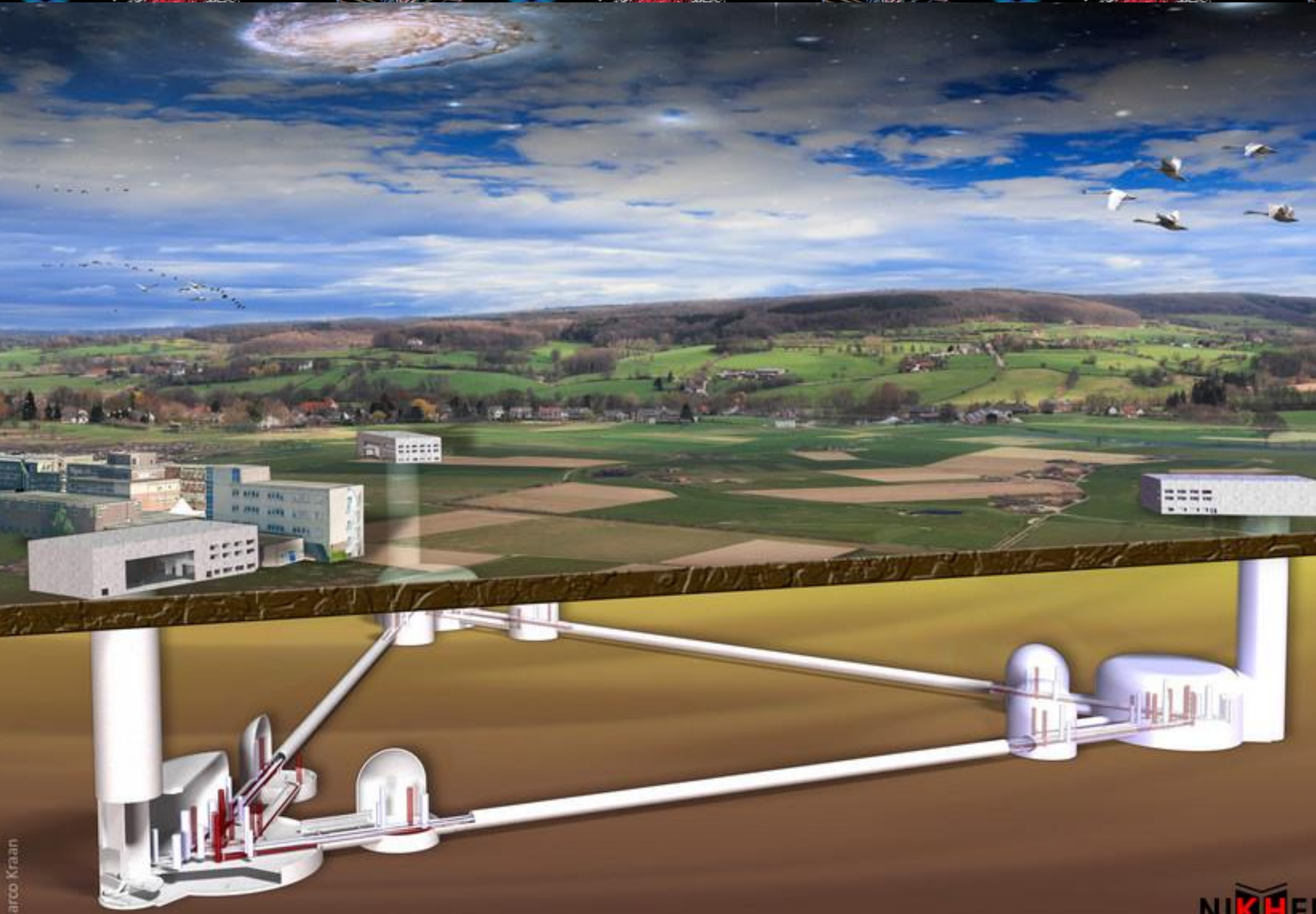




LSC LIGO Scientific Collaboration

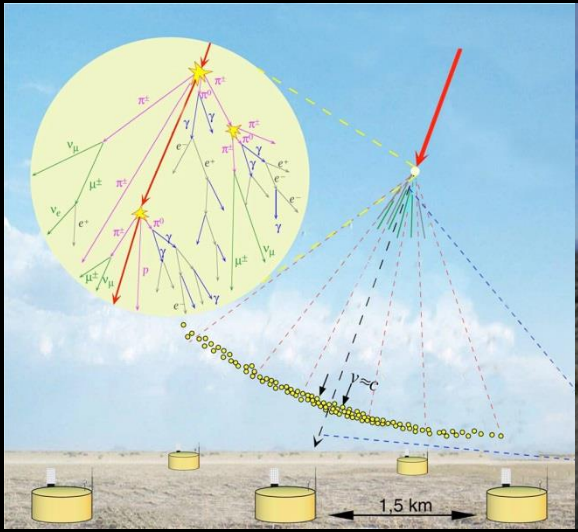
VIRGO

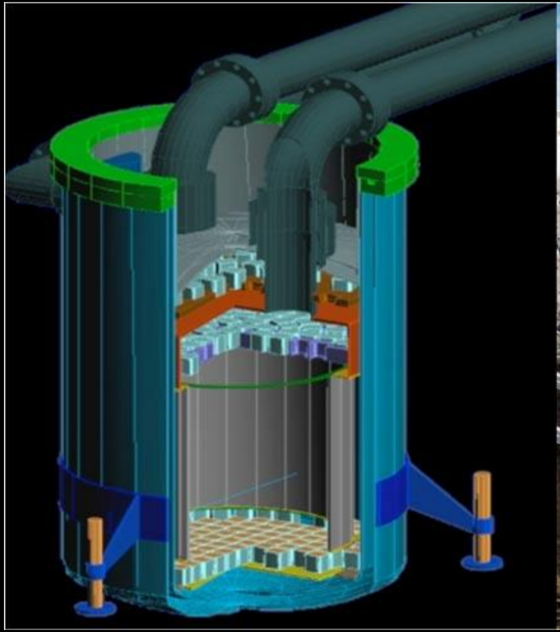
Imagery: gw-astronomy collaborations, LSC



© Marco Kraan

Einstein Telescope – projected onto Heuvelland Limburg, NL





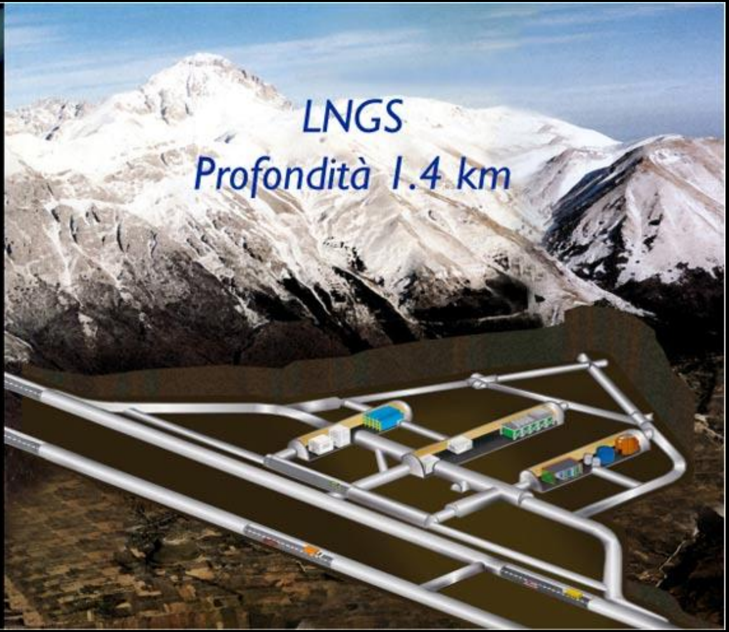
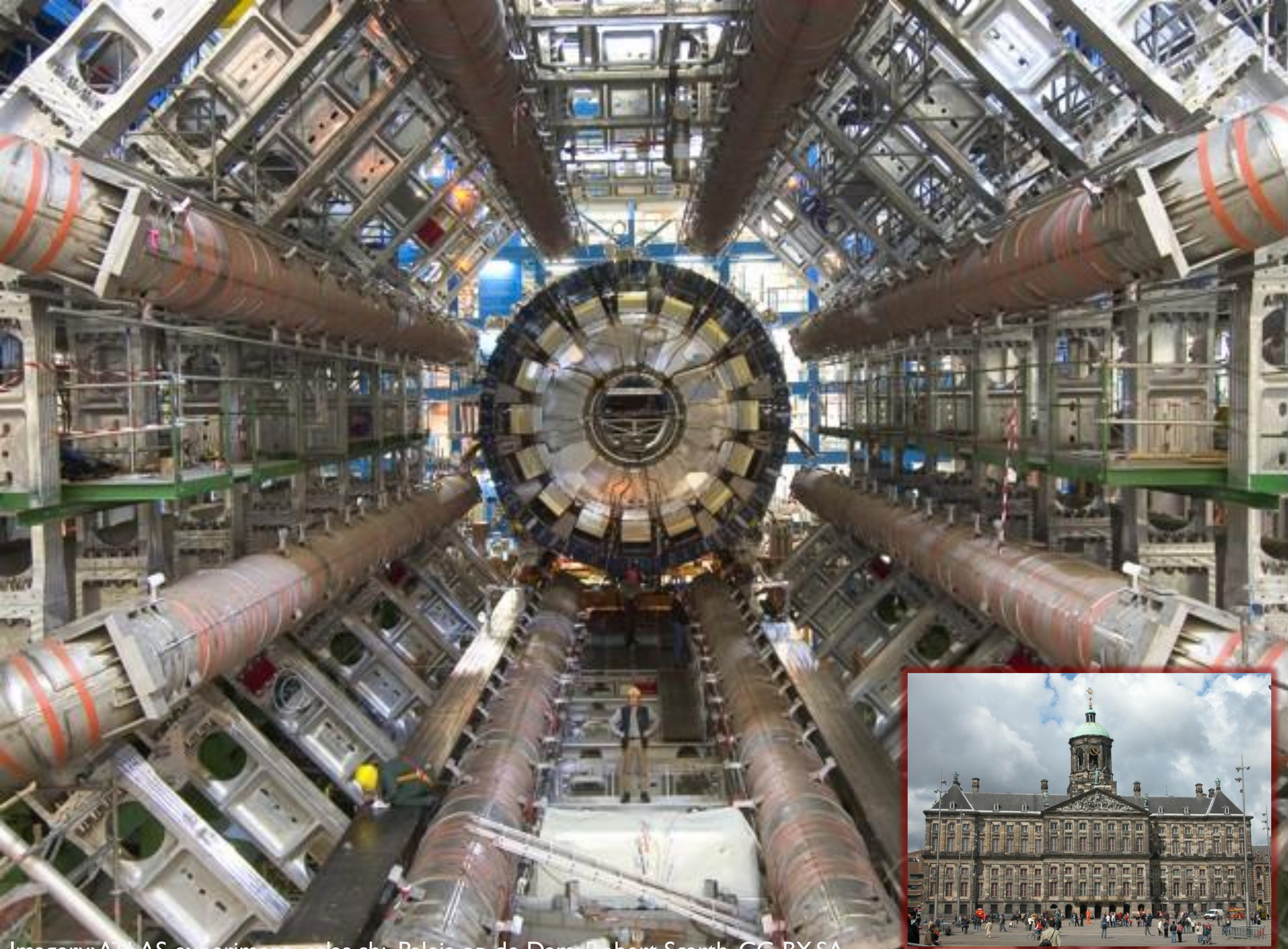


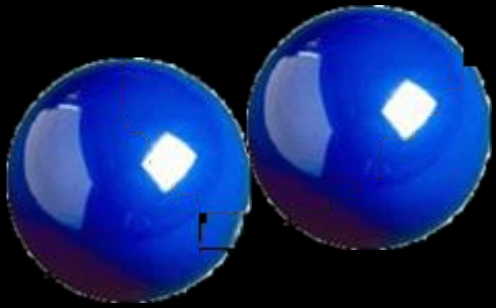
Image sources: LNGS/INFN, Xenon collaboration; Pierre Auger collaboration; Nikhef



Imagery:ATLAS experiment,, atlas.ch; Paleis op de Dam: Robert Scarth, CC-BY-SA

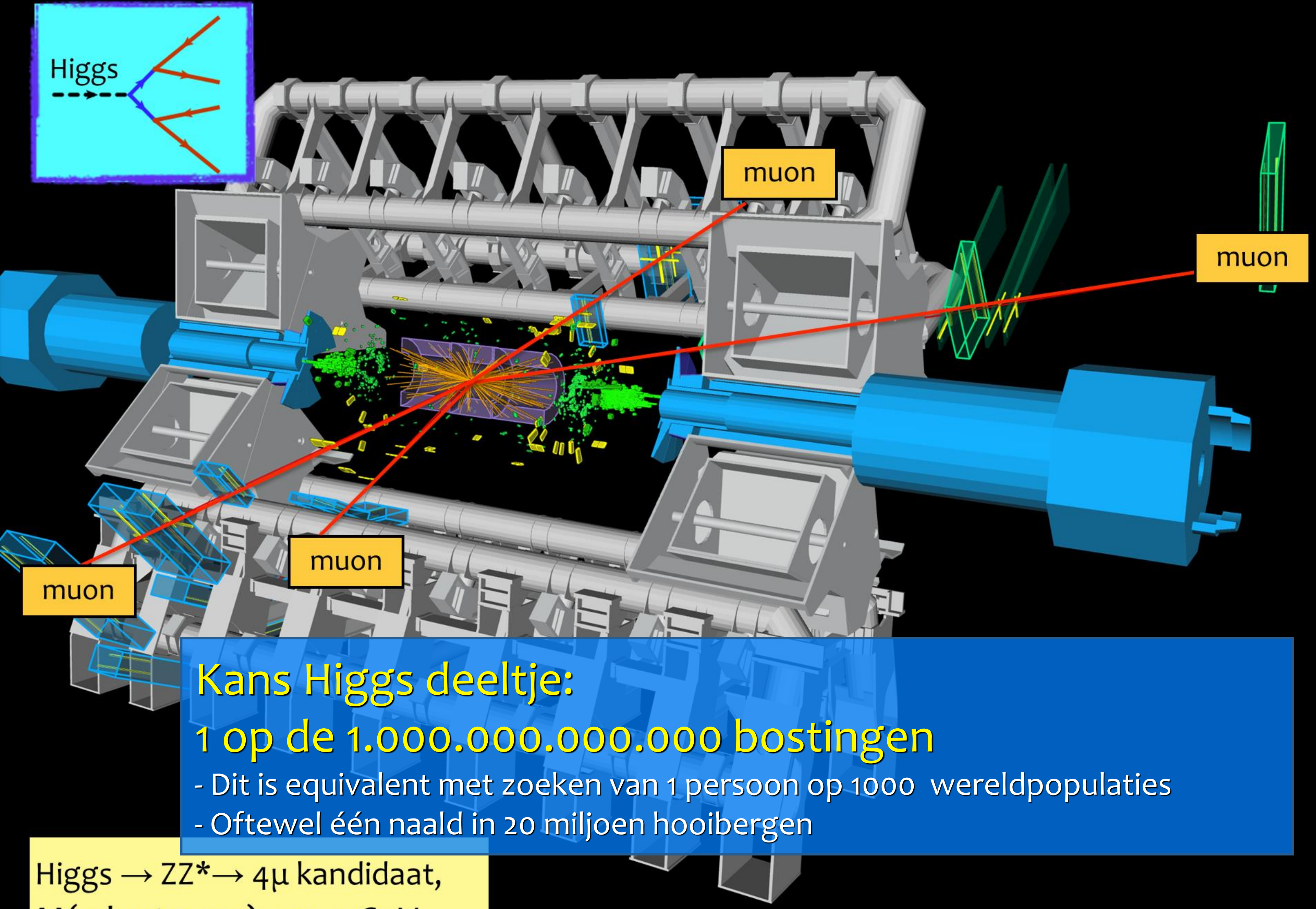
Deeltjes botsingen

Deeltjes botsingen



Deeltjes botsingen

$$E=mc^2$$



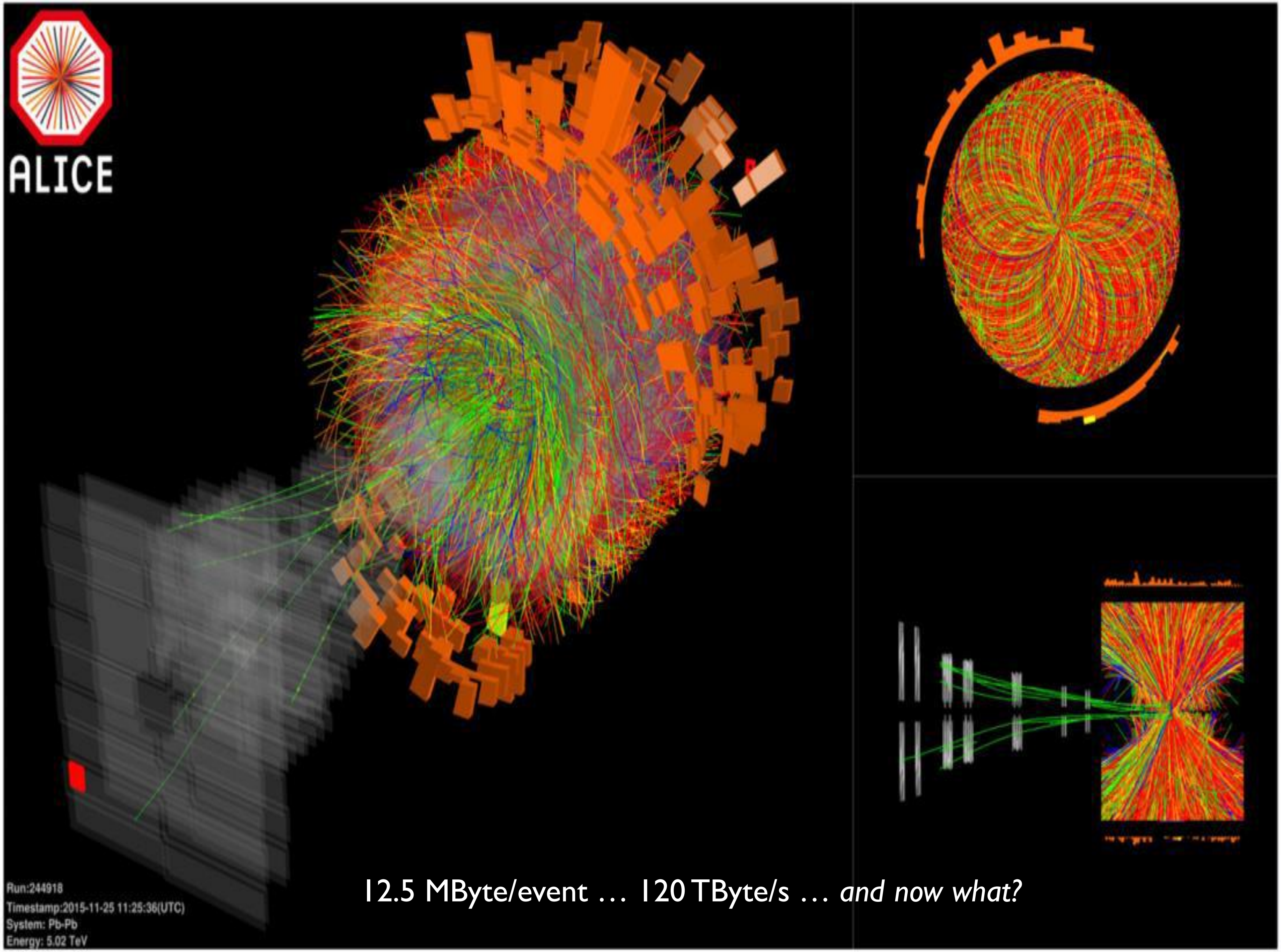
Kans Higgs deeltje:

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

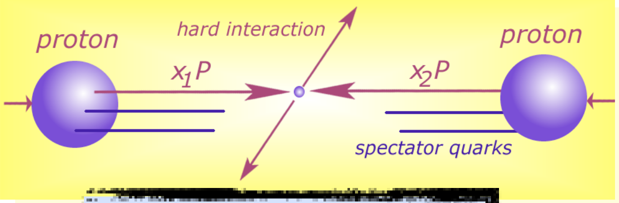
Higgs → ZZ*→ 4μ kandidaat,
M(4 leptonen)=125.1 GeV

Imagery: ATLAS experiment,, atlas.ch

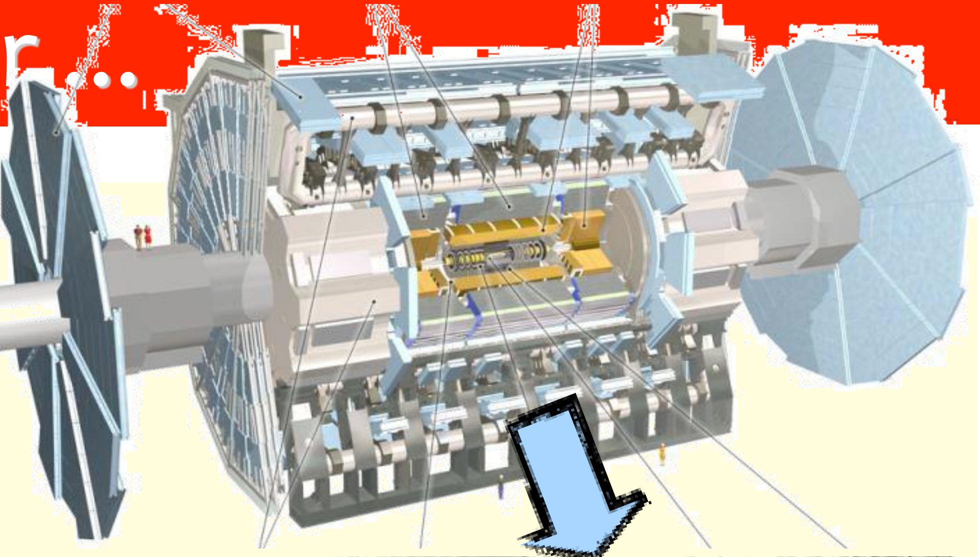


12.5 MByte/event … 120 TByte/s … *and now what?*

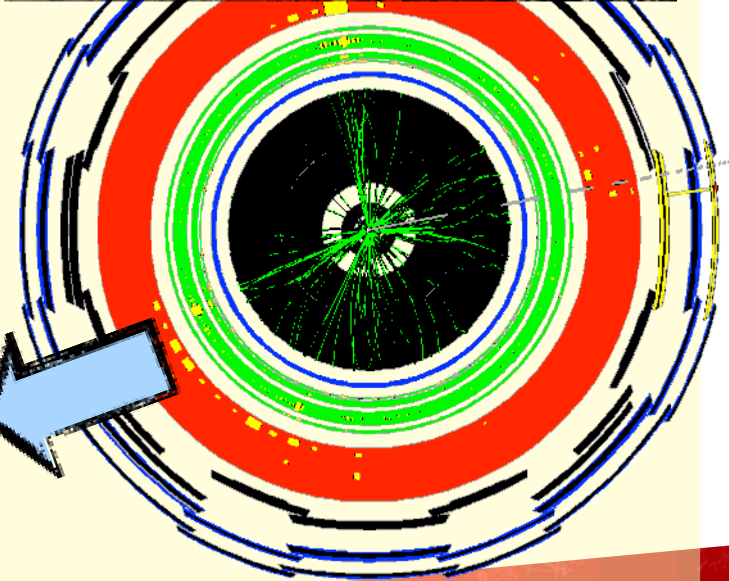
Detector to doctor ...



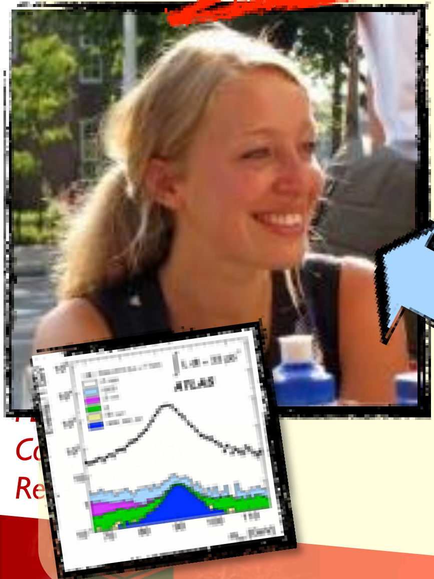
40 miljoen / seconde



Trigger systeem selecteert 600 Hz
~ 1 GB/s data



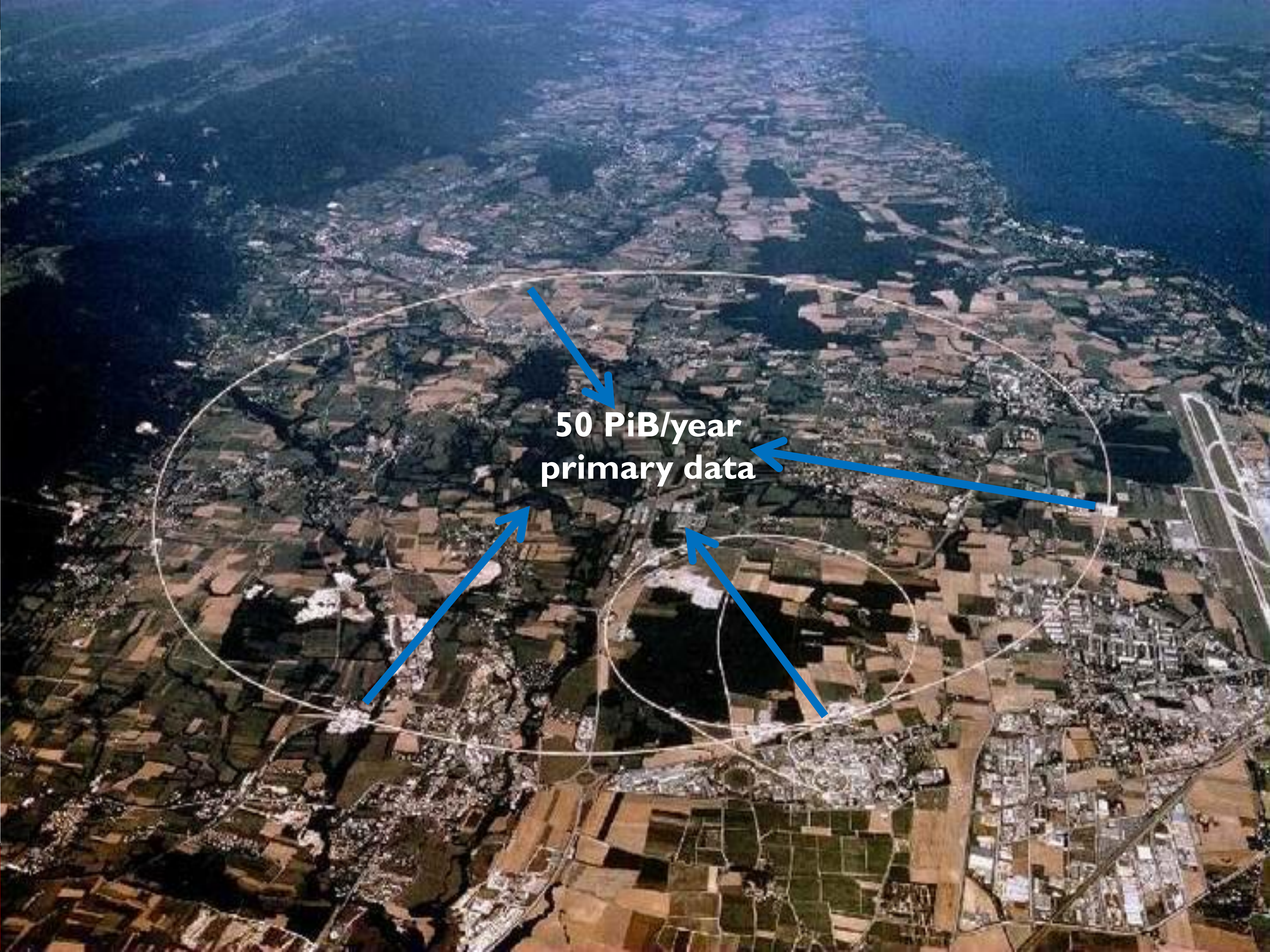
Analyse van botsingen door promovendi



Co
Re

and processing

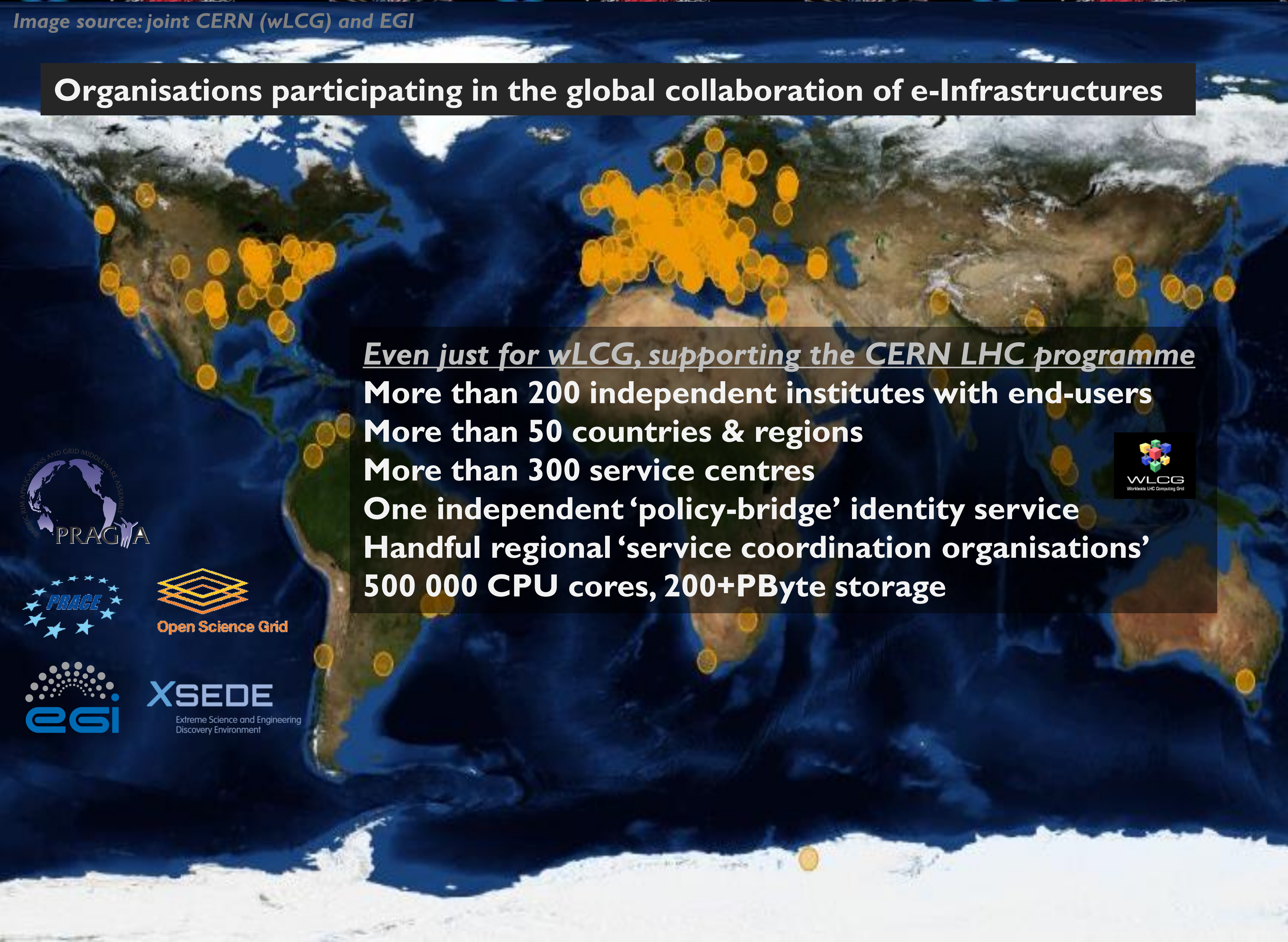
Data distributie met GRID computers



50 PiB/year
primary data

Image source: joint CERN (wLCG) and EGI

Organisations participating in the global collaboration of e-Infrastructures



Even just for wLCG, supporting the CERN LHC programme

More than 200 independent institutes with end-users

More than 50 countries & regions

More than 300 service centres

One independent 'policy-bridge' identity service

Handful regional 'service coordination organisations'

500 000 CPU cores, 200+PByte storage

WLCG
Worldwide LHC Computing Grid

Pacific Rim Applications and Grid Middleware Assembly
PRAGMA

PRACE

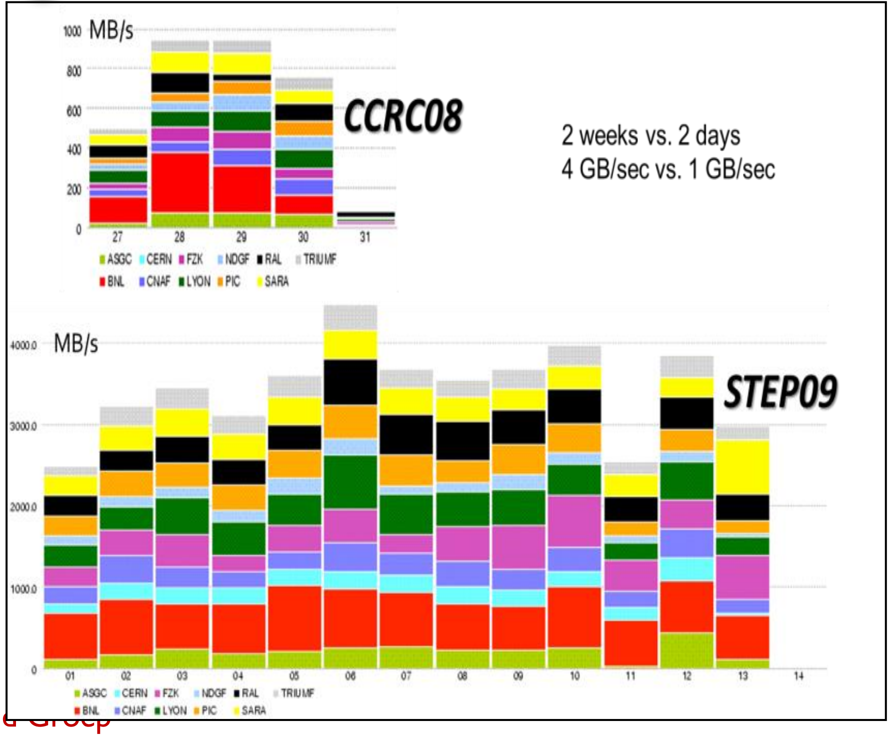
Open Science Grid

egi

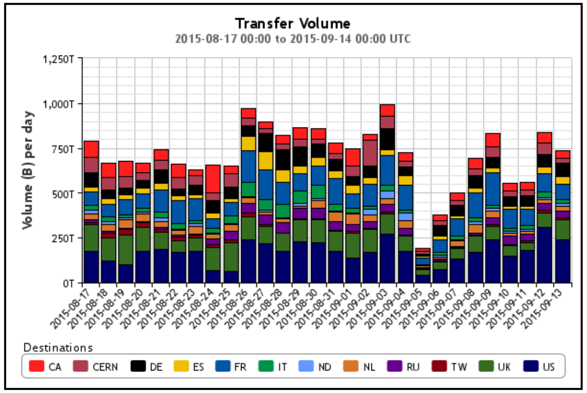
XSEDE
Extreme Science and Engineering
Discovery Environment

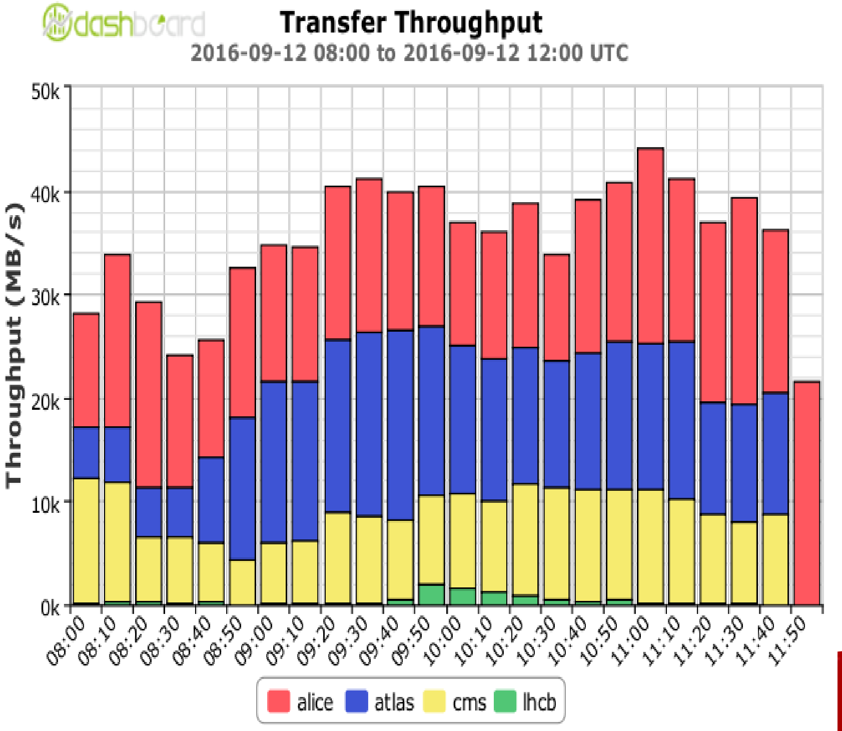
From SC04, CCRC08, STEP09, .. to today

Global transfer rates now > 40 GB/s – acquisition:10 PB/mo (~x2 derived data)



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


Transfer Volume
2015-08-17 00:00 to 2015-09-14 00:00 UTC


Transfer Throughput
2016-09-12 08:00 to 2016-09-12 12:00 UTC

David Groep
Nikhef
PDP - Advanced Computing for Research

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

Atlas: ~50 TByte/day raw data to tape; 1000 TByte/day processed data transfers

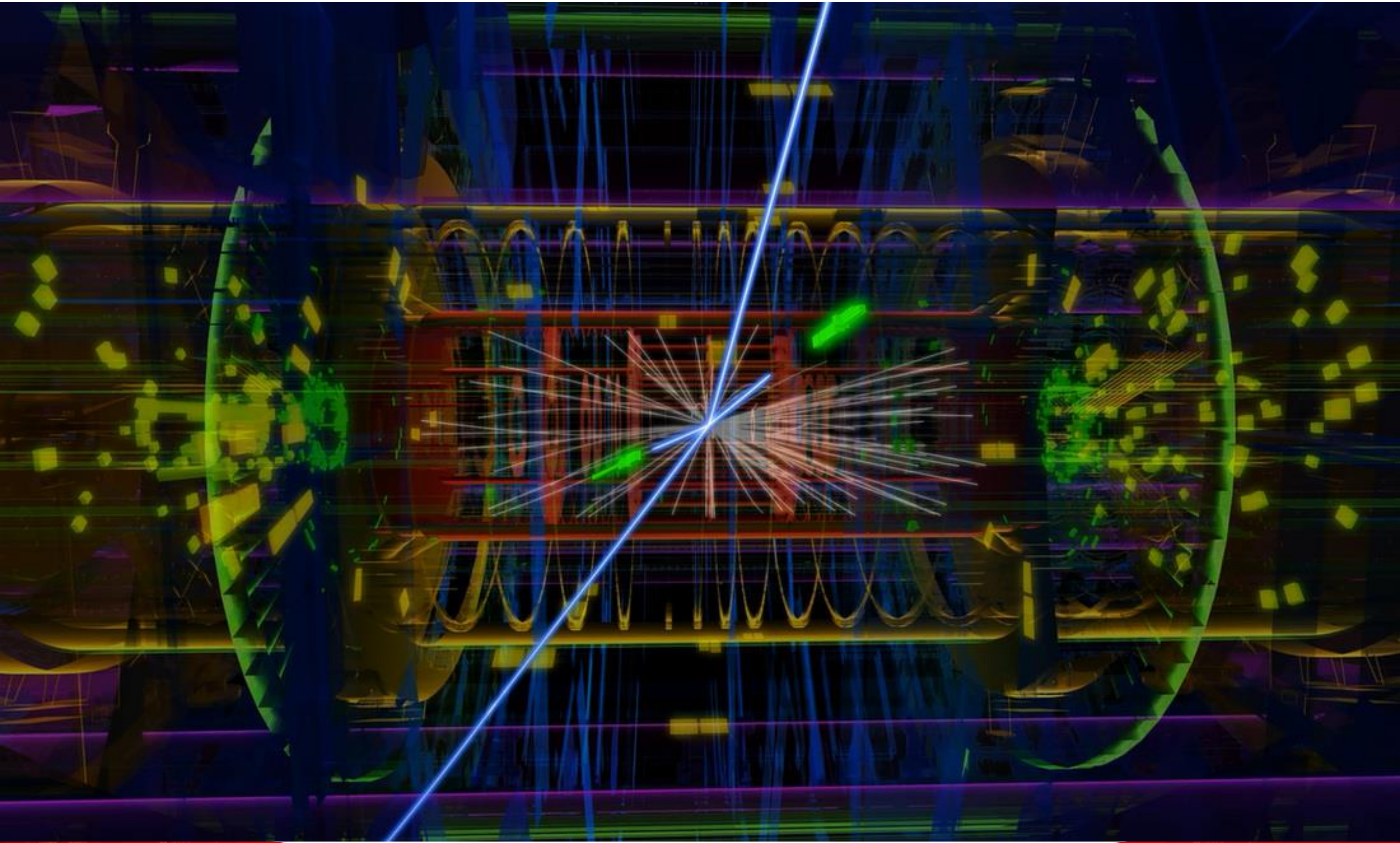
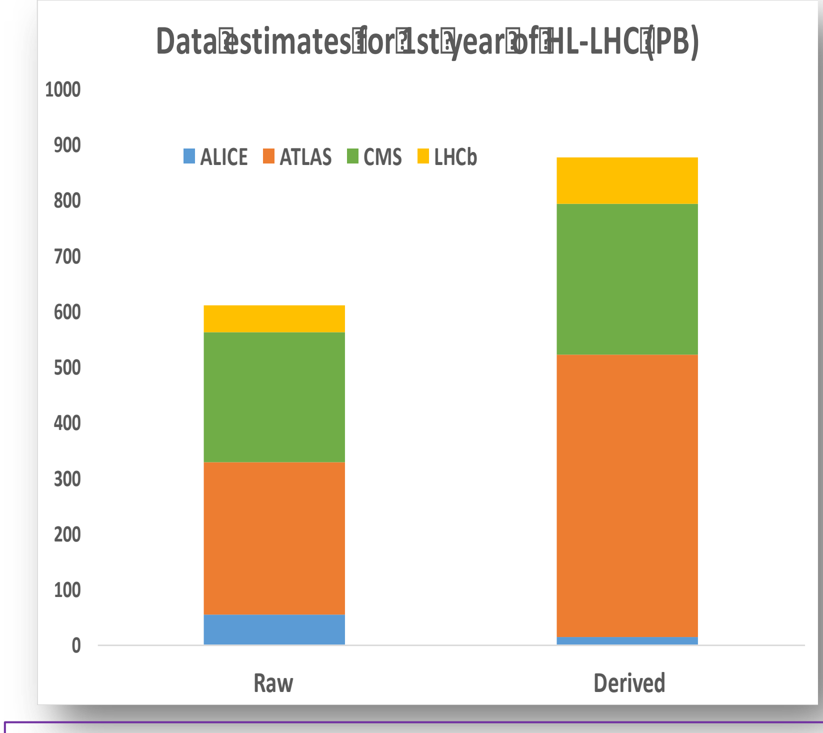
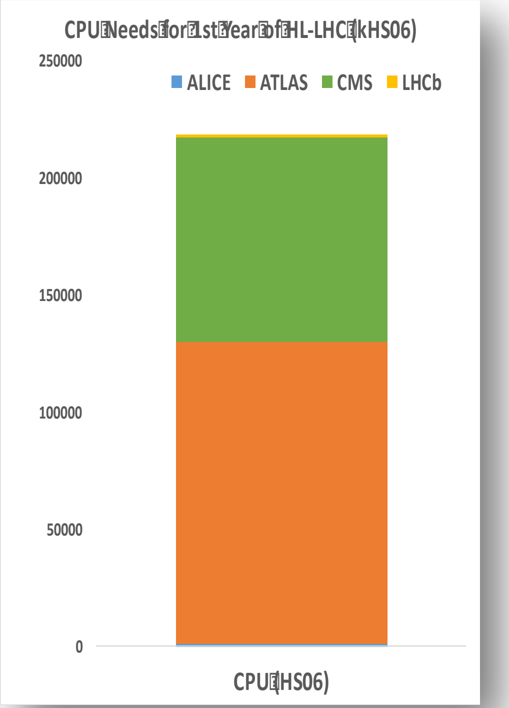


Image source: CERN, Atlas

… and tomorrow ?!





Data:

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

CPU:

- x60 from 2016

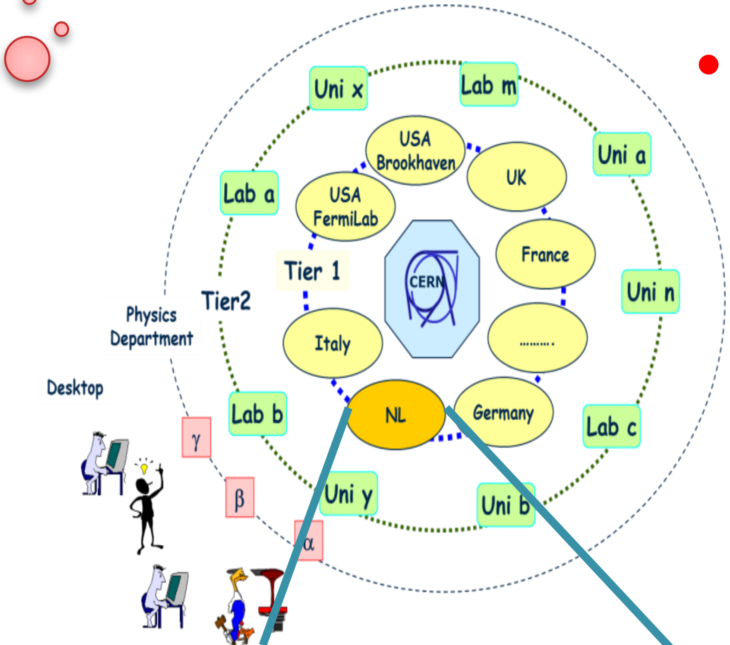
Technology at ~20%/year will bring x6-10 in 10-11 years

Infrastructure for research: balancing network, CPU, and disk

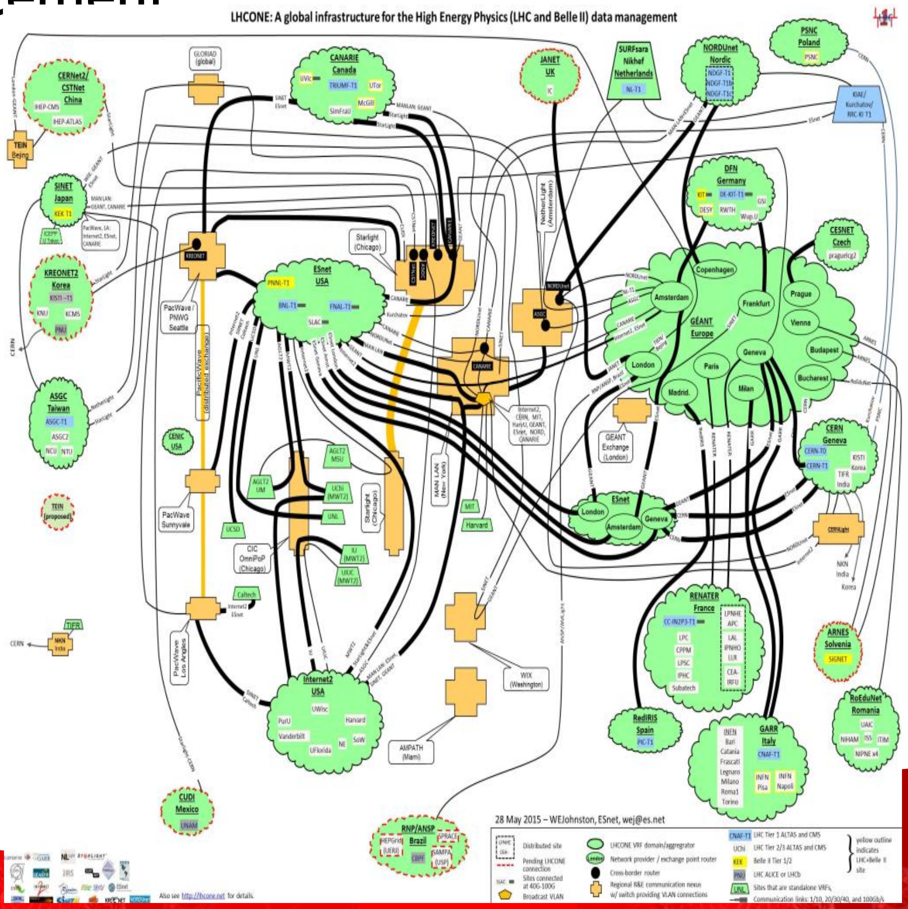
- CPU and disk both expensive, yet idling CPUs are ‘even costlier’
- architecture and performance matching averts any single bottleneck
- but requires knowledge of application (data flow) behaviour data pre-placement (local access), mesh data federation (WAN access)

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 …*

Building the infrastructure for the LHC data



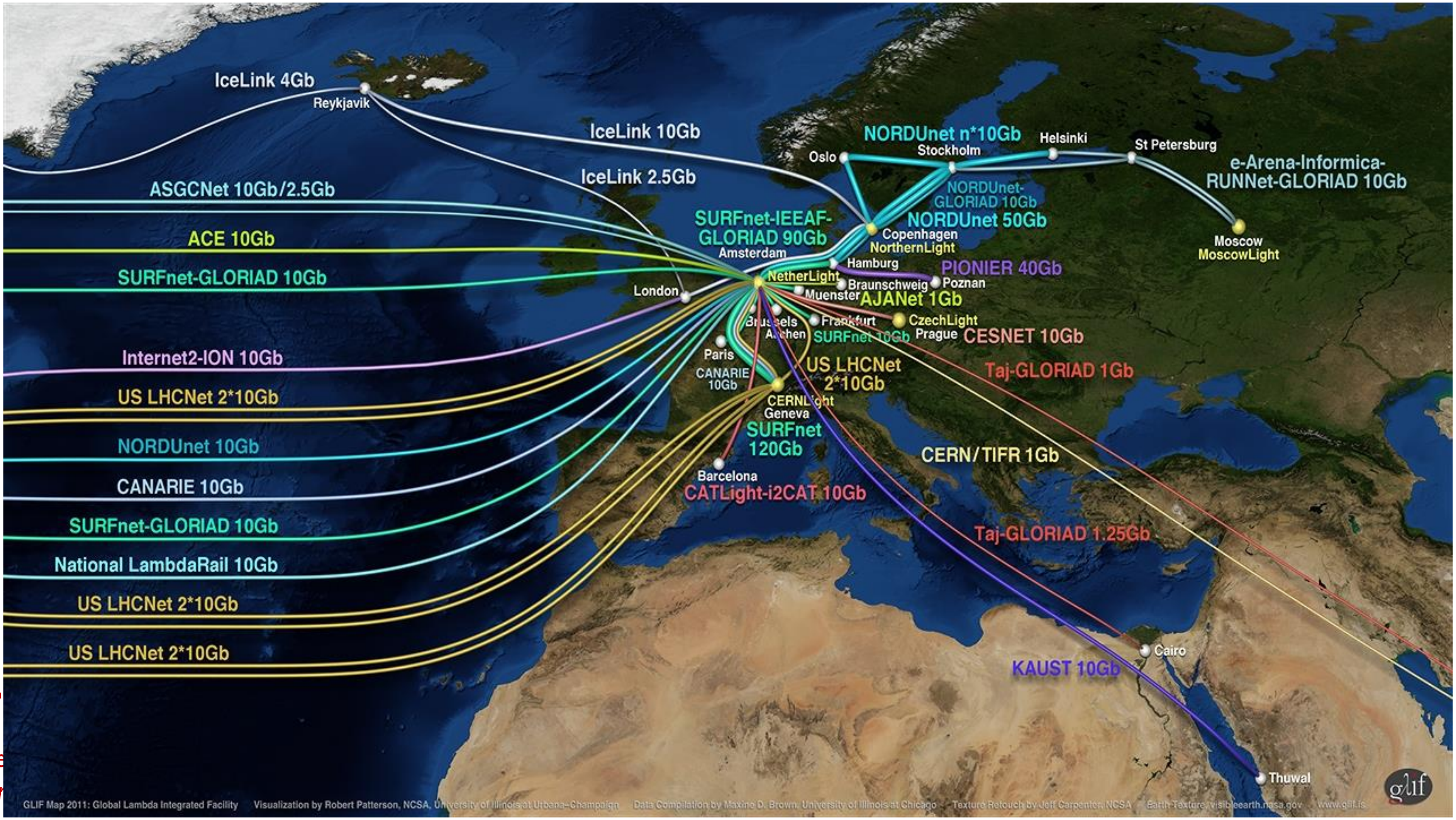
- From hierarchical data distribution to a full mesh and dynamic data placement


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

David Groep
Nikhef
PDP - Advanced Computing for Research

LHCOne graphic: lhcone.net

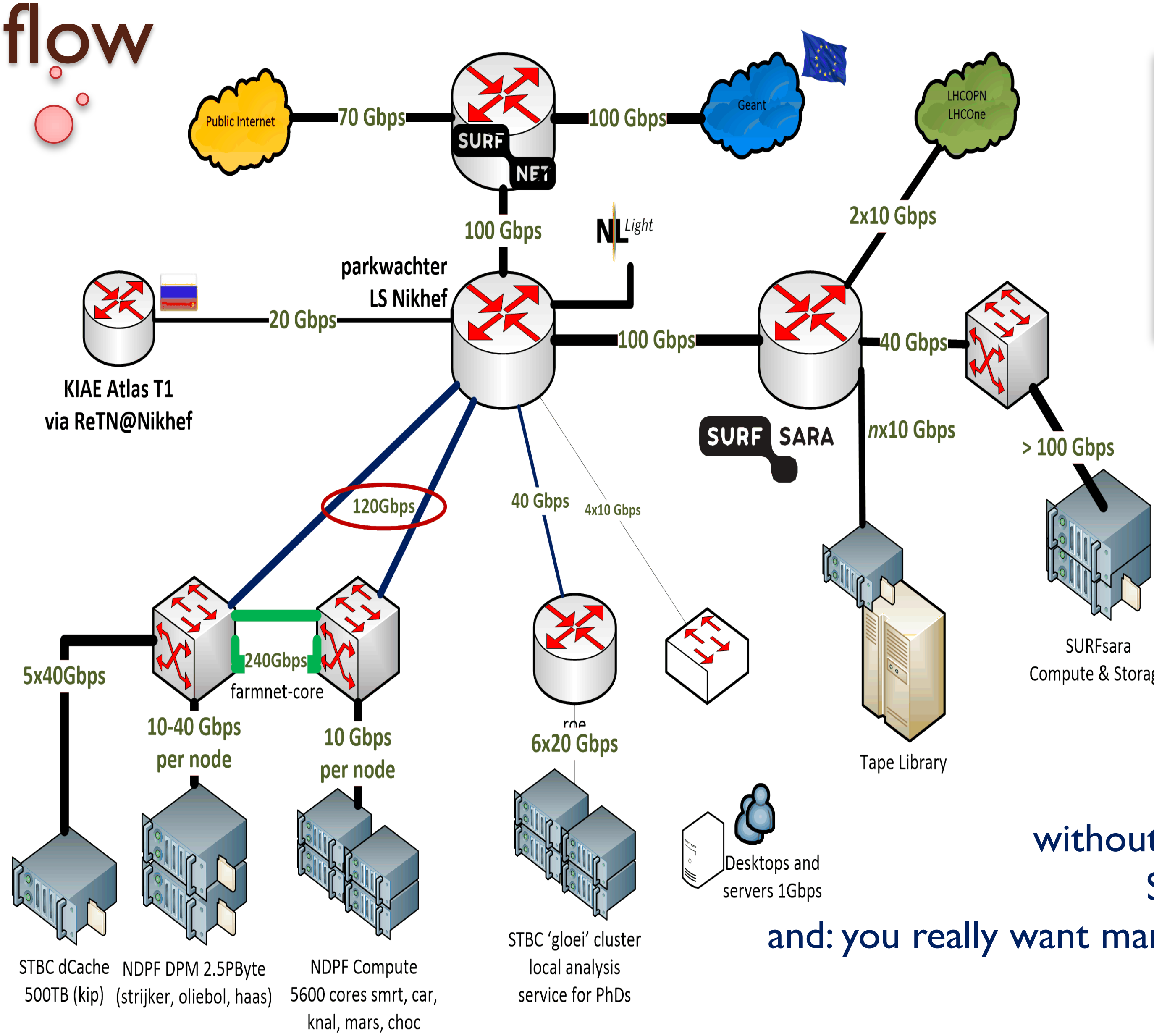
Connecting Science through Lambdas

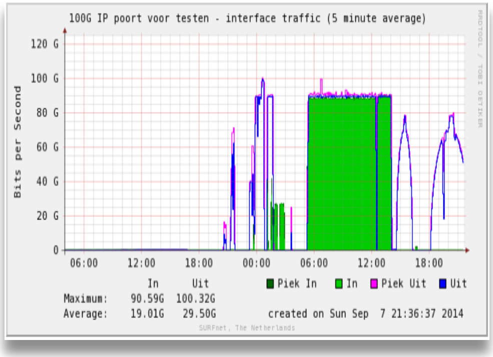


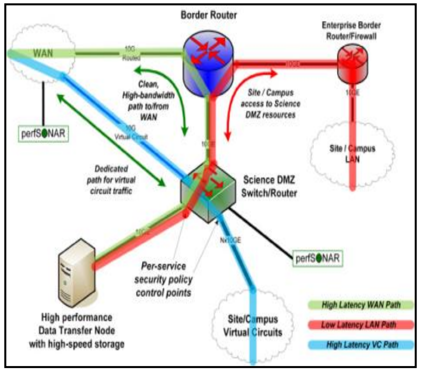
GLIF Map 2011: Global Lambda Integrated Facility   Visualization by Robert Patterson, NCSA, University of Illinois at Urbana-Champaign   Data Compilation by Maxine D. Brown, University of Illinois at Chicago   Texture Retouch by Jeff Carpenter, NCSA   Earth Texture, visibleearth.nasa.gov   www.glif.is

David Groep
Nikhef
PDP - Advance
Computing for
Research

Network built around application data flow







***Need** to work together!*

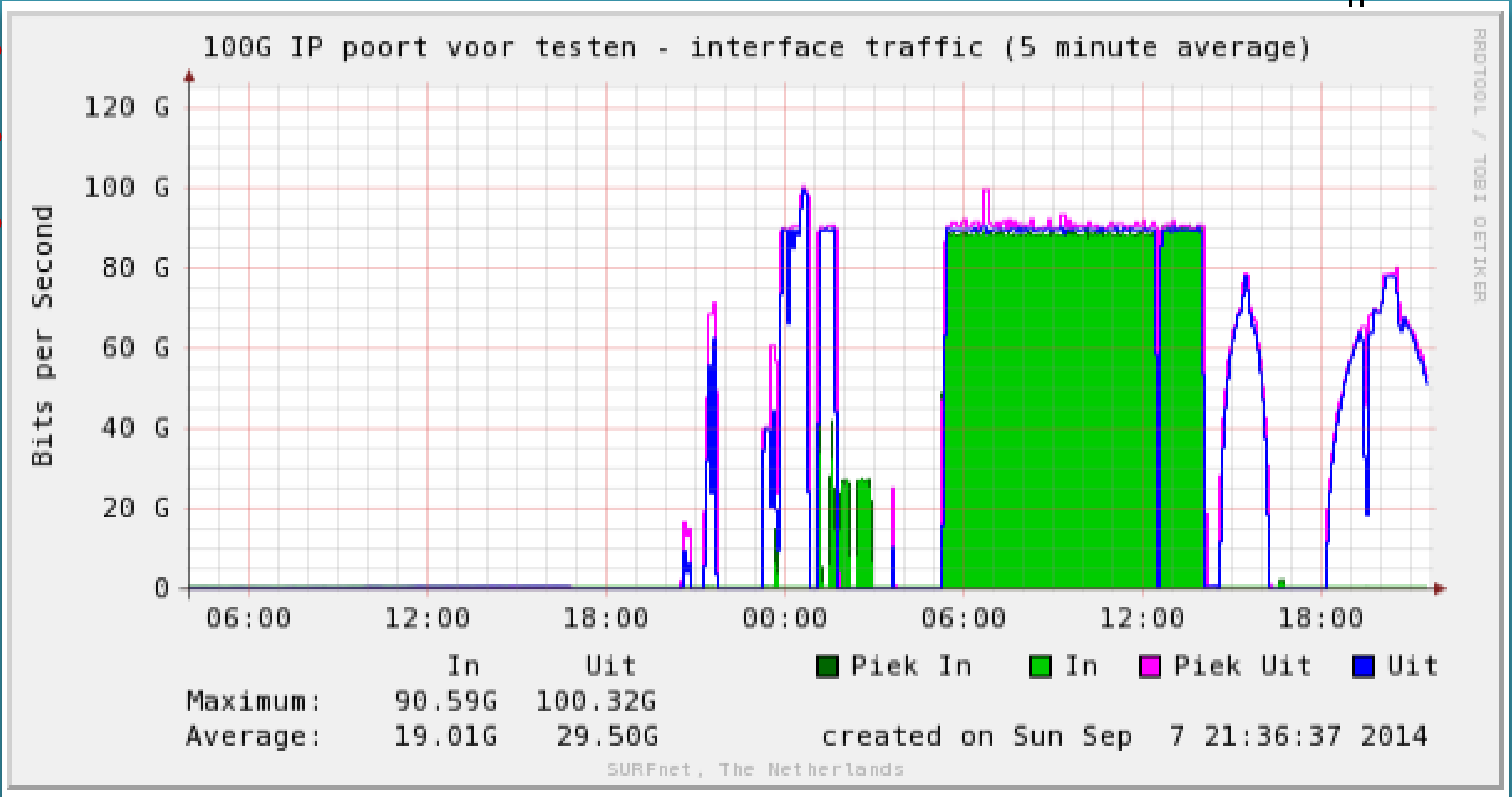
without our SURFsara peering,

SURFnet gets flooded ☺

and: you really want many of your own peerings

100Gbit

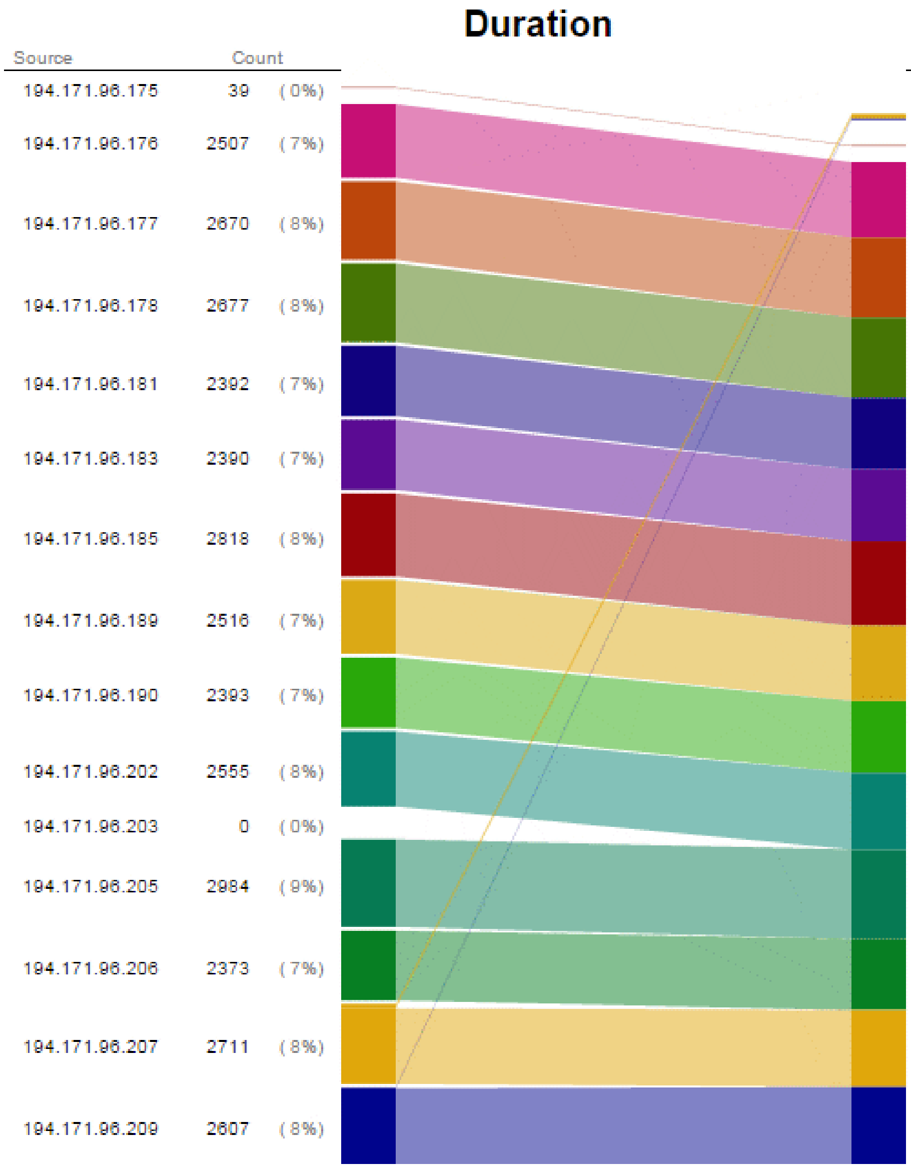
Nikhef → SURFnet → RUG-CIT||UvA



T Suerink
Nikhef
Amsterdam
PDP & Grid

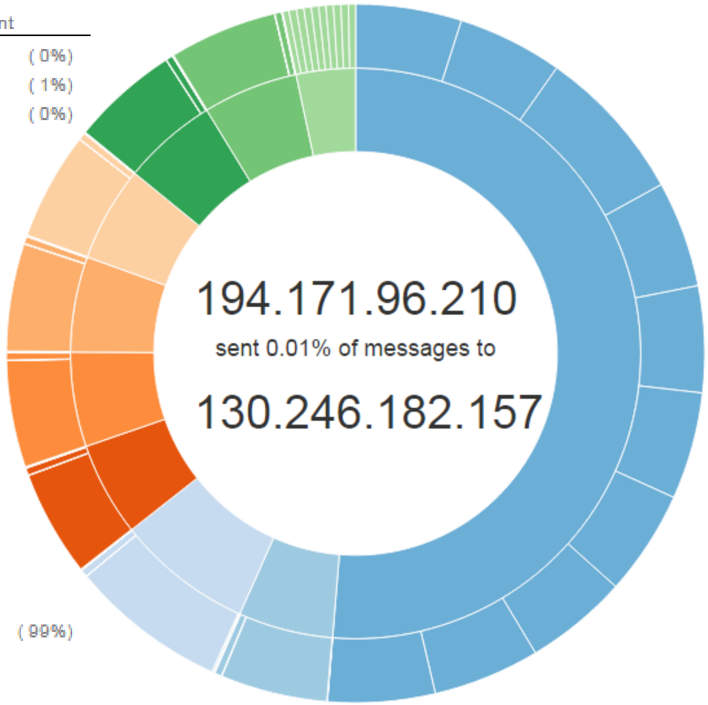
slides courtesy Tristan Suerink

Duration



Source	Count	
194.171.96.175	39	(0%)
194.171.96.176	2507	(7%)
194.171.96.177	2670	(8%)
194.171.96.178	2677	(8%)
194.171.96.181	2392	(7%)
194.171.96.183	2390	(7%)
194.171.96.185	2818	(8%)
194.171.96.189	2516	(7%)
194.171.96.190	2393	(7%)
194.171.96.202	2555	(8%)
194.171.96.203	0	(0%)
194.171.96.205	2984	(9%)
194.171.96.206	2373	(7%)
194.171.96.207	2711	(8%)
194.171.96.209	2607	(8%)

Destination	Count	
188.184.66.250	1	(0%)
192.108.46.8	189	(1%)
192.108.46.89	39	(0%)
192.5.19.10	33403	(99%)



Data flows: a user at UC Irvine reading a data set from Nikhef, with some background from CERN and KIT Karlsruhe

Graphics courtesy Jouke Roorda and Olivier Verbeek

Getting more bytes through?

- Power vs x64: 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

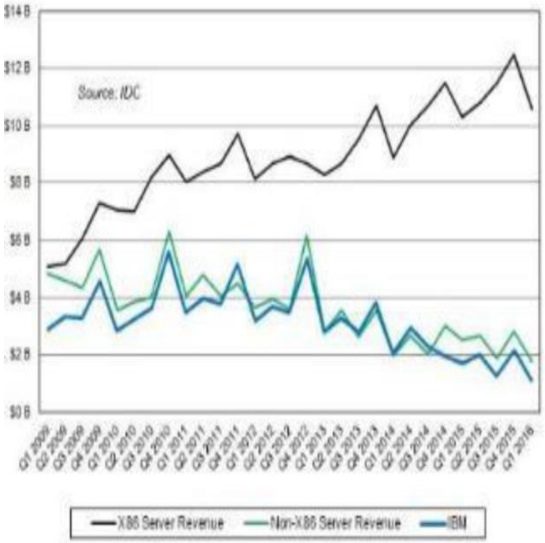
yet more is needed

HGST: 480 TByte gross capacity/4RU

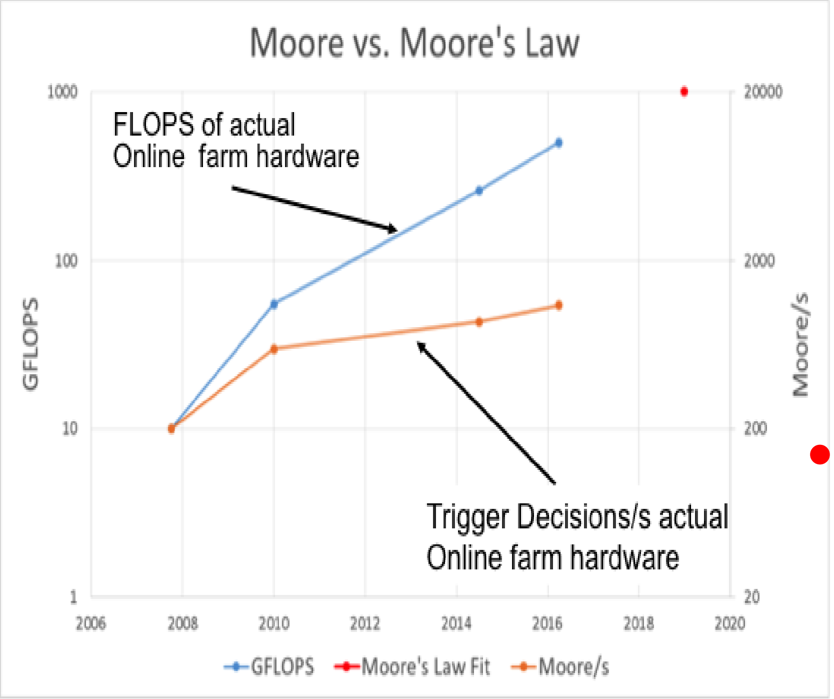
- RAID card are now a performance bottleneck
- JBOD changes CPU-disk ratio
- closer integration of networking to get >100Gbps

Matching systems architecture

- Most applications using x86 today, and probably will for a long time
- alternatives (GPGPU or Power) not quite viable
  … although for 'dedicated farms' FPGAs help,
  and KNH works better (we need the memory)



sales volume of different architectures

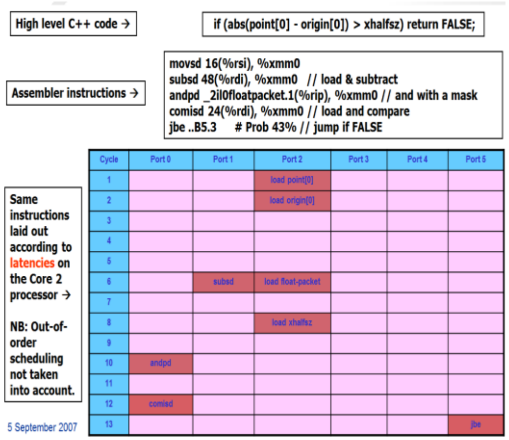


- Yet change must be:
  most gain to be had from SIMD vectorization and improved memory access patterns

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

Improvements at the application layer

- ‘traditional’ (1990’s) style HEP applications were ‘lean’, and fail to scale even in pipelining
- let alone vector instructions or multicore



High level C++ code →

```
if (abs(point[0] - origin[0]) > xhalfsz) return FALSE;
```

Assembler instructions →

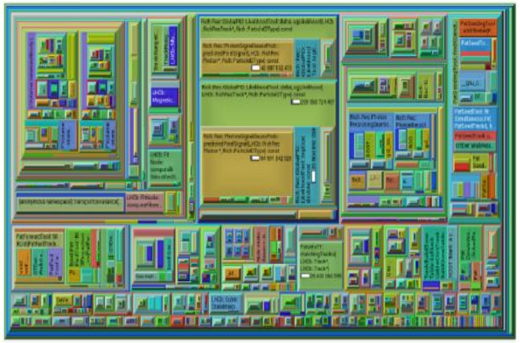
```
movsd 16(%rsi), %xmm0
subsd 48(%rdi), %xmm0   // load & subtract
andpd _2il0floatpacket.1(%rip), %xmm0 // and with a mask
comisd 24(%rdi), %xmm0 // load and compare
jbe ..B5.3     # Prob 43% // jump if FALSE
```

Same instructions laid out according to latencies on the Core 2 processor →

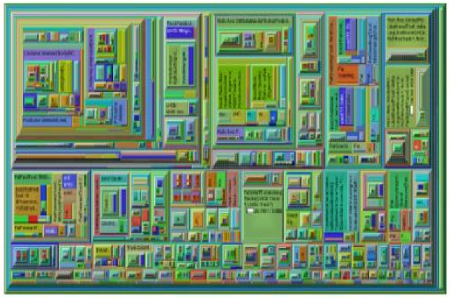
NB: Out-of-order scheduling not taken into account.

Cycle	Port 0	Port 1	Port 2	Port 3	Port 4	Port 5
1			load point[0]			
2			load origin[0]			
3						
4						
5						
6		subsd	load float-packet			
7						
8			load xhalfsz			
9						
10	andpd					
11						
12	comisd					
13						jbe

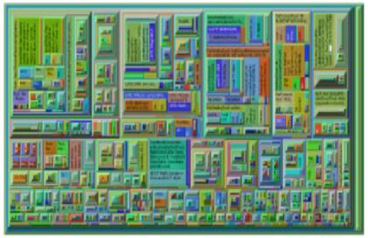
5 September 2007



2012 v45r1



v48r1



v48r1 (2015 reco)

review of algorithms gave overall +34% in LHCb – memory layout still to be done …

To use current processor generations, you need better – machine-aware! – code

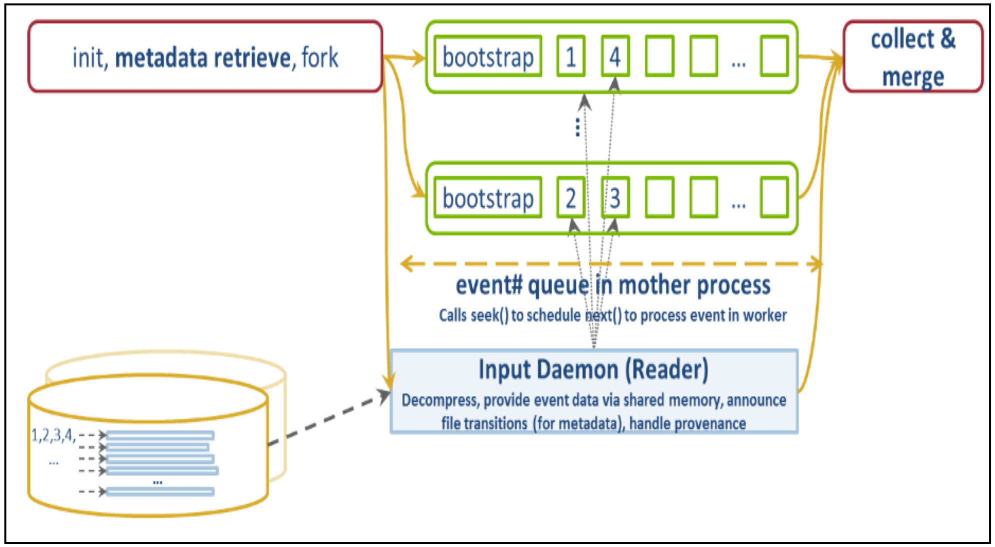
David Groep
Nikhef
PDP - Advanced Computing for Research

2007 Core 2 efficientcy: Sverre Jarp, CHEP2007
Performance data: M Schiller (CERN) et al. for LHCb, - profiling progress (update) 2015;; work by: Gerhard Raven, Nikhef

Systems architecture and your application

Many things you only find in production …

- When you're 'embarrassingly parallel' with a memory challenge *why not try 'priming' of memory for the first few events and then fork?*


init, **metadata retrieve**, fork
bootstrap
1
4
...
bootstrap
2
3
...
collect & merge
event# queue in mother process
Calls seek() to schedule next() to process event in worker
Input Daemon (Reader)
Decompress, provide event data via shared memory, announce file transitions (for metadata), handle provenance
1,2,3,4,
...

- Towards single-socket systems: cache coherence limits performance – there's a penalty to pay for massive multi-socket-big-memory hosts!

Graphic: P van Gemmeren et al. ATLAS collaboration, ATL-SOFT-PROC-2015-020

Systems for Research @ Nikhef

David Groep
Nikhef
PDP - Advanced Computing for Research

Statistics

Dutch National e-Infrastructure coordinated by SURF

"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

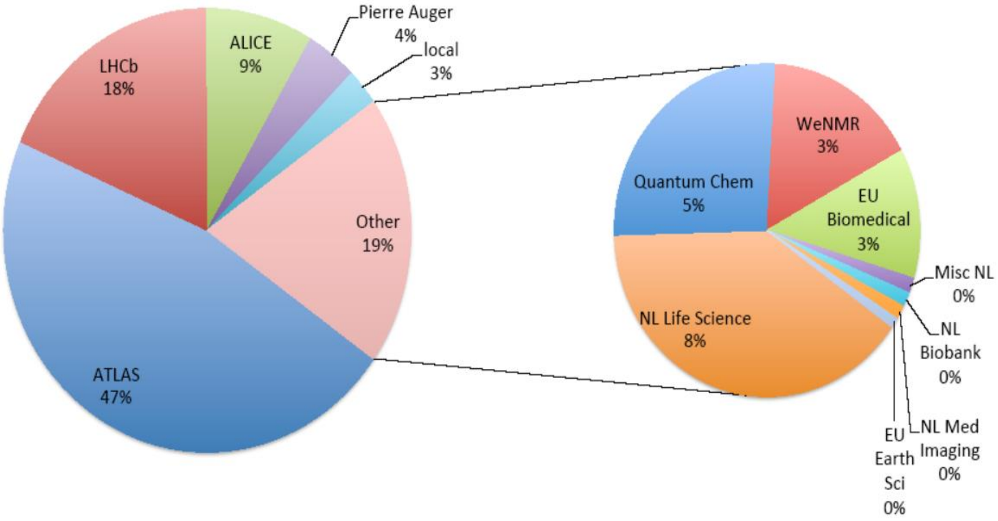
focus on large/many-core systems

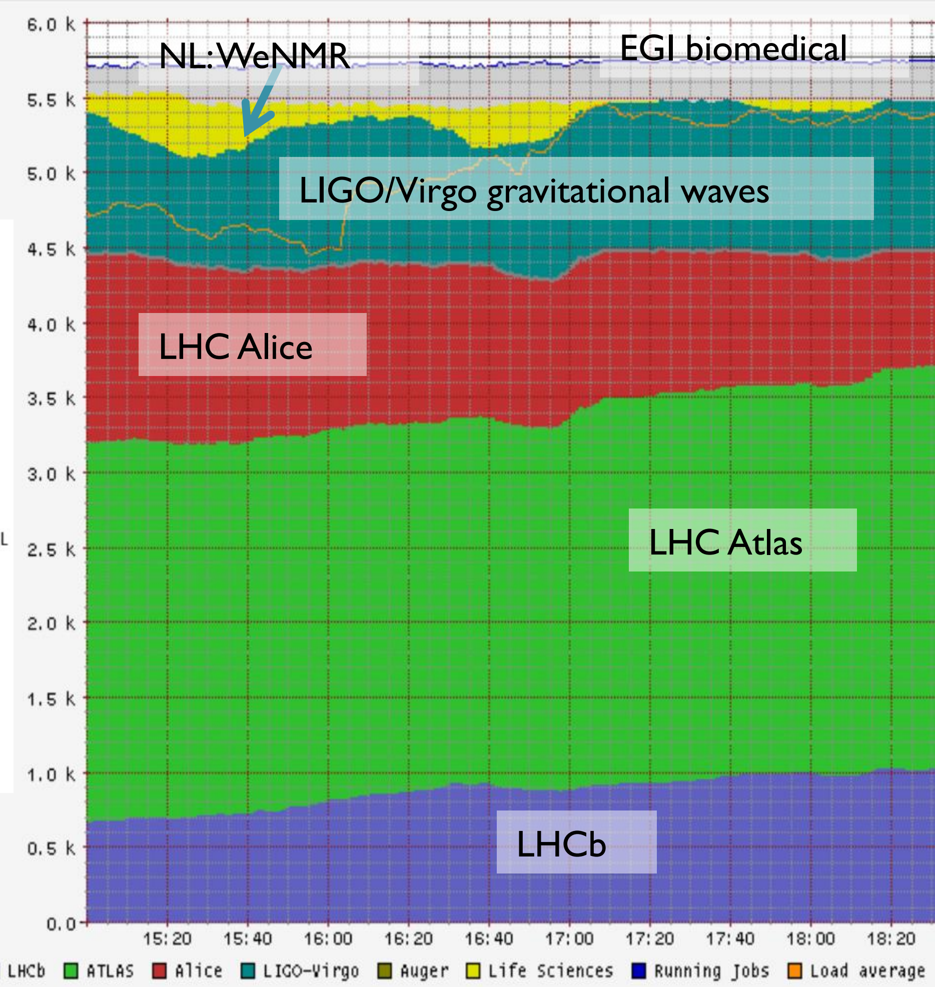
> 45 install flavours (service types)

and a bunch of one-off systems

Shared infrastructure, efficient infrastructure!

- >98% utilisation, >90% efficiency

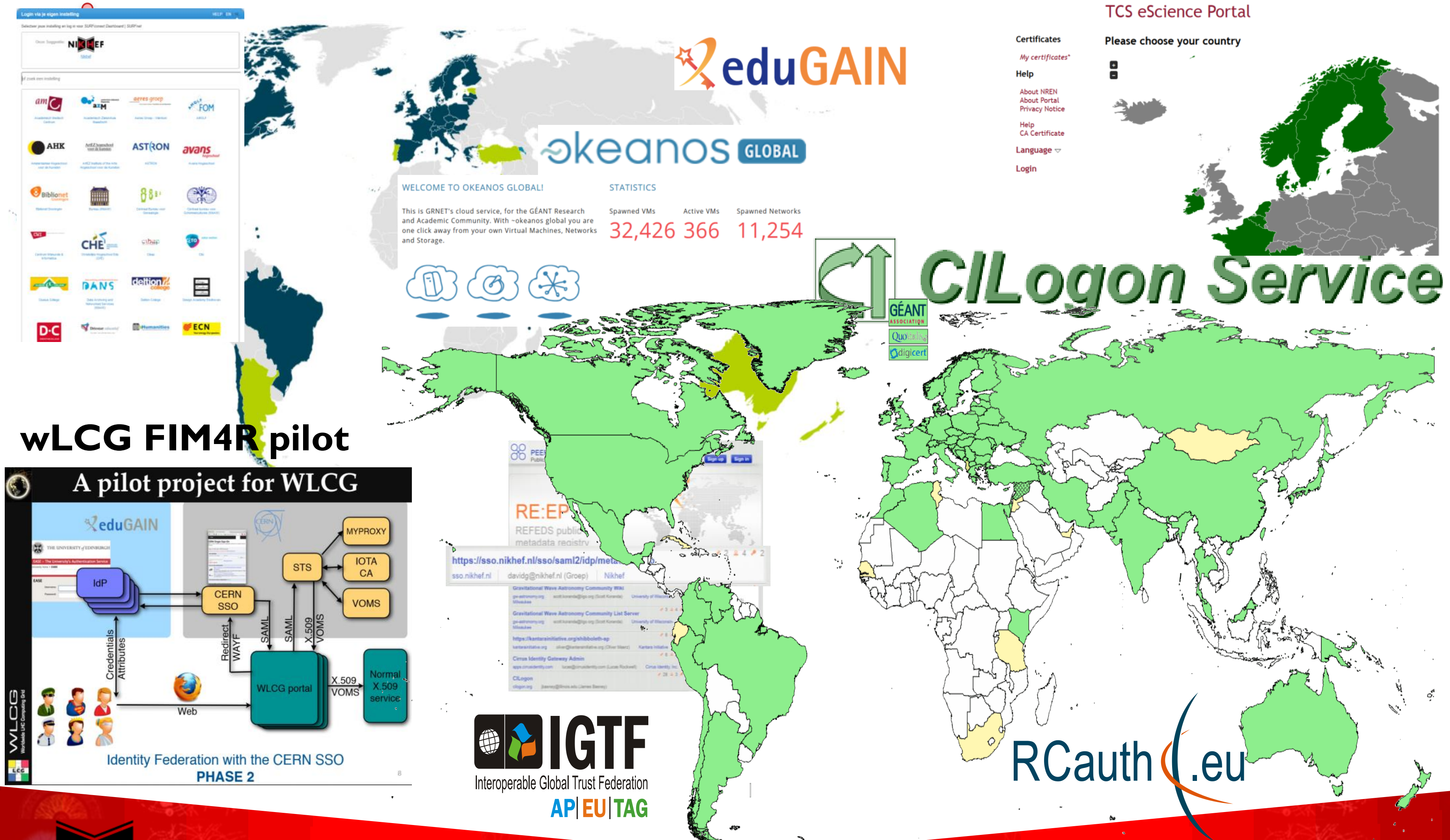




David Groep
Nikhef
PDP - Advanced Computing for Research

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

Federation of high-throughput services



Login via je eigen instelling

Selecteer jouw instelling en log in voor SURFconext Dashboard | SURFnet

of zoek een instelling

TCS eScience Portal

Please choose your country

Certificates
- My certificates

Help
- About NREN
- About Portal
- Privacy Notice
- Help
- CA Certificate

Language

Login

WELCOME TO OKEANOS GLOBAL!

This is GRNET's cloud service, for the GÉANT Research and Academic Community. With ~okeanos global you are one click away from your own Virtual Machines, Networks and Storage.

STATISTICS

Spawned VMs	Active VMs	Spawned Networks
32,426	366	11,254

RE:EP

REFEDS public metadata registry

https://sso.nikhef.nl/sso/saml2/idp/meta

sso.nikhef.nl | davidg@nikhef.nl (Groep) | Nikhef

Gravitational Wave Astronomy Community Wiki

Gravitational Wave Astronomy Community List Server

https://kantarainitiative.org/shibboleth-sp

Cirrus Identity Gateway Admin

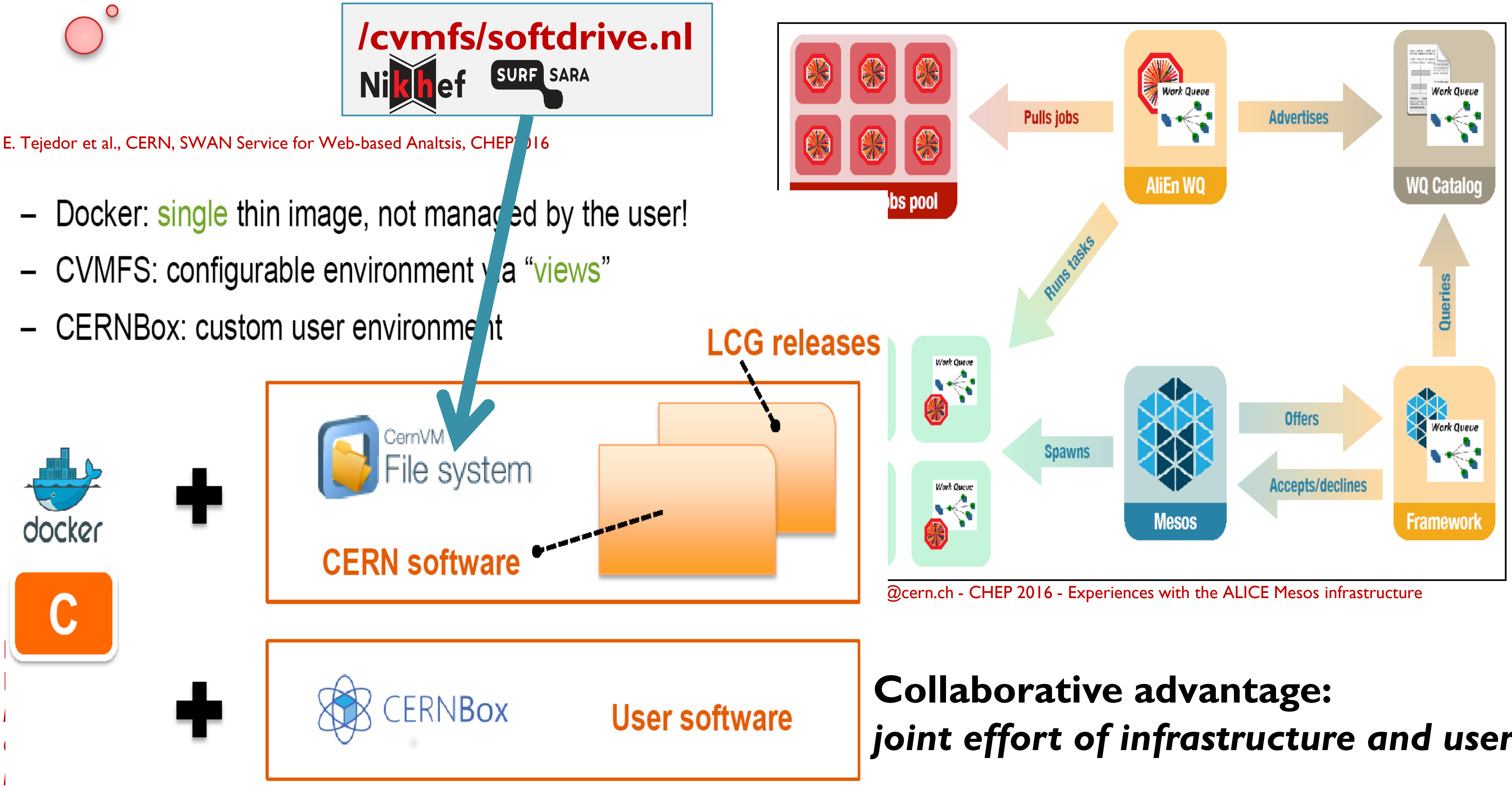
CILogon

wLCG FIM4R pilot

A pilot project for WLCG

Identity Federation with the CERN SSO
PHASE 2

‘cloud’ is a means, not an end-all solution



/cvmfs/softdrive.nl

E. Tejedor et al., CERN, SWAN Service for Web-based Analtsis, CHEP2016

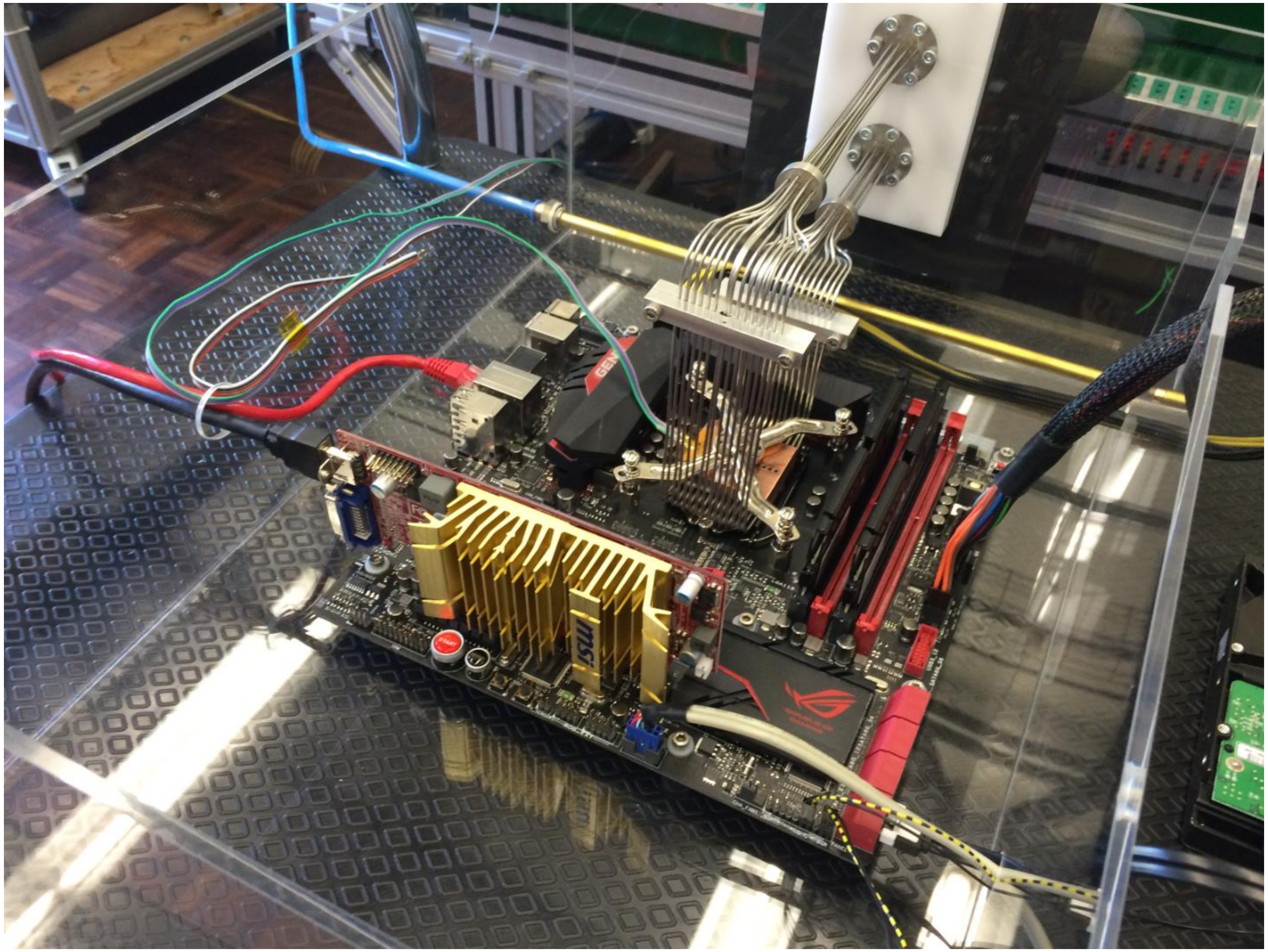
- Docker: single thin image, not managed by the user!
- CVMFS: configurable environment via “views”
- CERNBox: custom user environment

@cern.ch - CHEP 2016 - Experiences with the ALICE Mesos infrastructure

Collaborative advantage:
joint effort of infrastructure and users

http://doc.grid.surfsara.nl/en/latest/Pages/Advanced/grid_software.html#softdrive

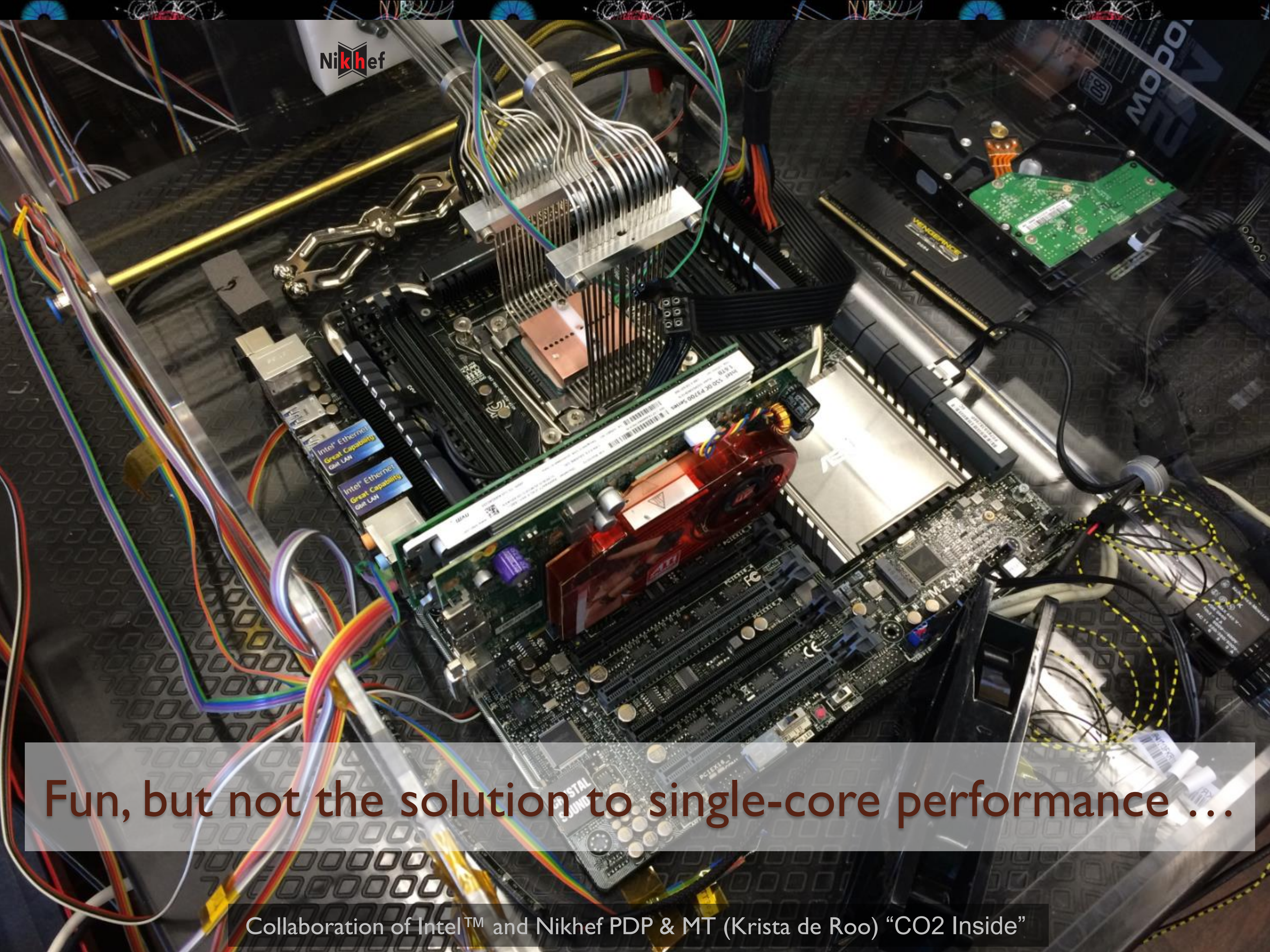
For (informed) fun & testing – some random one-off systems …



David Groep
Nikhef
PDP - Advanced Computing for Research

CO2-cooled Intel CPUs @6.2GHz

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



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

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