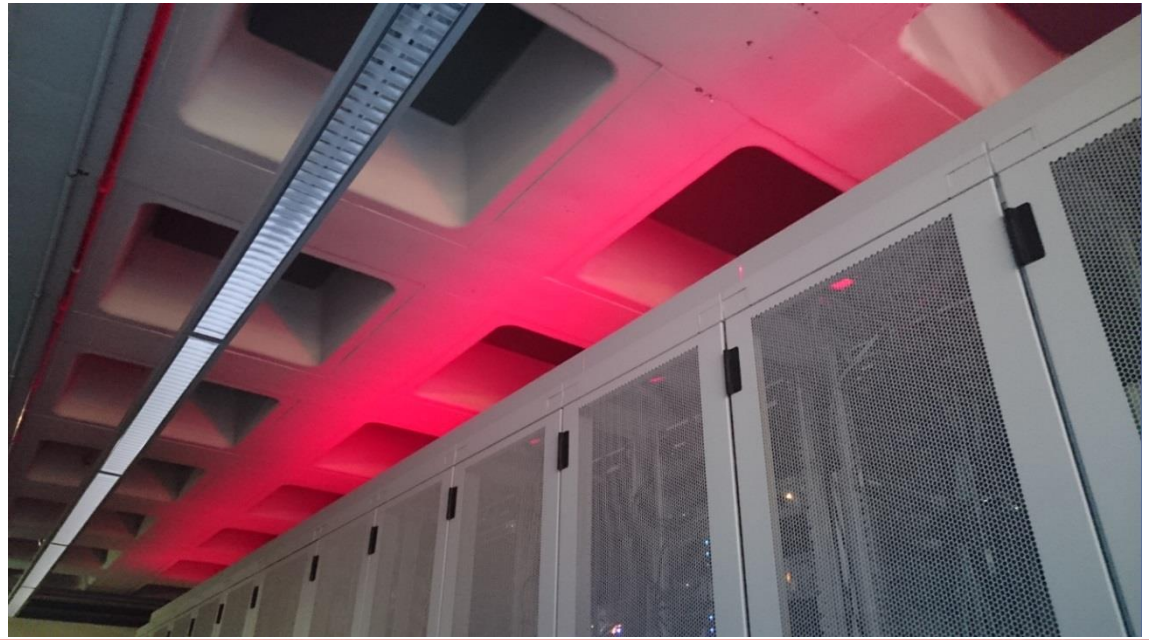
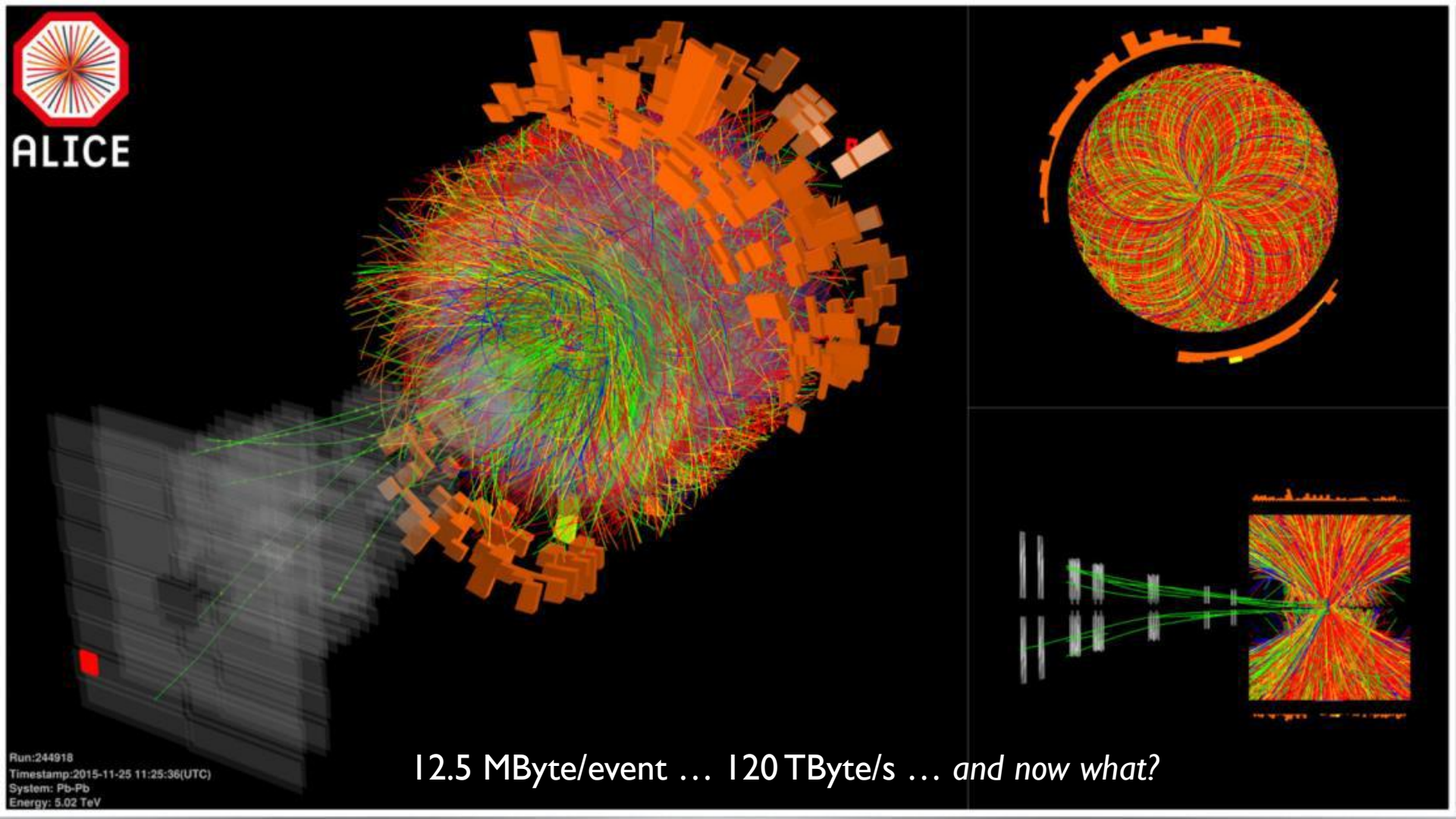
Accelerating Throughput – from the LHC to the World

David Groep

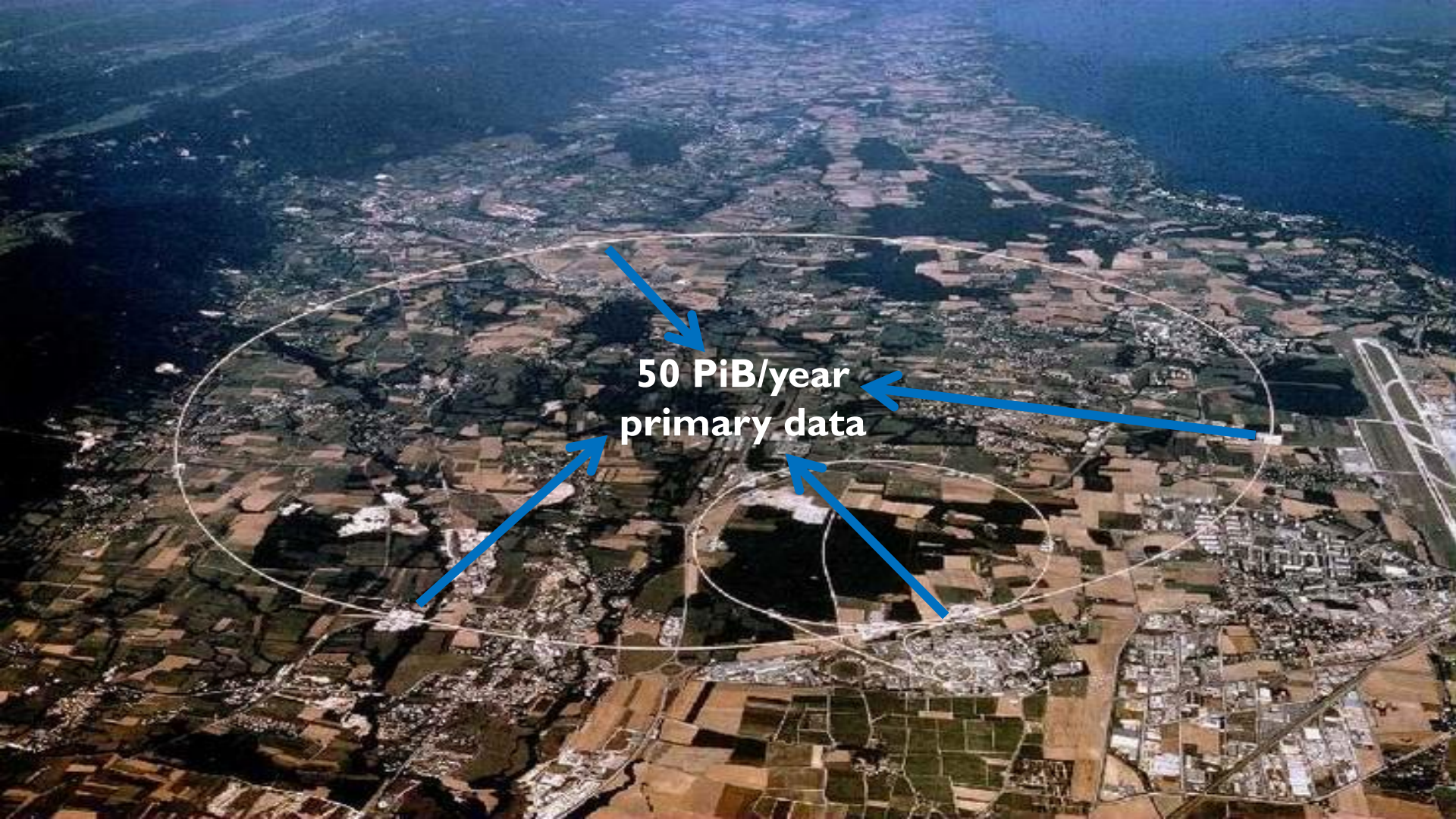
Nikhef



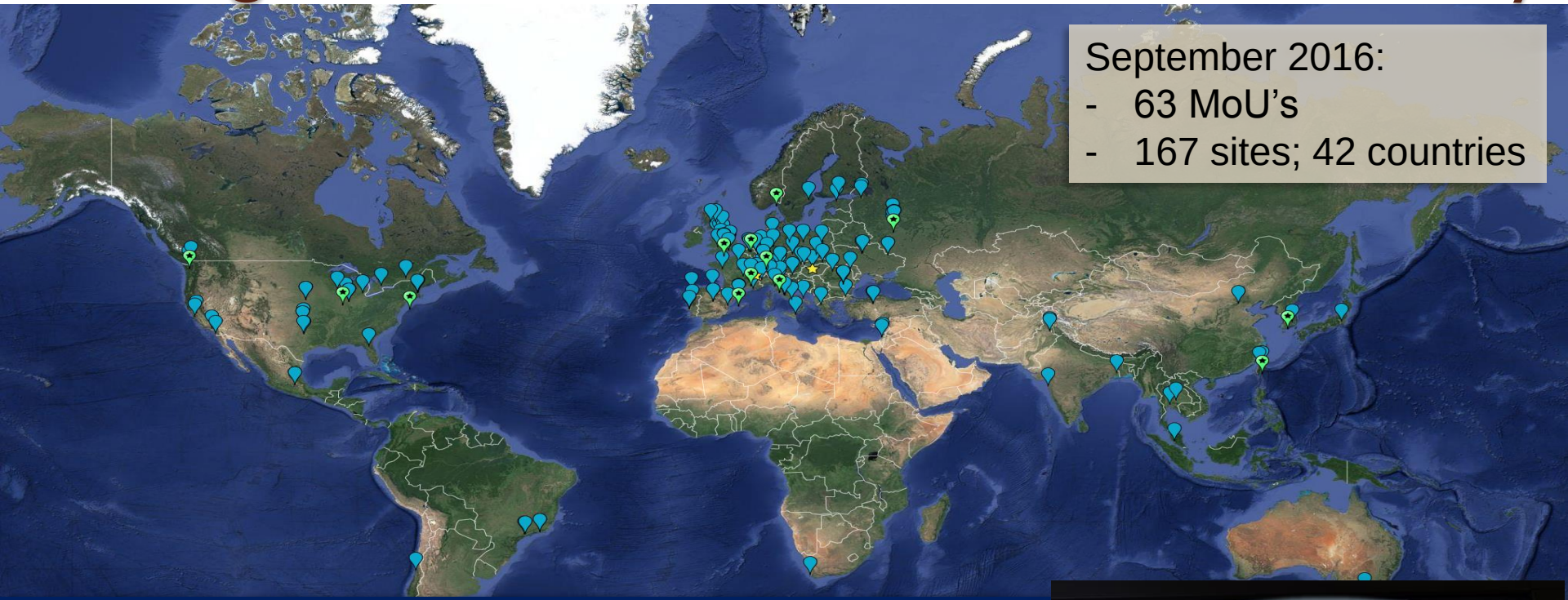
David Groep
Nikhef
PDP – Advanced Computing for Research



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

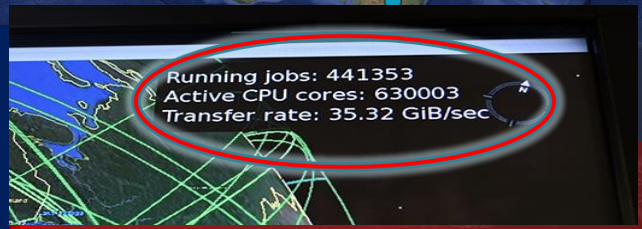

50 PiB/year
primary data

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



~300 resource centres
~250 communities
Federated infrastructure

Imperial College London

Global collaboration – in a secure way

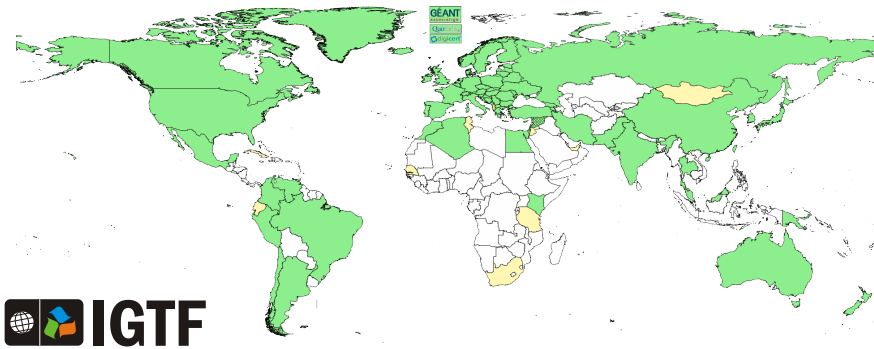
Collaboration is people as well as (or even more than) systems

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

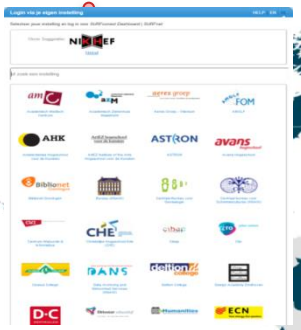
- Common baseline assurance (trust) requirements
- Persistent and globally unique

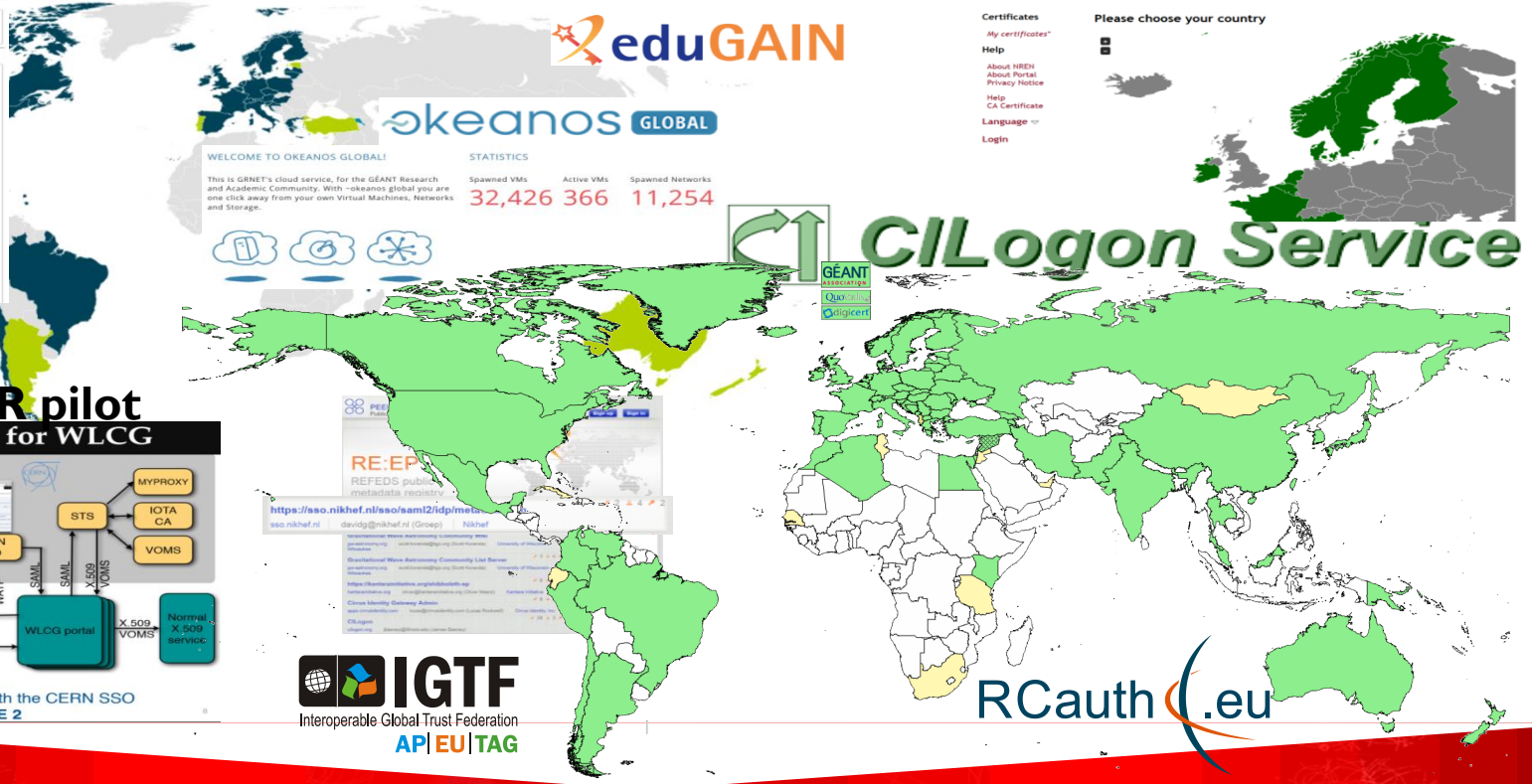
needs a global scope – so we built the Interoperable Global Trust Federation

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

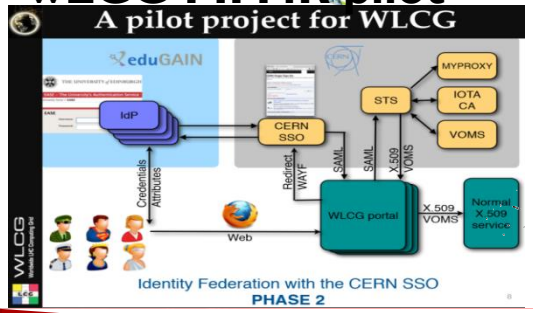


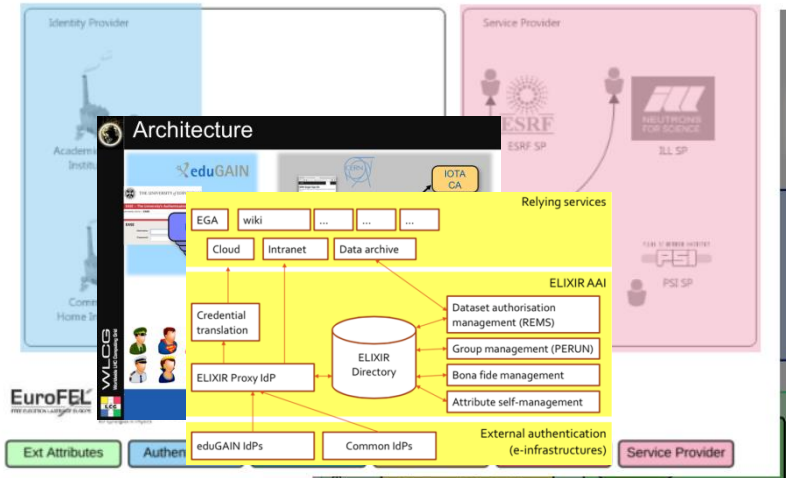
But federation is much more …





wLCG FIM4R pilot





AAI: The e-Infrastructure view

What is happening on top of existing Federation infrastructures today

User

National federations (eduGAIN)

CA

OIDC IdP

Social IdP

eGovID

Other

Authenticated User

The authenticated user can go
- Directly to the service
- Authenticate via Proxy
- Use Token Translation

Account linking?

Attribute Certificates

addressed via Attribute Authorities
This is still unclear

AA:Grp

AA:LoA

Attribute Authorities

Enrichment

Possible flows of additional attributes:
- Directly to the service
- via the Proxy / TTS

Translation

Proxy

DS

Token Translation Service

Attribute Flow

Unauthenticated User

Authenticated User

Three options to authenticate at a service:
- Directly
- Via Proxy
- Via Token Translation

End Service

SP

NonWeb

Site Token services

ssh

S3

X.509

Username Password

Federation

same pattern

Umbrella

ELIXIR

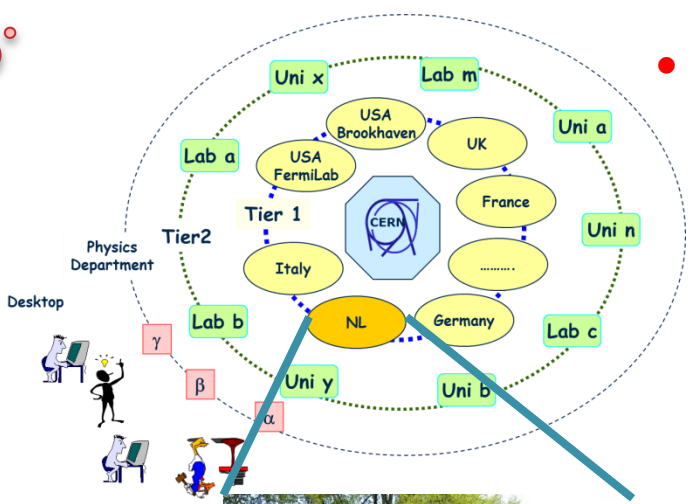
BBMRI

WLCG

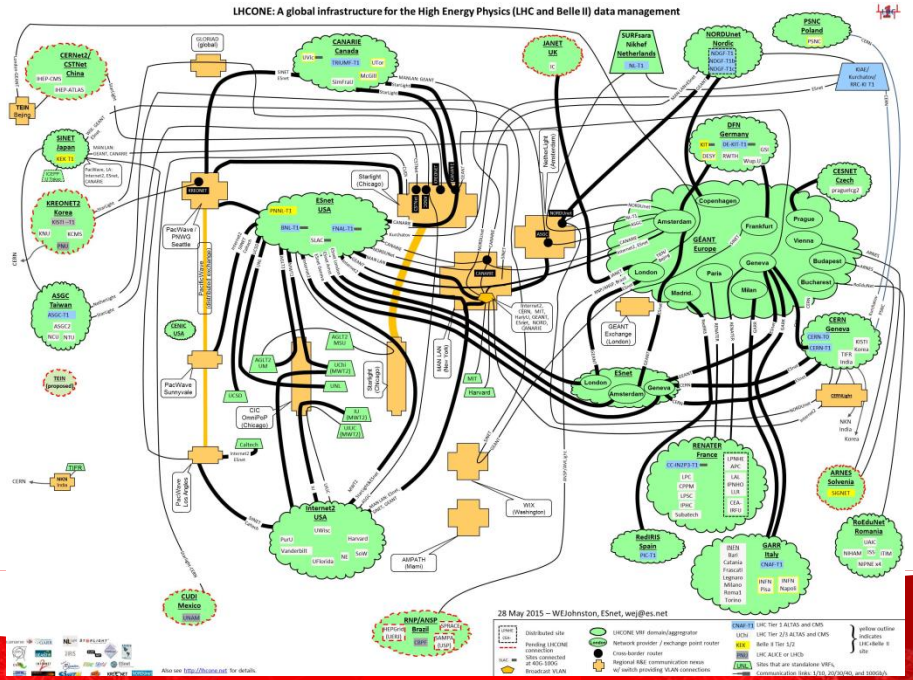
…

https://aarc-project.eu/wp-content/uploads/2015/11/20151102-JRA1.2-Blueprint-Status-Hardt.pdf

Building the infrastructure for the LHC data

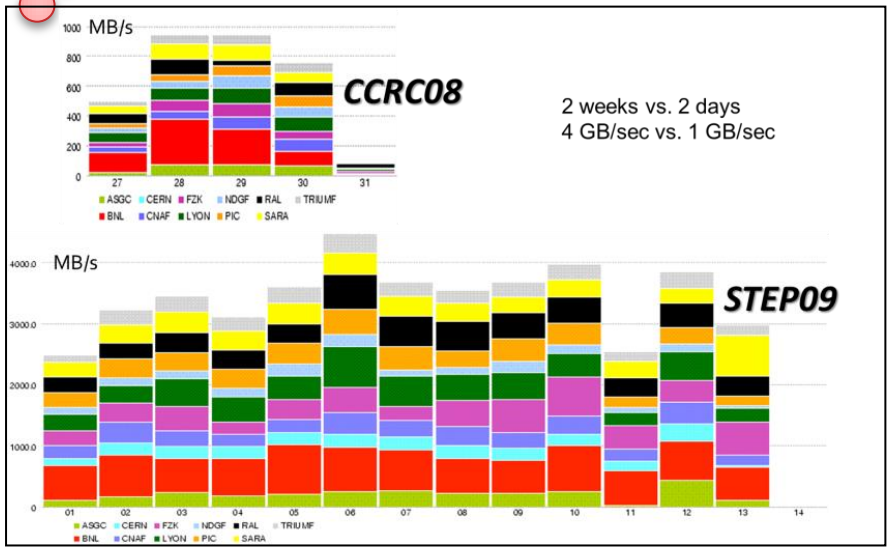


- From hierarchical data distribution to a full mesh and dynamic data placement

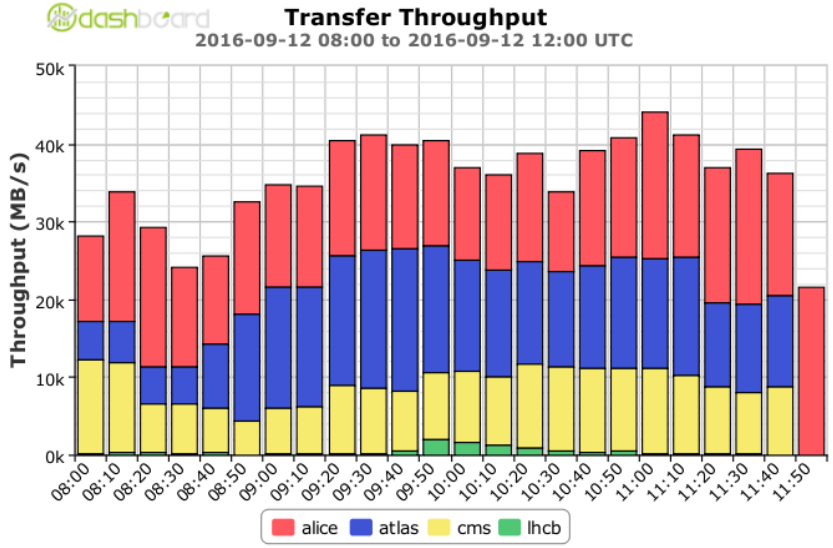


LHCOne graphic: lhcone.net

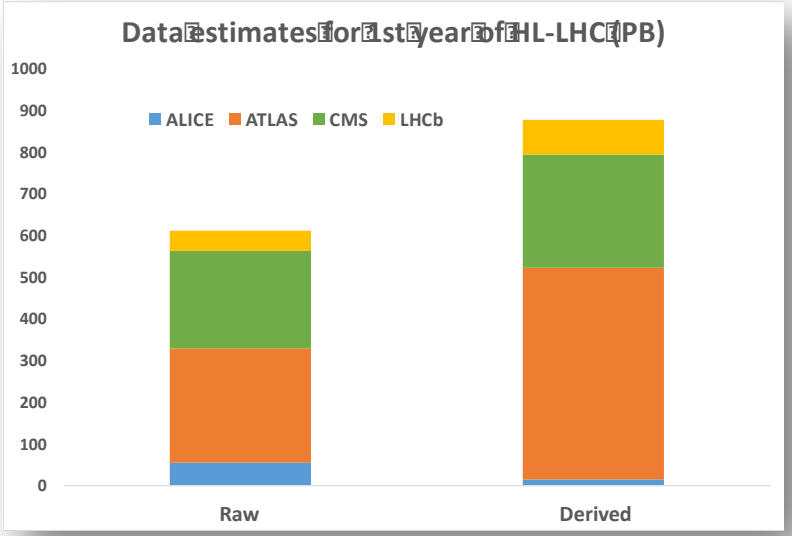
From SC04, CCRC08, STEP09, .. to today …

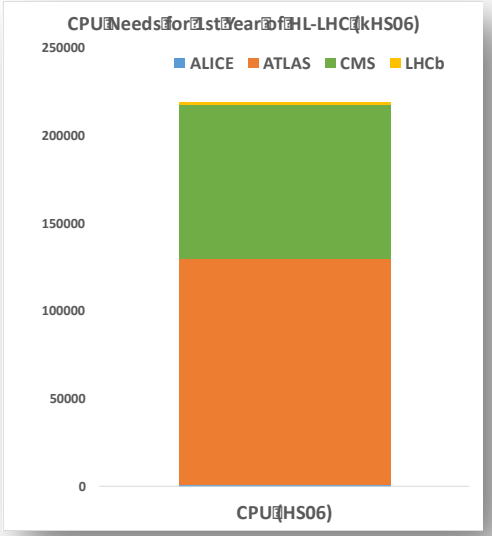


Global transfer rates increased to > 40 GB/s
Acquisition: 10 PB/mo (~x2 for physics data)



… 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

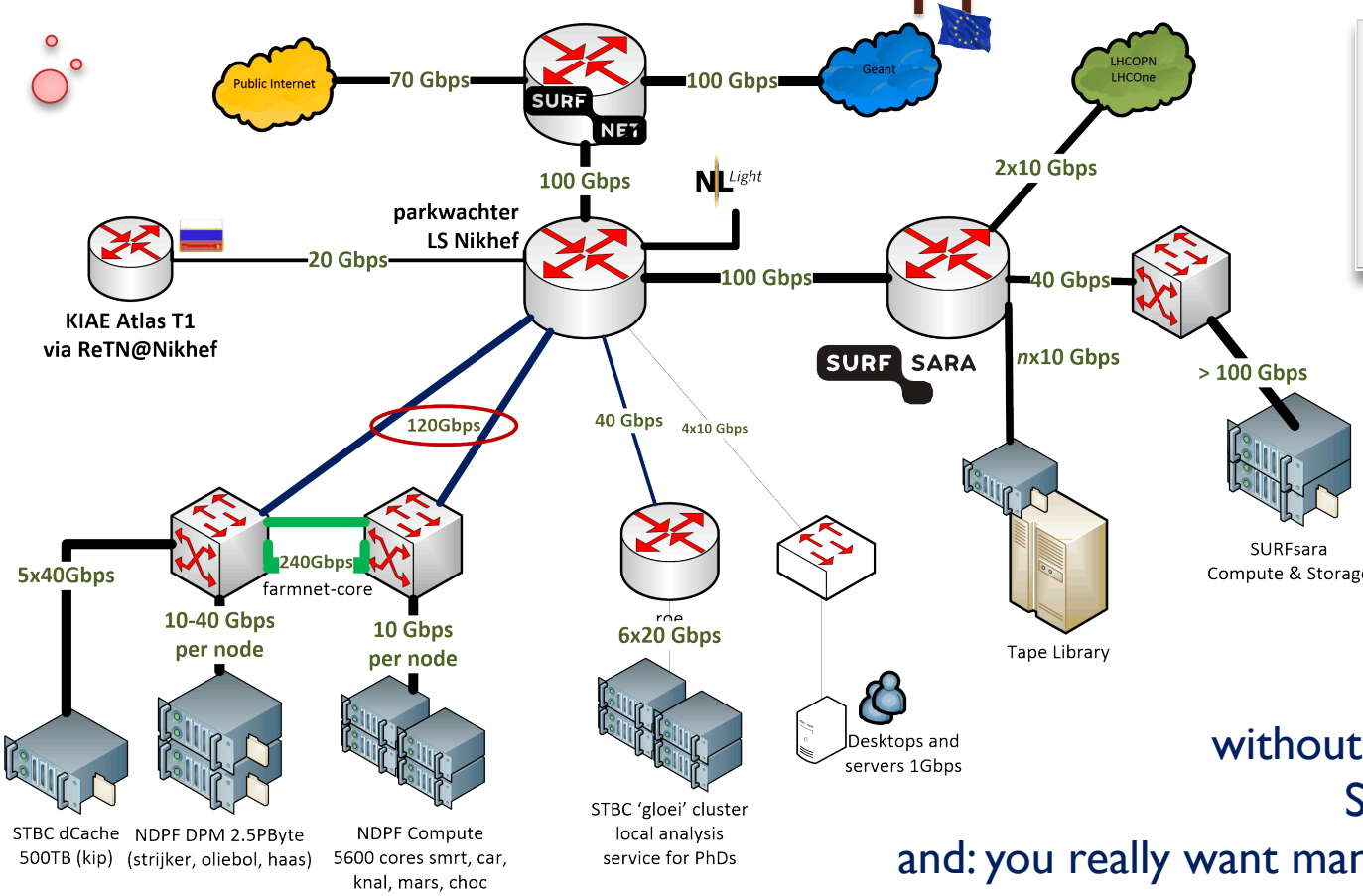
Nikhef

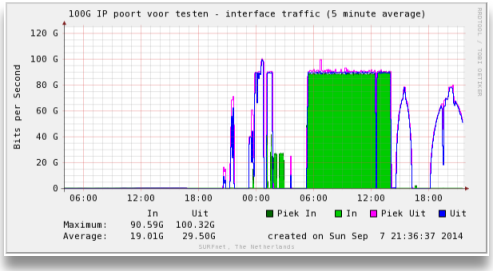
Infrastructure for research: balancing network, CPU, and disk

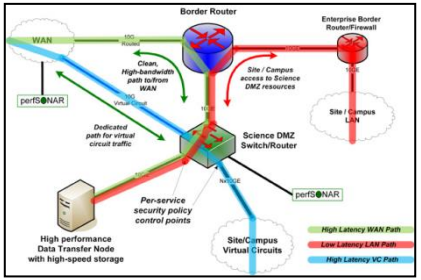
- CPU and disk both expensive, yet idling CPUs are 'even costlier'
- architecture and performance matching averts any single bottleneck
- but requires knowledge of application (data flow) behaviour
  data pre-placement (local access), mesh data federation (WAN access)

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 …

Network built around application data flow





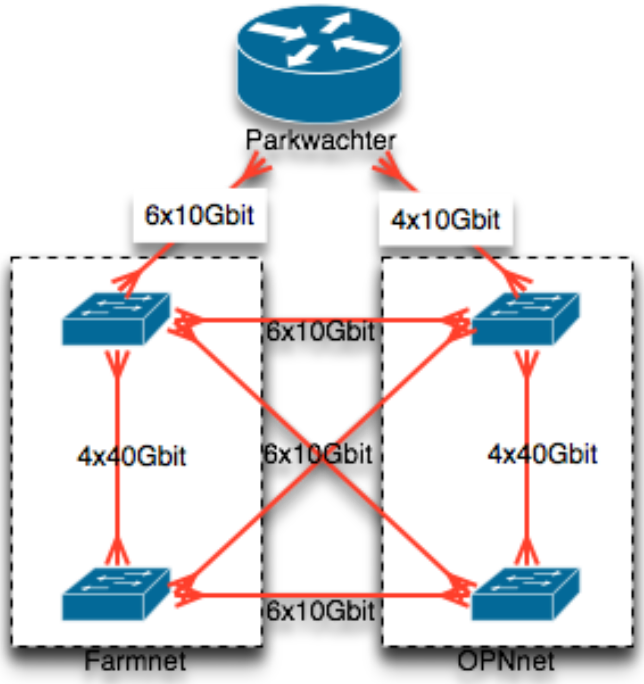


***Need** to work together!*

without our SURFsara peering, SURFnet gets flooded ☺

and: you really want many of your own peerings

‘Homebrew’ SDN … from the ground up



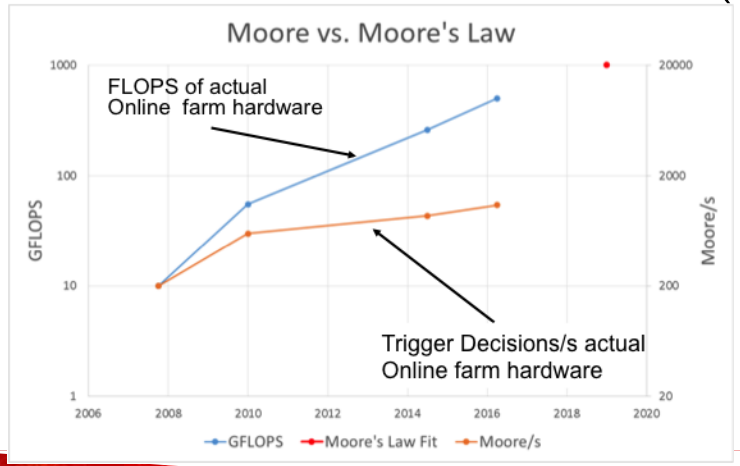
Software Defined Networking (SDN)
real-time re-programming
of switches to follow
connected topology

“DIY SDN” using
switch-native
python capability

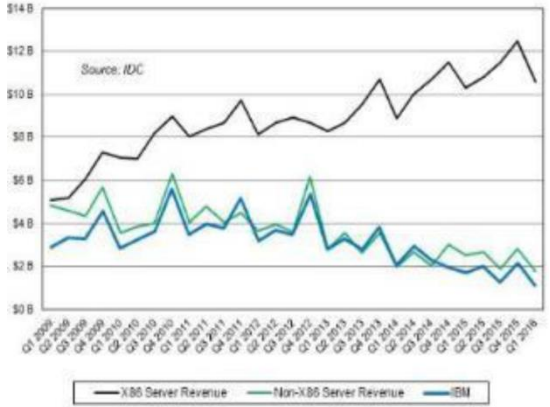
Giving in-switch reprogramming
to support LHCOPN/LHCOne
policy based routes

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)

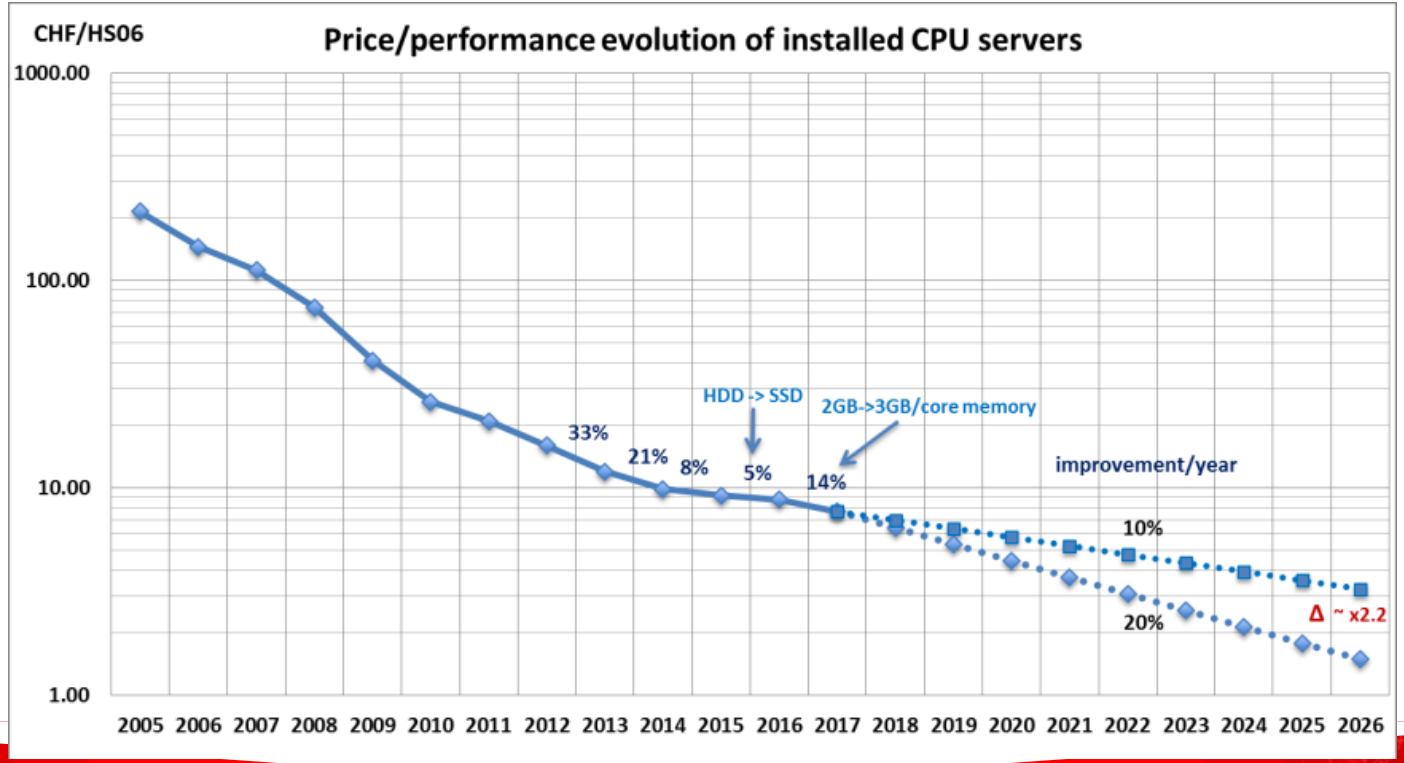


sales volume of different architectures



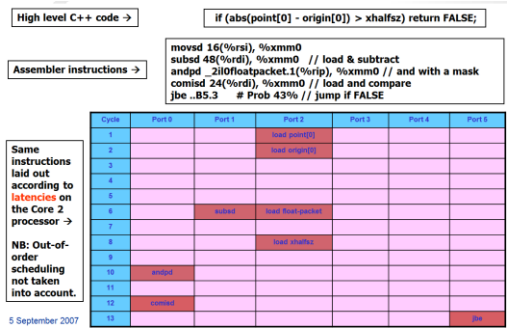
- Yet change must be: most gain to be had from SIMD vectorization and improved memory access patterns

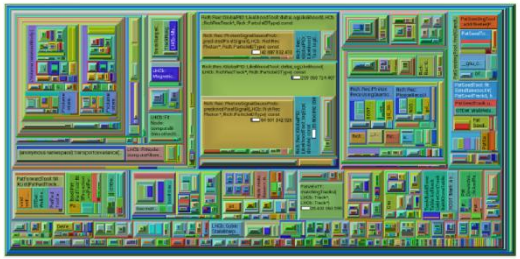
Waiting will not help you any more …

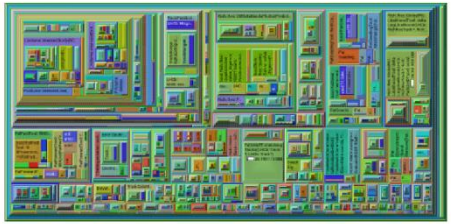


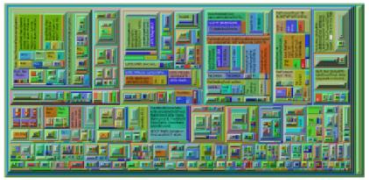
Improvements at the application layer

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









2012 v45r1 v48r1 v48r1 (2015 reco)

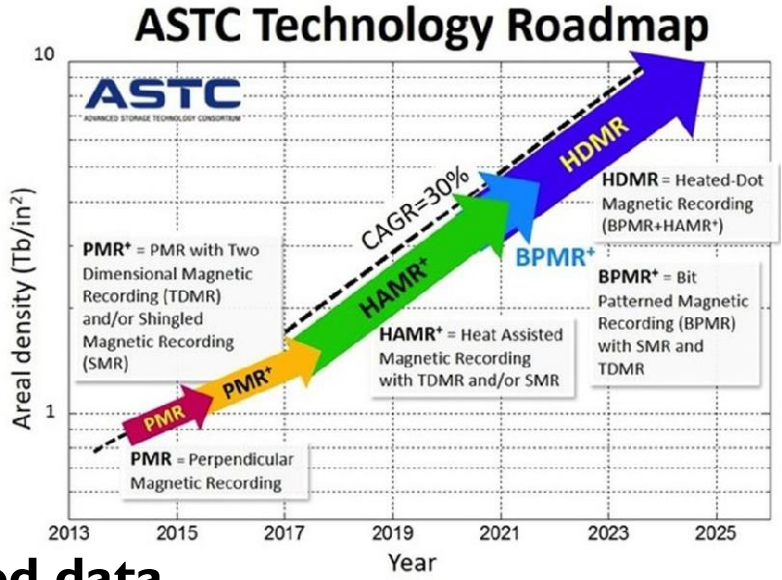
review of algorithms gave overall +34% in LHCb – memory layout still to be done …

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

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

Surviving in the multi-petabyte world

- IO characteristics at 20-30TByte HDD bring unpredictable latencies
  - for new disk: IOPS/platter ~ constant
  - any re-writing uses on-disk caches, those must not be trashed lightly …
  - with disk vendors, exploring massive JBOD disk arrays, keeping ~12 MiB/s/TiB



Don't leave the CPU idle: write ordered data

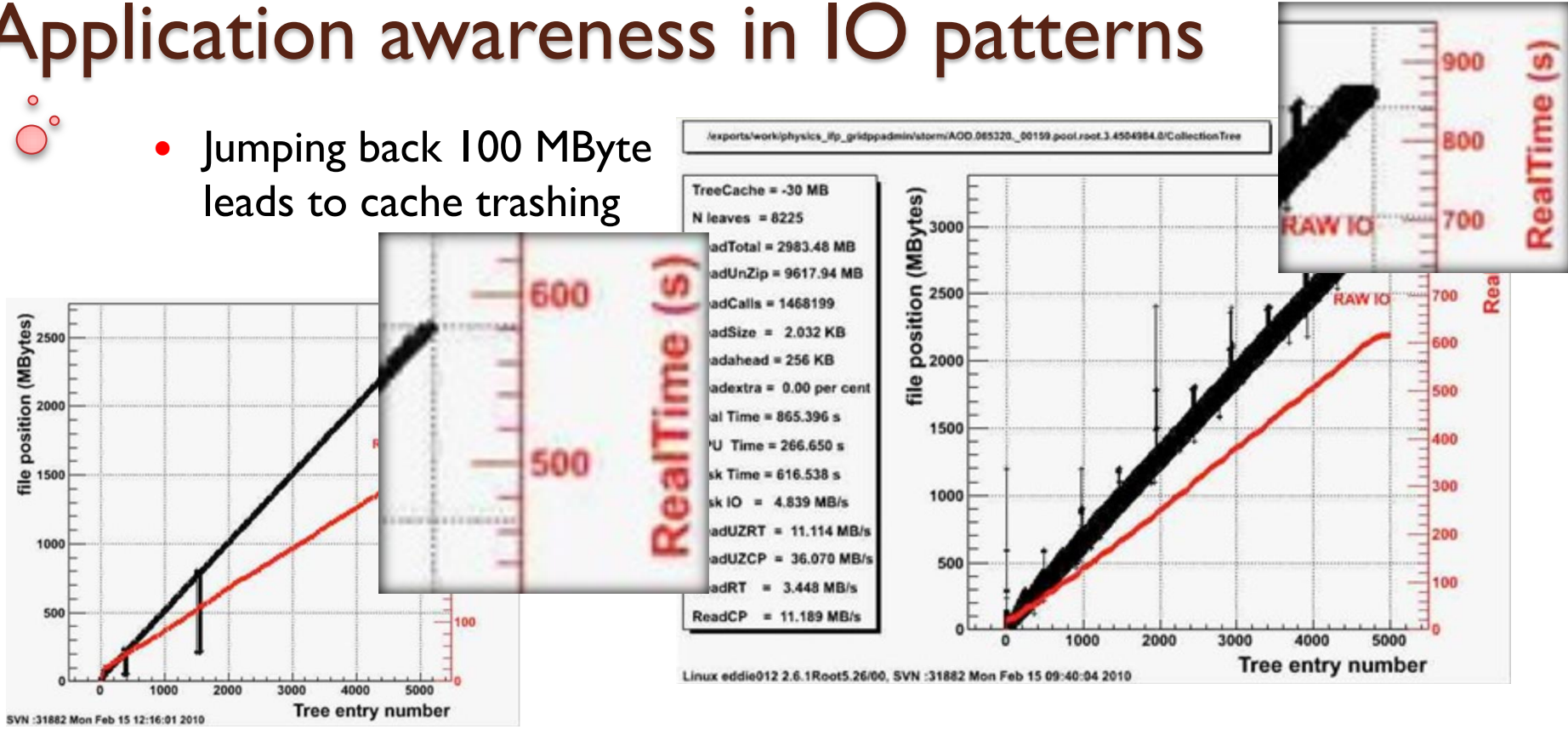
- SSDs will not help us … we would outstrip supply and it's hard to get more: initial Fab investments start at $100-200 B

total (all techs) 2016: ~ 600 EB shipped

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

Application awareness in IO patterns

- Jumping back 100 MByte leads to cache trashing



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

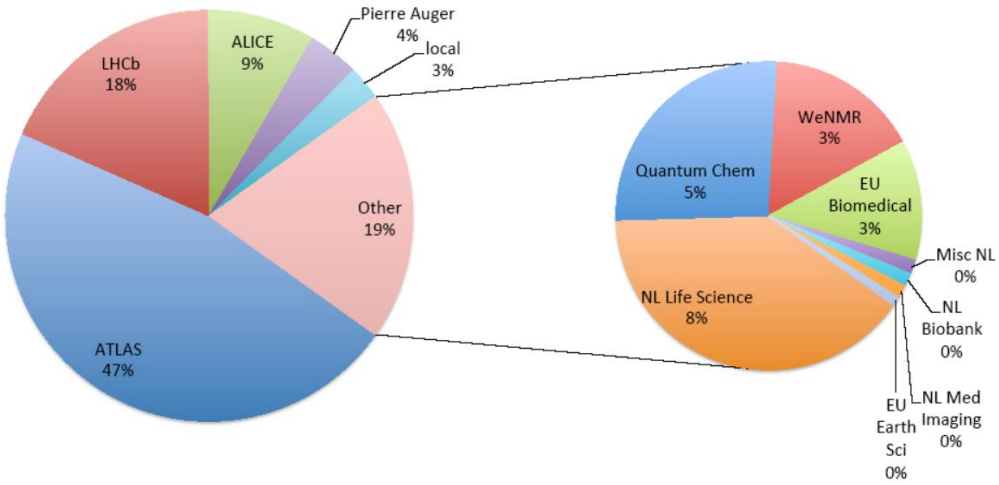
HGST: 480 TByte gross capacity/4RU

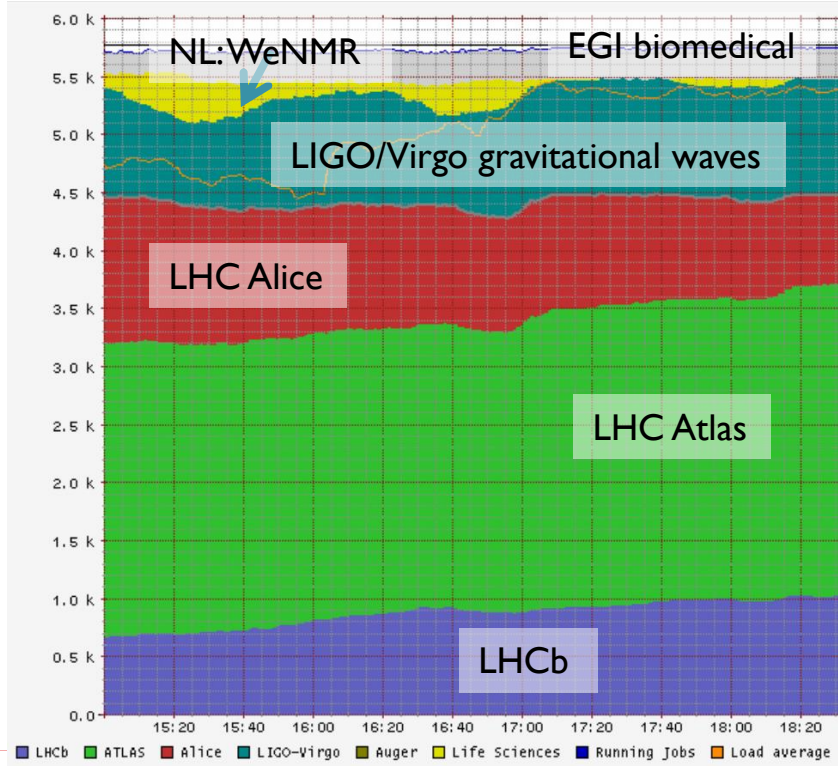
yet more is needed

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

Shared infrastructure, efficient infrastructure!

- >98% utilisation, >90% efficiency

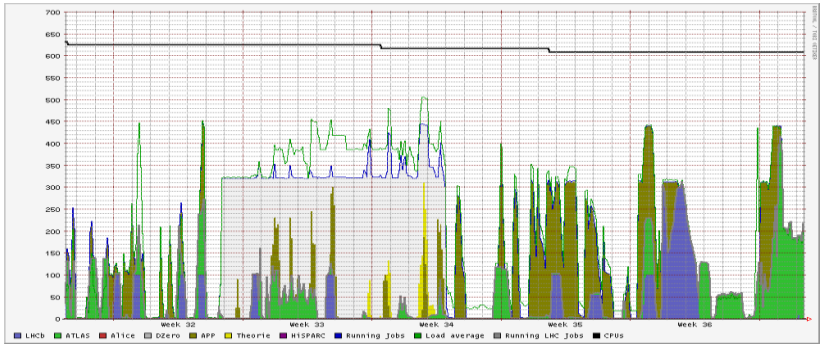




Right:: NIKHEF-ELPROD facility, Friday, Dec 9th, 2016
Left: annual usage distribution 2013-2014

Improve utilisation: towards ‘cloudification’ @Nikhef

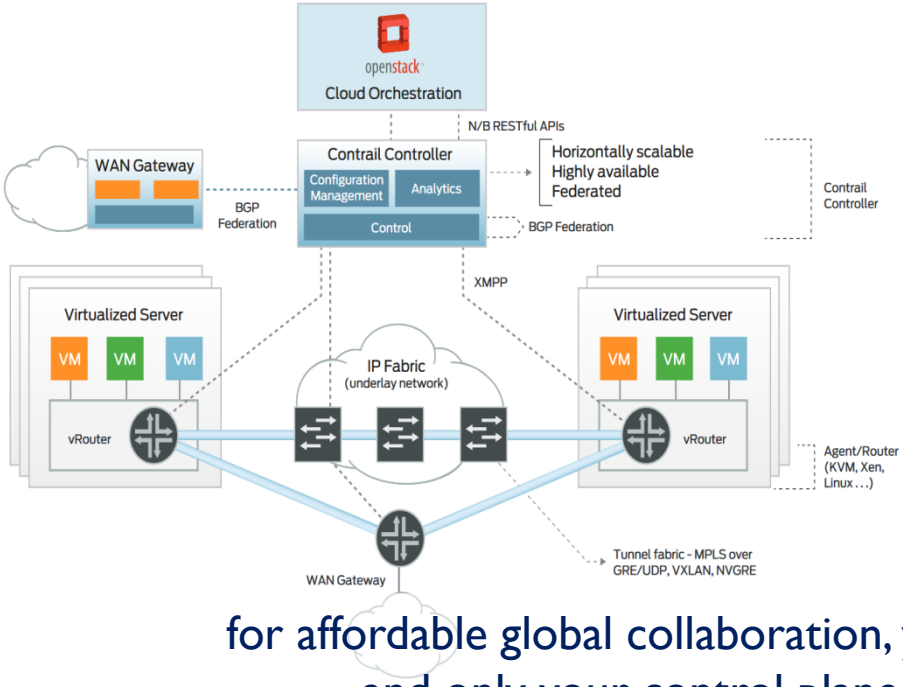
- attract new HTC use cases beyond WLCG and traditional WeNMR/LS
  new communities prefer a different OS distribution and diverse software suites … although they still like a platform service and indulge in orchestration …
- dynamic scaling between DNI nodes, ex-DNI nodes,
  and local computing (‘stoomboot’) to allow short-term bursting



- easier multi-core scheduling and keep >95% ocupancy

Nikhef

Networking from Datacentre to WAN
Virtualising a high-throughput network



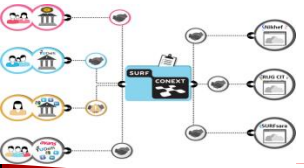
for affordable global collaboration, your network path needs to be 'dumb' and only your control plane virtual (and don't start *NAT*ing)

contrail graphic: http://www.juniper.net/assets/us/en/local/pdf/datasheets/1000521-en.pdf

‘Cloudification’: operations and advanced use cases

- *High Throughput* cloud, a quite different beast
  - from the HPC cloud because of the bias of throughput over memory (and there’s a perfectly great HPC cloud already at SURFsara!)
  - from public cloud offerings, because of unlimited and unmetered bandwidth and data transport
  - Burstring at the application and at the network layer over lightpaths

- Empowering users, but protect them at the same time
  - researchers ‘ill suited’ for system hardening & engineering (we’ve seen that!)
  - use offered collaboration and support – *services* are better than *machines*



SURF CONEXT

Nikhef

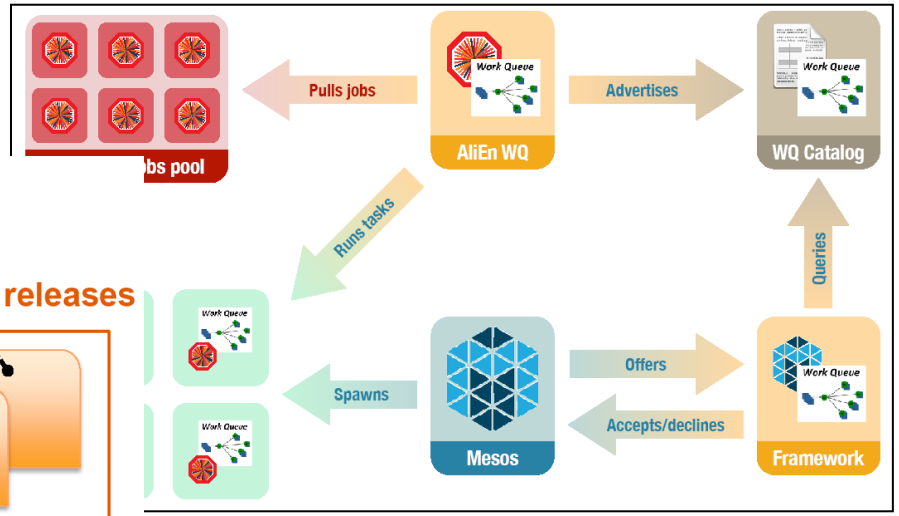
‘cloud’ is a means, not an end-all solution …

/cvmfs/softdrive.nl

Nikhef SURF SARA

E. Tejedor et al., CERN, SWAN Service for Web-based Analtsis, CHEP 2016

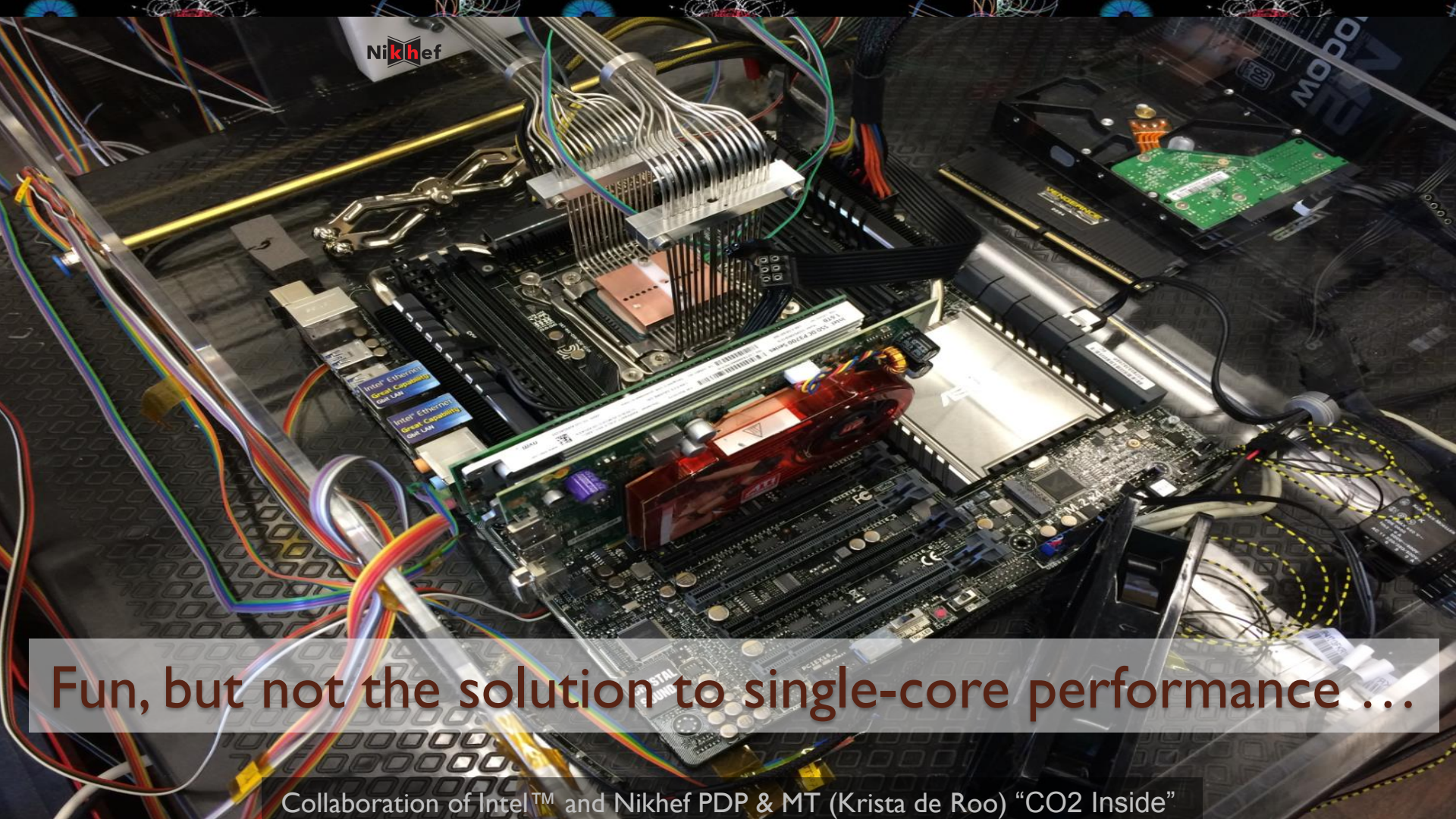
- Docker: single thin image, not managed by the user!
- CVMFS: configurable environment via “views”
- CERNBox: custom user environment



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

Advantage of collaboration: *joint effort of infrastructure and experiments*

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



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

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

Processing of really voluminous data requires more than just a set of disks with a processor glued on top. It needs global networks, continuous performance tuning of storage models, and tight integration with the application framework design to build an efficient data processing system that can span the globe. Using the Netherlands Tier-1 facility for the LHC Computing Grid as an example, we explore how the e-Infrastructure evolves and why the national collaboration is crucial for Nikhef to build such large facilities in a cost-effective way.