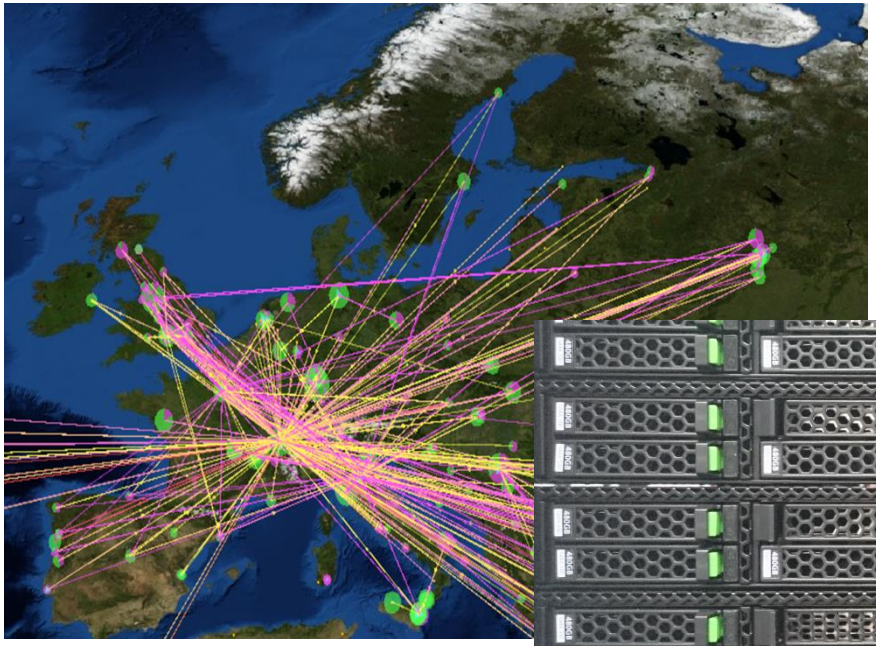
Nikhef –
Advanced Computing for Research

Beyond just LHC Computing



David Groep
Nikhef
PDP - Advanced Computing for Research

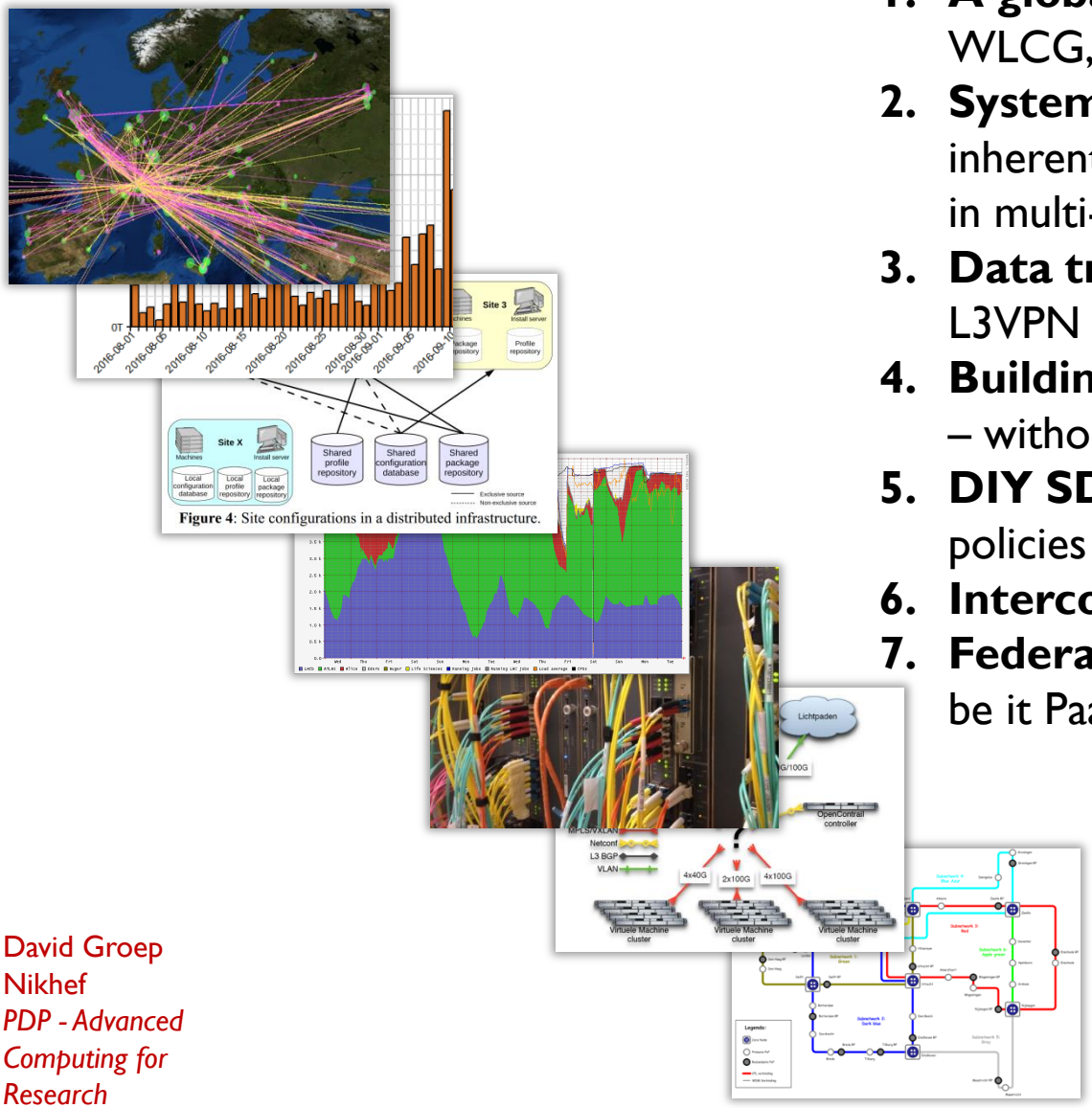


Figure 4: Site configurations in a distributed infrastructure.

1. **A global e-Infrastructure**
   WLCG, federation, and collaboration
2. **Systems for Science**:
   inherently coherent management
   in multi-domain diverse infrastructures
3. **Data transfers via LHCOPN** and LHCone
   L3VPN and the Science DMZ
4. **Building a data intensive data centre**
   – without breaking the bank
5. **DIY SDN**: managing LHCOPN switching
   policies in the network layer
6. **Interconnect**, at least 100G please …
7. **Federated cloud bursting** – at L7 and L2,
   be it Paas-over-IaaS or just IaaS.

David Groep
Nikhef
PDP - Advanced Computing for Research

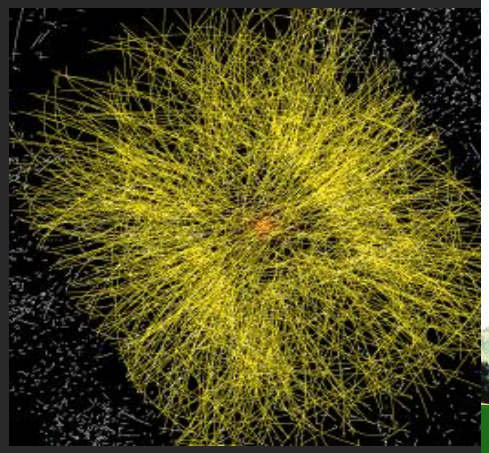
Nikhef

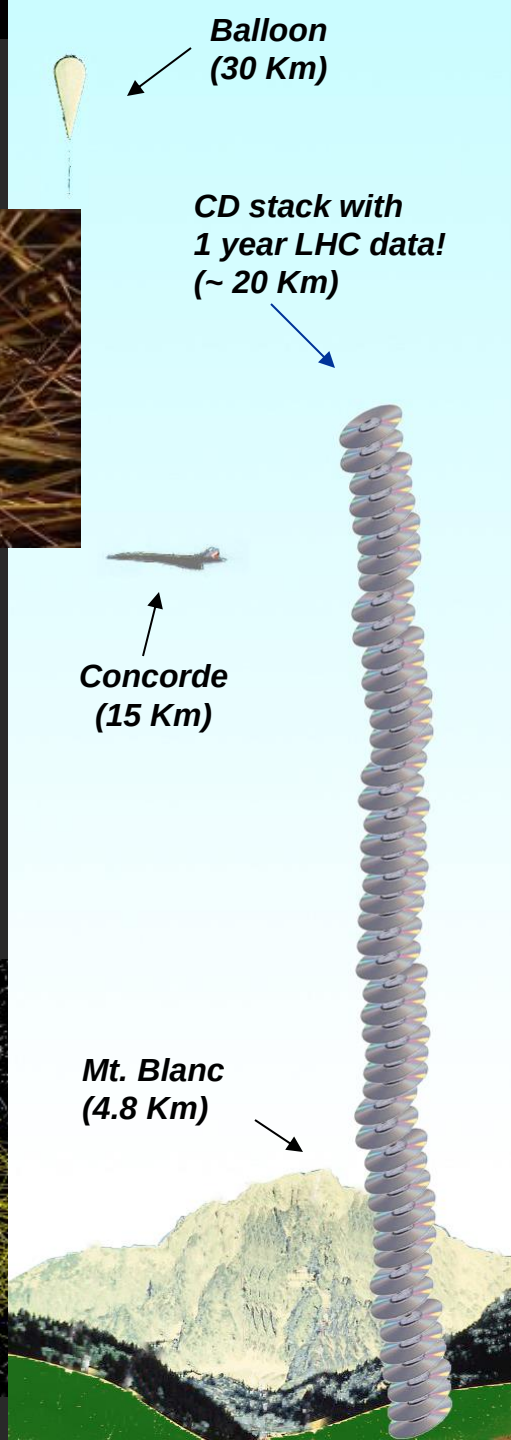
Data from the LHC

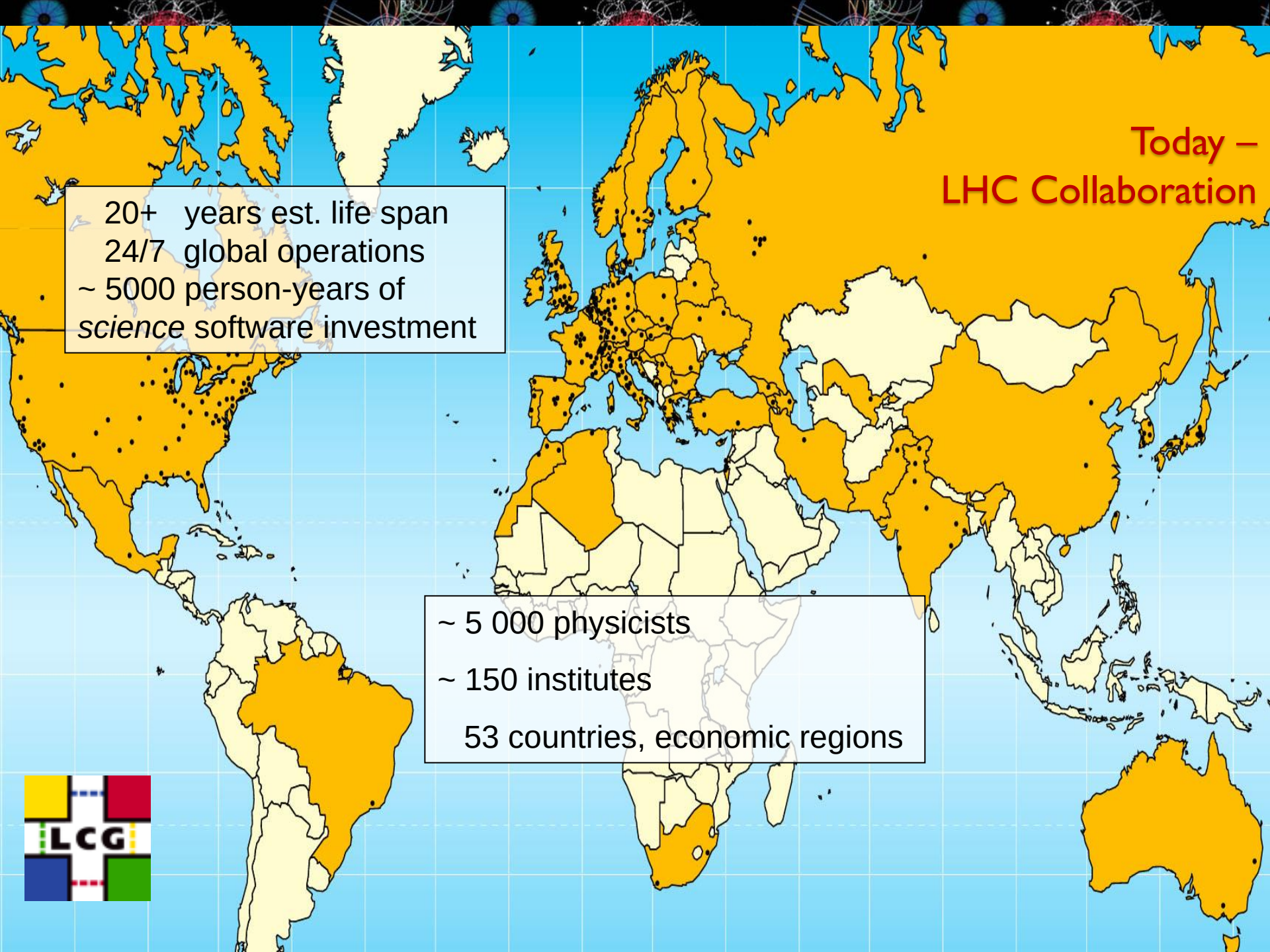
- Signal/Background 10^{-9}

- Data volume
  - (high rate) **X**
    (large number of channels) **X**
    (4 experiments)
  - ➔ **20 PetaBytes of new data each year**

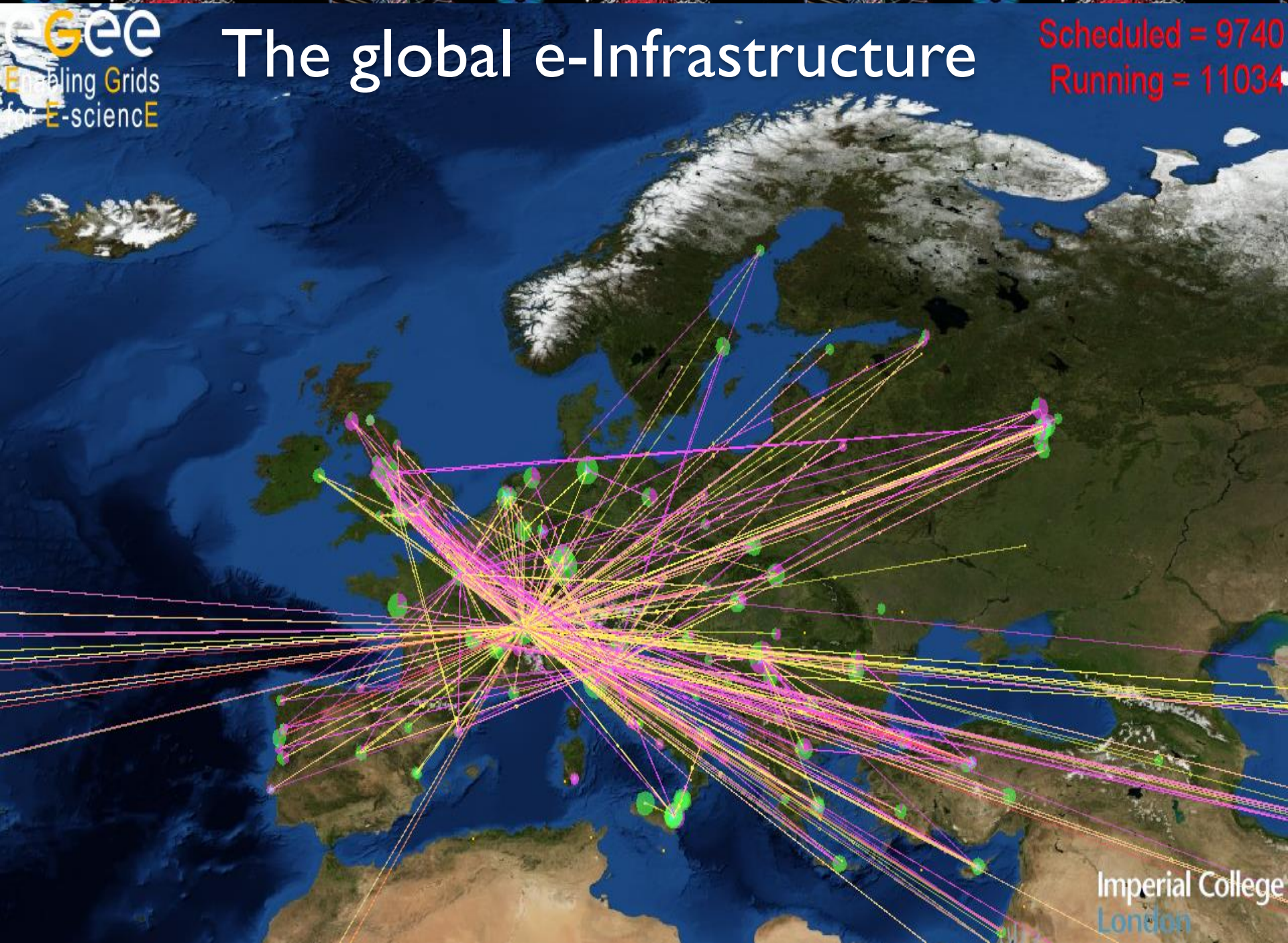
- Compute power
  - (event complexity) **X**
    (number of events) **X**
    (thousands of users)
  - ➔ **60'000 of (today's) fastest CPUs**






Today –
LHC Collaboration
20+ years est. life span
24/7 global operations
~ 5000 person-years of
science software investment
~ 5 000 physicists
~ 150 institutes
53 countries, economic regions
LCG

The global e-Infrastructure



BiG Grid, the Dutch e-Science Grid

- Started @Nikhef in 1999 Nikhef with WTCW VL
  - 2001: European DataGrid and EGEE projects joint with SARA (and KNMI)
- Started BiG Grid in 2005/2007 to consolidate e-Science infrastructure & production support
- Initiative lead by the science domains
  - NCF and its scientific user base
  - NBIC, Netherlands BioInformatics Center
  - Nikhef, where you are now
- With SARA as main operational partner

  Resulting in SURF today coordinating the national e-Infra

BiG Grid
the dutch e-science grid

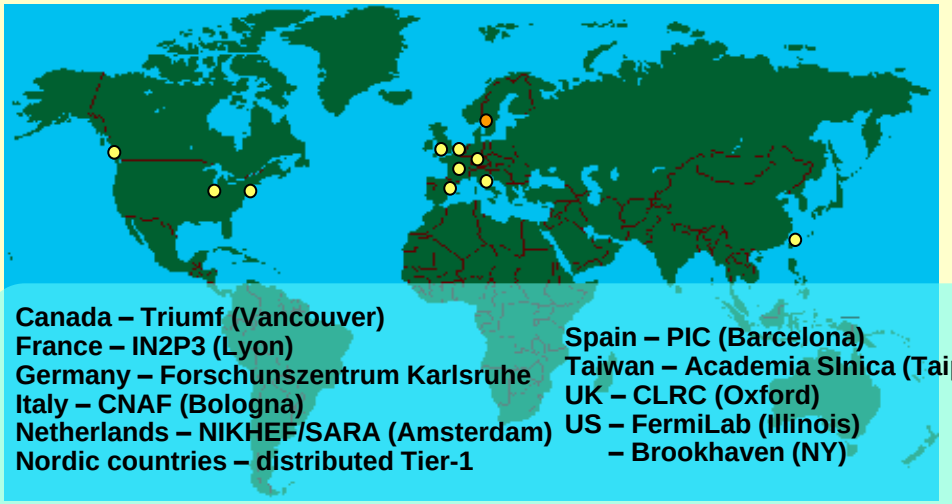
David Groep
Nikhef
PDP - Advanced Computing for Research

Nikhef

LCG Services

Tier-0 – the accelerator centre

- Data acquisition & initial processing
- Long-term data curation
- Distribution of data → Tier-1 centres



Canada – Triumf (Vancouver)
France – IN2P3 (Lyon)
Germany – Forschunszentrum Karlsruhe
Italy – CNAF (Bologna)
Netherlands – NIKHEF/SARA (Amsterdam)
Nordic countries – distributed Tier-1
Spain – PIC (Barcelona)
Taiwan – Academia SInica (Taipei)
UK – CLRC (Oxford)
US – FermiLab (Illinois)
– Brookhaven (NY)

Tier-1 – "online" to the data acquisition process → high availability

- Managed Mass Storage – → grid-enabled data service
- Data-heavy analysis
- National, regional support

Tier-2 – ~120 centres in ~35 countries

- **End-user (physicist, research group) analysis** – where the discoveries are made
- Simulation

Wide Area Network

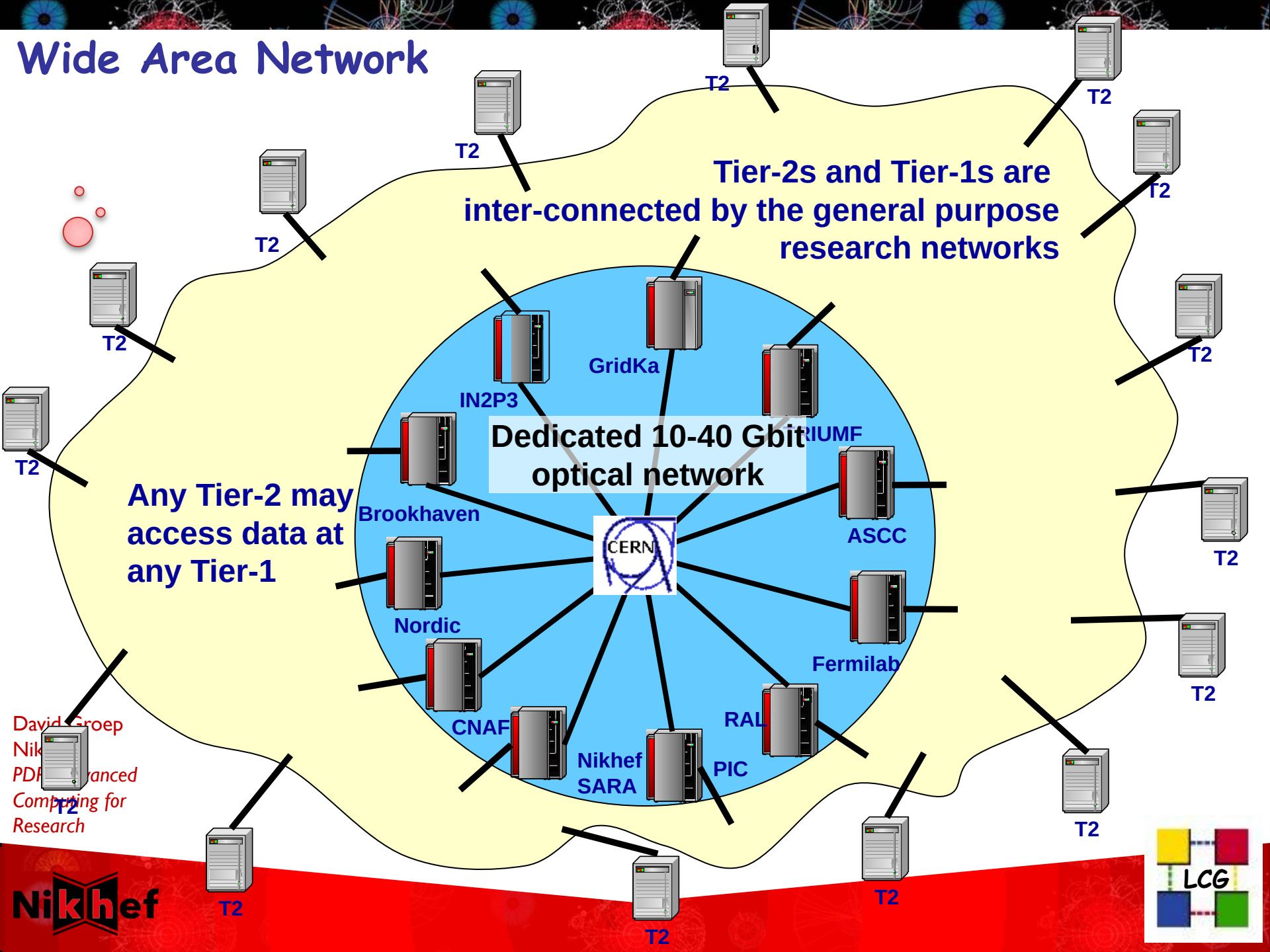


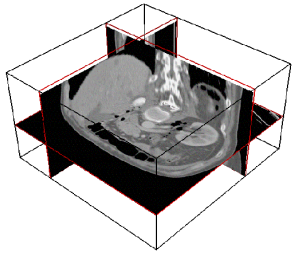
Image sources: VL-e Consortium Partners

e-Science and National e-Infrastructure

vl·e

Data integration for genomics, proteomics, etc. analysis

Timo Breit et al.
Swammerdam Institute of Life Sciences

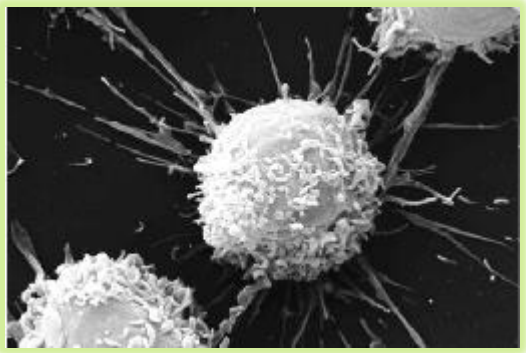


Medical Imaging and fMRI

Silvia Olabarriaga et al.
AMC and UvA IvI

Avian Alert and FlySafe

Willem Bouten et al.
UvA Institute for Biodiversity Ecosystem Dynamics, IBED

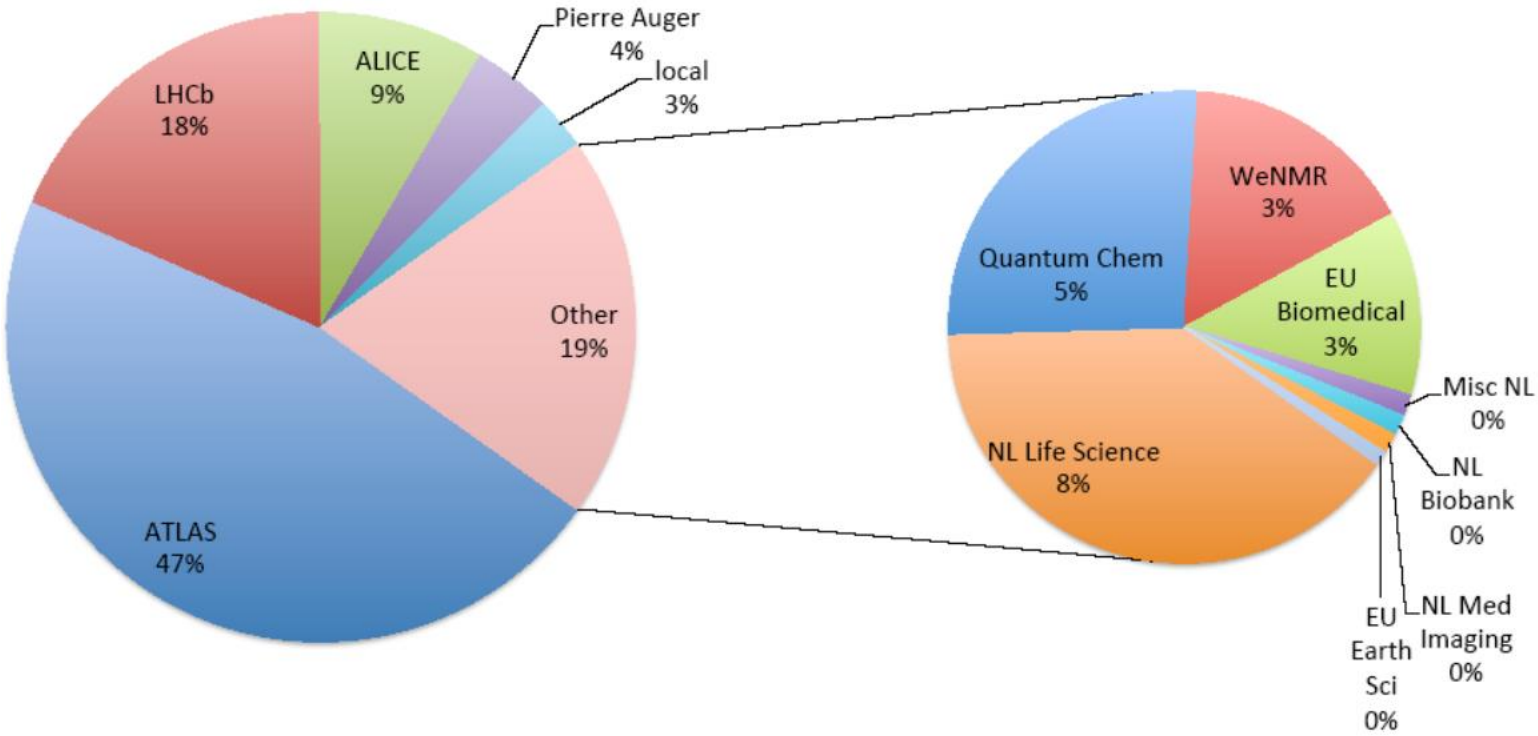


Bram Koster et al.
LUMC
Microscopic Imaging group

Molecular Cell Biology and 3D Electron Microscopy

Many user communities

- HTC Grid Computing service



Global collaboration (in a secure way)

Why would I trust you? How do I know who you are?

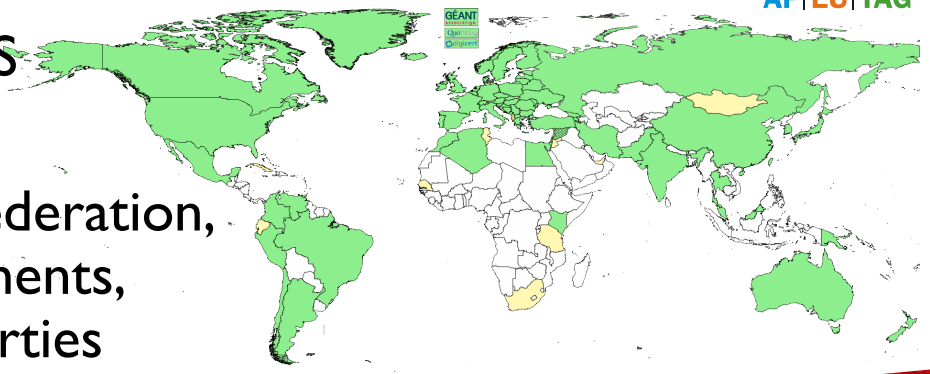
A global identity federation for e-Infra and cyberinfrastructures

- Common baseline assurance requirements
- Persistent and globally unique

For the Grid a truly global identity is needed
— so we built the International Grid Trust Federation

- over 80 member Authorities
- Including, e.g., the TCS

IGTF
Interoperable Global Trust Federation
AP | EU | TAG



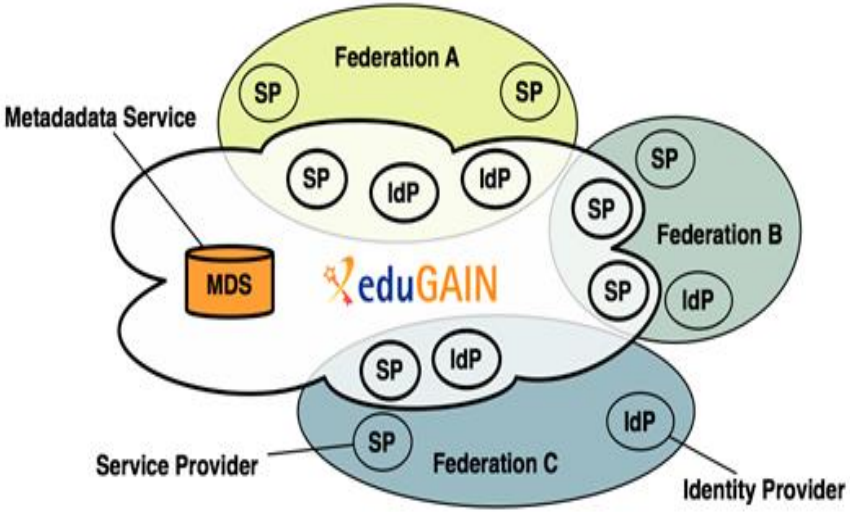
- And it works in a global federation, with harmonized requirements, driven by actual relying parties

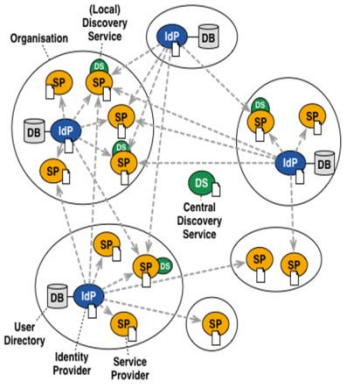
David Groep
Nikhef
PDP - Advanced Computing for Research

Nikhef

R&E federations – providing multi-domain services

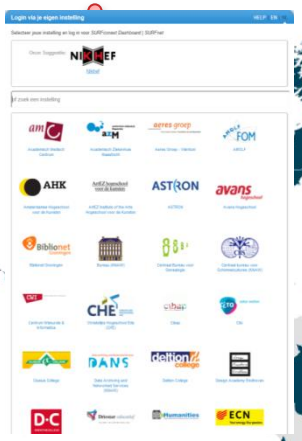
- Multiple technologies (SAML2, OIDC, PKIX) allow cross-organisational 'single sign-on' – and trust between user home organisations and service providers created a global service ecosystem





Graphic courtesy Alessandra Sciccitano, GEANT

And now we have this!



eduGAIN

okeanos GLOBAL

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

TCS eScience Portal

Certificates

My certificates

Help

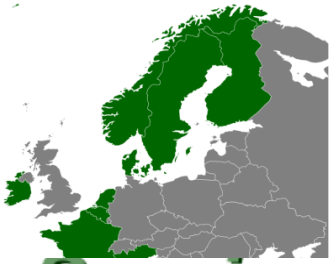
About NREN
About Portal
Privacy Notice

Help
CA Certificate

Language

Login

Please choose your country



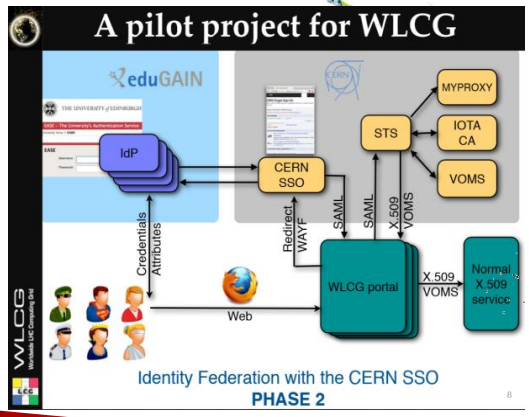
CILogon Service

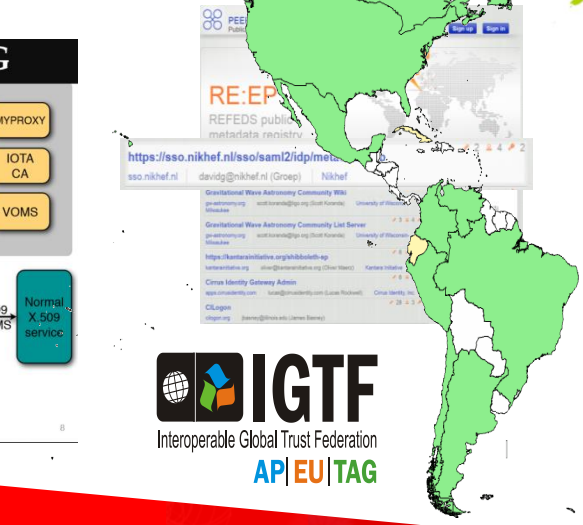
GÉANT association

QuoVadis

digicert

wLCG FIM4R pilot





IGTF
Interoperable Global Trust Federation
AP | EU | TAG

RCauth.eu

Nikhef

Systems for Science

David Groep
Nikhef
PDP - Advanced Computing for Research

Nikhef

Statistics

BiG Grid
the dutch e-science grid

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

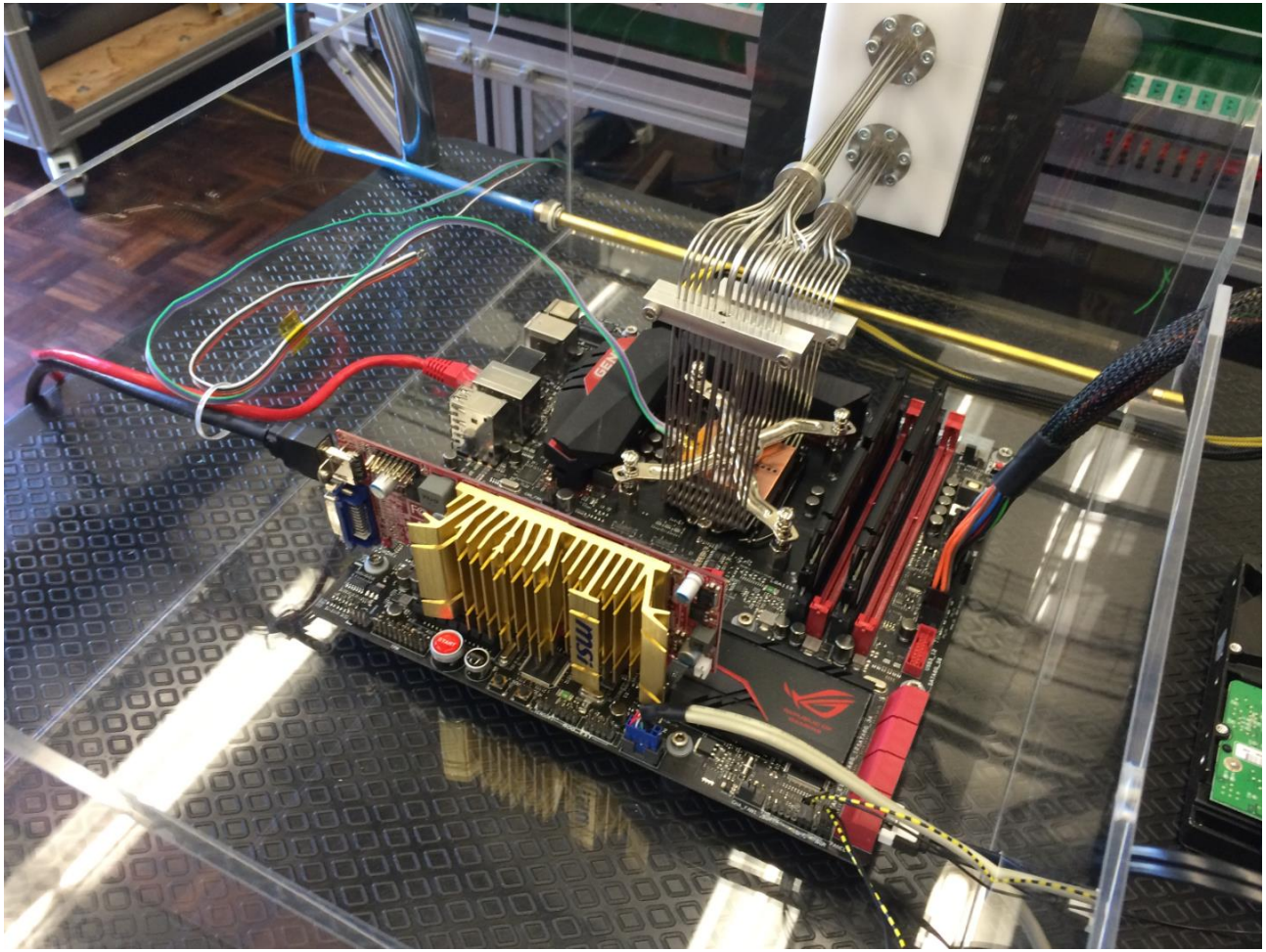
~ 5500 cores and 3.5 PiB

~ 400 systems installs (focus on large/many-core systems)

> 45 install flavours (service types)

and a bunch of one-off systems

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



David Groep
Nikhef
PDP - Advanced Computing for Research

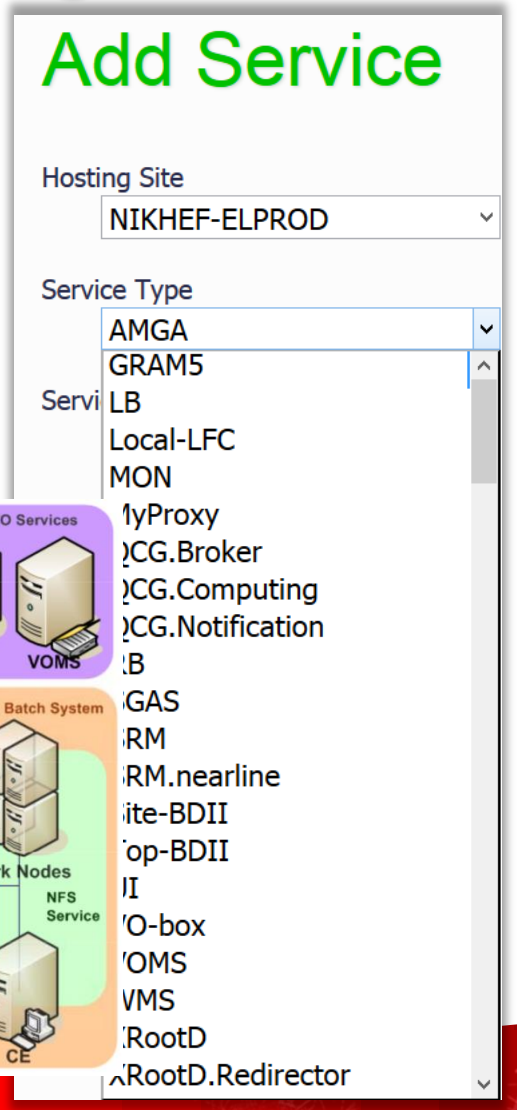
CO2-cooled Intel CPUs @6.2GHz

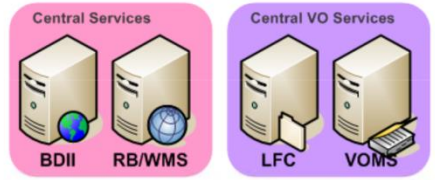
For (informed) fun & testing – some random one-off systems …

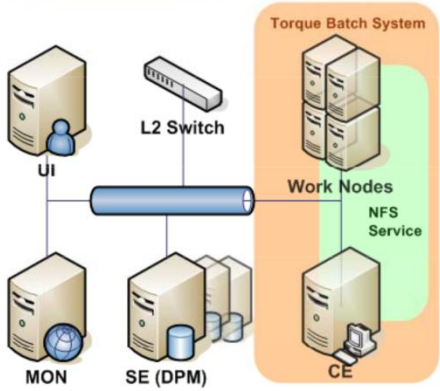
David Groep
Nikhef
PDP - Advanced Computing for Research

Heterogeneous systems management

- Globally federated compute and storage
- ‘hourglass’ model service specifications
  yet unlike TCP, the neck is thick
  compute service, file staging, storage, monitoring, brokering, information system, cataloging, authorization data, online-nearline-offline stage transitions, …







- Service and system variety
- Fabric management: Quattor

David Groep
Nikhef
PDP - Advanced Computing for Research

Nikhef Data Processing Facility – phased deployment model

CTB

- Software beta versions
- Assessment and suitability
- One-off installations permitted

ITB

quattor

- Pre-production validation
- Mirror of production system at *N+1*
- *Configuration debugging permitted*

PRD

quattor

- Production system
- Configuration exclusively through version management

HTC Computing Platform 'Grid' configuration

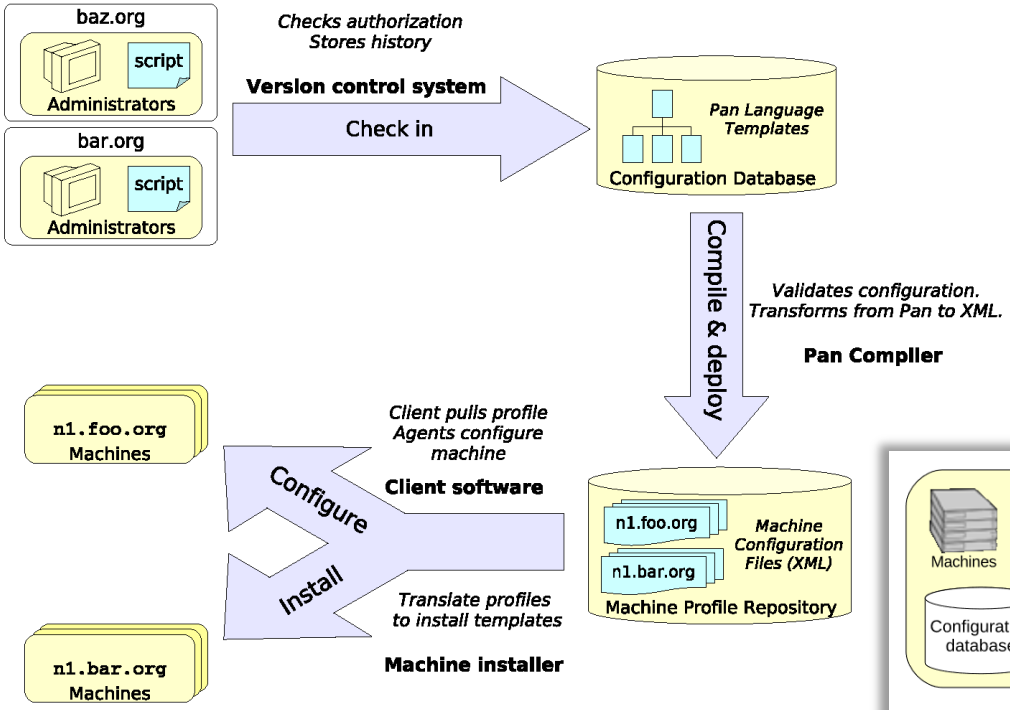
Uses **Quattor** configuration management suite

quattor

- **Federated management**: modular tool-kit for installation, configuration, and state management
- **Shared configuration** and specialisation: Quattor encourages re-use of configuration information across nodes (and across installation in a federated multi-domain service)
- **Inherent coherency**: based on site model used to manage range of different resources, such as real machines, virtual machines and cloud resources, and **ensures configuration compliance** (any install always revert to its intended state)

David Groep
Nikhef
PDP - Advanced Computing for Research

Configuration workflow for systems



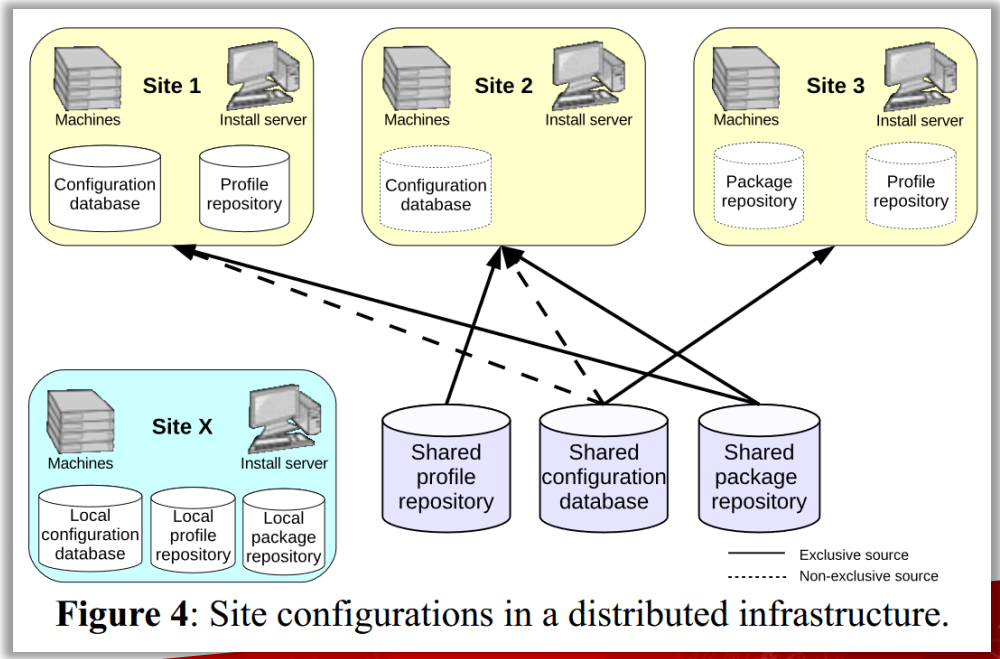
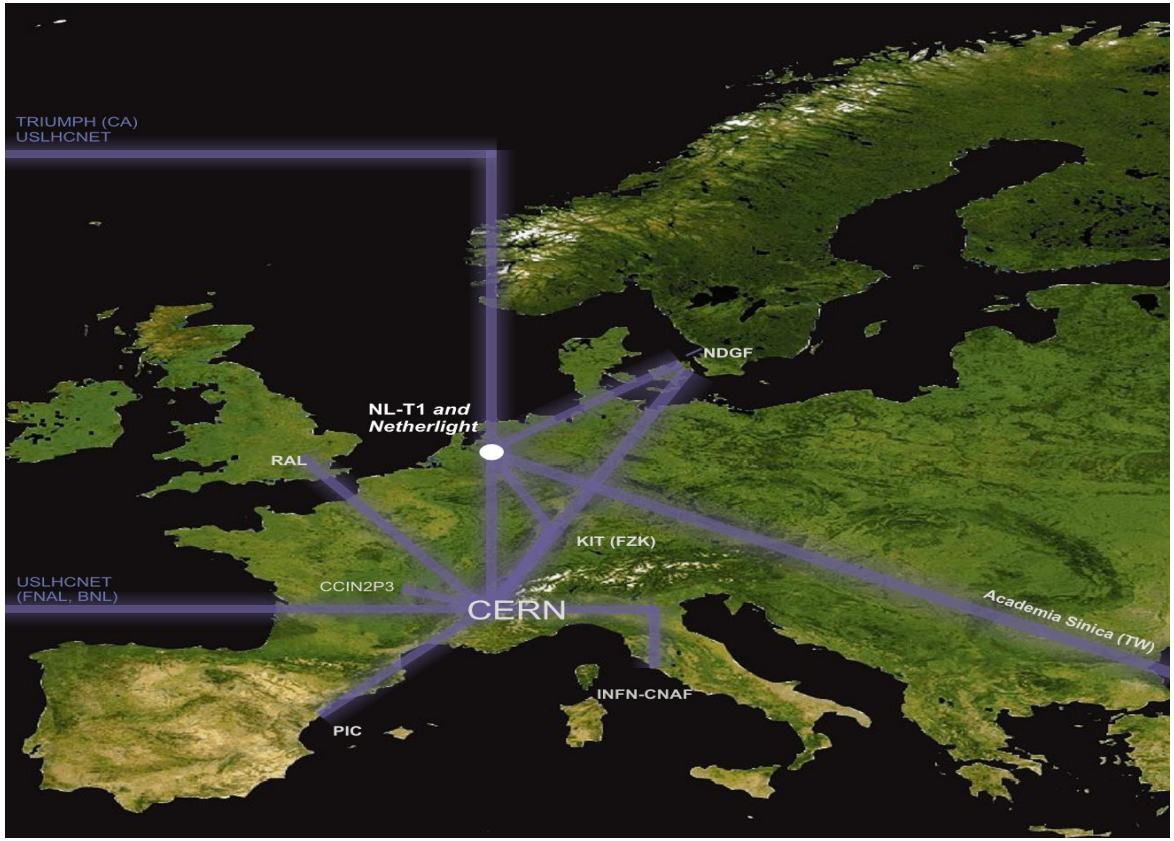


Figure 4: Site configurations in a distributed infrastructure.

David Groep
Nikhef
PDP - Advanced Computing for Research

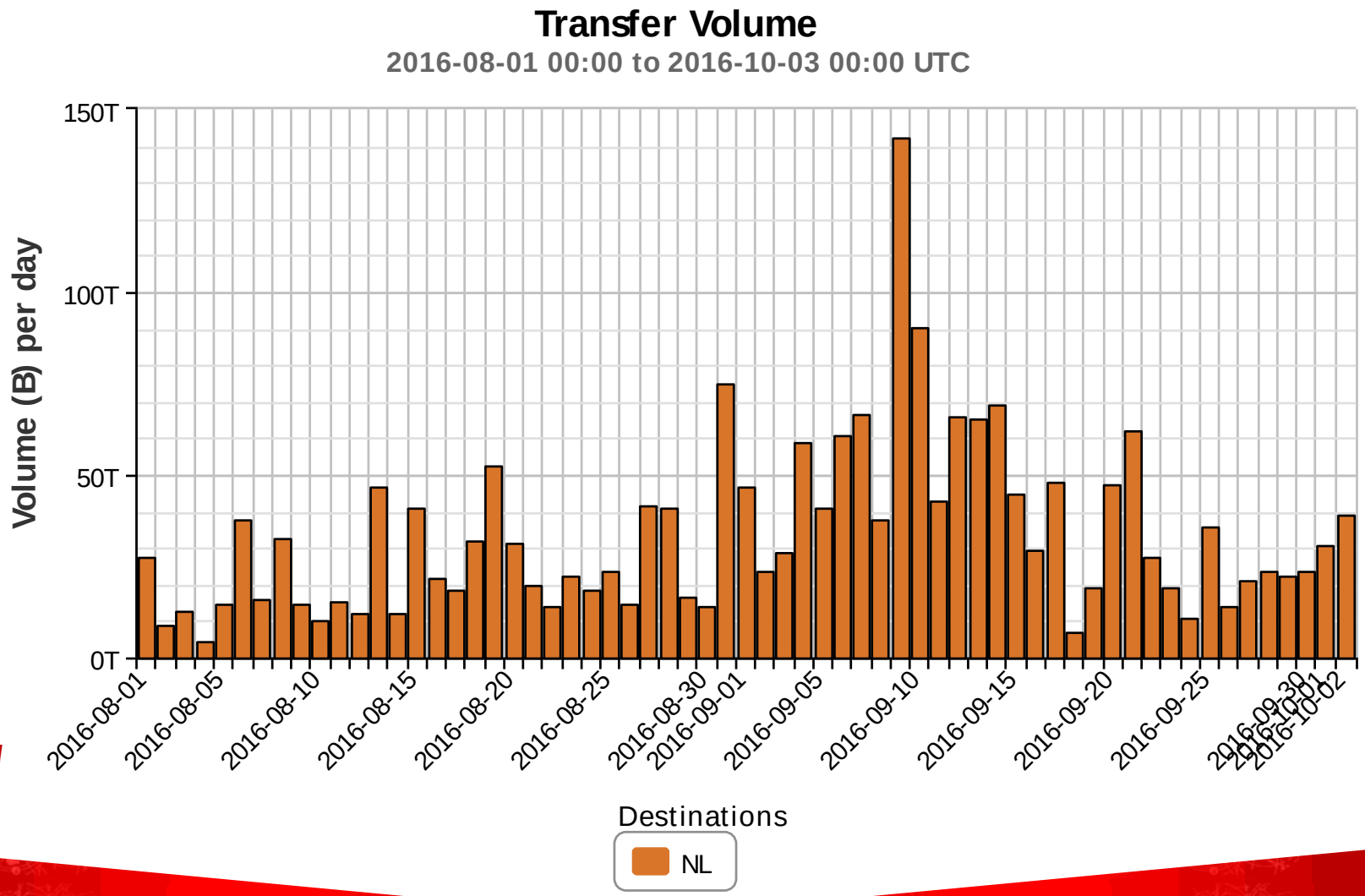
Graphics: Childs et al., Devolved Management of Distributed Infrastructures
22nd Large Installation System Administration Conference



LHCOPN and global data transfers

David Groep
Nikhef
PDP - Advanced Computing for Research

ATLAS transfers to Nikhef & NL-T1



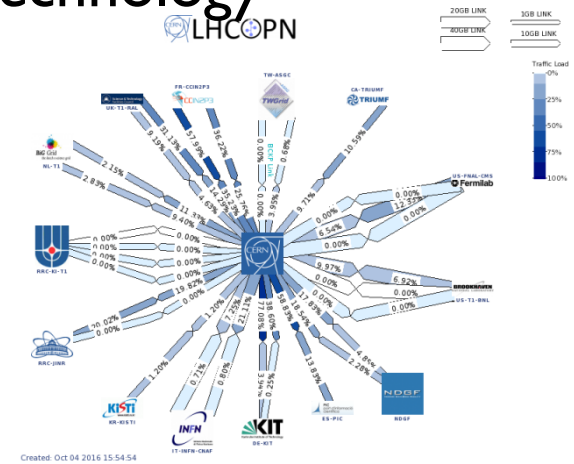
David Groep
Nikhef
PDP - Advanced Computing for Research

source: http://dashb-atlas-ddm.cern.ch/ddm2

LHC OPN – policy based private net

Aim: offload T0-T1 data transfers over protected links

- Plugs into the 'ScienceDMZ' concept – separating the 'elephant flows'
- bypassing stateful firewalls if any
- Restricted policy: allow only T0-T1-T1 traffic
- Scalable model using (primarily) VRF technology
- RS-LHCOPN today holds 38 prefixes



~5% of total traffic to NL

we are ~4.5% of the CERN membership

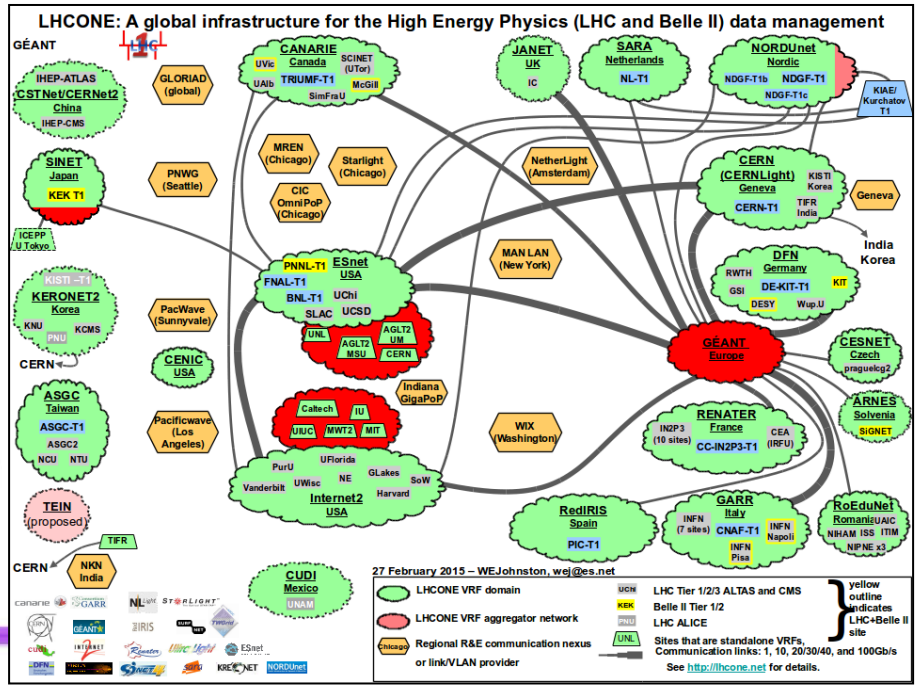
perfSONAR powered

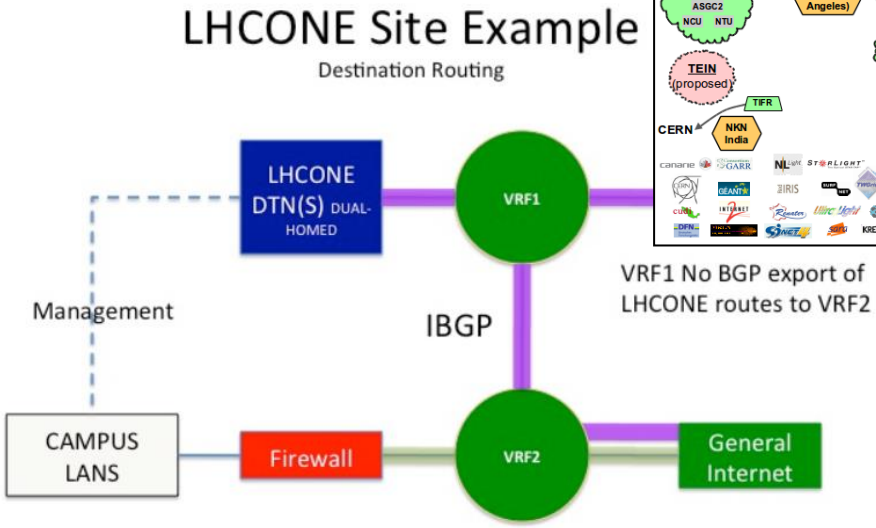
David Groep
Nikhef
PDP - Advanced Computing for Research

LHCone – dynamic networking

towards full SDN capability: the L3VPN service for LHC

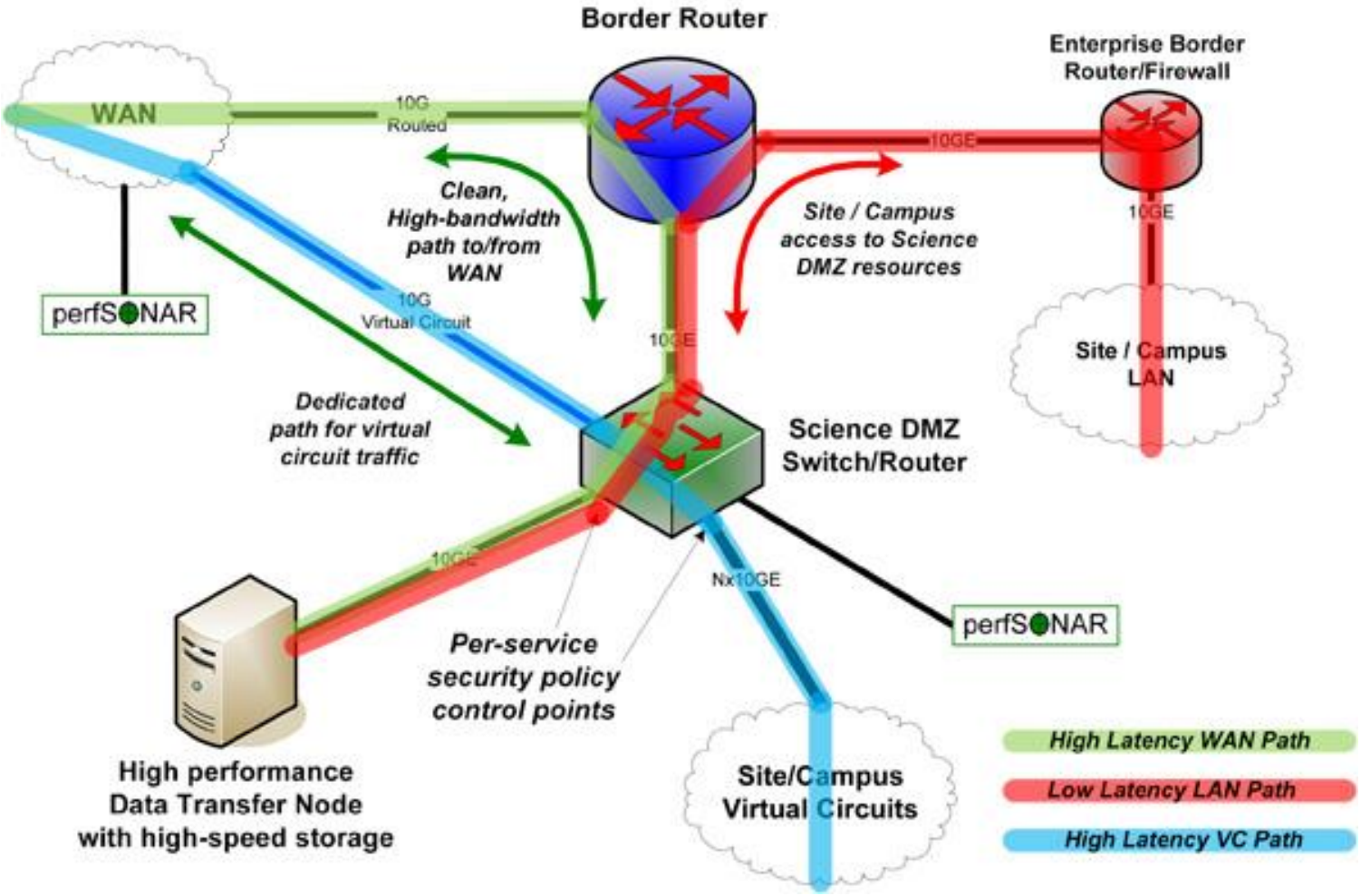
- NSI
- OpenFlow
- TRILL/SPB
- VRF





David Groep
Nikhef
PDP - Advanced Computing for Research

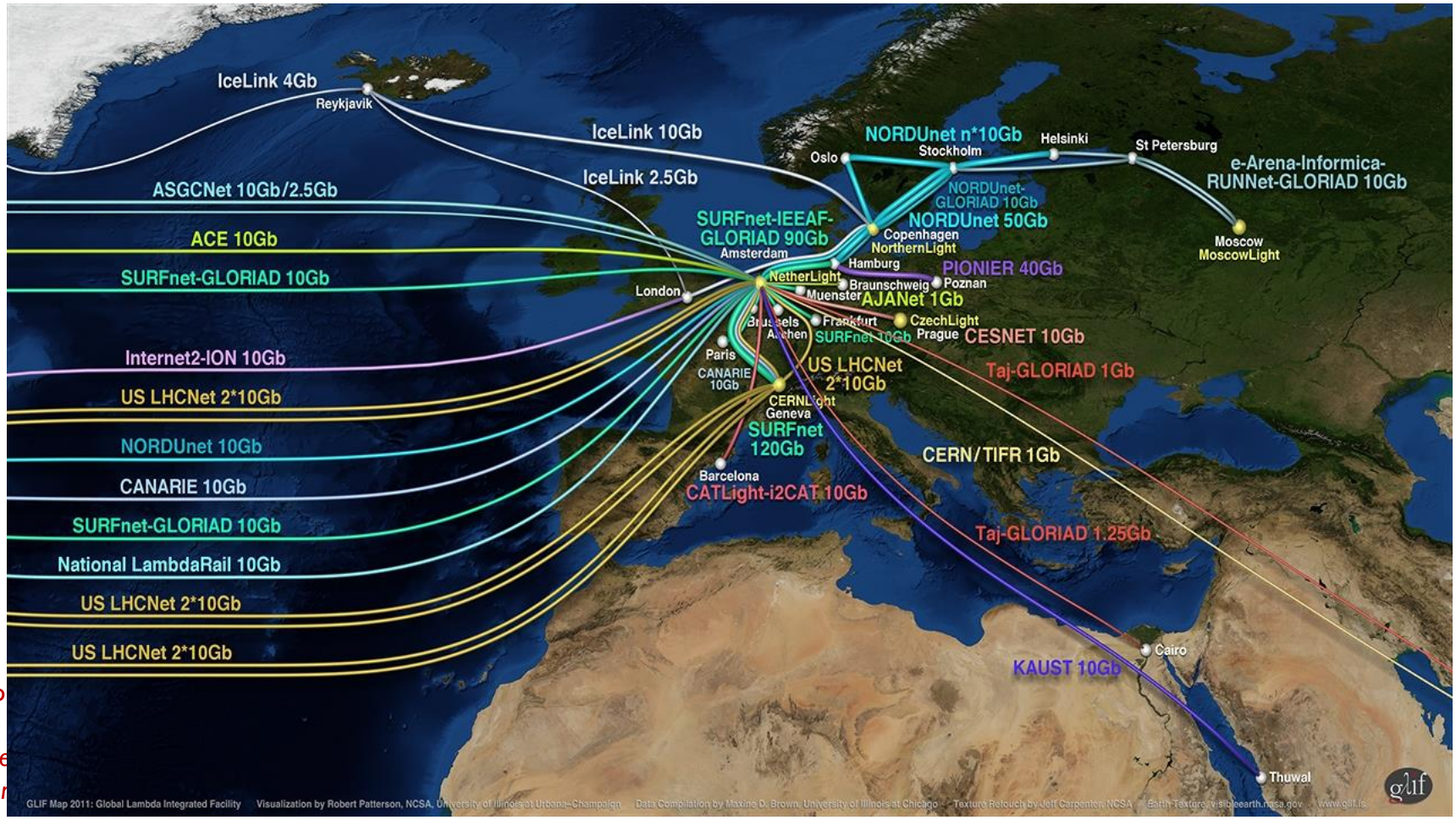
Science DMZs – separating network flows



David Groep
Nikhef
PDP - Advanced Computing for Research

image and concept courtesy: Eli Dart, Brian Tierney et al, ESnet

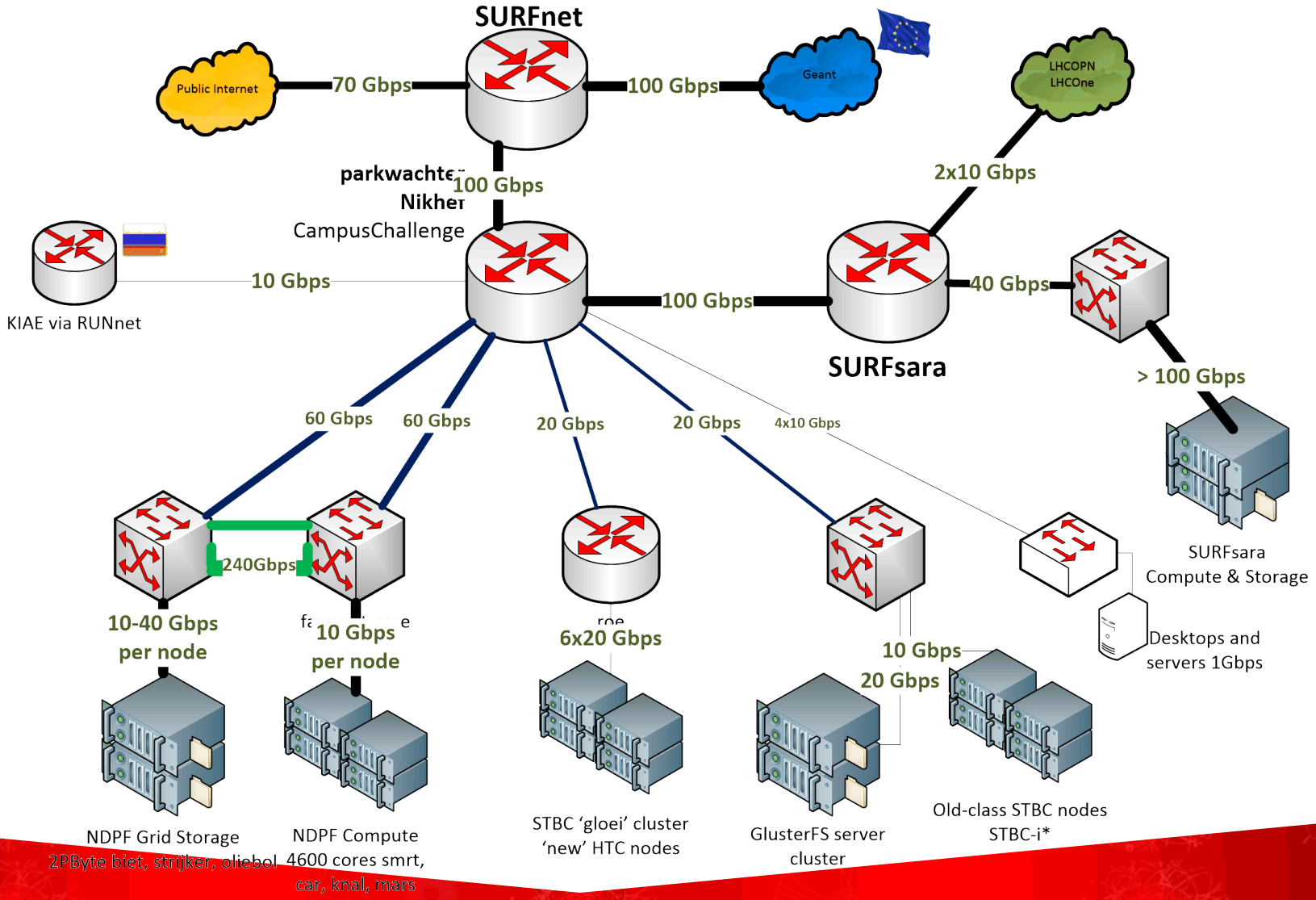
Connecting Science through Lambdas



David Groep
Nikhef
PDP - Advance
Computing for
Research

Nikhef

Nikhef and the world at the moment



David Groep
Nikhef
PDP - Advanced Computing for Research

Showing production links only – Netherlight, MSPs, and testbed not shown

Building a high throughput 'data processor' … without breaking the bank …

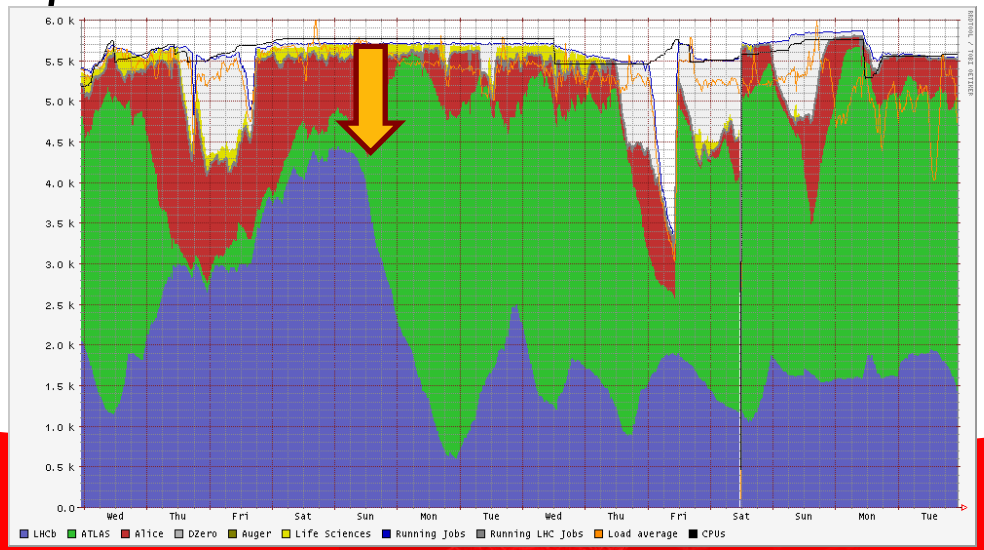
David Groep
Nikhef
PDP - Advanced Computing for Research

Nikhef

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



David Groep
Nikhef
PDP - Advanced Computing for Research

IO performance means network

- 100-400 TByte storage nodes now standard
- Processing is done on 'new' data, spread over the last ~1 PiB of storage
  so ~ 5 storage blocks serve > 16GB/s

- Realistic: up to ~ 3.2 GiB/s per node, so: ask 12 MiB/s/TiB
- It also means at least 40 Gbps NICs in each box

- And >> 160 Gbps sustained interconnect between all

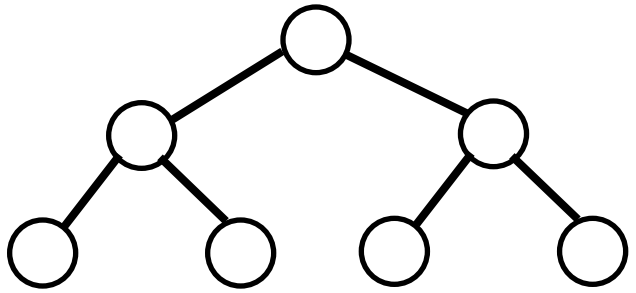
Dilemma

- 4 rekenclusters met totaal 50x10Gb/s
- 4 opslagclusters met totaal 40x10Gb/s
- Voldoende bandbreedte ertussen
- Gescheiden routing engines verplicht

- Hoe dit op te lossen?

Tristan Suerink
Nikhef
Amsterdam
PDP & Grid

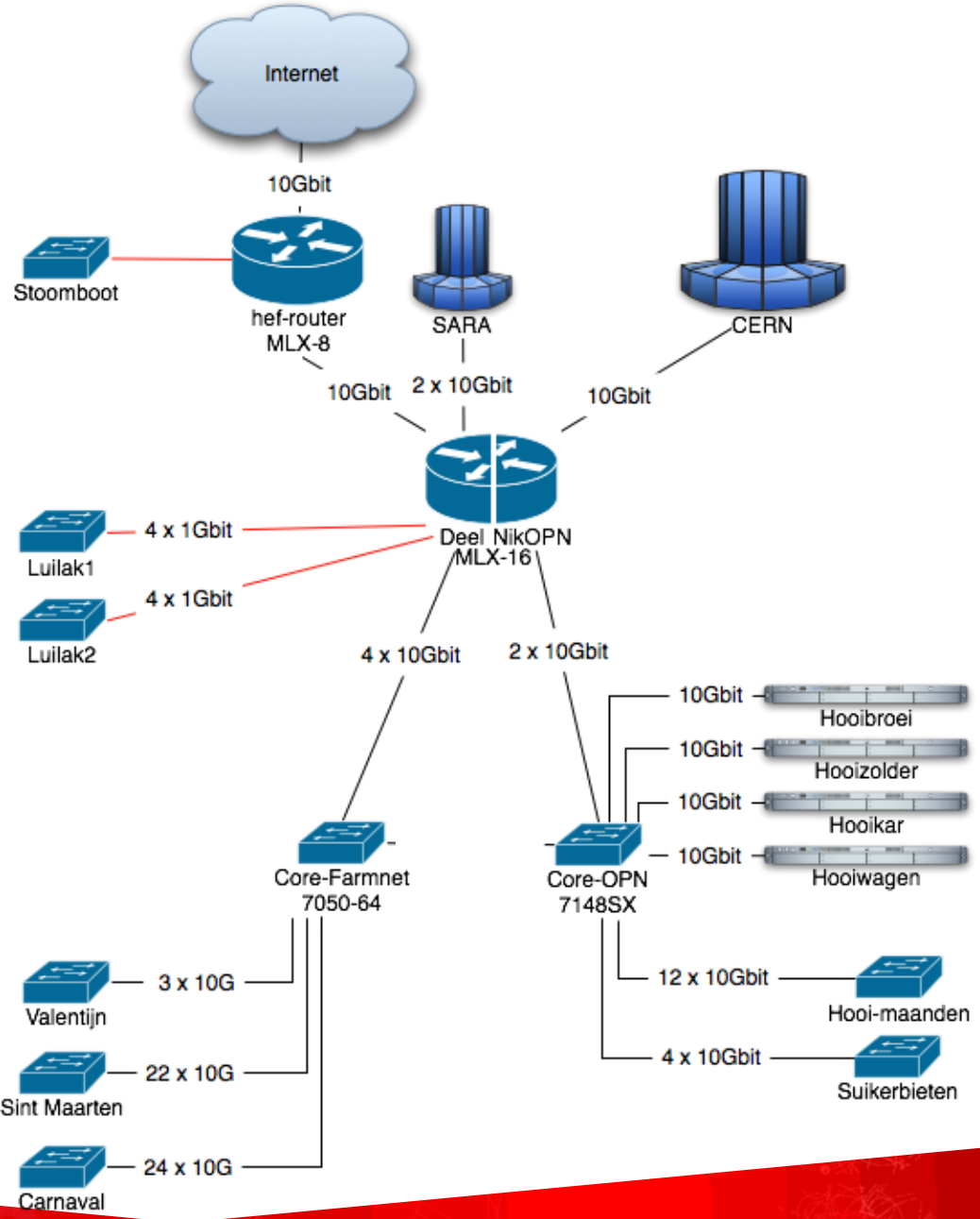
Optie #1: Bouw een boom



- Voordelen:
  - Makkelijk ontwerp
  - Makkelijk uit te breiden
- Nadelen:
  - Duurder in onderhoud
  - Veel poorten onbruikbaar
  - single point of failure

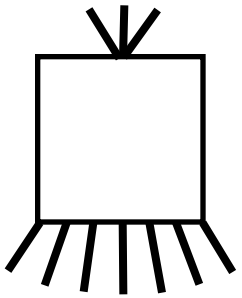
Tristan Suerink
Nikhef
Amsterdam
PDP & Grid

Traditional OPN implementation



David Groep
Nikhef
PDP - Advanced Computing for Research

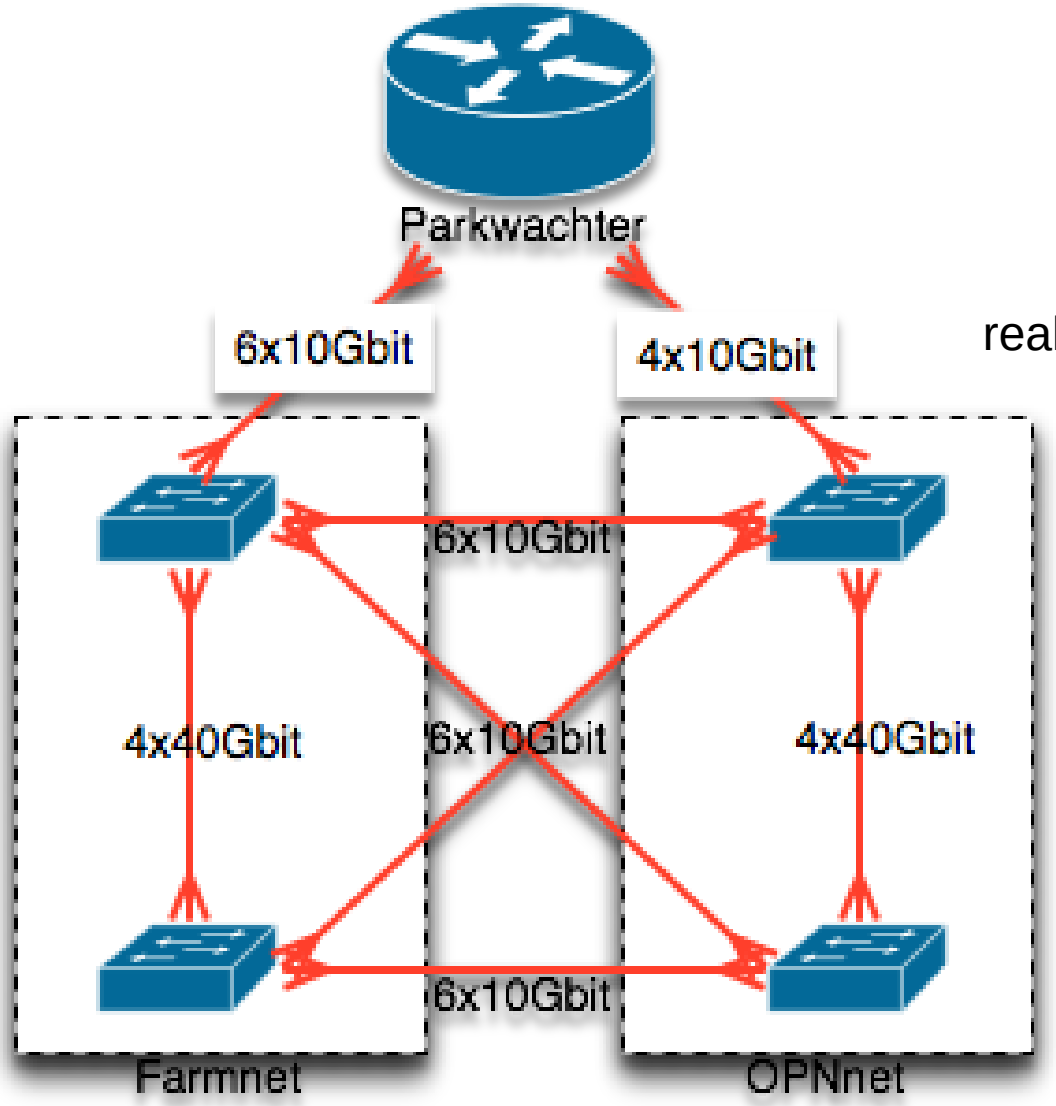
Optie #2: De monoliet



- Voordelen:
  - 1 switch chassis
- Nadelen
  - Hoge aanschaf prijs
  - Heel veel features die niet nodig zijn

Tristan Suerink
Nikhef
Amsterdam
PDP & Grid

Storage/Workernode netwerk – our choice



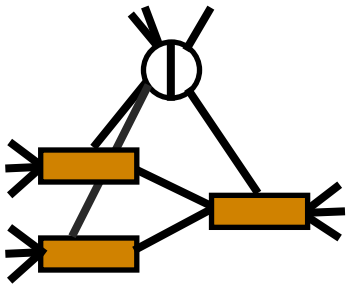
real-time re-programming of switches to follow connected topology: "DIY SDN" using switch-native python capability

In-switch reprogramming to support LHCOPN policy based routes

Tristan Suerink
Nikhef
Amsterdam
PDP & Grid

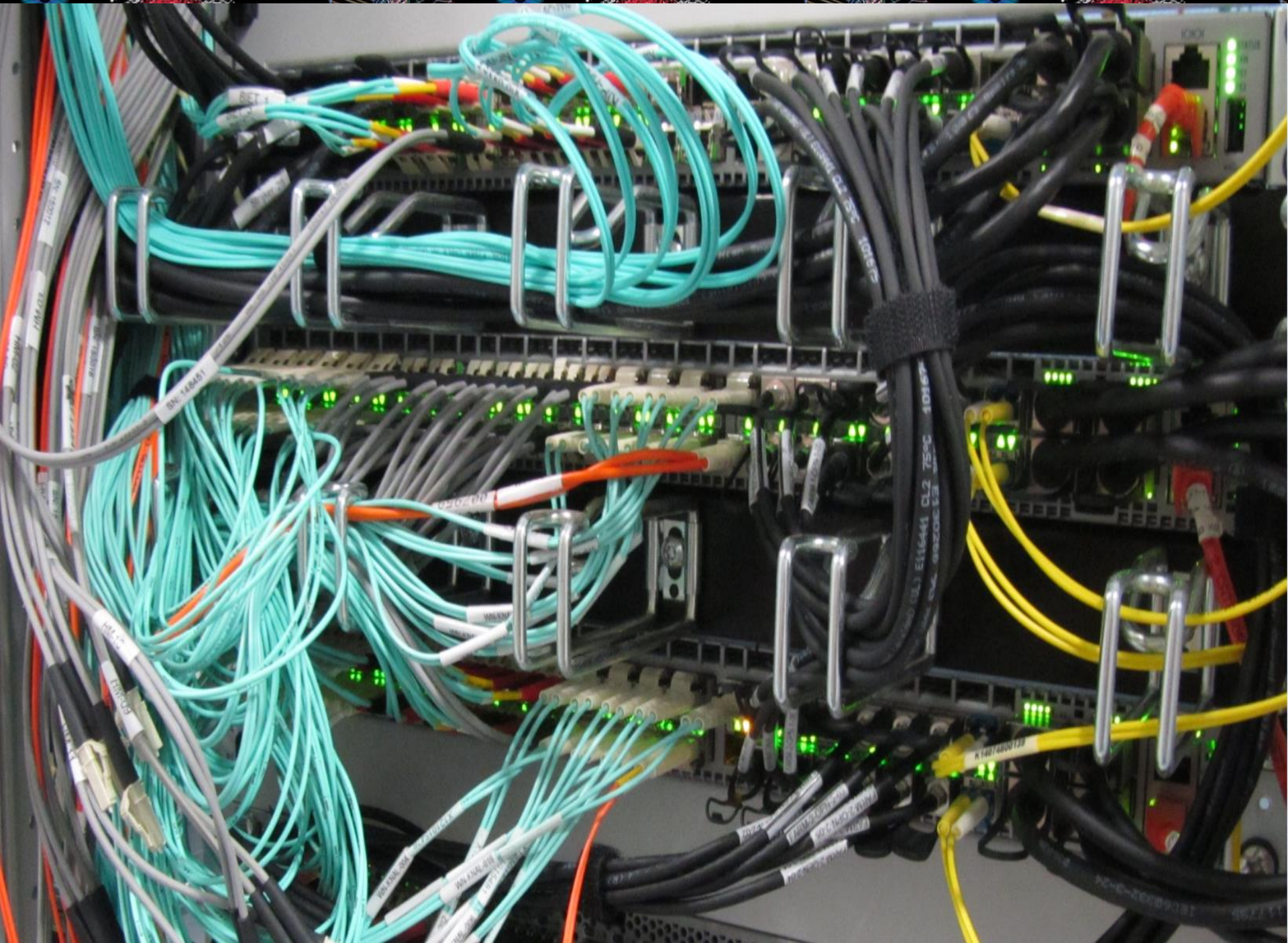
imagery courtesy Tristan Suerink

Grid netwerk



- Voordelen:
  - Meer bandbreedte
  - Alle capaciteit in gebruik
  - Redundante opstelling naar rekenclusters
  - OSPF+ECMP
  - Flexibel functionaliteiten uitbreiden
- Nadeel
  - Grote uitbreidingen (veel meer poorten) zijn complex

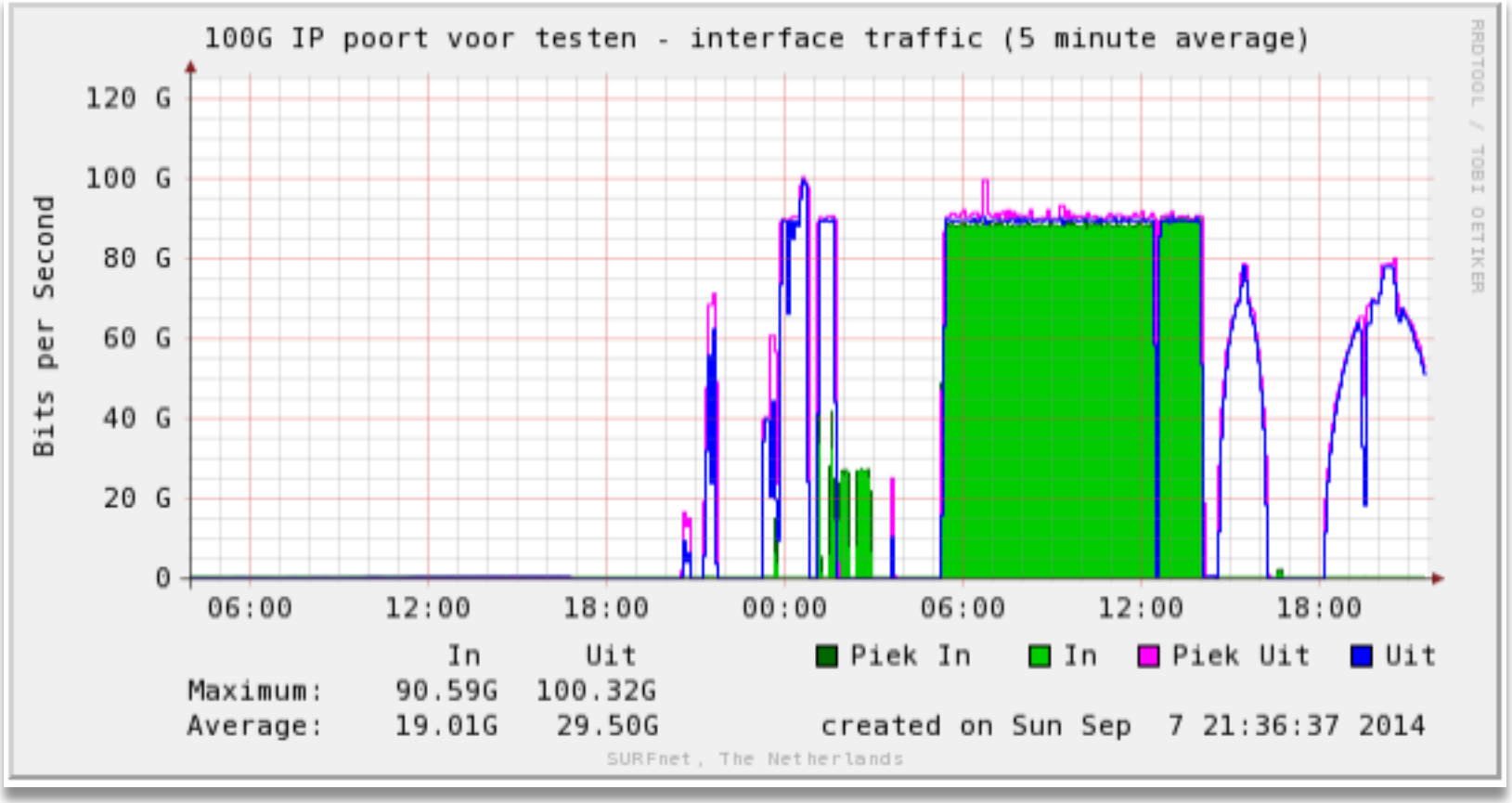
Tristan Suerink
Nikhef
Amsterdam
PDP & Grid



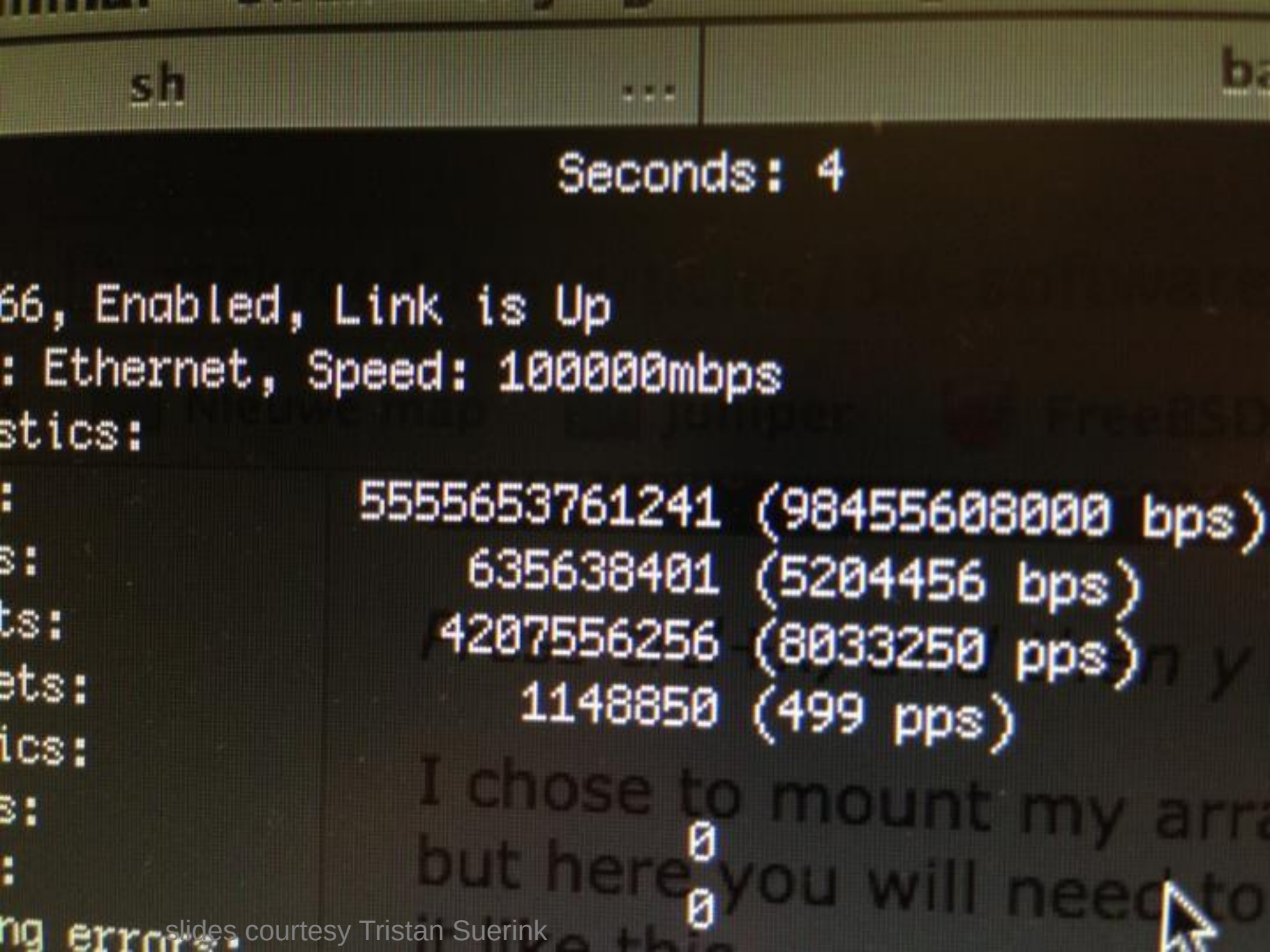
slides courtesy Tristan Suerink

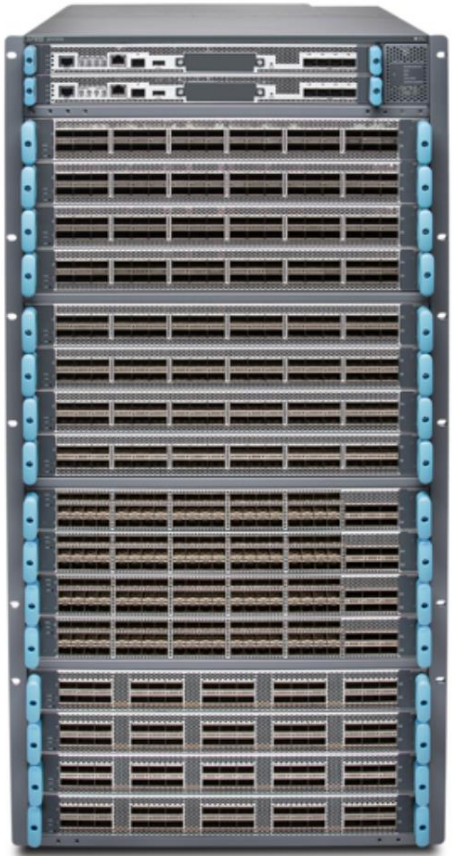
100Gbit

Nikhef → SURFnet → RUG-CIT||UvA



T Suerink
Nikhef
Amsterdam
PDP & Grid

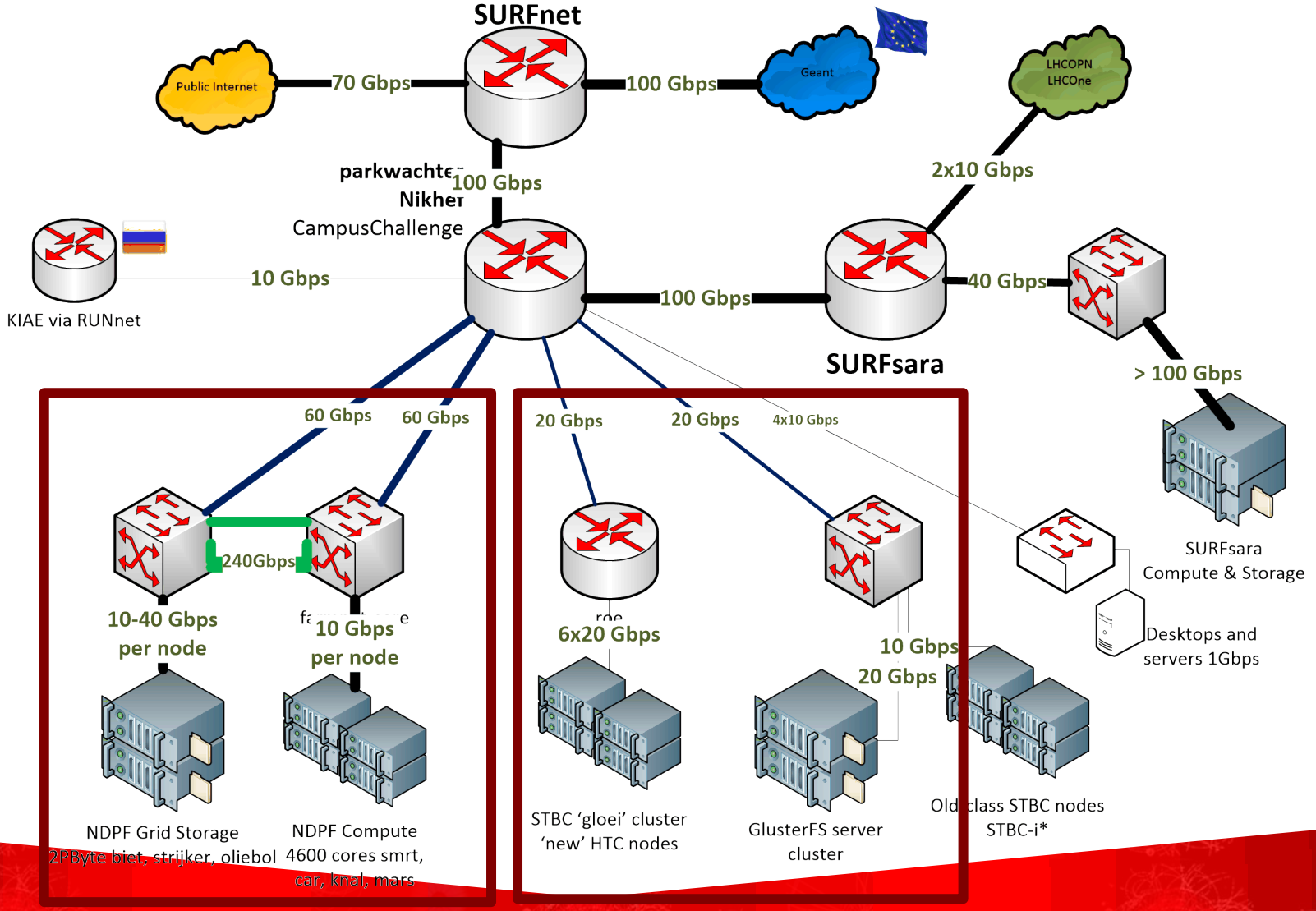

sh
Seconds: 4
66, Enabled, Link is Up
: Ethernet, Speed: 100000mbps
stics:
5555653761241 (98455608000 bps)
635638401 (5204456 bps)
4207556256 (8033250 pps)
1148850 (499 pps)
0
0
I chose to mount my arra
but here you will need to
slides courtesy Tristan Suerink



Towards a scalable high-throughput infrastructure and platform service

David Groep
Nikhef
PDP - Advanced Computing for Research

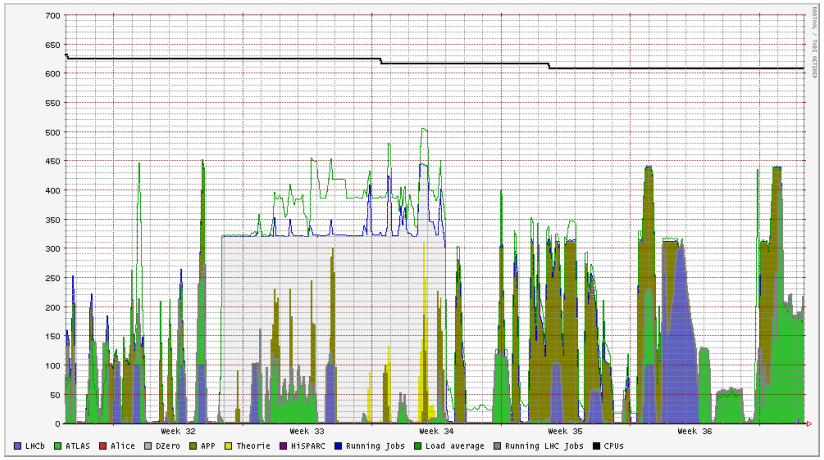
NDPF systems distribution



David Groep
Nikhef
PDP - Advanced Computing for Research

Incentives for cloudification

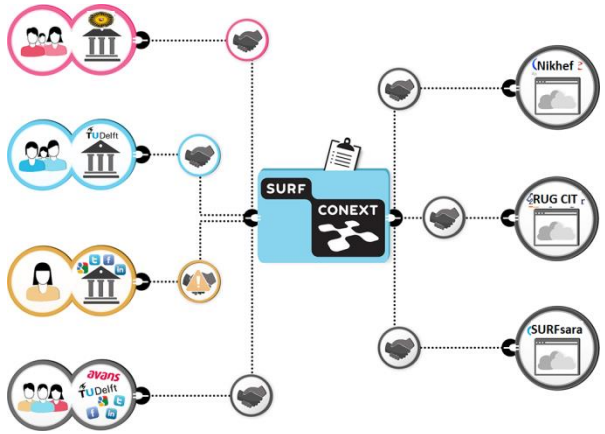
- attract more HTC use cases beyond WLCG
  these communities prefer different OS and software suites … although they still like a platform service!
- dynamic scaling between DNI nodes, ex-DNI nodes, and 'Stoomboot' computing to allow short-term bursting
- easier multi-core scheduling at >95% ocupancy



David Groep
Nikhef
PDP - Advanced Computing for Research

Requirements

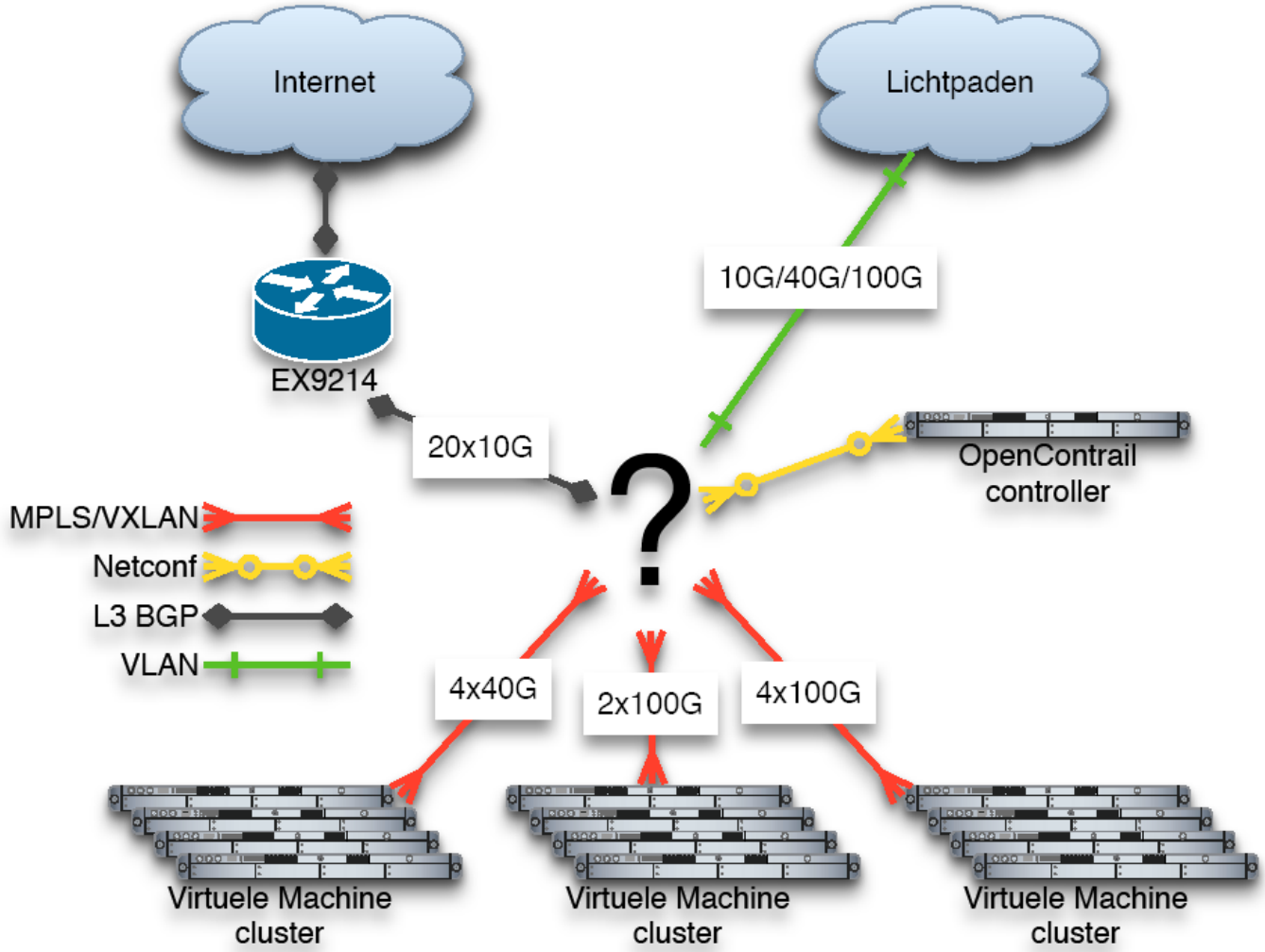
- high-bandwidth interconnect between CPU-disk >240Gbps
- true multi-tenant security & isolation
- near-native node IO performance for disk and network (say, no les than 95%) at ~400 MByte/s and 10Gbps
- public and on-demand (elastic) IPv4+v6 connectivity
- keep dynamicity in the system (resource sharing)
- permit cross-site transparent cloud bursting
- hide infrastructure differences and latency where possible *between SARA, RUG, Nikhef*



Plans for network

- Network: use host-local network endpoints and distributed packets translation
  - Packet copying is a major bottleneck in performance (NAT is not a sustainable solution due to overhead)
  - Scalable per-tenant networking (beyond ~4000 tenants) using MPLS-EVPN and SDN integrated in the hypervisor & cloud stack
  - Central termination of tenant networks in a single dedicated box (which will need MPLS-EPVN support on multiple 100G+ links)
- Leave room for some 100G or DWDM to interconnect at L2 (with MPLS?) to e.g. RUG & SARA over a light path
  - Tenant should see a single cluster – the network extended at L2 – quite contrary to e.g. the EGI Federated Cloud). Cloud bursting across the e-Infra will then be truly seamless

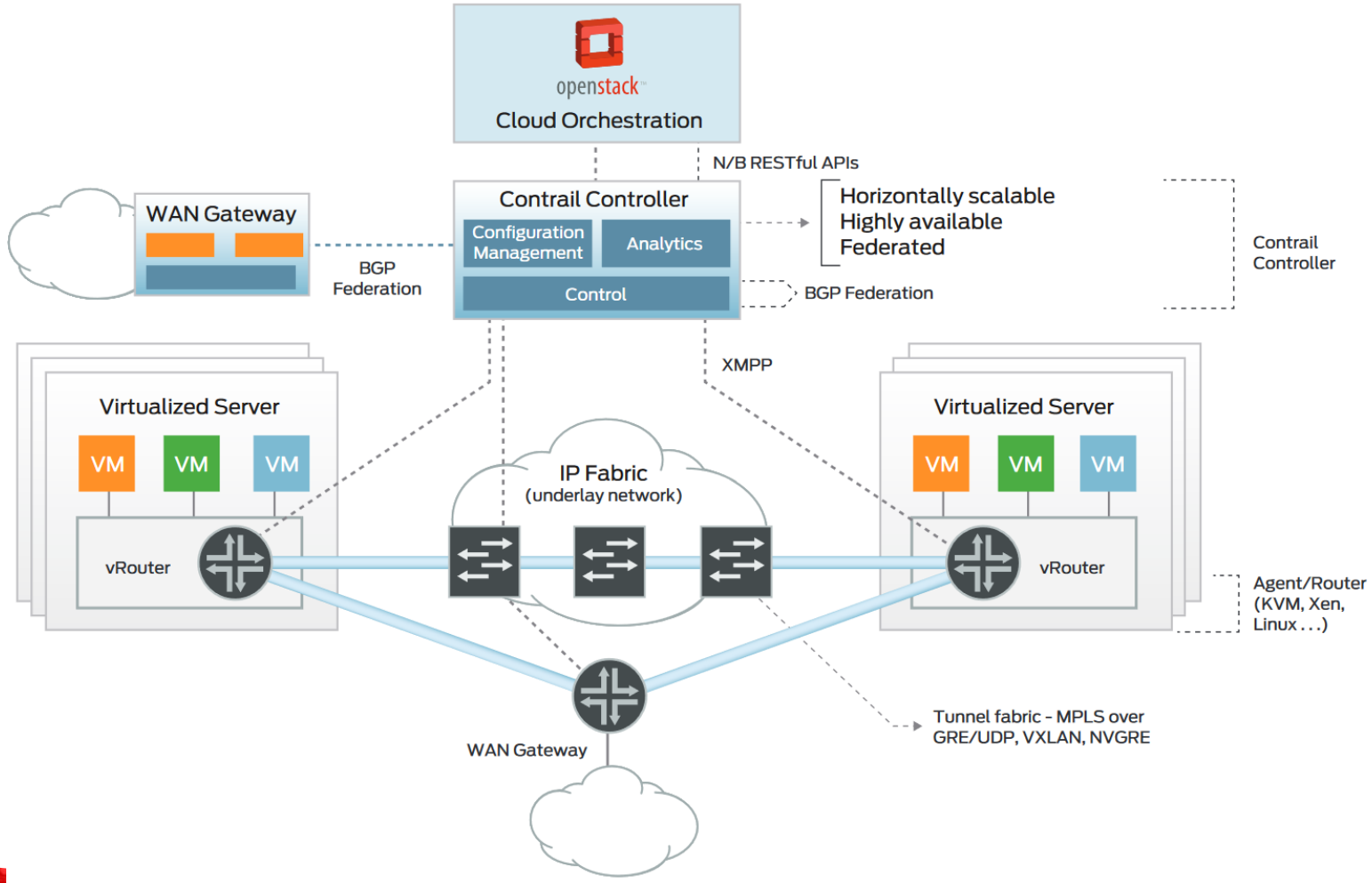
Voorbeeld


Internet
Lichtpaden
10G/40G/100G
EX9214
20x10G
?
OpenContrail
controller
MPLS/VXLAN
Netconf
L3 BGP
VLAN
4x40G
2x100G
4x100G
Virtuele Machine
cluster
Virtuele Machine
cluster
Virtuele Machine
cluster

David Groep
Nikhef
PDP - Advanced Computing for Research

David Groep
Nikhef
PDP - Advanced Computing for Research

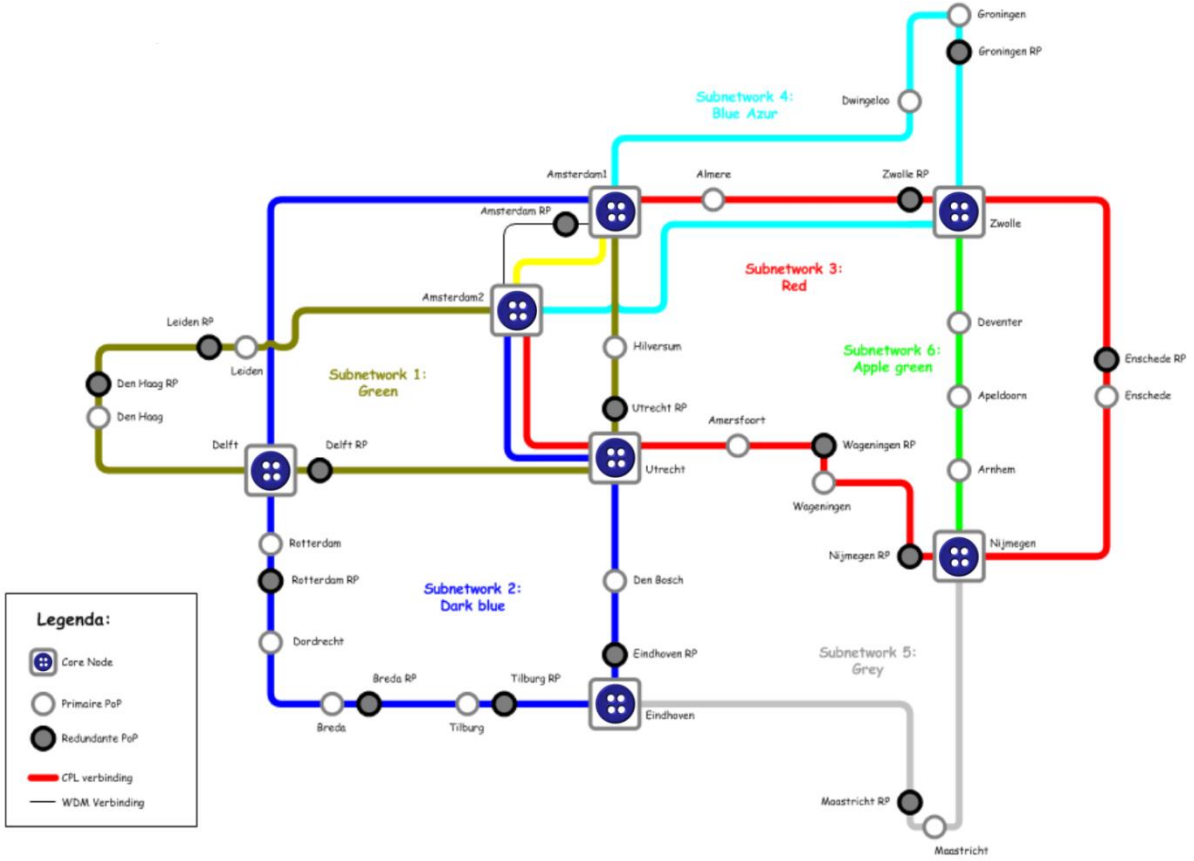
Contrail Networking – DC to WAN



David Groep
Nikhef
PDP - Advanced Computing for Research

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

L2 cloud bursting: connecting services with MSPs and WDM



David Groep
Nikhef
PDP - Advanced Computing for Research

Extending the MPLS fabric across SURFnet MSPs, Netherlight, or Alien Waves

Graphic: SURFnet 7 DWM, SURFnet

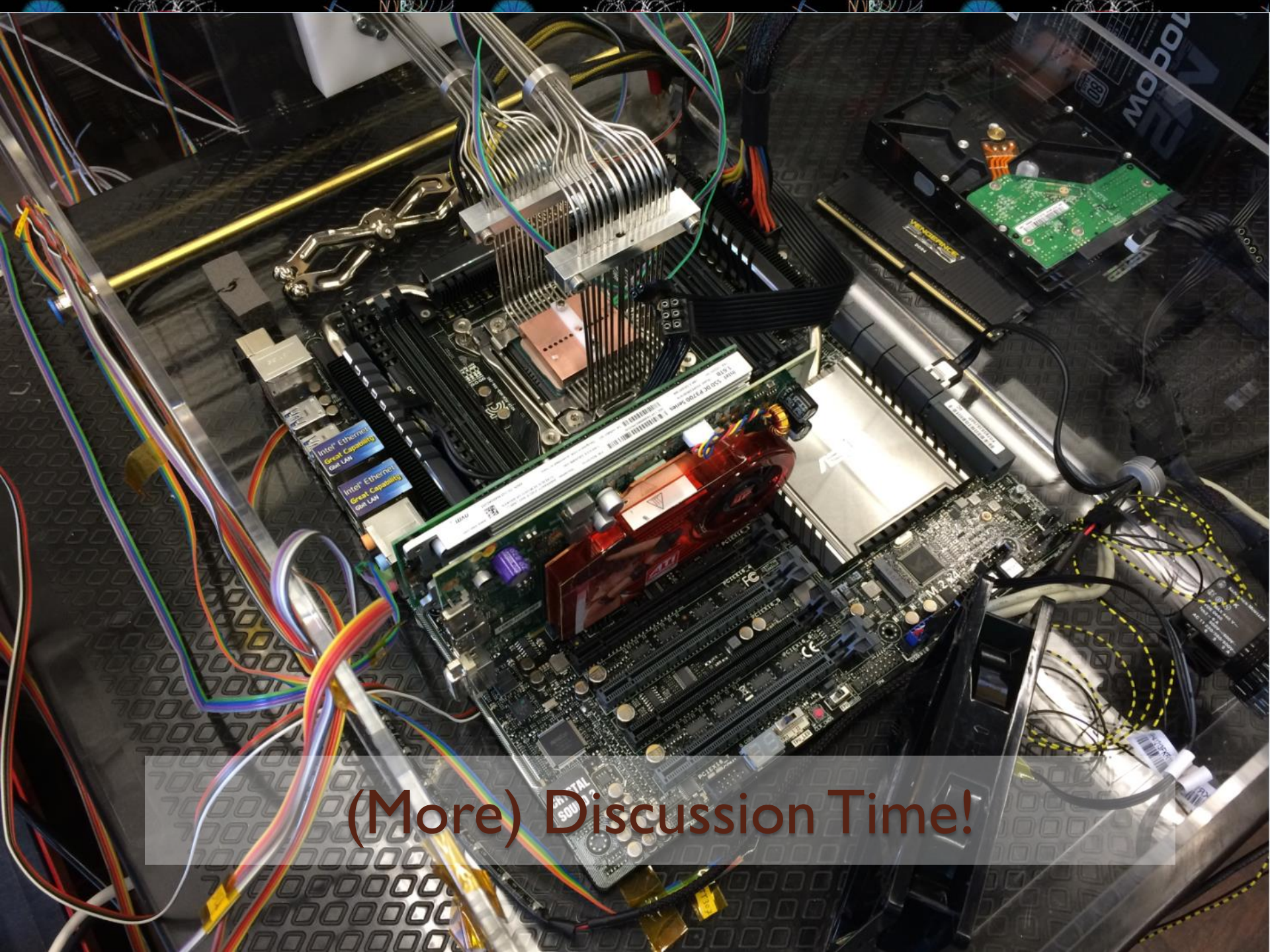
‘NiKloud’ –
a DNI service in coordination with SURF

- Hybride cluster, opslag en netwerk omgeving
- IP Fabric
- Overlay op basis van VXLAN/MPLS
- 10/25Gbit aansluiting per rekennode
- 40/50Gbit aansluiting per opslagnode
- meerdere 100Gbit per cluster; en multi-Tbit/s basisnetwerk
- Hardware offloading d.m.v. DPDK op de rekennodes
- ‘Helicopter’ aansturing via OpenContrail (NFV)
- Stricte isolatie van tenants - maar onbeperkte connectivity
- ‘Gebruiker krijgt de macht’

David Groep
Nikhef
PDP - Advanced Computing for Research

SURF CONEXT

Nikhef


(More) Discussion Time!