“Fancy” ;-) Networking

Tristan Suerink

IT Architect

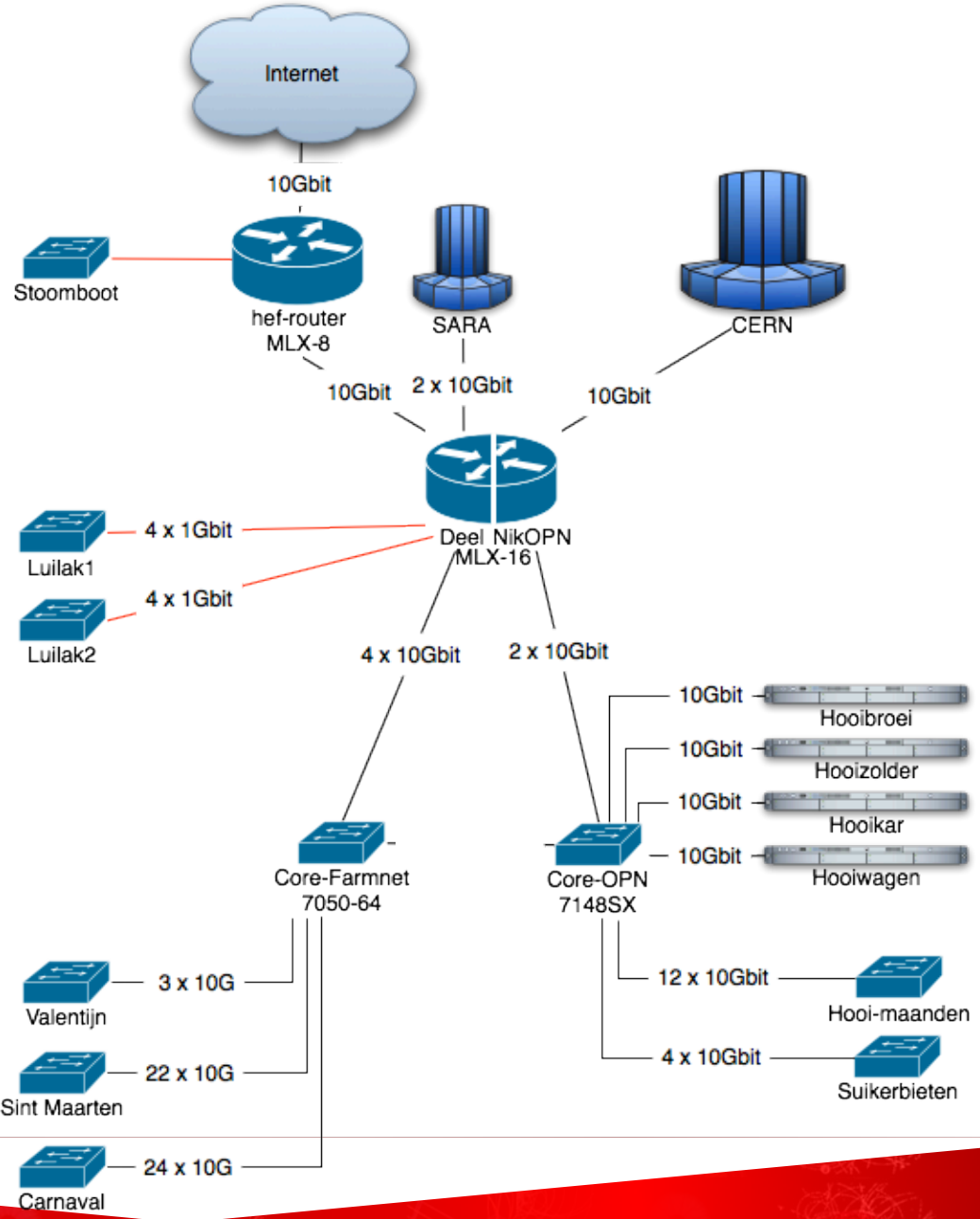
Nikhef

Nationaal instituut voor subatomaire fysica

We needed to do something

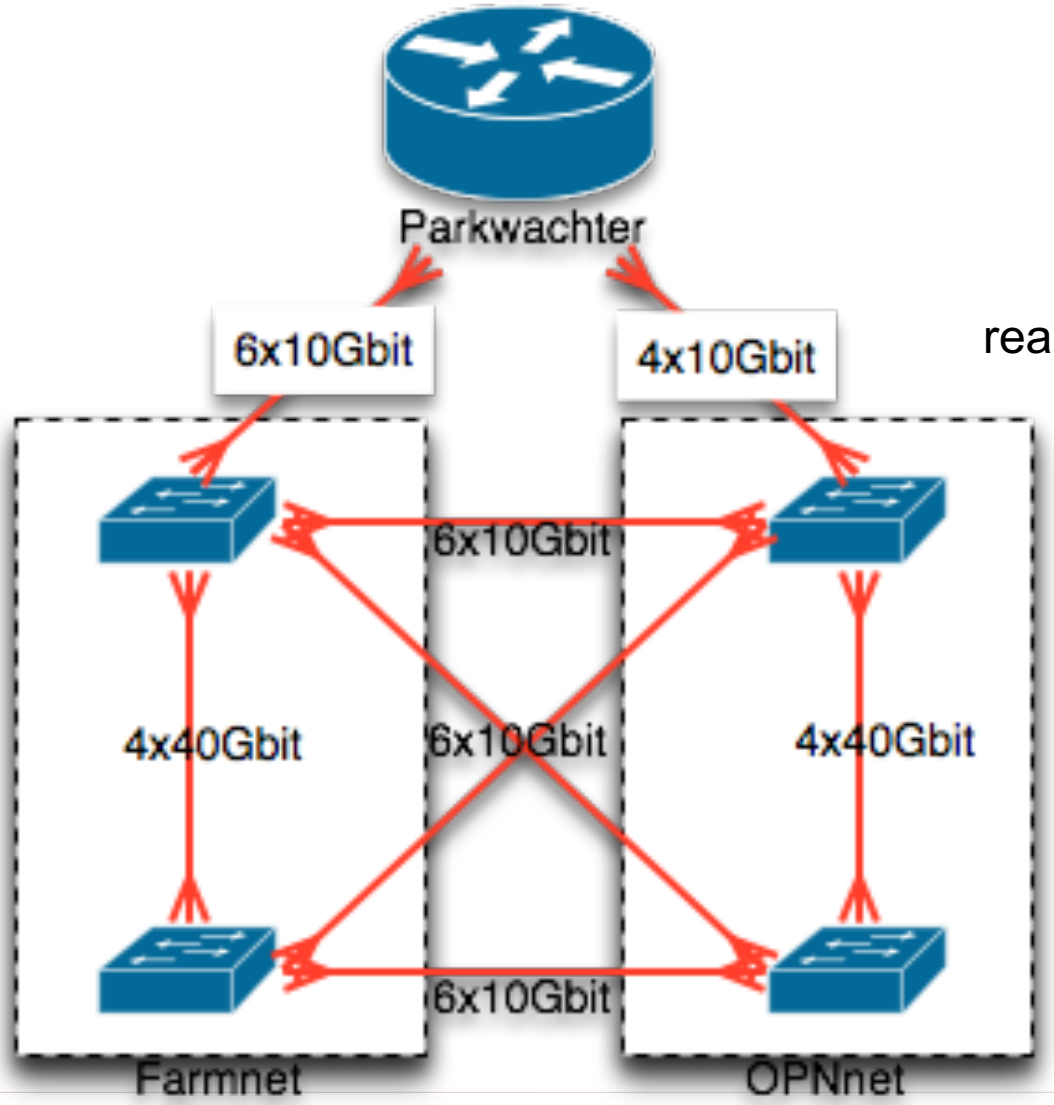
- Previous network designed in 2009
- In 2012 upgraded the network
- We've reached the physical limits of the design
- No support for new technologies
- Building an HTC Cloud environment
- Time to replace the equipment
- Investigate long distance network technology

Traditional OPN implementation



Slide courtesy of David Groep Nikhef

Storage/Worker node network – our choice



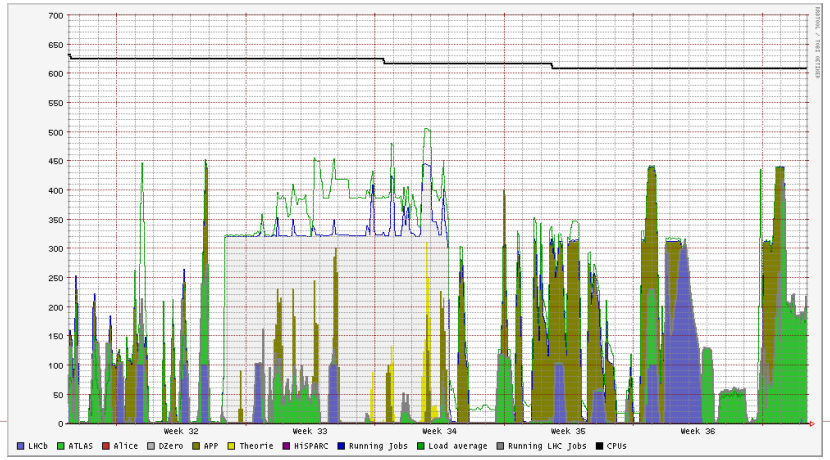
real-time re-programming of switches to follow connected topology: "DIY SDN" using switch-native python capability

In-switch reprogramming to support LHCOPN policy based routes

Slide courtesy of David Groep Nikhef

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 GRID nodes, ex-GRID nodes, and local computing to allow short-term bursting
- easier multi-core scheduling at >95% occupancy

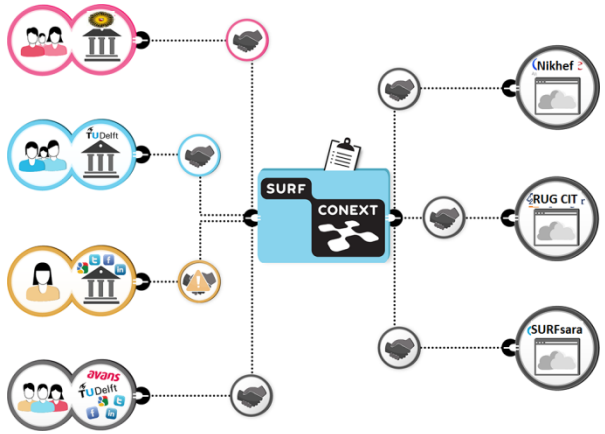


Slide courtesy
of David Groep
Nikhef

Nikhef

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*



Slide courtesy of David Groep Nikhef

Network design requirements

- Lots of 100Gbit/s ports
- 400Gbit/s ready per port
- Chassis based (8 slots)
- Deliverable in 2016
- Support for:
  - MPLS over UDP/GRE
  - L3VPN
  - EVPN
  - OpenContrail
  - VRFs with route-leaking
  - VXLAN (as nice to have)

Possible candidates

- Arista 7500R
- Brocade SLX
- Juniper QFX10000

Arista 7500R

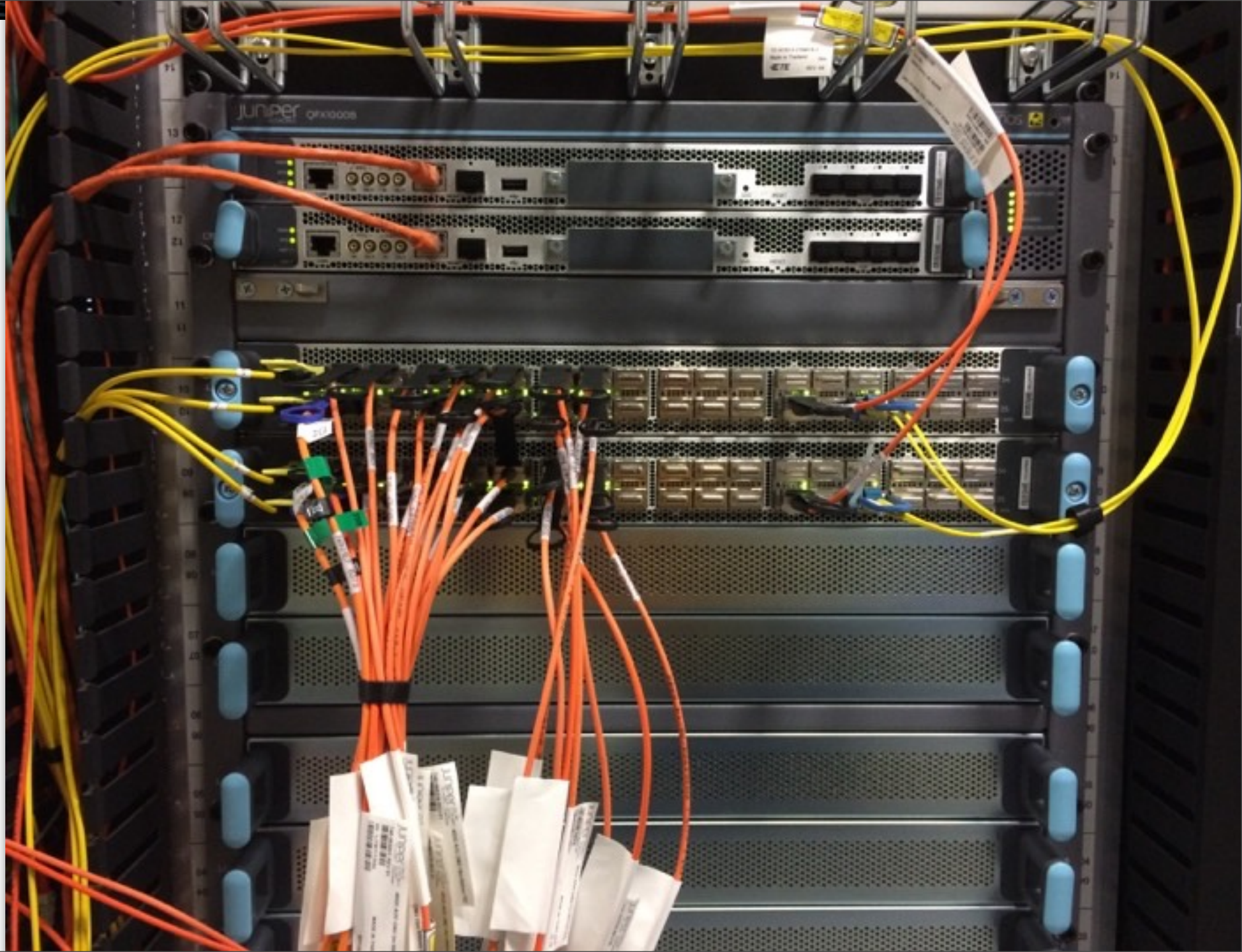
- Pros:
  - One image for all Arista switches
  - Easy to configure
- Cons:
  - Very expensive
  - No real MPLS features
  - Very limited VRF features
  - Extremely small ACL table

Brocade SLX

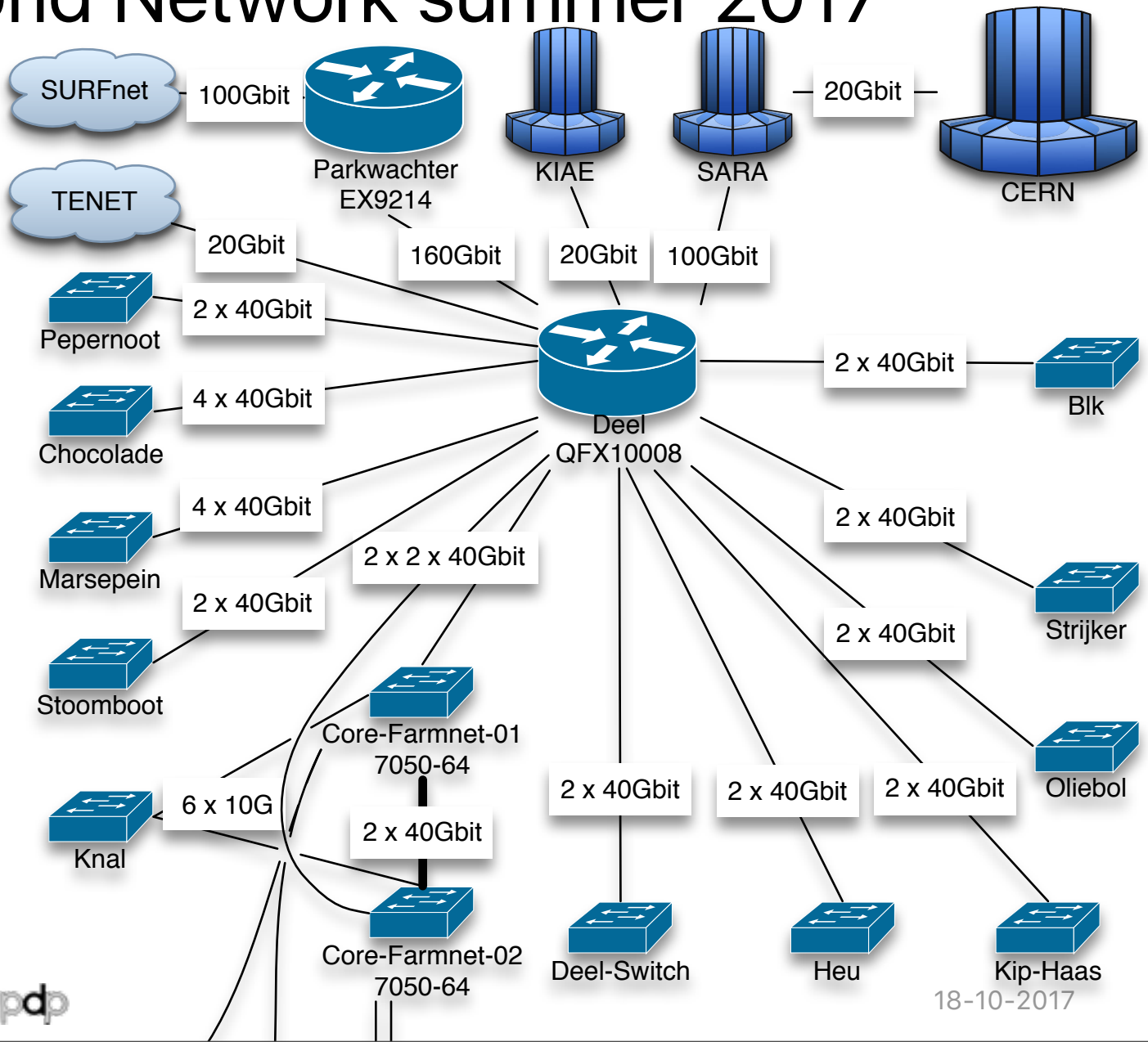
- Pros:
  - Not a pure Broadcom HW platform (more flexible)
  - Complete refreshed software (compared with MLX)
- Cons:
  - Very expensive
  - Focus was on L2 and L2.5 at that time
  - Missing too many features at that time
  - Too late for us

Juniper QFX10000

- Pros:
  - Juniper's own very flexible ASIC
  - Running JunOS
  - Available since 2015
  - Big tables for L2, L3 and ACL's
- Cons:
  - Less dense than the other two at the moment
  - Boot time could be faster


juniper

Grid Network summer 2017

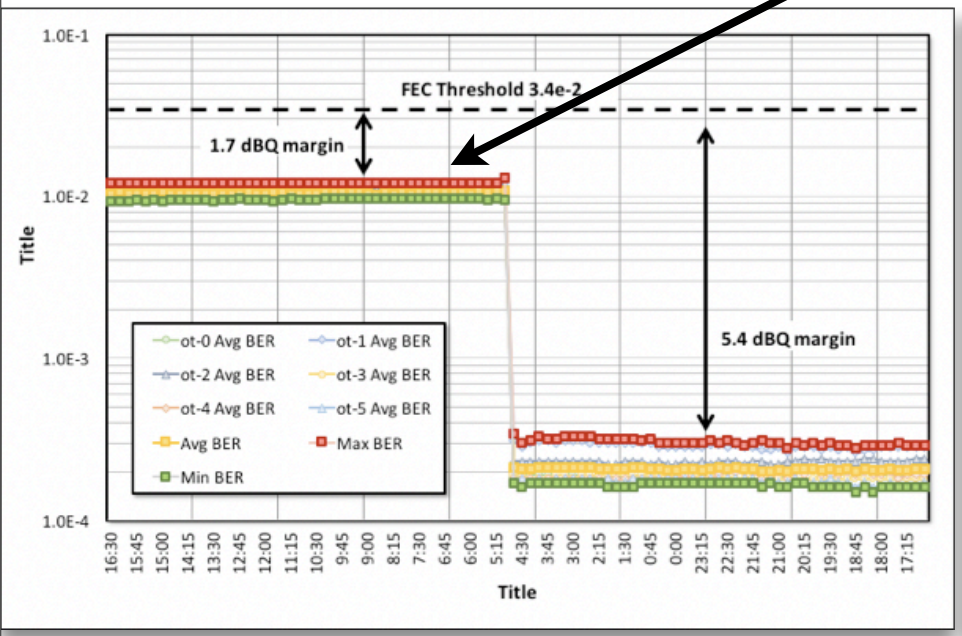


Long distance DWDM test

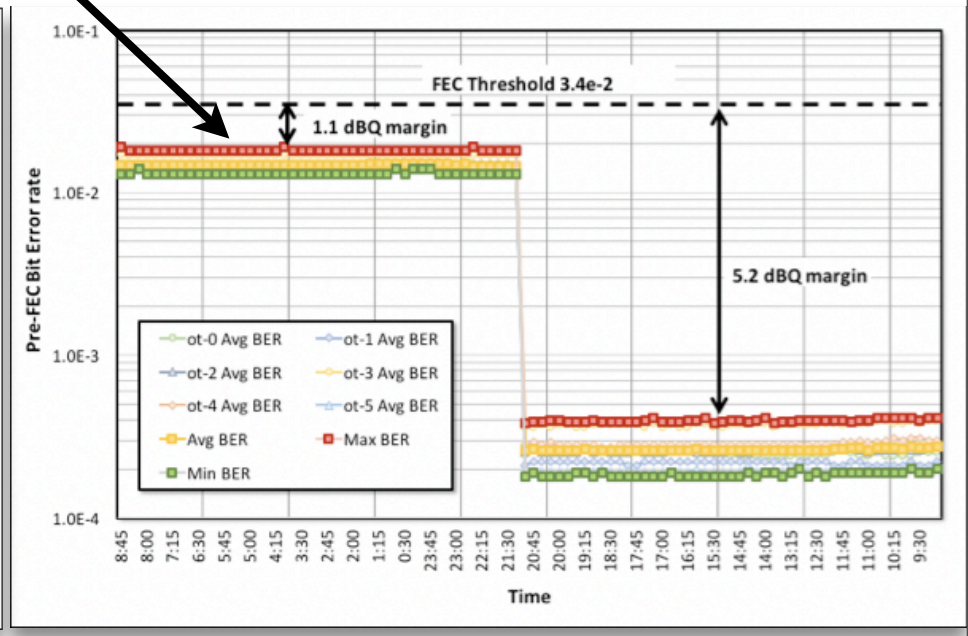
- Between Amsterdam and Geneva
- Experimental DWDM equipment from Juniper
- 1618KM of fiber from SURFnet
- Using 6 wavelengths
- QPSK (100G), 8QAM (150G) and 16QAM (200G)
- From March until May 2017

Difference between QPSK and 8QAM

This is the 6x150Gbit/s



Amsterdam > Geneva



Geneva > Amsterdam

Things to know

- Long distance DWDM isn't trivial
- Really clean your fibers! And double check them!
- We've missed ±3dB for 16QAM
- Up to 4000KM reach using QPSK
- The cards have the same functionality as the rest
- Separate configuration for DWDM and Ethernet side
- 8QAM mode combines 2 front ports
- The ethernet side works like multiple 100G's

Questions?

- Couldn't do the DWDM tests without the help from:

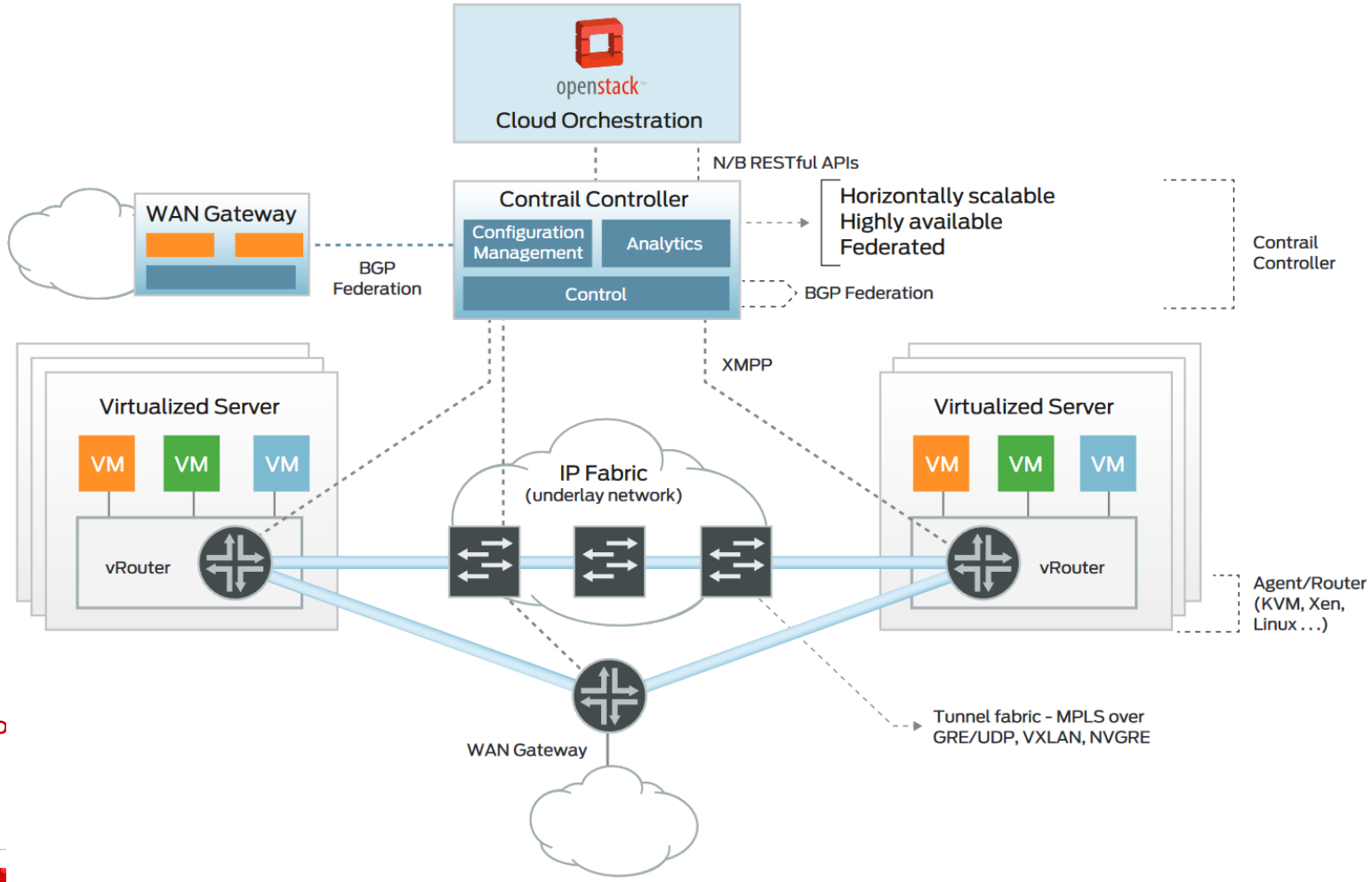
- CERN: Eduardo and John
- SURFnet: Rob, Marcel, Pieter and Lucas
- Juniper: Dirk, Vincent, Washid and Roberto
- NIKHEF: Erwin, David, Dennis and Floris

- Thank you all!

Backup slide

- We want to be flexible with our resources
- Keep our high speed interconnect
- Tenant cloud based networks
- Stateless networking
- Office enclave integrated with HTC
- Technology shift within the market
- Overlay networking into the hypervisor
- Using standard network technology
- ScienceDMZ is not enough
- Neutron and Openflow doesn’t work in production

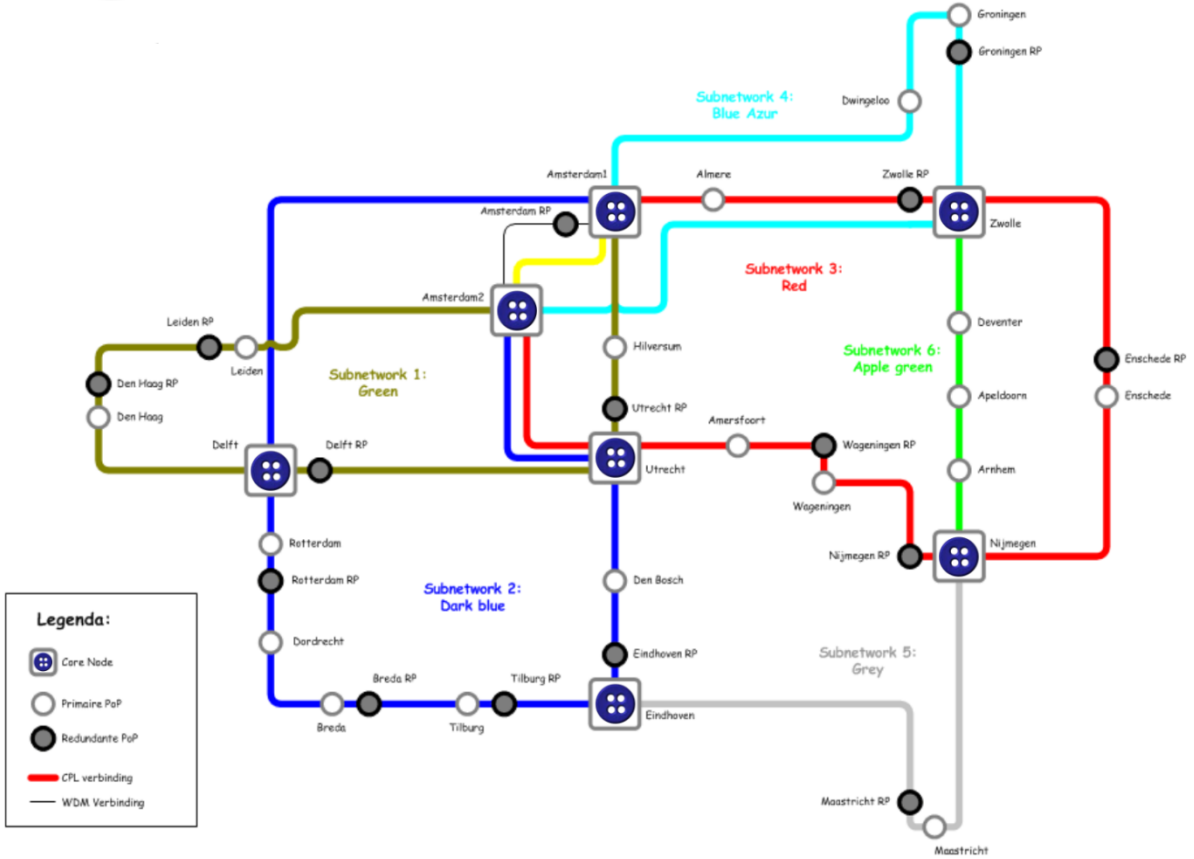
Contrail Networking – DC to WAN



Slide courtesy of David Groep Nikhef

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

L2 cloud bursting: connecting services with MSPs and WDM



Slide courtesy of David Groep Nikhef

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

‘NiKloud’ –
a DNI service in coordination with SURF

- Hybrid cluster, storage and network omgeving
- IP Fabric
- Overlay using VXLAN/MPLS
- 10/25Gbit connection per worker node
- 40/50Gbit connection per storage node
- multiple100Gbit per cluster; and multi-Tbit/s basenetwork
- Hardware offloading d.m.v. DPDK on the worker nodes
- ‘Helicopter’ control via OpenContrail (NFV)
- Strict isolation of tenants – but unlimited connectivity
- ‘The power to the user’

Slide courtesy of David Groep Nikhef

SURF CONEXT

Nikhef