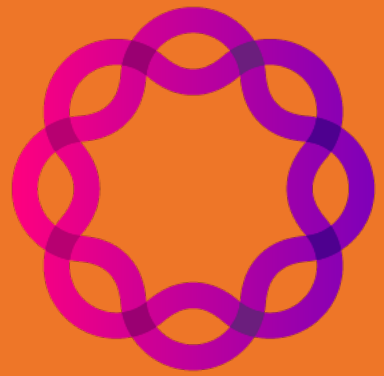
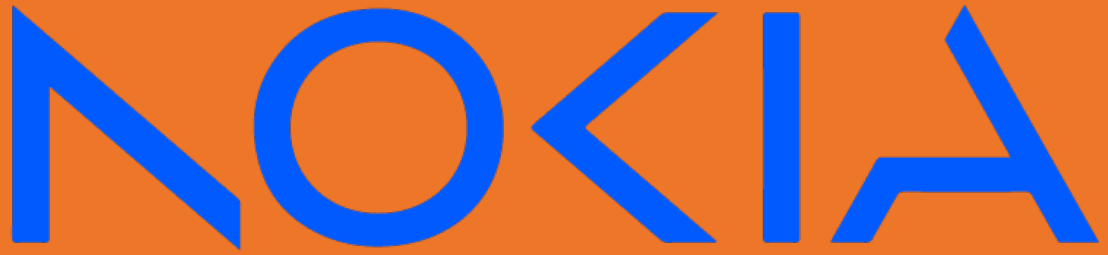




800G for LHC: physics beat ethernet

RIPE88

Joachim Opdenakker
Network Engineer
SURF

21-05-2024

Who am I?

Joachim Opdenakker

- Network engineer at SURF (Dutch NREN) since aug. 2022
- Specialised in:
  - Peering
  - BGP and BGP analysis
  - Service development
  - Architecture

SURF?

- Dutch National Research & Education network
- National network of apporx. 14000 km fiber and 2000 km internationally
- ~470 backbone routers
- International PoP’s in London, Brussels, Hamburg, Geneva
- Involved with LHC networks
- Developing the Workflow Orchestrator for vendor Agnostic SDN
- Consortium of trans-atlantic research network capacity

SURF

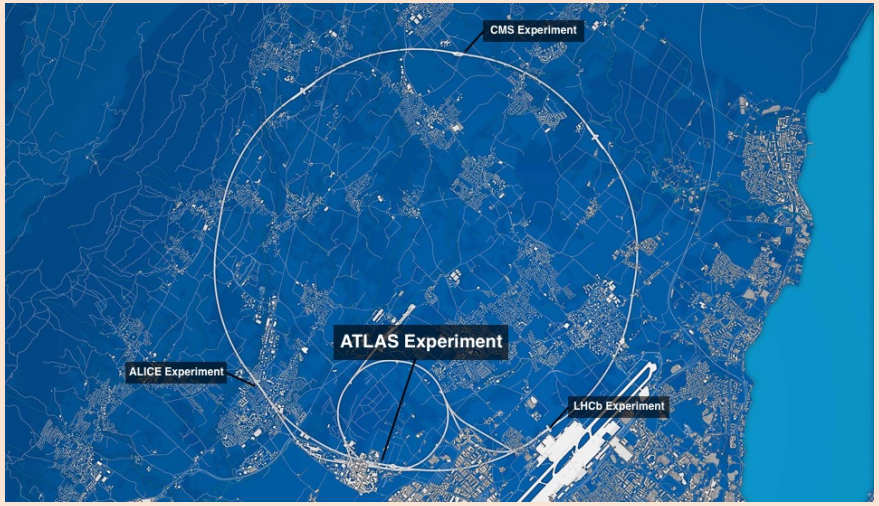
Why 800Gbps testing?

LHC

World’s largest and highest-energy particle collider. Located across the border of France and Switzerland.

Experiments

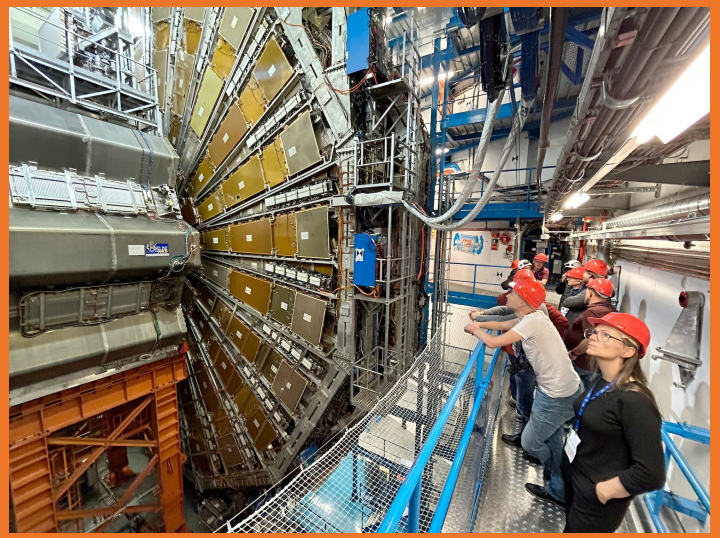
Alice, LHCb, ATLAS and CMS

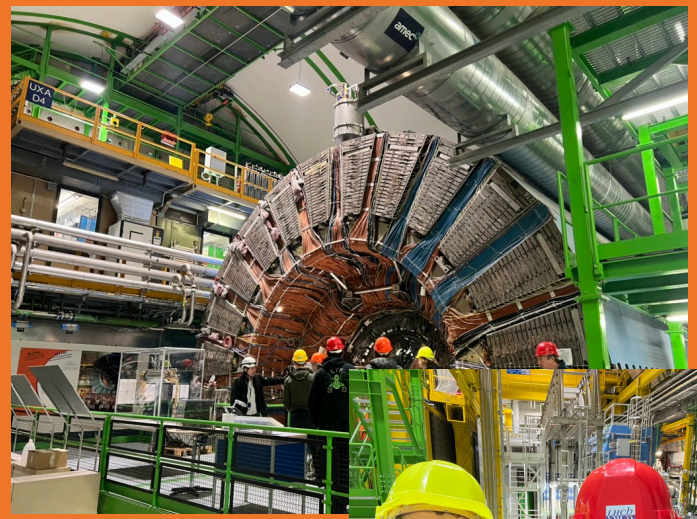


SURF

Experiments

ATLAS





LHCb

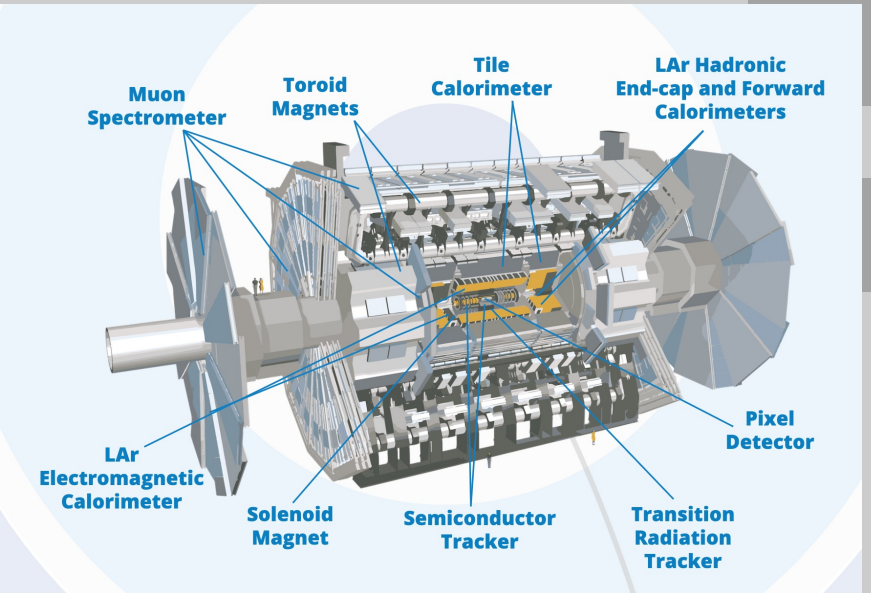
Alice

SURF

ATLAS detector

Event rates:

- At a beam luminosity of 10 34 cm-2 s-1, there will be about 20 collisions per bunch crossing.
- 40 million bunch crossings per second.
- Yields about 1 billion collisions per second.
- Level 1 trigger filters that down to about 75 000 events per second.
- Level 2 trigger reduces it to about 2 000 events per second.
- The Event Filter then selects for permanent storage about 200 “interesting” events per second.



Trigger and Data Acquisition (TDAQ)

TDAQ has a 3 level Trigger system (reduction in three steps).
Total event reduction factor by the trigger system: 200 000.

- 1st level trigger: Hardware, level 1 is done using special-purpose processors.[1]
- 2nd level trigger: Software, large computing farms with ~ 500 dual pc processors.
- 3rd level trigger: Software, large computing farms with ~ 1700 dual pc processors.

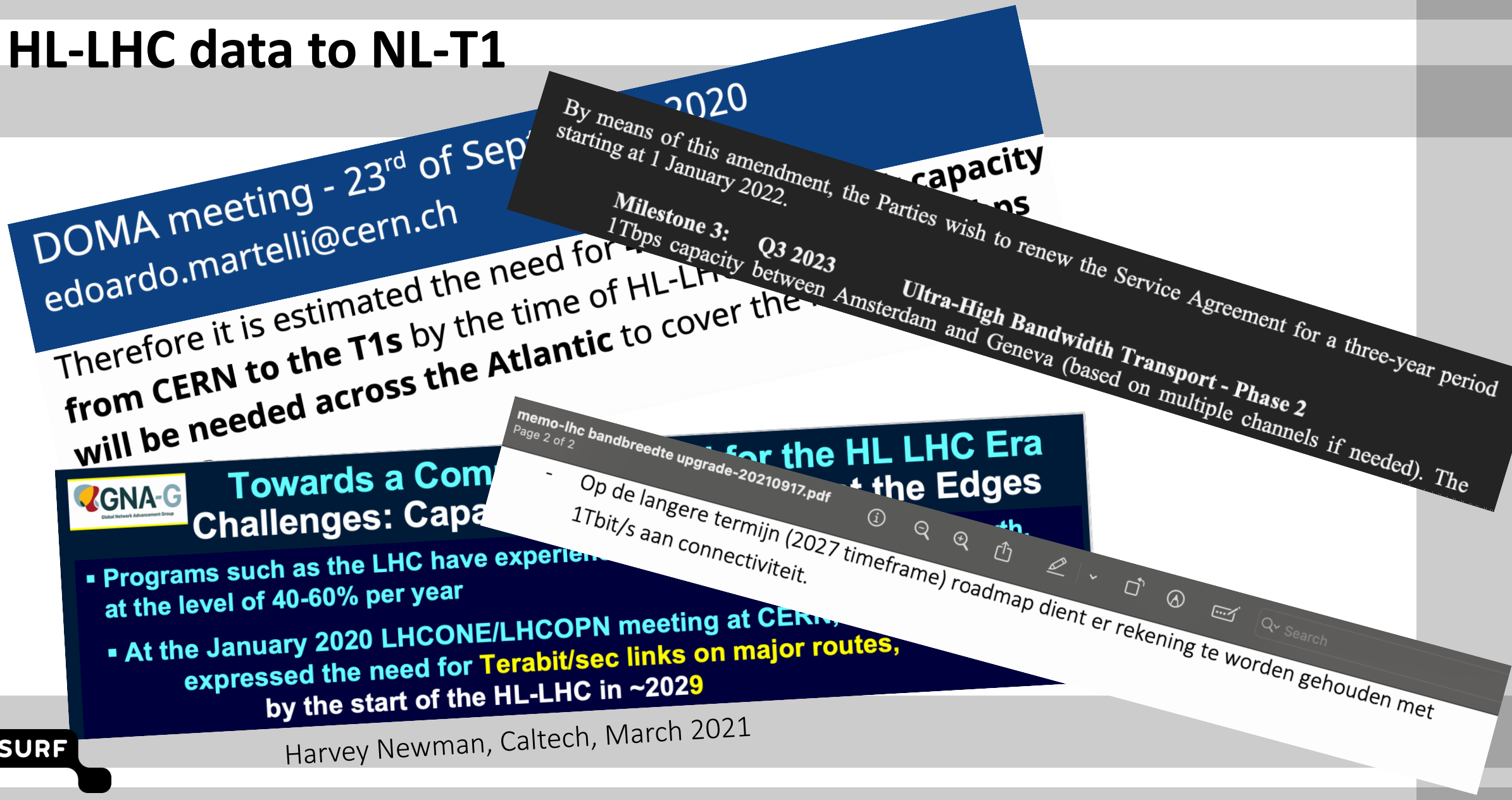
The rates and reduction factors at 14 TeV are summarized as:

	Incoming event rate per second	Outgoing event rate per second	Reduction factor
Level 1	40 000 000	100 000	400
Level 2	100 000	3 000	30
Level 3	3 000	200	15

TDAQ records 320 Mbytes per second, which would fill more than 27 CDs per minute.

From the ATLAS fact sheet (pre HL-LHC)

HL-LHC data to NL-T1



Harvey Newman, Caltech, March 2021

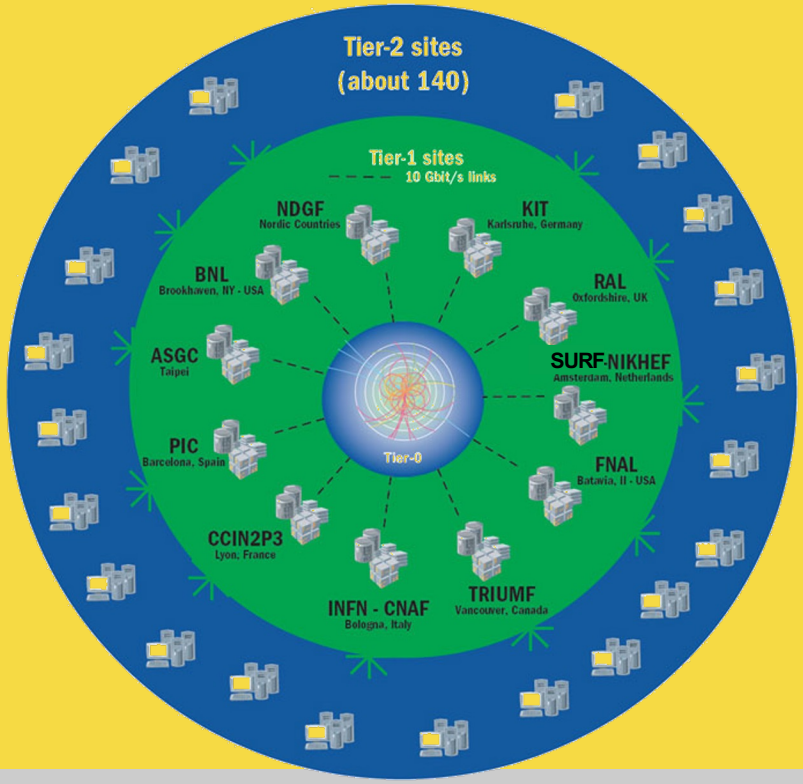
SURF

Data transport: tiered

Tiered storage locations

need to be connected

CERN is T0 storage site



Data transport: T0 -> T1: LHCOPN

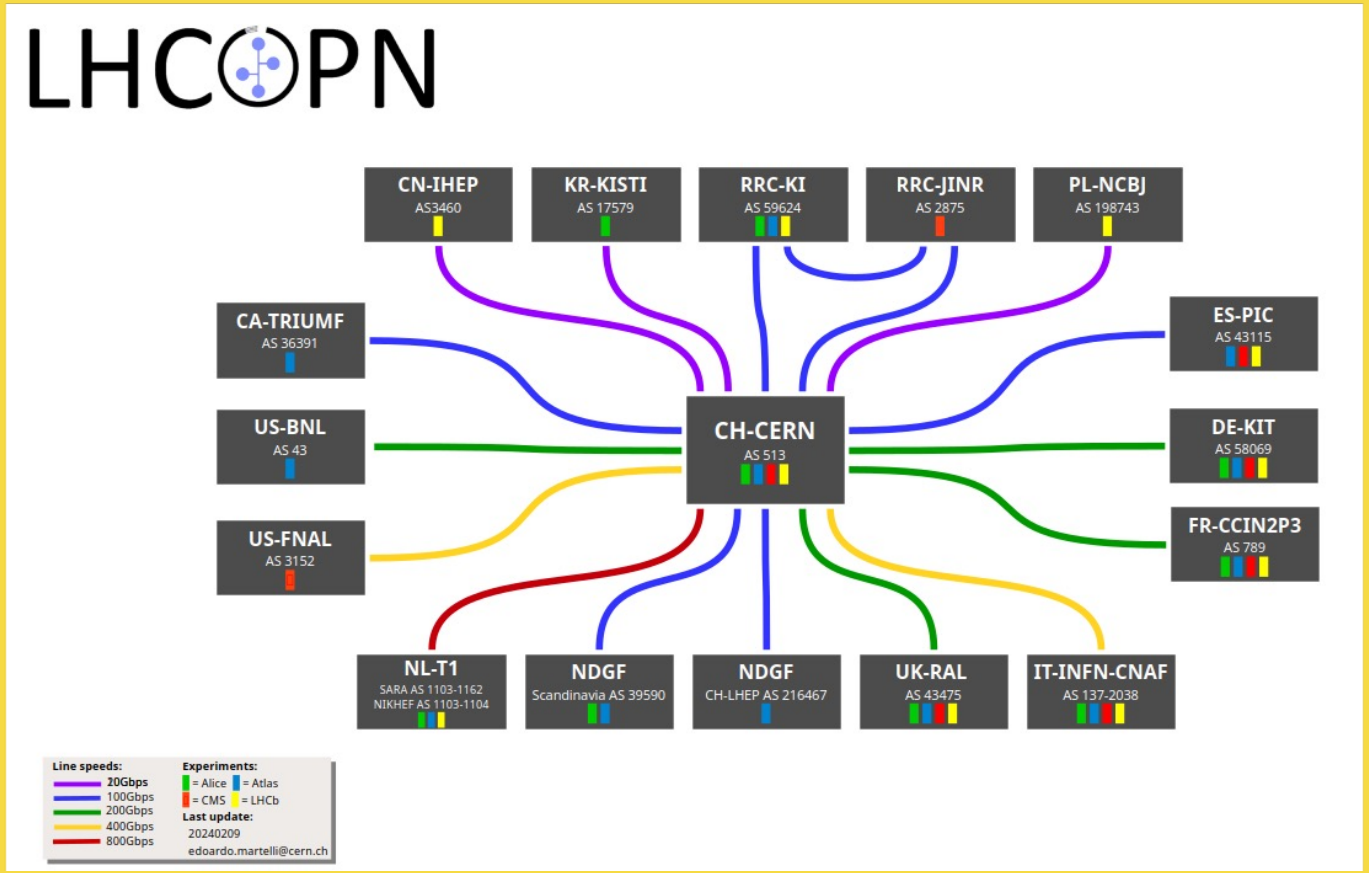
LHCOPN – Large Hadron Collider Optical Private Network

Private network between T0 and T1 storage sites.

Netherlands has a shared T1 storage facility:

- SURF (formerly SURFsara)
- Nikhef (Dutch National Institute

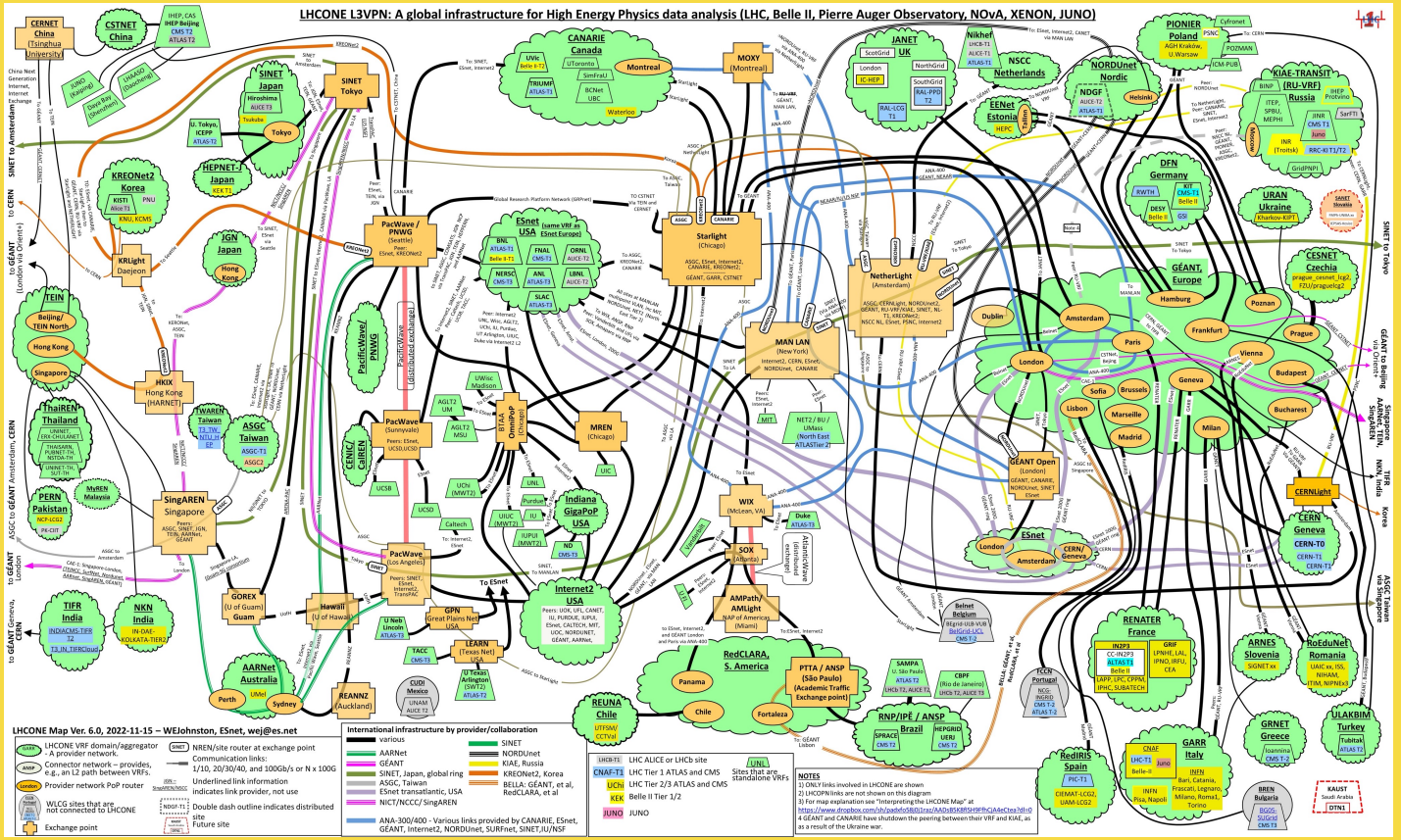
for Subatomic Physics)



Data transportation: T1 -> T2: LHCONE

LHCONE – Large Hadron Collider Open Network Environment

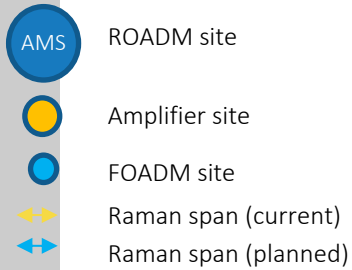
Private network between T1 and T2 storage sites.



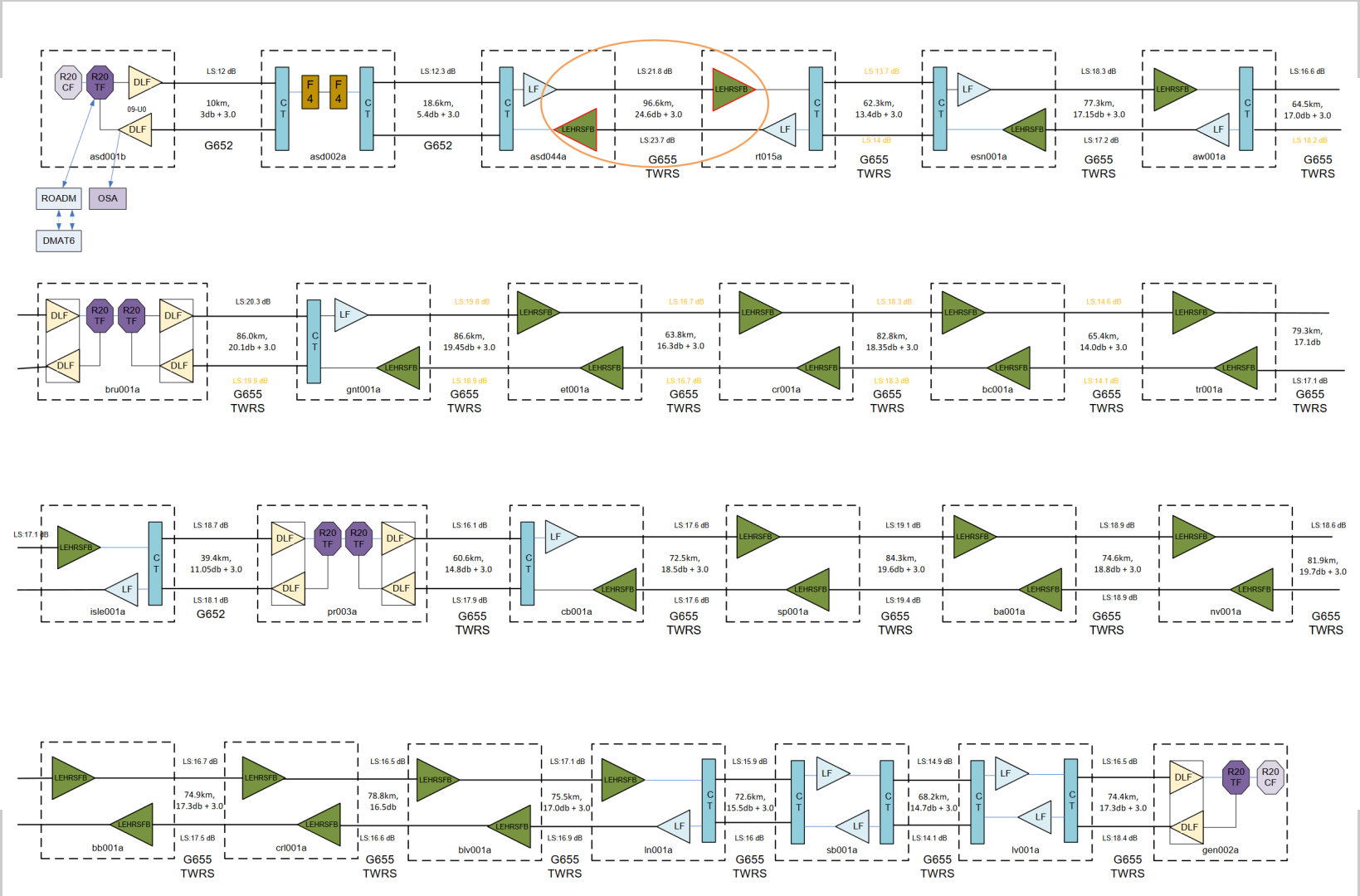
SURF's line system on a map

Amsterdam – Geneva

- Total fiber distance (one-way) is 1648 km.



SURF's line system between Amsterdam and Geneva



Preparing…

Q1/Q2 2023 - New software release for NMS and amplifiers

- Minor improvement of signal quality

Q3 2023 – RAMAN upgrades

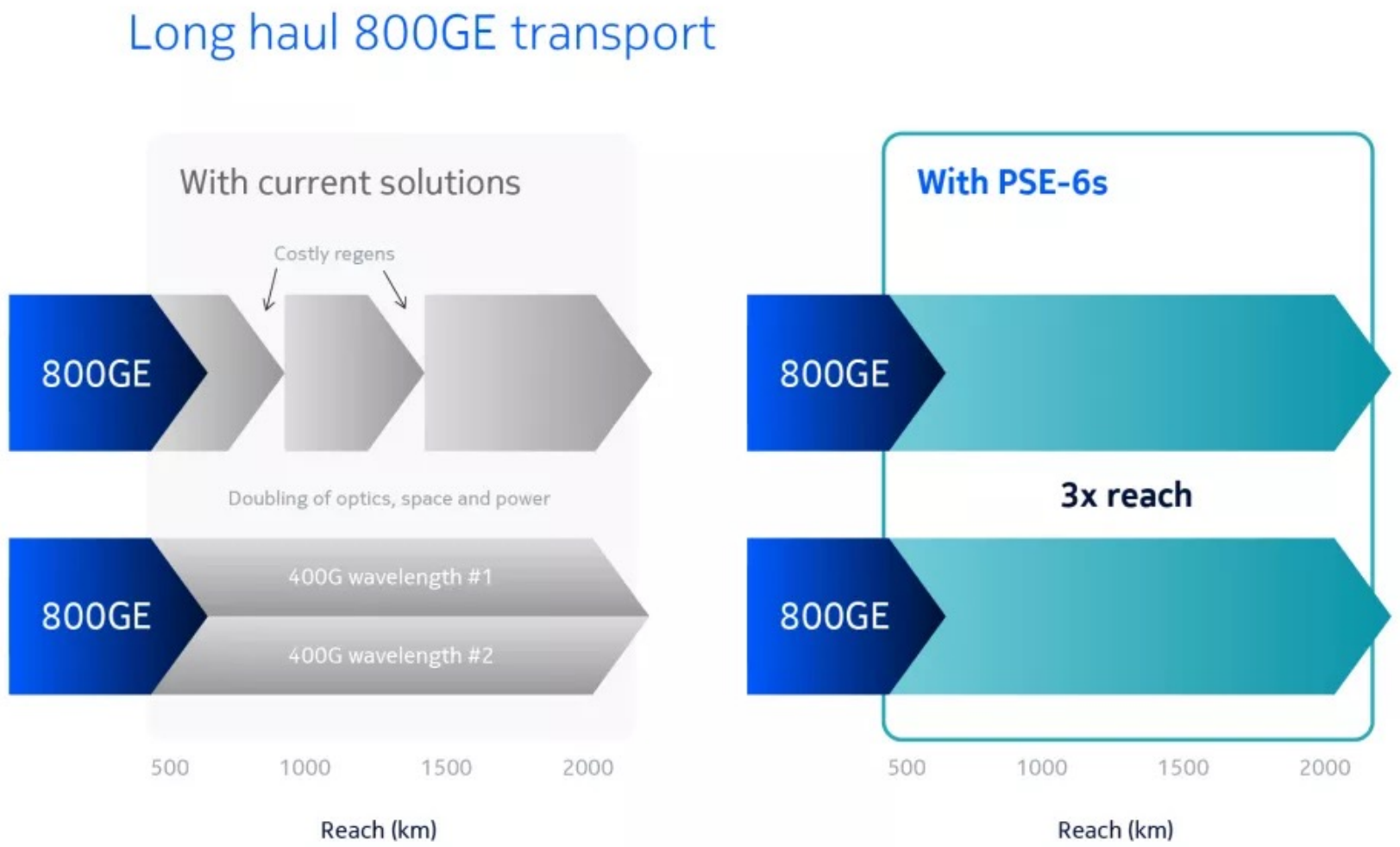
- Replacing Raman for Raman with VOA and optimise OSC in scope, reducing tilt.

Q4 2023 – RAMAN installations

- Replacing EDFA type for Raman on some spans, improve link performance.

Week 8 / 2024 – Trial with Nokia transponder and IP equipment

Hardware used: optical transport (Nokia)

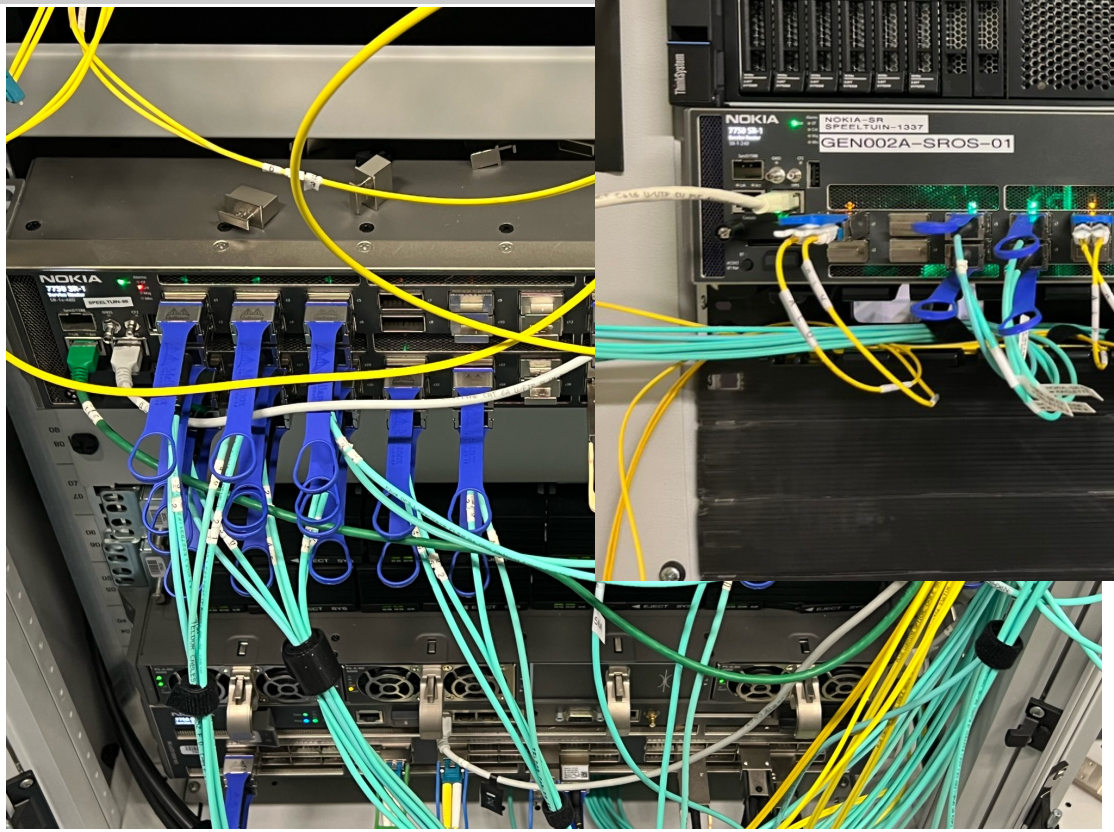


Hardware used: optical transport (Ribbon)

- Ribbon Apollo 96xx chassis
- Roadm’s
- Amplifiers
- Transponders

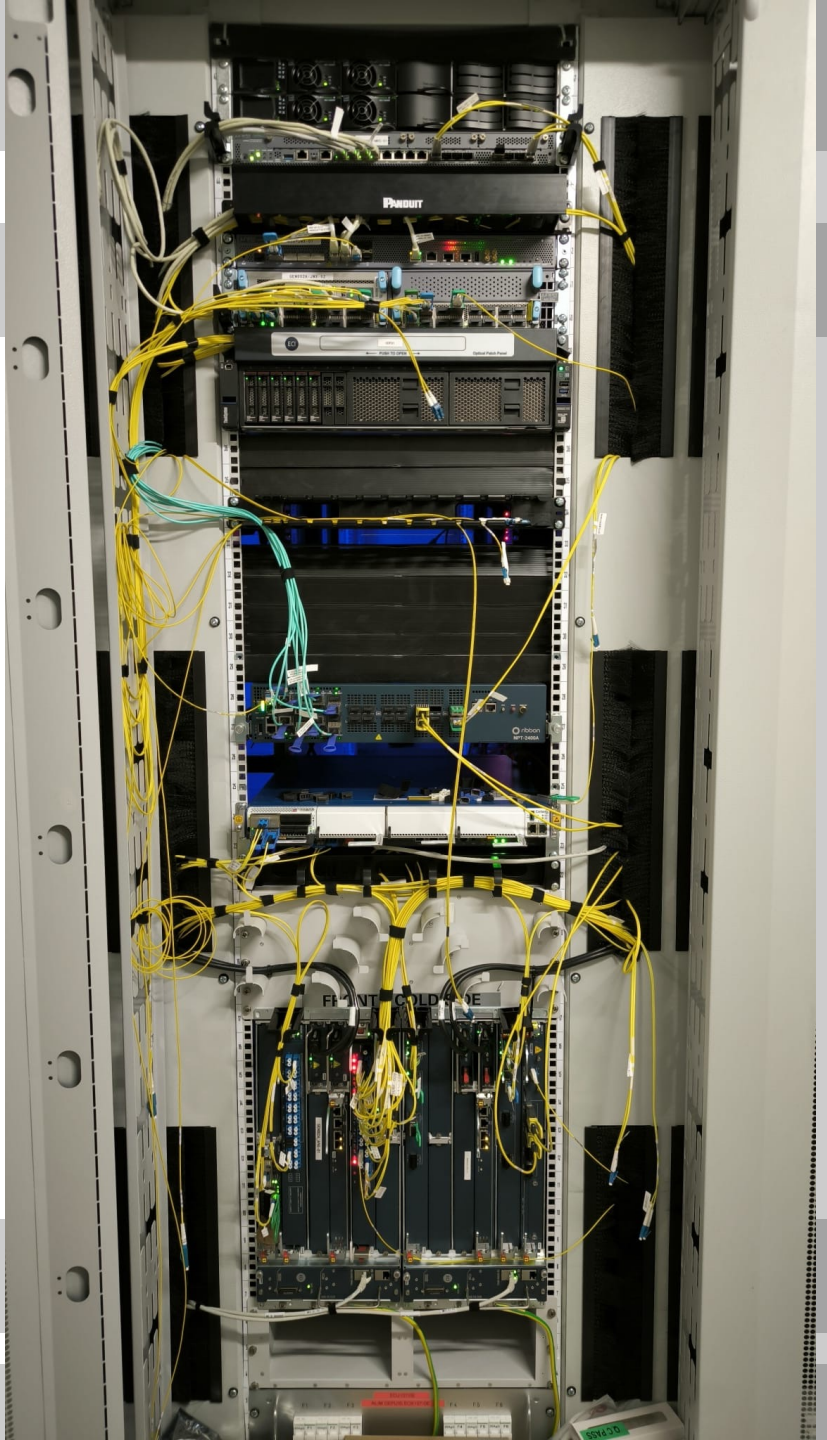
Hardware used: IP

Nokia FP5 based SR-1 routers



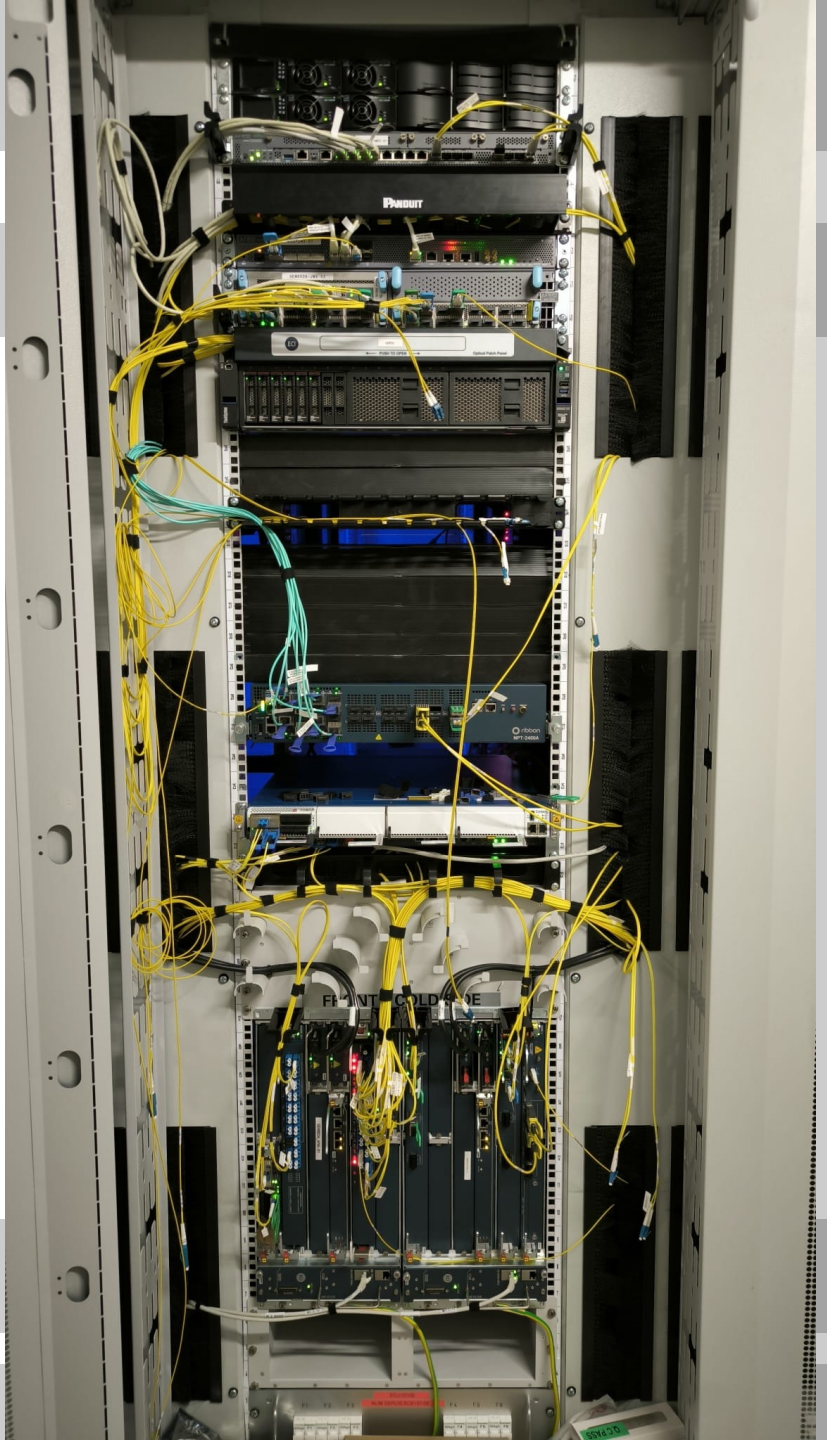
Hardware used: IP

Ribbon NPT-2400 Broadcom J2C 4.8T

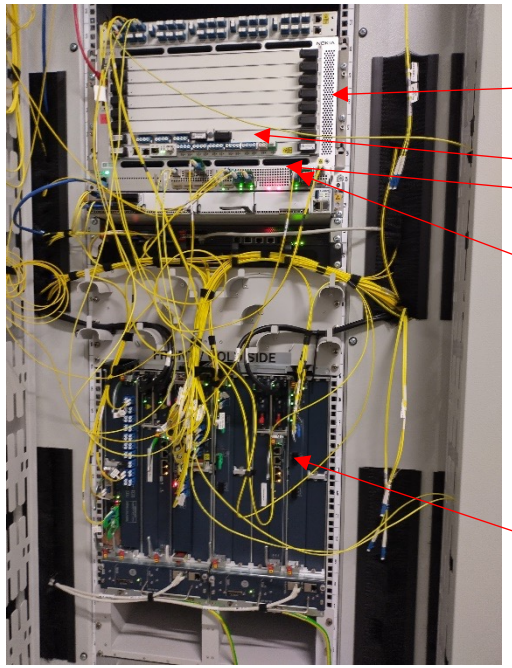


Hardware used: IP

- Ribbon NPT 2400: Broadcom 4.8Tbit/s Jericho2c



SURF's LHCOPN 800G test: CERN Pictures



PSI-L

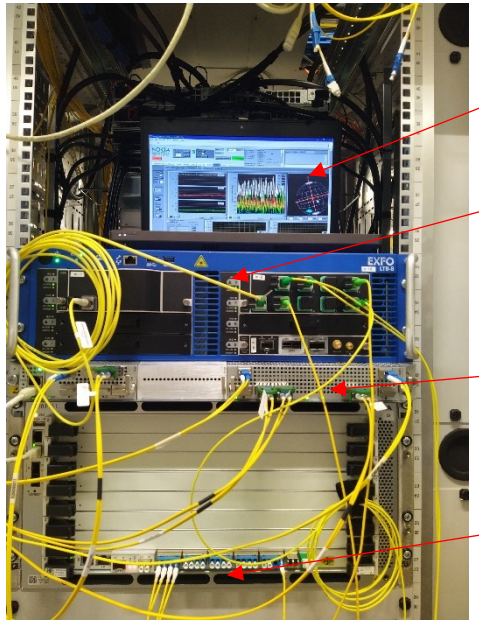
IR9-LP

PSIM
+DMAT6
+SFM6

Ribbon
ROADM

EXFO OSA+OSW

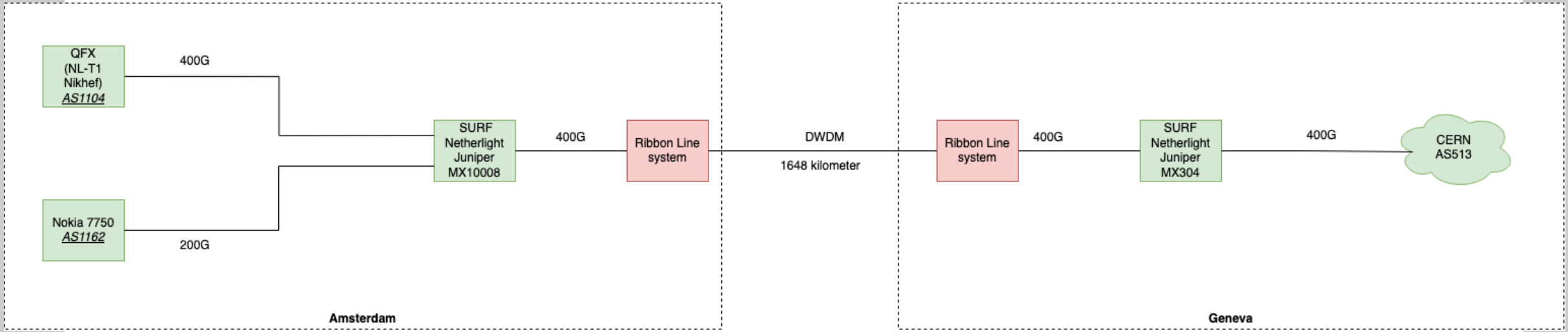
SURF's LHCOPN 800G test: Pictures AMSTERDAM



SURF's LHCOPN Production topology (as of end 2023)

400GE capable transport system

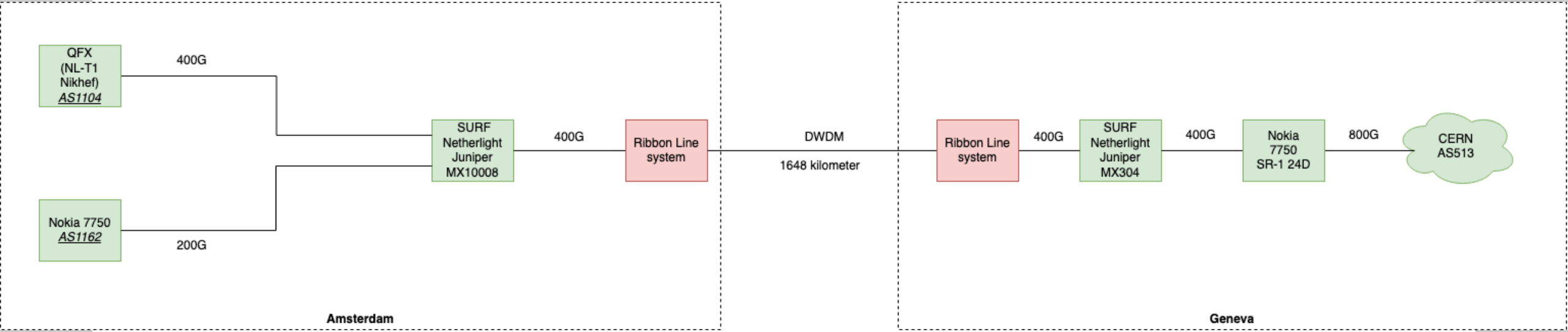
- 1648km fiber trajectory long between Amsterdam & Geneva
- 400GE capable routers on each end
- 2x 200GE lambda between linesystem ends



SURF's LHCOPN intermediate topology (as of 07/02/2024)

400GE capable transport system

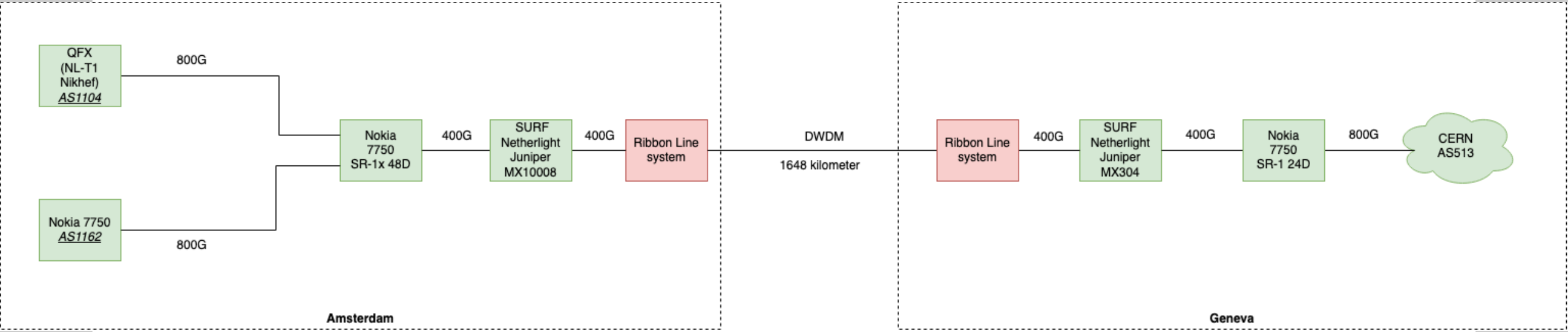
- Add in Nokia in between MX304 and CERN
- Moved service attach point from MX304 to Nokia



SURF's LHCOPN intermediate topology (as of 14/02/2024)

400GE capable transport system

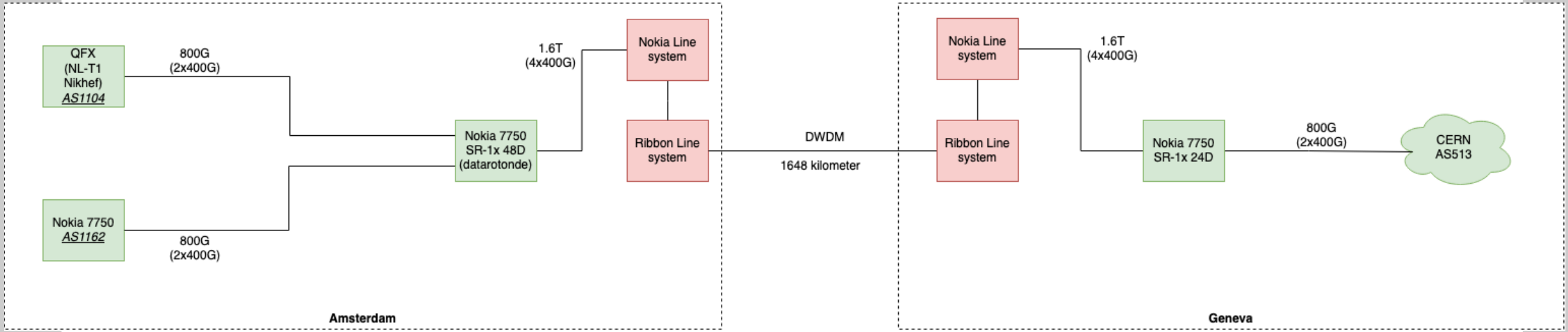
- Add in Nokia between SURF and Nikhef in Amsterdam
- Add extra service attach point to Nokia

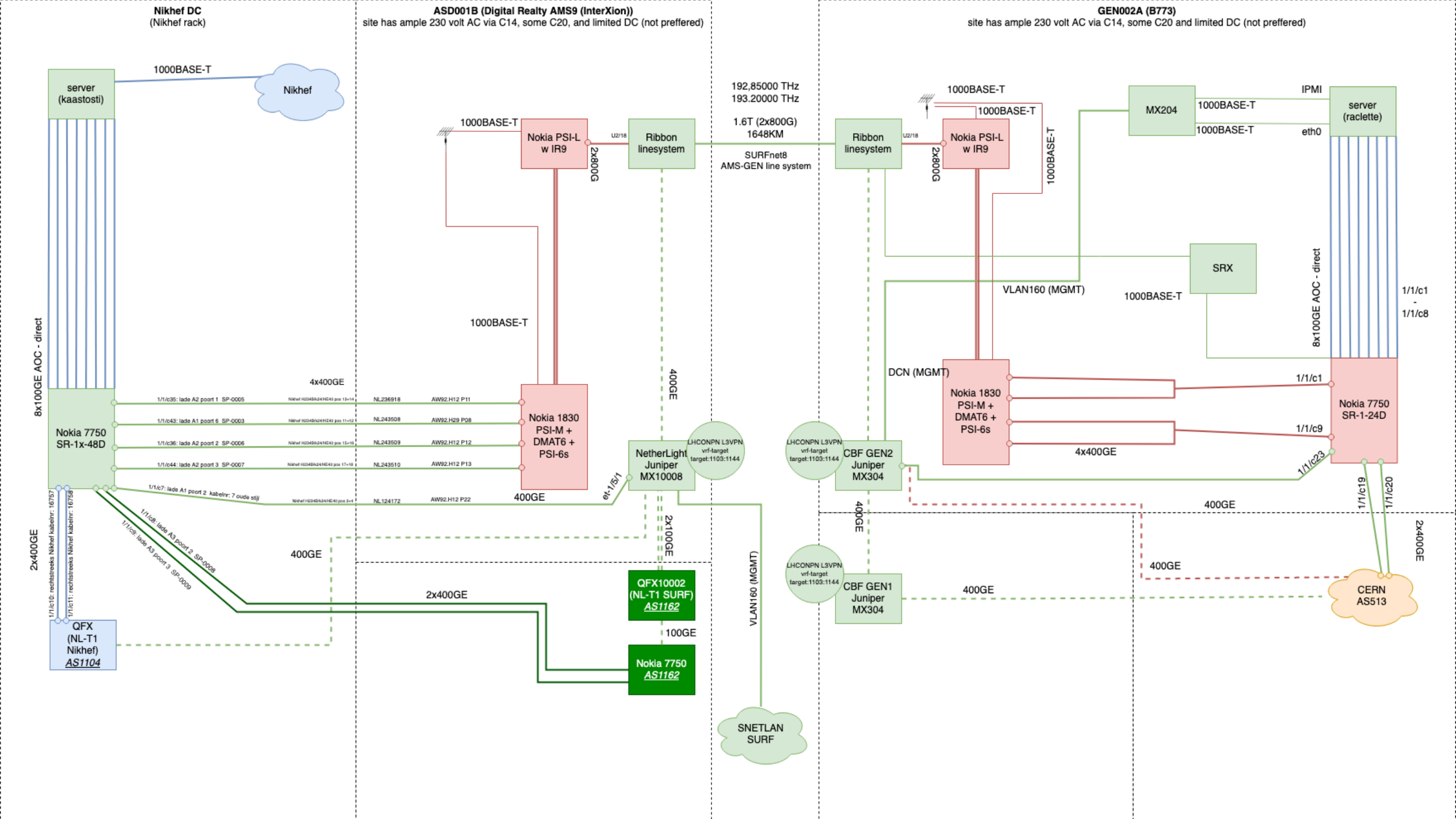


SURF's LHCOPN Test topology

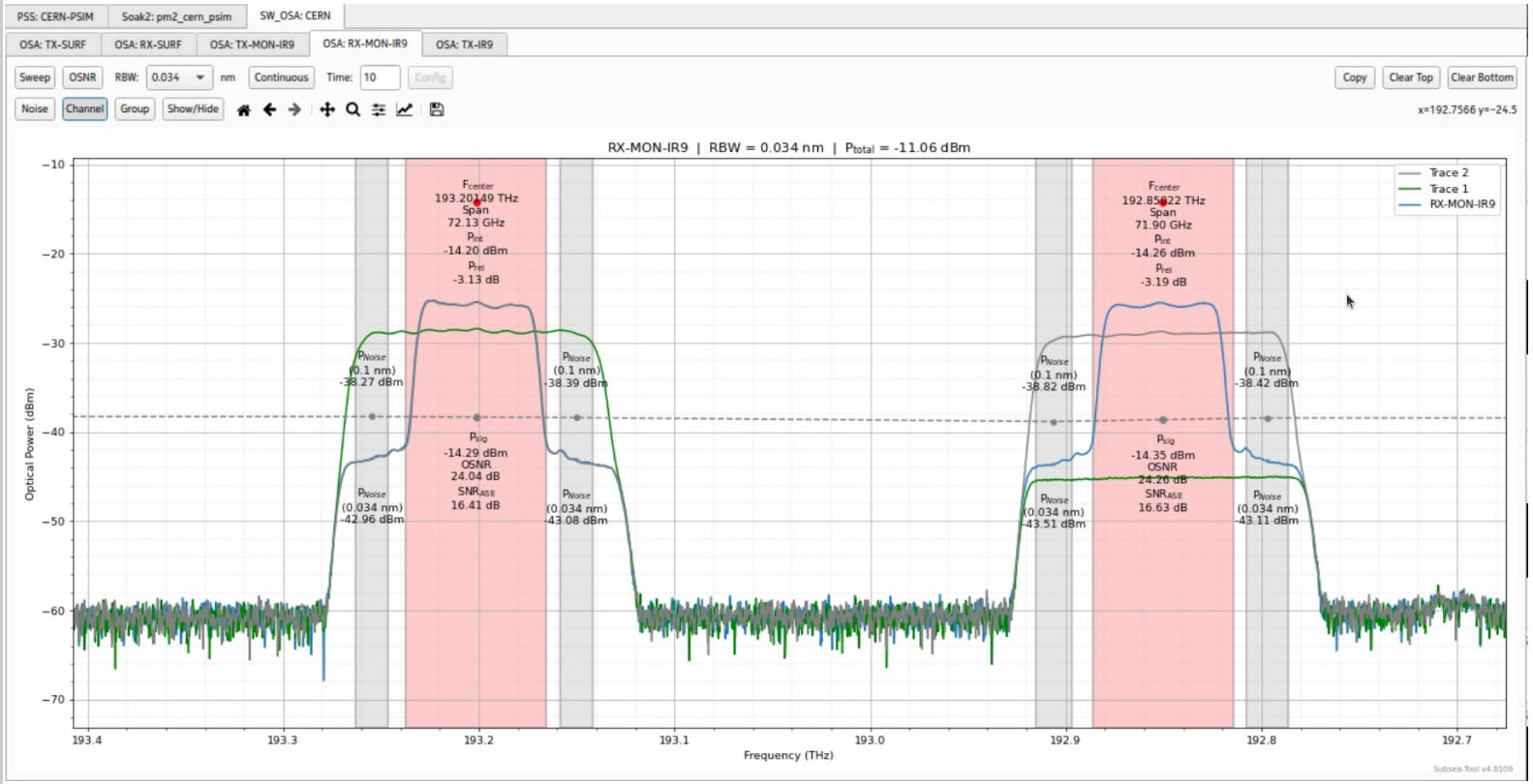
800GE capable transport system

- 1648km fiber trajectory long between Amsterdam & Geneva
- 800GE capable routers on each end (stacking 400G towards sites)

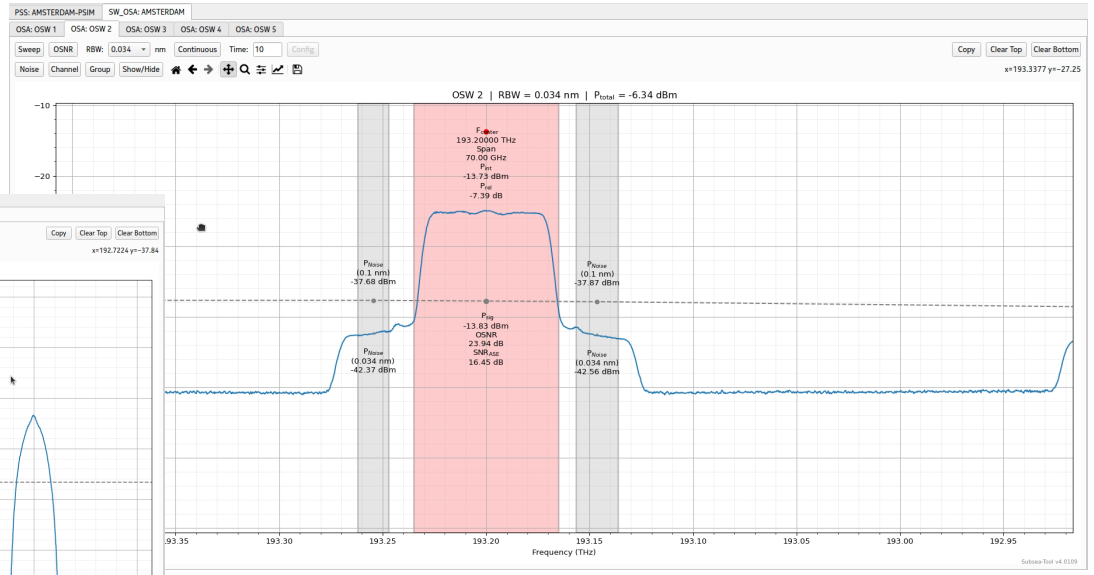



Nikhef DC
(Nikhef rack)
ASD001B (Digital Realty AMS9 (InterXion))
site has ample 230 volt AC via C14, some C20, and limited DC (not preffered)
GEN002A (B773)
site has ample 230 volt AC via C14, some C20 and limited DC (not preffered)
server (kaastosti)
1000BASE-T
Nikhef
8x100GE AOC - direct
Nokia 7750 SR-1x-48D
2x400GE
1/1/c10: rechtstreeks Nikhef kabelnr: 16757
1/1/c11: rechtstreeks Nikhef kabelnr: 16758
QFX (NL-T1 Nikhef) AS1104
4x400GE
1/1/c35: lade A2 poort 1 SP-0005
1/1/c43: lade A1 poort 6 SP-0003
1/1/c36: lade A2 poort 2 SP-0006
1/1/c44: lade A2 poort 3 SP-0007
1/1/c7: lade A1 poort 2 kabelnr: 7 oude stijl
1/1/c8: lade A3 poort 2 SP-0008
1/1/c9: lade A3 poort 3 SP-0009
400GE
1000BASE-T
Nokia PSI-L w IR9
U2/18
2x800G
Ribbon linesystem
1000BASE-T
Nokia 1830 PSI-M + DMAT6 + PSI-6s
400GE
et-1/5/1
NetherLight Juniper MX10008
LHCONPN L3VPN vrf-target target:1103:1144
400GE
2x100GE
QFX10002 (NL-T1 SURF) AS1162
100GE
Nokia 7750 AS1162
2x400GE
VLAN160 (MGMT)
SNETLAN SURF
192,85000 THz
193.20000 THz
1.6T (2x800G)
1648KM
SURFnet8
AMS-GEN line system
Ribbon linesystem
U2/18
2x800G
1000BASE-T
Nokia PSI-L w IR9
1000BASE-T
MX204
IPMI
1000BASE-T
1000BASE-T
eth0
server (raclette)
SRX
VLAN160 (MGMT)
1000BASE-T
8x100GE AOC - direct
1/1/c1 - 1/1/c8
DCN (MGMT)
Nokia 1830 PSI-M + DMAT6 + PSI-6s
1/1/c1
1/1/c9
4x400GE
Nokia 7750 SR-1-24D
1/1/c23
1/1/c19
1/1/c20
2x400GE
LHCONPN L3VPN vrf-target target:1103:1144
CBF GEN2 Juniper MX304
400GE
400GE
LHCONPN L3VPN vrf-target target:1103:1144
CBF GEN1 Juniper MX304
400GE
400GE
CERN AS513

Spectrum analyzer: CERN



Spectrum analyzer: AMSTERDAM



SURF

SURF's LHCOPN 800G test: Protocol stack

No IGP between Nokia routers

- iBGP with short BFD timers
- Used our testbed AS 1125
- "pretend" to SURFsara, Nikhef and CERN to be AS 1103
- Hiding global AS

Shifting traffic

- During migration phase, used LOCAL-PREF and AS-prepend
- In production, short AS path always results in test path being used

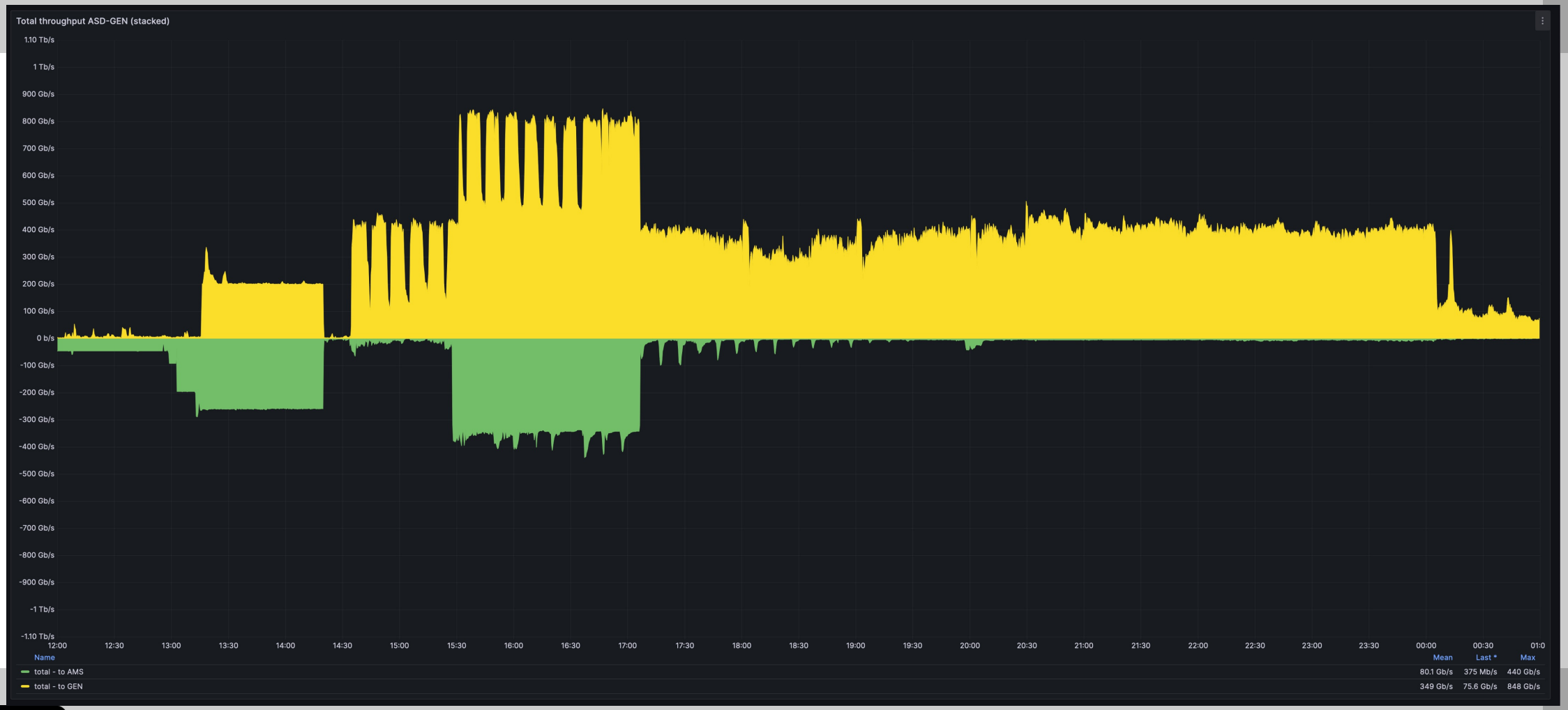
SURF

Hardware used: Test servers

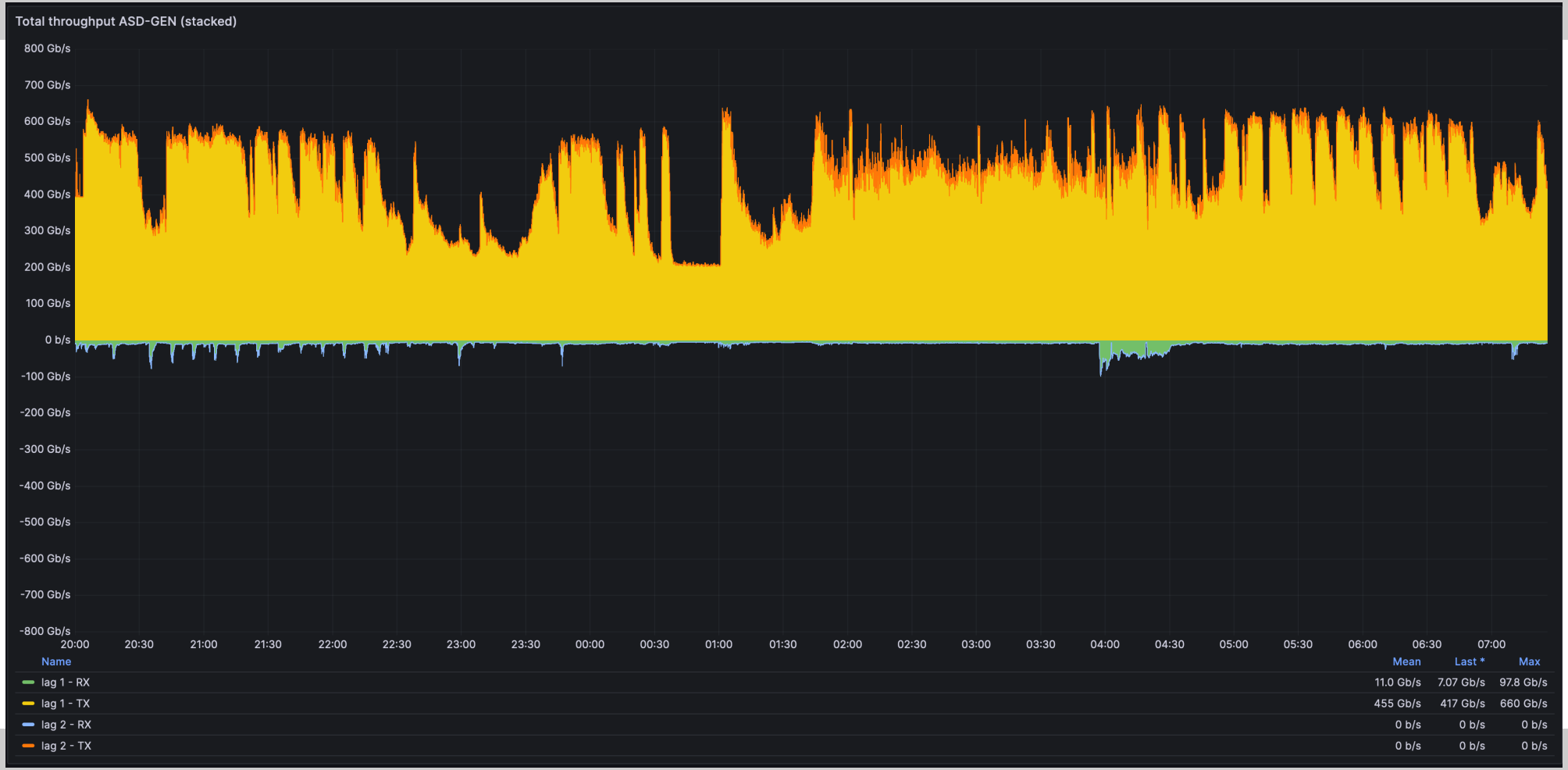
- AMD EPYC 9554P 64C 128T @3,76 Ghz
- 768GB DDR5 4800 MT/s
- 4x dual port Mellanox ConnectX-7
- 6x 3.84T Intel/Solidigm P5520 Nvme
- DPDK pkt-gen

SURF

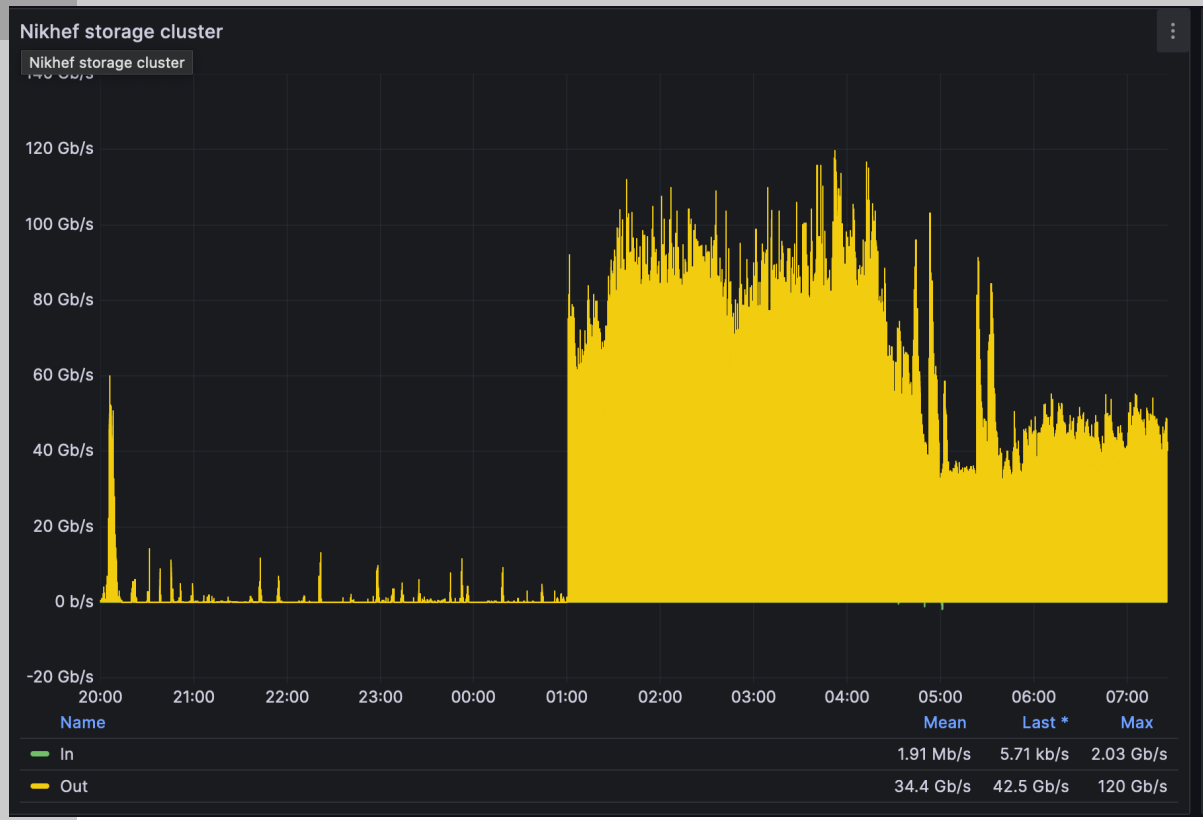
Total throughput Amsterdam <-> Geneva



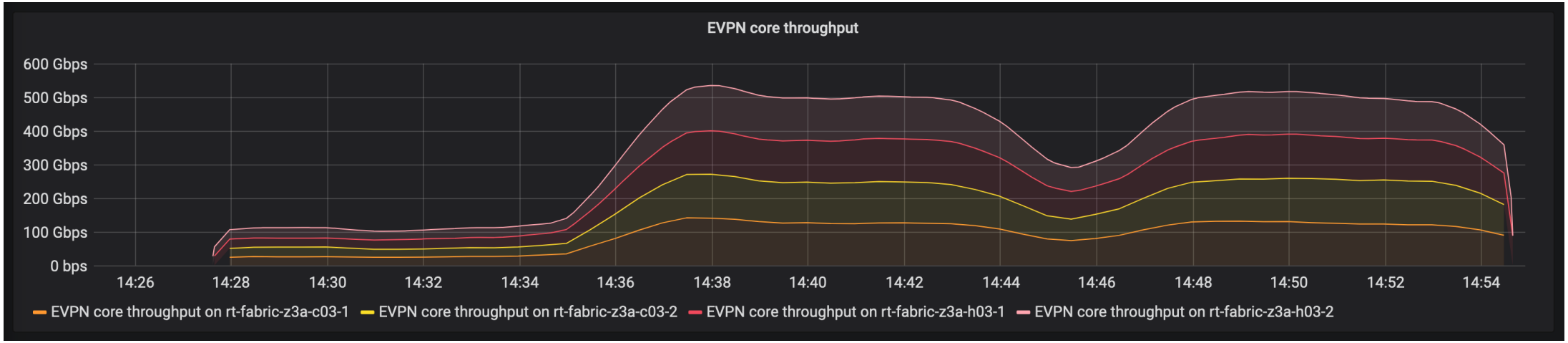
Last day of only production testing



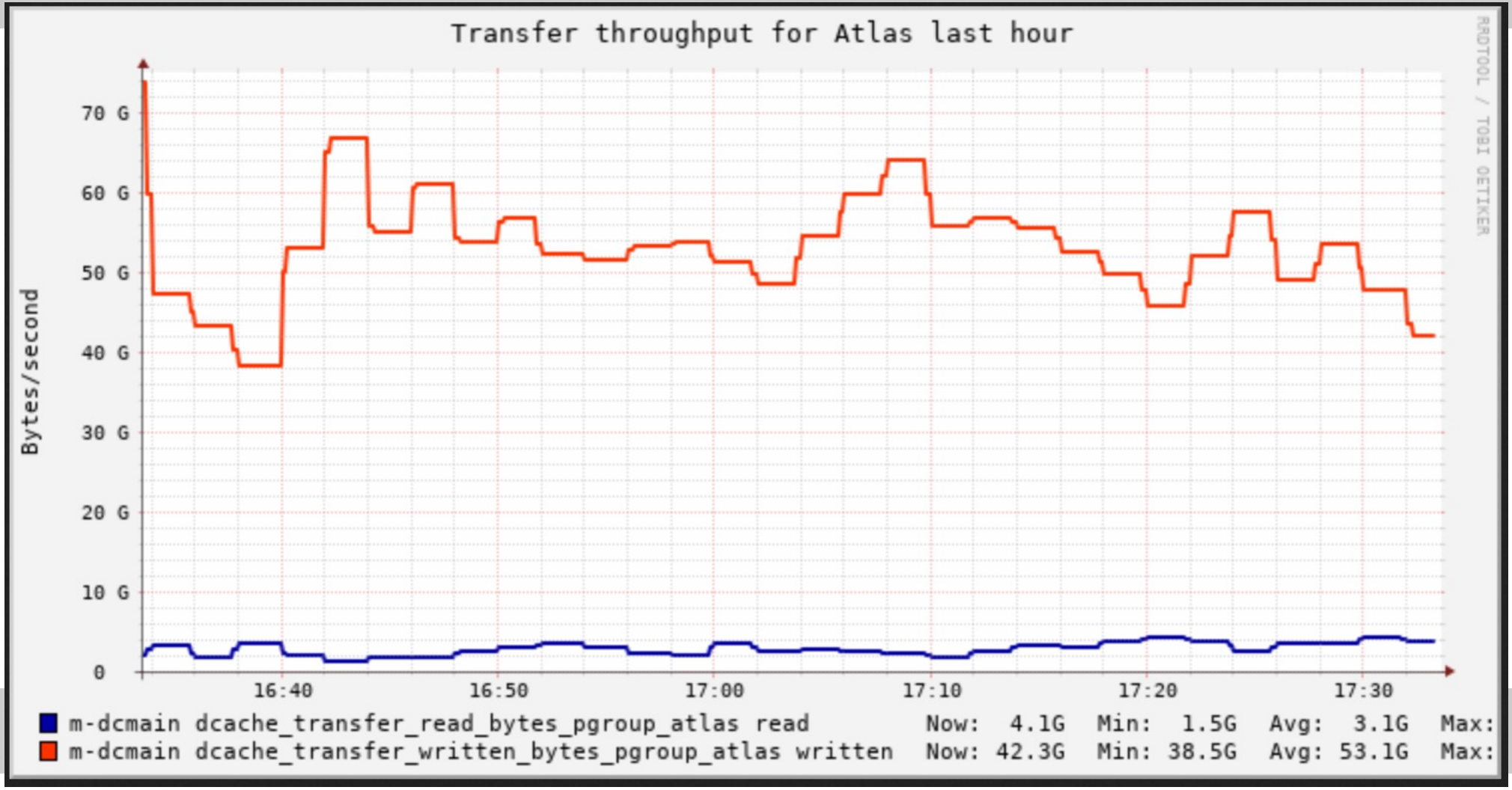
Last day of only production testing



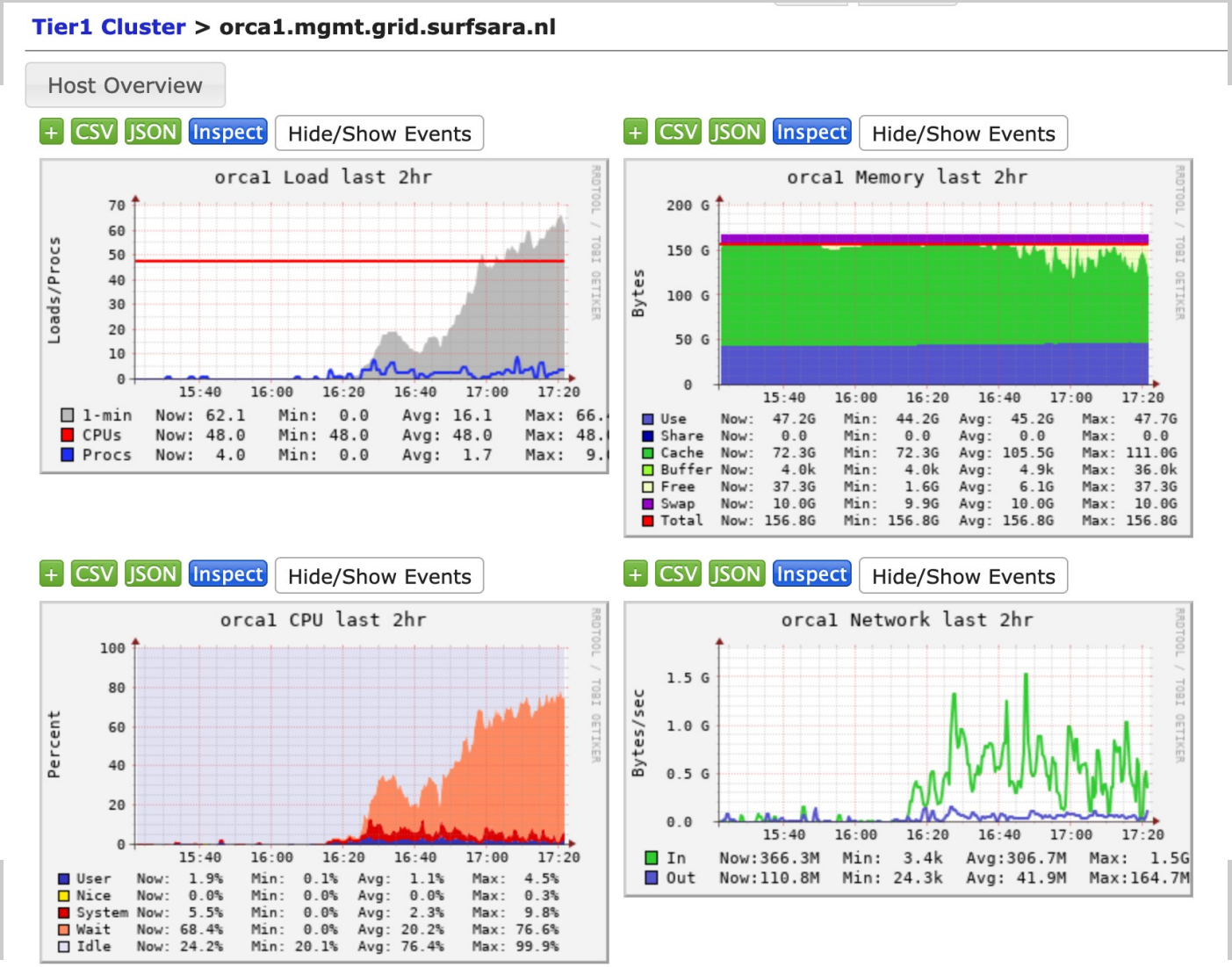
SURF datacenter network fabric load



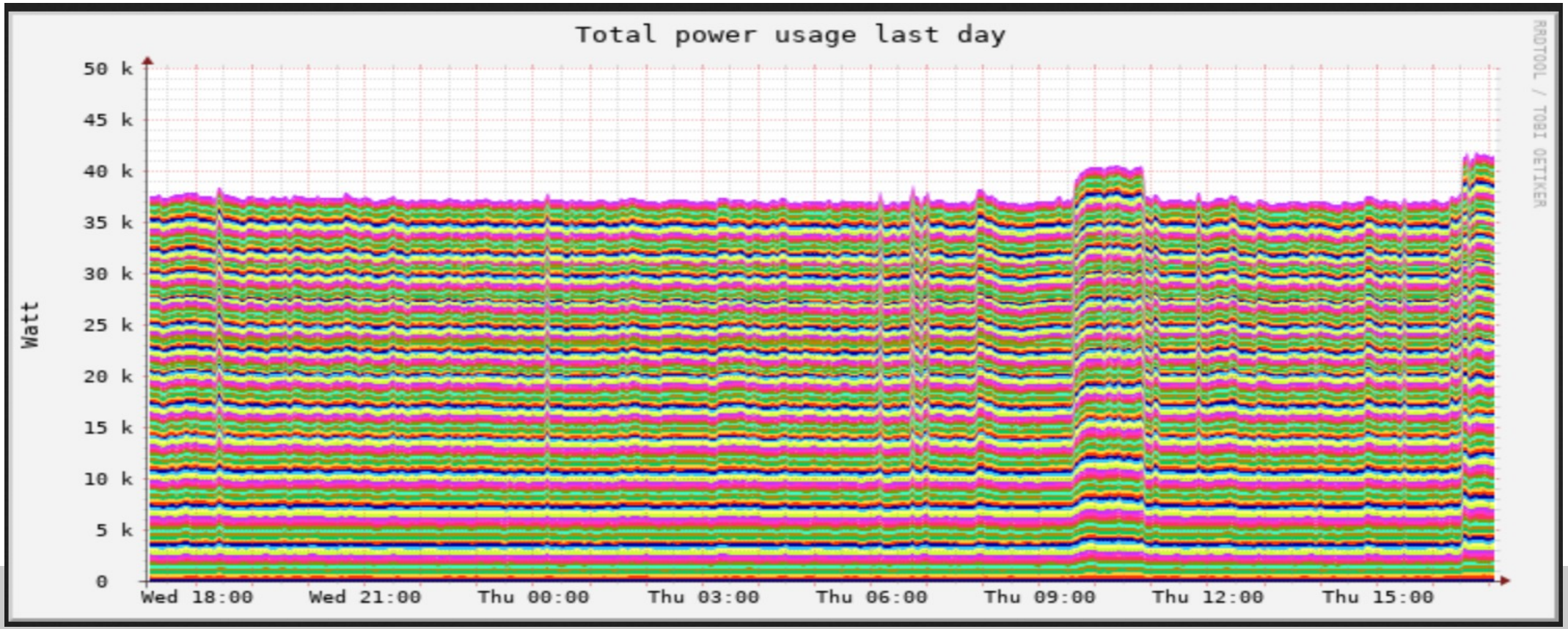
dCache storage throughput



dCache CPU load



dCache power consumption



Lessons learned

- Fiber AMS-GEN is of poor quality. In trial environments we might reach 800G per channel, but in production environments there is insufficient margin
- Different vendors on the line system side and transponder side seem to work quite well.
- IP hardware used in trial worked as expected. (no 200G Ethernet available yet)
- L3VPN interop between Nokia and Juniper worked great
- Packet canons can easily reach 800G with 8x100G
- We are the fastest LHC T1 both on network and storage throughput!! Reached up to 661Gbit/s
- Continued testing at 2 400G Channels to reach 800Gb/s Connectivity to NL-T1 for storage troughput testing

SURF


SURF