LHC networking - update

4th Asia Tier Centre Forum
Bangkok, Thailand

21st November 2018

edoardo.martelli@cern.ch

Contents

- LHCOPN
- LHCONE
- On-going and future developments

LHCOPN update

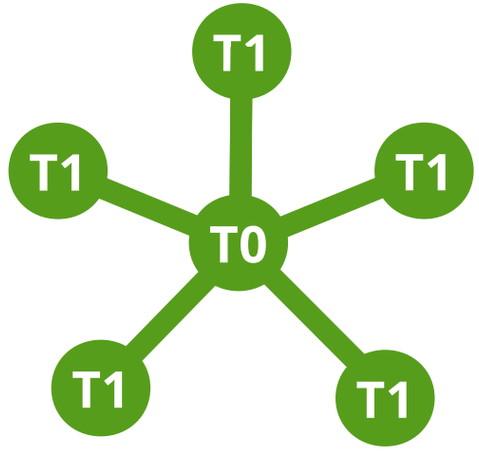
LHCOPN

Private network connecting Tier0 and Tier1s

- Dedicated to LHC data transfers and analysis

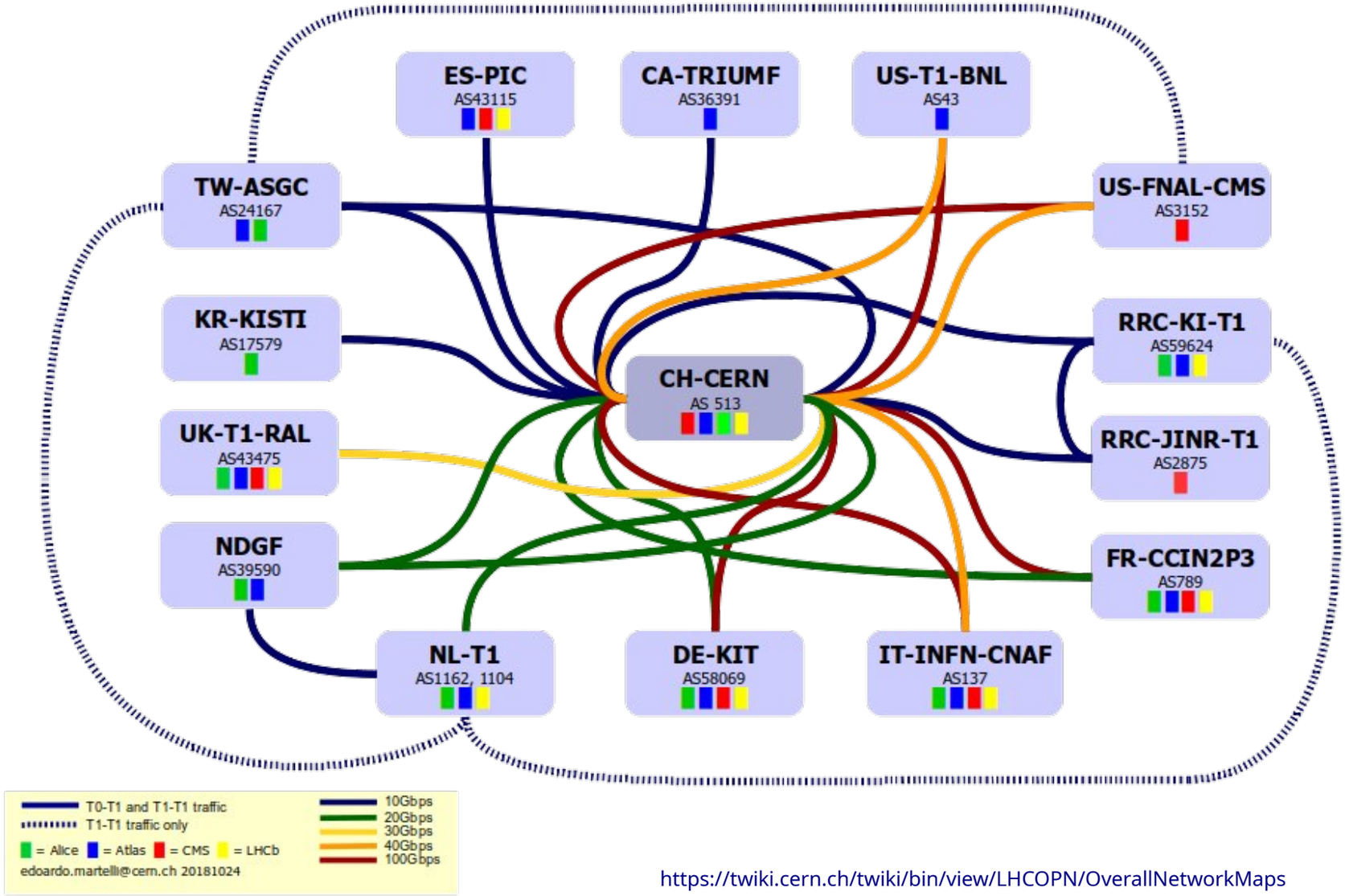
- Secured: only declared IP prefixes can exchange traffic

- Advanced routing: communities for traffic engineering, load balancing.



LHCOPN

LHCOPN



Numbers

- 14 Tier1s + 1 Tier0
- 12 countries in 3 continents
- Dual stack IPv4-IPv6
- 660 Gbps to the Tier0
- Moved ~160 PB in the last year

https://twiki.cern.ch/twiki/bin/view/LHCOPN/OverallNetworkMaps

LHCOPN latest deployments

- **IT-INFN-CNAF**: primary link upgraded to 100Gbps
- **DE-KIT**: primary link upgraded to 100Gbps
- **IT-CCIN2P3**: primary link upgraded to 100Gbps
- **US-FNAL-CMS**: using LHCOPN for protoDUNE data transfers
- **BelleII** and **protoDUNE** data transfers allowed on LHCOPN
- **All links** are dual-stack IPv4 and IPv6

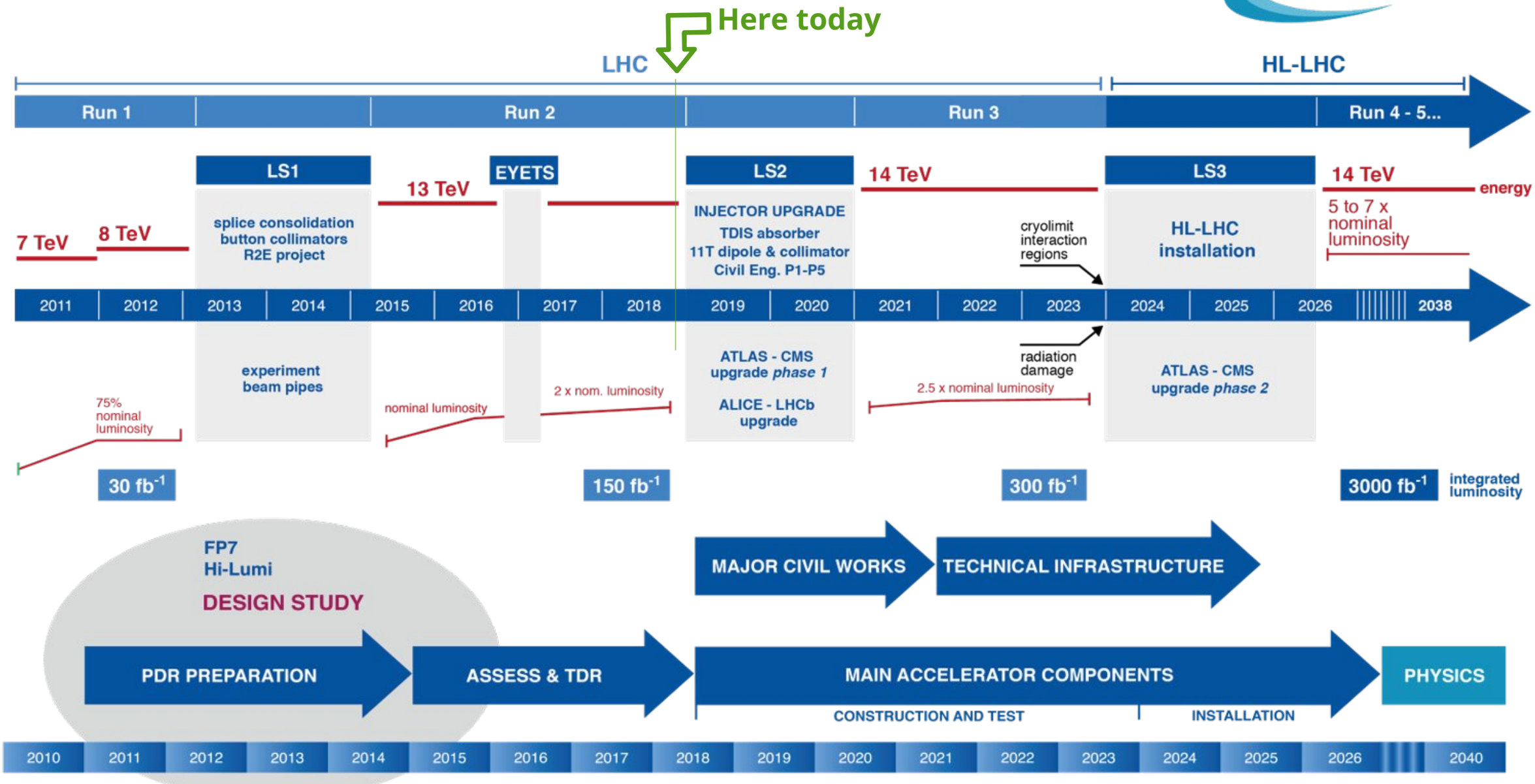
Tier0's latest deployments

CH-CERN:

- Started deployment of Juniper QFX1000x in datacentre (replacing Brocade MLXE)
- LHCOPN border routers being migrated to QFX10002
- 2nd network hub in Prevessin site (FR) completed
- Wigner data-centre will be decommissioned at the end of 2019
- Using LHCb datacentre-in-containers as datacentre extension during LS2
- Data-centre extension in Prevessin may be built for Run4 (2025)

LHCOPN

HL-LHC plan


Here today
LHC
HL-LHC
Run 1
Run 2
Run 3
Run 4 - 5...
LS1
EYETS
LS2
LS3
7 TeV
8 TeV
13 TeV
14 TeV
14 TeV
energy
splice consolidation
button collimators
R2E project
INJECTOR UPGRADE
TDIS absorber
11T dipole & collimator
Civil Eng. P1-P5
cryolimit
interaction
regions
HL-LHC
installation
5 to 7 x
nominal
luminosity
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2038
experiment
beam pipes
ATLAS - CMS
upgrade phase 1
ALICE - LHCb
upgrade
radiation
damage
ATLAS - CMS
upgrade phase 2
75%
nominal
luminosity
nominal luminosity
2 x nom. luminosity
2.5 x nominal luminosity
30 fb-1
150 fb-1
300 fb-1
3000 fb-1
integrated
luminosity
FP7
Hi-Lumi
DESIGN STUDY
MAJOR CIVIL WORKS
TECHNICAL INFRASTRUCTURE
PDR PREPARATION
ASSESS & TDR
MAIN ACCELERATOR COMPONENTS
PHYSICS
CONSTRUCTION AND TEST
INSTALLATION
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2040

Long Shutdown 2: 2019-2020

Maintenance period of two years, starting on December 2018

- Perform major maintenance and consolidation
- Increase injector reliability and lifetime to cover HL-LHC run until ~2035
- Increase intensity in the injectors to match HL-LHC requirements
- Anticipate HL-LHC work

LHCONE update

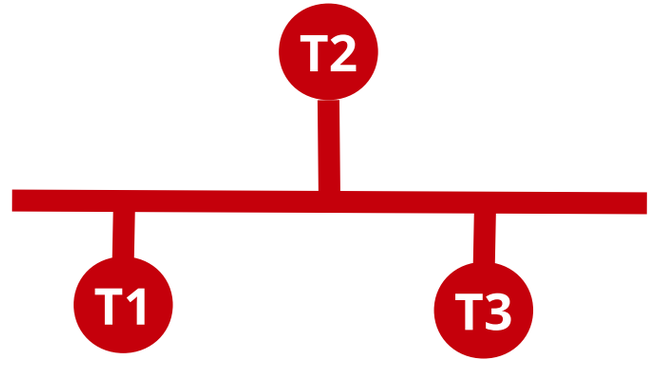
LHCONE L3VPN service

Layer3 (routed) Virtual Private Network:

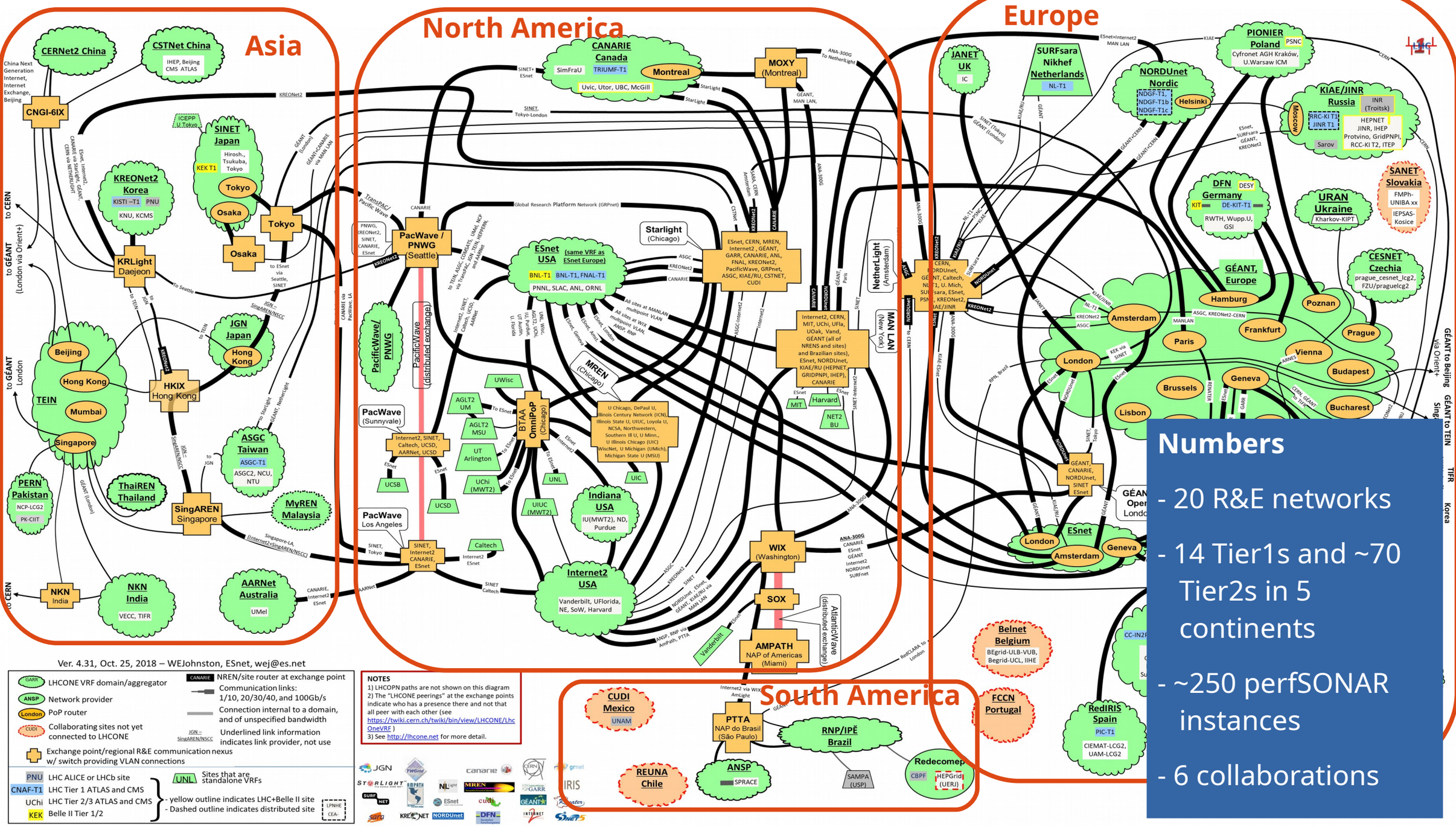
- Worldwide network backbone **connecting Tier1s, T2s and T3s** at high bandwidth

- Bandwidth **dedicated to HEP** data transfers

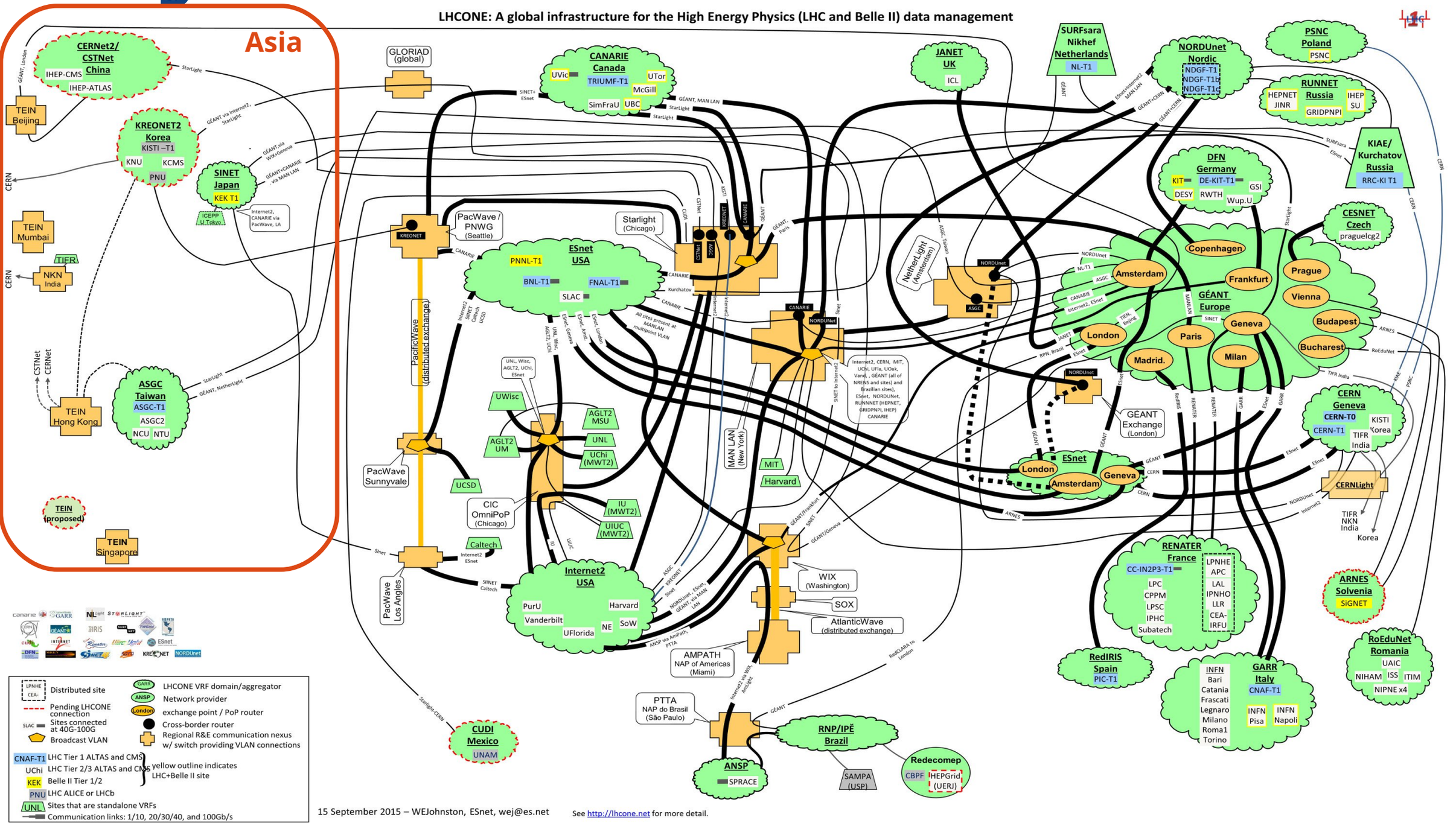
- Trusted traffic that **can be allowed to bypass slow perimeter firewalls**



L3VPN Current infrastructure



Progresses made in Asia since ATCF1



LHCONE last year deployments

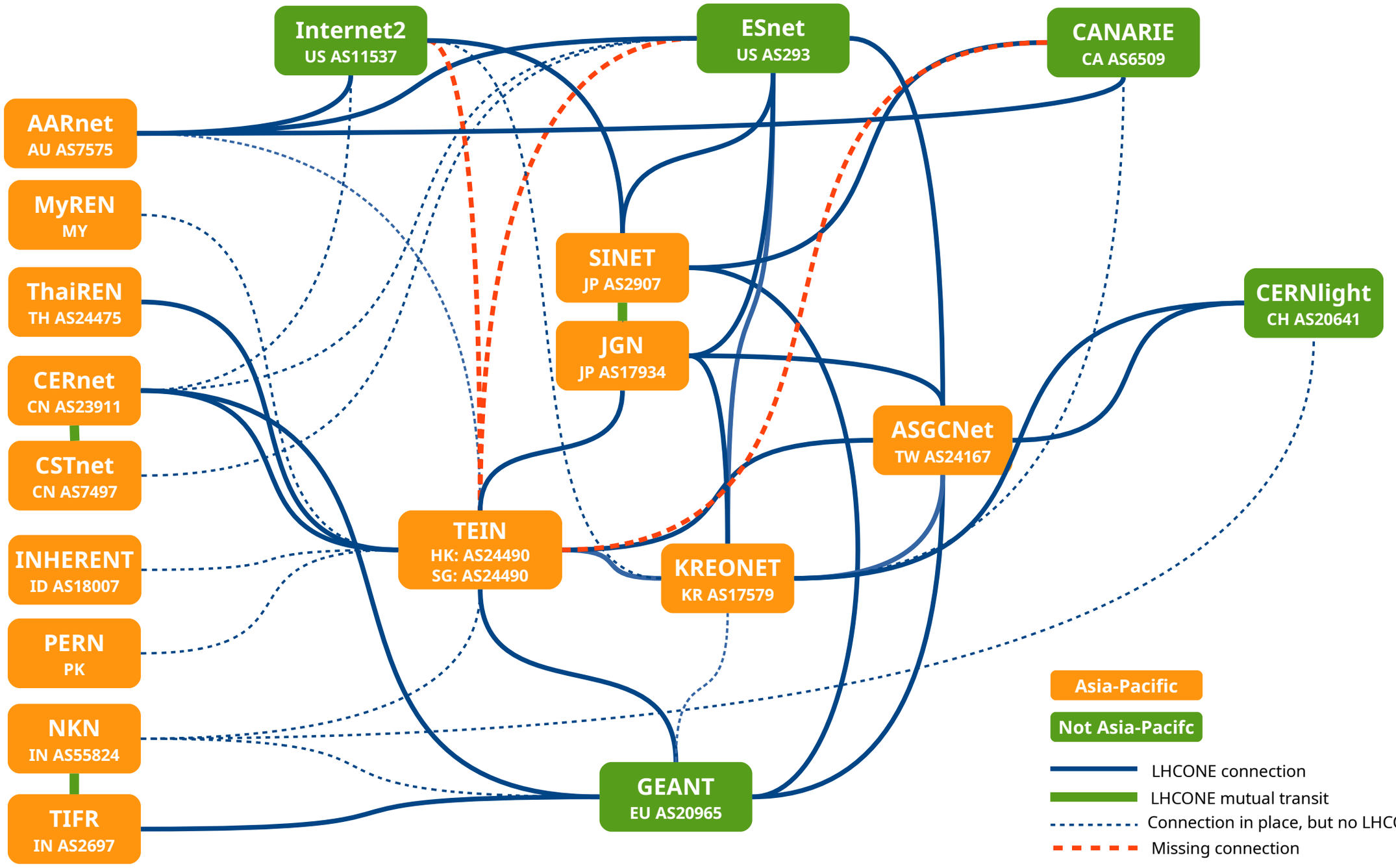
L3VPN latest changes

- UK-T1-RAL will soon connect (only Tier1 not yet connected)
- Chile started the process to implement LHCONE. Will connect to GEANT first

In Asia:

- IHEP T2 in China connected to LHCONE thanks to CSTNet, CERNet and TEIN
- ASGC reaches Europe directly via Singapore (shorter RTT)
- ASGC connected to NORDUnet , TWAREN
- KREONET connected to CERNlight, TEIN, ASGC,JGN, ESnet
- KREONET planning 100G links to Seattle, Chicago and Hong Kong
- Sinet now connected to ESnet and Internet2
- JGN now connected to TEIN and ESnet
- CAE100G: considering a 100G ring
  Asia-Europe via Siberia and Indian Ocean

Asia-Pacific VRFs – Current Status



Chasing improper packets in LHCONE

DE-KIT and ESnet have run an investigation on improper packets coming from LHCONE connections, i.e. **packets with source addresses that don’t have a return path inside LHCONE**

Sources are notified and prompted to fix the situation

Fewer LHCONE unroutable source packets are being detected by ESnet after they were reported at the LHCONE meeting in March 2018

Still room for improvement, particularly in the private IP ranges (RFC 1918)

Monitoring will continue and progresses will be reported at upcoming LHCONE meetings

MTU size recommendation

At the LHCONE meeting of March 2018, several issues were reported being caused by the use of different MTU sizes in LHCONE connections.

A working group has analyzed the problem and have come up with this recommendation:

LHCONE/LHCOPN network paths should allow MTU size up to 9000 bytes and not block PMTUD packets (RFCs 1911, 1981 and 4821)

- In practice this means that the frame size (Layer2) should be at least 9080 bytes for all devices on the path
- ICMP “Fragmentation Needed” (Type 3, Code 4) should not be blocked by any devices on the path

The working group will produce a document with configuration examples for different Operating Systems

Digging in the LHCONE routing table

NORDUnet has run an analyses of the routing tables of most of the LHCONE VRFs.

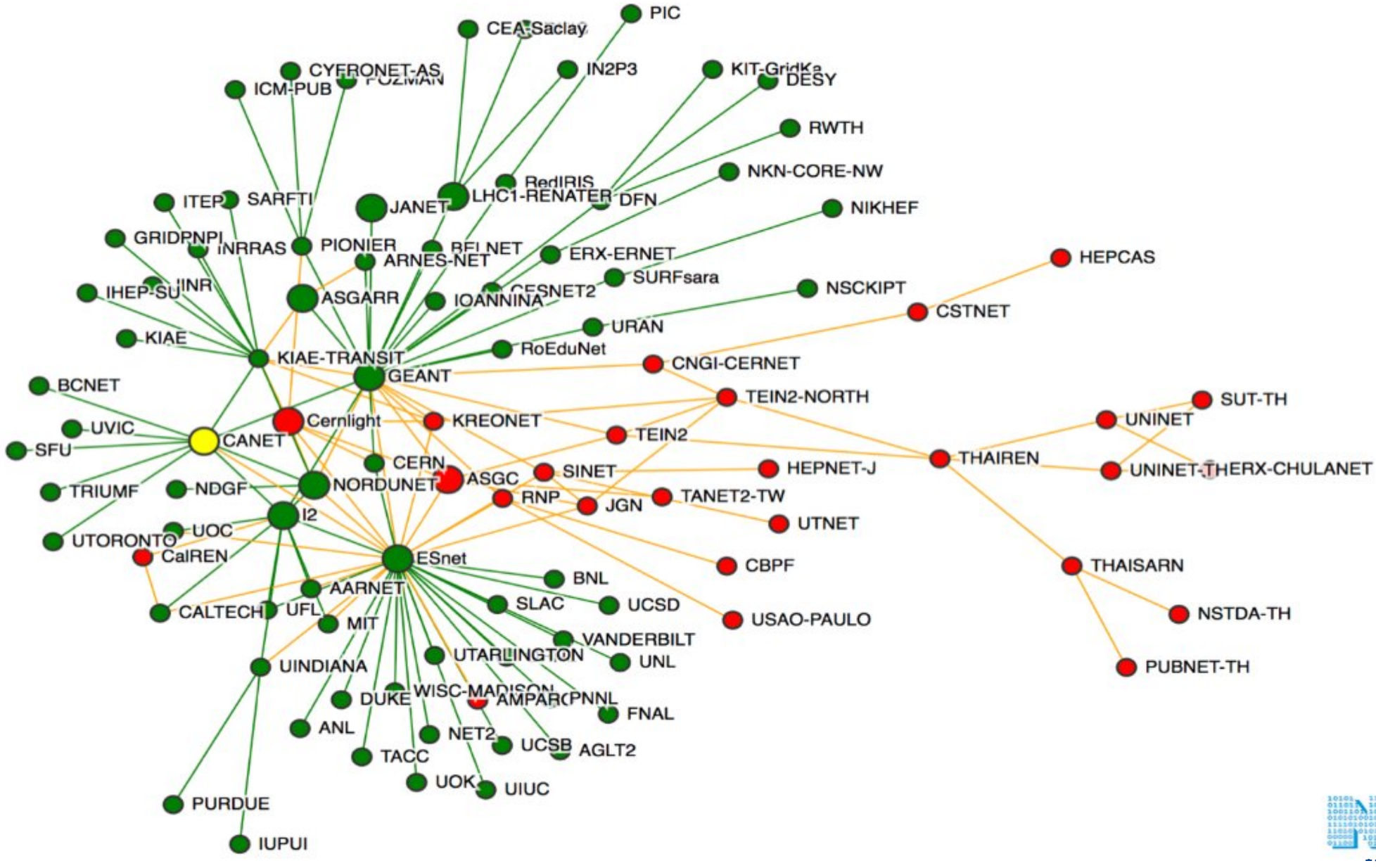
It resulted that **reachability is fragmented**, especially on IPv6:
- Only GEANT has a full view of all LHCONE destinations
- Especially the sites behind TEIN cannot reach a fraction of LHCONE

IPv4 reachability map:
https://indico.cern.ch/event/725706/contributions/3149436/attachments/1744301/2823447/go

IPv6 reachability map:
https://indico.cern.ch/event/725706/contributions/3149436/attachments/1744301/2823448/go

The community will work on improving this situation, which will be followed-up in the future meetings

Example: Canarie IPv4 reachability



LHCONE Looking Glass

At the LHCONE meeting of March 2018 it was decided to implement a **tool to analyze the routing tables of the different VRFs**

For this, a Route Server and a Looking Glass web interface have been set up at CERN

The looking glass can show the prefixes known in the different VRFs and the attributes of the routes

At the moment the route-server peers with CERNlight and NORDUnet VRFs only. Other peerings will be added in the coming months

The looking glass is accessible at http://lhcone-lg.cern.ch/

LG example

Show BGP route detail IPv6 <address>
Shows details of a route received
from multiple peers

LHCONE Looking Glass Results - ex2j.cern.ch

Date: Thu Oct 18 06:21:36 2018 CEST

Query:
Argument(s): 2001:6a0:fff:ff0a::/64

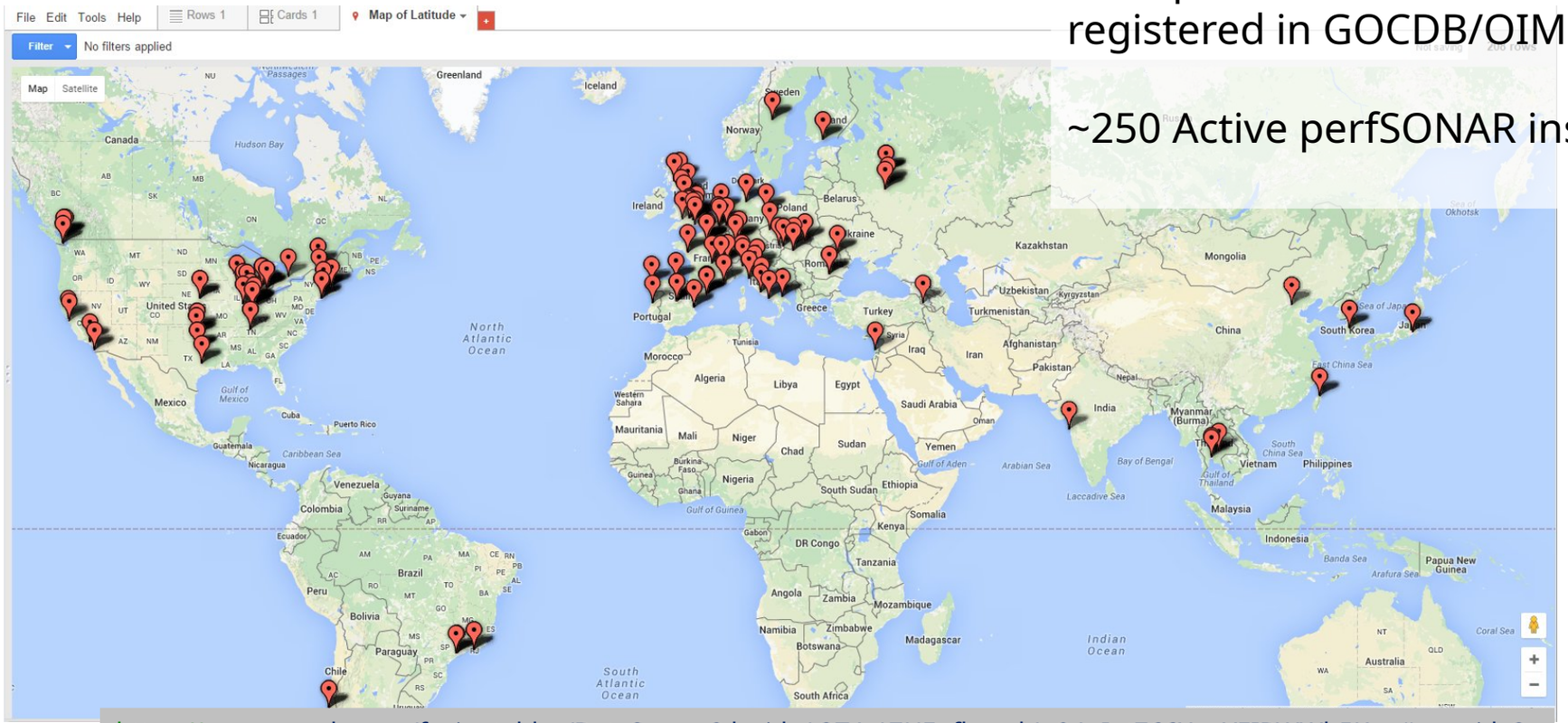
```
LG.inet6.0: 117 destinations, 181 routes (117 active, 0 holddown, 0 hidden)
2001:6a0:fff:ff0a::/64 (2 entries, 1 announced)
        *BGP    Preference: 170/-101
                Next hop type: Indirect, Next hop index: 0
                Address: 0xb2d4f30
                Next-hop reference count: 211
                Source: 192.65.183.30
                Next hop type: Router, Next hop index: 1708
                Next hop: 2001:1458:0:1d::1 via irb.183, selected
                Session Id: 0x0
                Protocol next hop: ::ffff:192.65.183.30
                Indirect next hop: 0xd1cf980 131071 INH Session ID: 0x0
                State:
                Peer AS:  2603
                Age: 13:04:54   Metric2: 0
                Validation State: unverified
                Task: BGP_2603_61339.192.65.183.30
                Announcement bits (2): 0-KRT 3-Resolve tree 2
                AS path: 2603 20965 8501 I
                Communities: 2603:2112 20965:155 target:2603:434300001 target:20965:111
                Accepted
                Localpref: 100
                Router ID: 192.65.183.30
         BGP    Preference: 170/-101
                Next hop type: Router, Next hop index: 1740
                Address: 0xb2cff50
                Next-hop reference count: 71
                Source: 2001:1458:a00:b::2
                Next hop: 2001:1458:a00:b::2 via irb.503, selected
                Session Id: 0x0
                State:
                Inactive reason: AS path
                Peer AS: 20641
                Age: 14:29:00
                Validation State: unverified
                Task: BGP_20641_61339.2001:1458:a00:b::2
                AS path: 20641 17579 20965 8501 I
                Communities: 20965:155
                Accepted
                Localpref: 100
                Router ID: 192.65.183.1

{master:0}
```

LHCONE perfSONAR: status

~280 perfSONAR instances registered in GOCDB/OIM

~250 Active perfSONAR instances



https://www.google.com/fusiontables/DataSource?docid=1QT4r17HEufkvnqhJu24nIptZ66XauYEIBWWh5Kpa#map:id=3

- Initial deployment coordinated by WLCG perfSONAR TF
- Commissioning of the network followed by WLCG Network and Transfer Metrics WG

Slide credit: Shawn McKee, University of Michigan

perfSONAR update

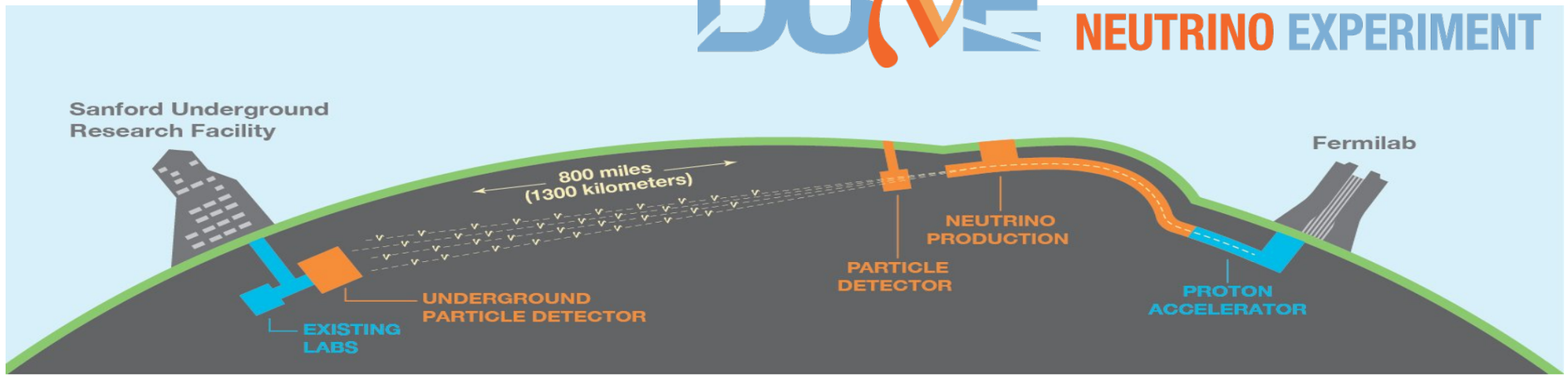
LHCONE MaDDash:
- improved situation
- New MaDDash v2.0: changed color coding, several improvements

PerfSONAR 4.1 released in August 2018:
- dropped SLC6 support
- new central web interface for configuration
- new pScheduler that replaces BWCTL. It requires TCP port 443 to be opened
- support for TWAMP (two way measurement)
- support for Docker

OSG/WLCg now storing perfSONAR measurements in Elasticsearch
OSG working on alarming for network issues

LHCOPN measures now visible on CERN monit-grafana

The DUNE collaboration



Fermilab presented the DUNE experiment:
- 2018: prototyping with protoDUNE at CERN
- 2019-2025 construction
- 2026 First neutrino beams from Fermilab to Sanford

Due to large overlap between LHC and DUNE sites, use of LHCONE may be envisaged

Expected 30PB/yr of raw data (similar to LHC Run2).
Primary storage at Fermilab, data replication at CERN

On-going and future developments

Use of LHCOPN/ONE by other collaborations

Following the protoDUNE and DUNE requests, it was discussed **whether allowing any type of collaboration to use LHCOPN and LHCONE**

No consensus was reached, but these points were noted:

- Different private networks for the different collaborations would be easy to create in backbone networks. They would allow a clearer separation of traffic for better statistics, security, billing
- However, end-sites serving multiple collaborations may have difficulties in separating the traffic to use those different private networks. **If funding agencies agree, sharing the same network instance would be preferable**

Networking activities around WLCG

Survey of networking research activities happening around WLCG:

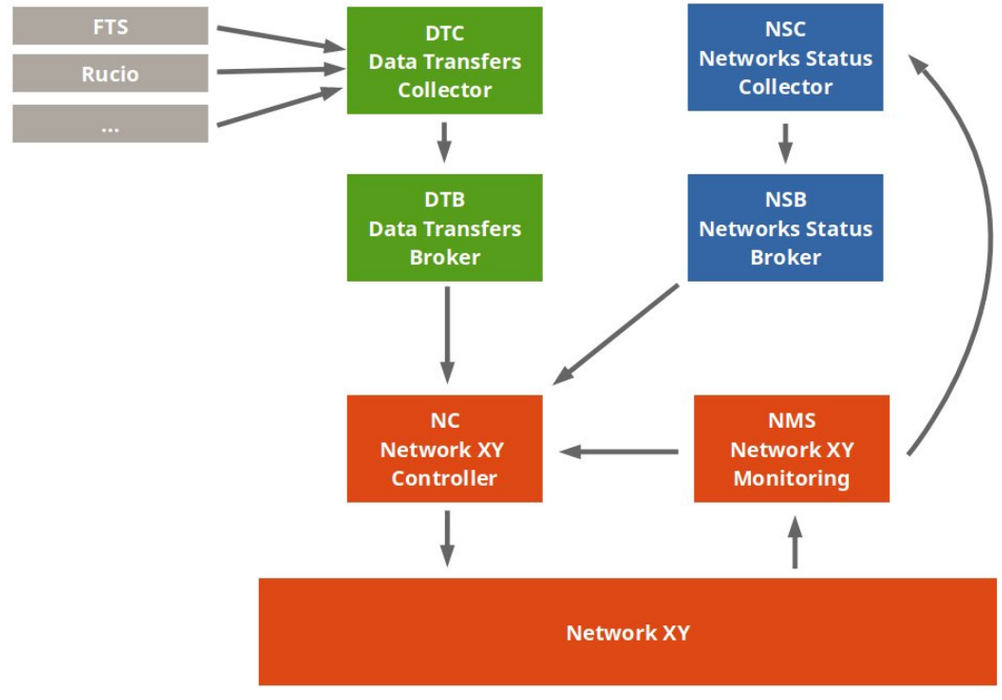
- DTN Nodes (a-la ESnet) and Test Nodes (a-la GEANT)
- High level protocol alternatives (DOMA)
- Low level protocol (TCP) alternatives (AENEAS SKA)
- Efficient use of WAN connections (NOTED)
- Adding additional bandwidth with Bandwidth on Demand and P2P (NOTED, LHCONE-P2P)
- Network Function Virtualization (HEPiX NFV Working group)
- Connectivity for commercial service providers (LHCONE)

NOTED activity

Network activity in the WLCG DOMA contest

Exploring options to **select outgoing network path from a site to load balance traffic across links**, to smooth peaks and increase usable bandwidth

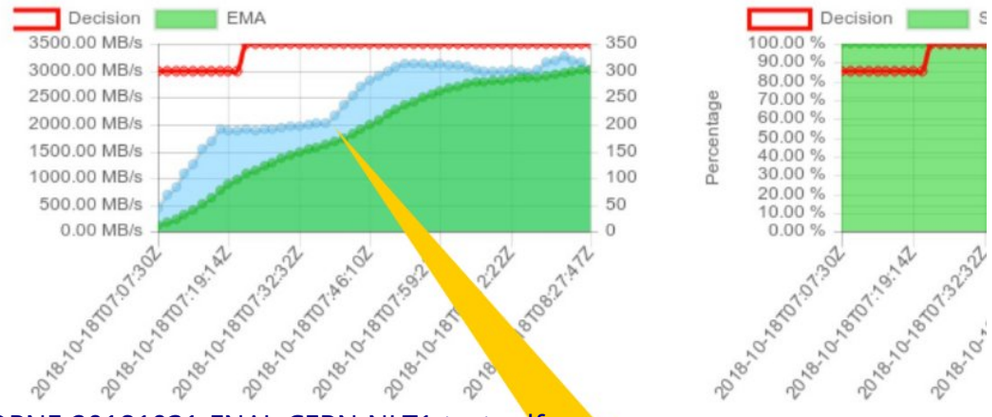
Useing information from transfer services (FTS, Rucio …) to identify “significant” data transfers



NOTED: CERN-NLT1 PoC

Proof of concept of one NOTED proposal:
successfully demonstrated the **possibility to off-load traffic over LHCONE when LHCOPN link to a Tier1 is congested**, by tweaking routing metrics

Routing tweaking could be automated and triggered by FTS signaling of on-going major transfer



Full report:
https://indico.cern.ch/event/725706/contributions/3169200/attachments/1744659/2824103/LHCOPNE-20181031-FNAL-CERN-NLT1-test.pdf

Conclusion

Summary

LHCOPN:
- Increasing utilization
- Increasing capacity: already 5 Tier1s connected at 100Gbps

LHCONE:
- Expanding in Asia-Pacific
- Working on improving reachability and cleanness

Future developments:
- Several on-going activities to improve network efficient utilization

More information

Next LHCOPN/ONE meeting:
Date: 4-5 of June 2019
Location: University of Umeå, Sweden
https://indico.cern.ch/event/772031/

Previous LHCOPN/ONE meetings:
RAL UK, March 2018: https://indico.cern.ch/event/681168/
FNAL US, October 2018: https://indico.cern.ch/event/725706/

Websites:
LHCOPN: https://twiki.cern.ch/twiki/bin/view/LHCOPN/WebHome
LHCONE: https://twiki.cern.ch/twiki/bin/view/LHCONE/WebHome

Questions?

edoardo.martelli@cern.ch