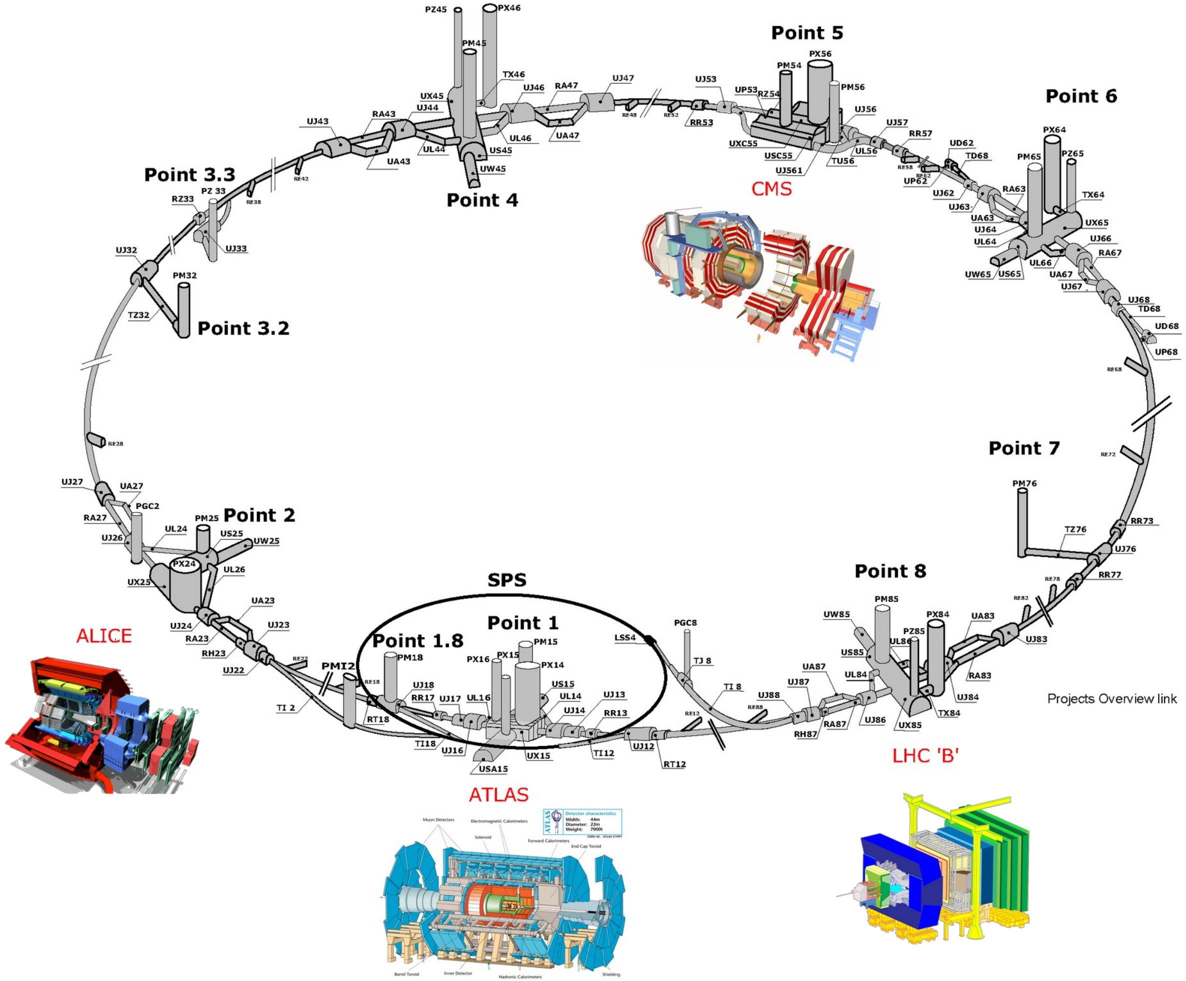
Evolution of LHC networking, future perspective

GEANT SIG-NGN Catania - 9th of April 2024

edoardo.martelli@cern.ch

Lately at CERN

The LHC and its experiments



The Large Hadron Collider (LHC) is the world's largest and most powerful particle accelerator.

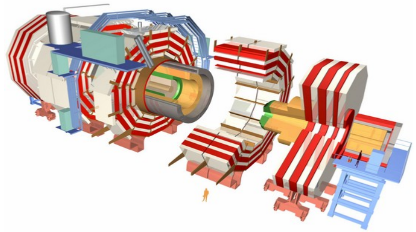
It first started up on 10 September 2008.

Beams inside the LHC are made to collide at four locations around the accelerator ring, corresponding to the positions of four particle detectors ALICE, ATLAS, CMS, and LHCb.

Data moving from Detectors to Computing…

Tier 0

Data source
Full data on Tape
Data reconstruction



Tier 1s

Distributed 2^{nd} copy on Tape
Simulations, Data analyses

Tier 2s

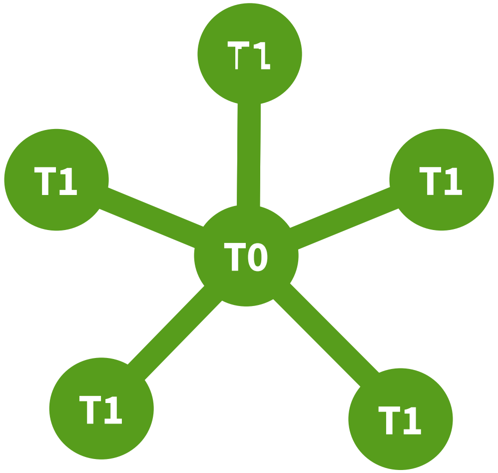
Data caches
Data analyses

using LHCOPN...

Private network connecting Tier0 and Tier1s
- Direct links from the Tier0 to all the Tier1s
- Dedicated to LHC data transfers

Secure:
- Only declared IP prefixes can exchange traffic
- Can connect directly to Science-DMZ at sites,
  to bypass slow perimeter firewalls

Advanced routing:
- BGP communities for traffic engineering



LHCOPN

...and LHCONE

Private network connecting Tier1s and Tier2s

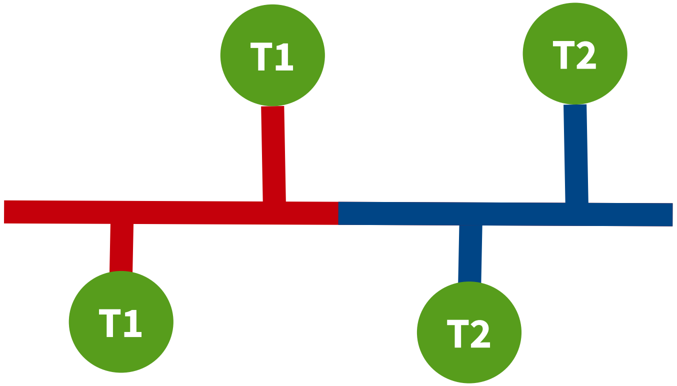
- Layer3 VPN implemented by National and International Research and Education Network operators
- Dedicated to LHC data transfers

Secure:

- Only allowed sites can exchange traffic
- Can connect directly to Science-DMZ at sites, to bypass slow perimeter firewalls

Advanced routing:

- Multi domain L3 VPN
- BGP communities for traffic engineering



LHC traffic keeps growing

Y-Axis: Gbps – Average yearly bandwidth in LHCOPN

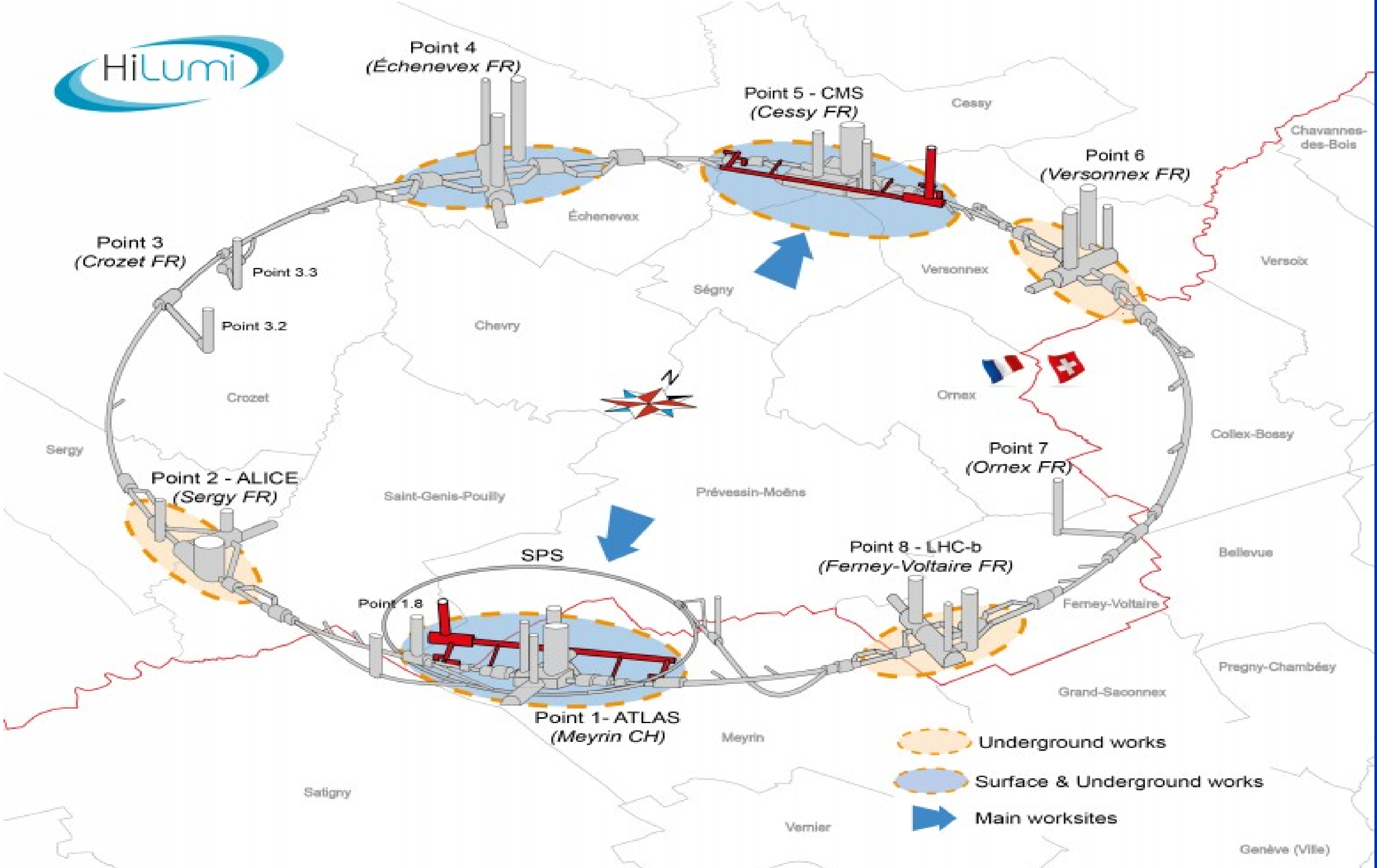
Ref: https://twiki.cern.ch/twiki/bin/view/LHCOPN/LhcopnStats

LHC runs and shutdowns:

Run1: 2010-12
LS1: 2013-14
Run2: 2015-18
LS2: 2019-21
Run3: 2022-25
LS3: 2026-28
Run4: 2029-32

What’s coming next?

The High Luminosity upgrade

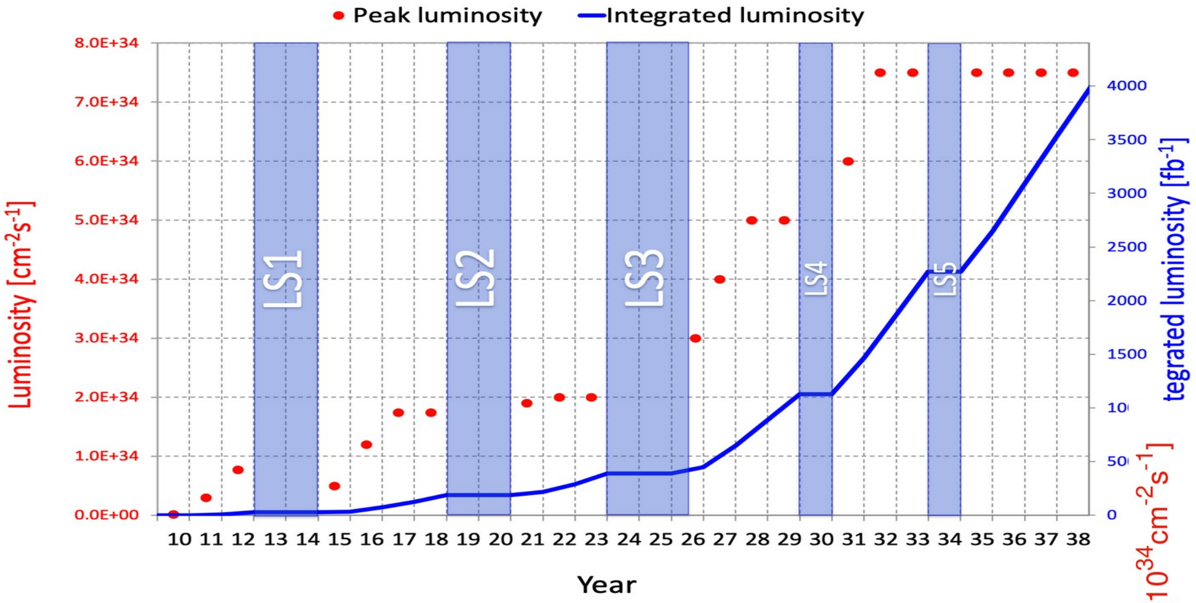


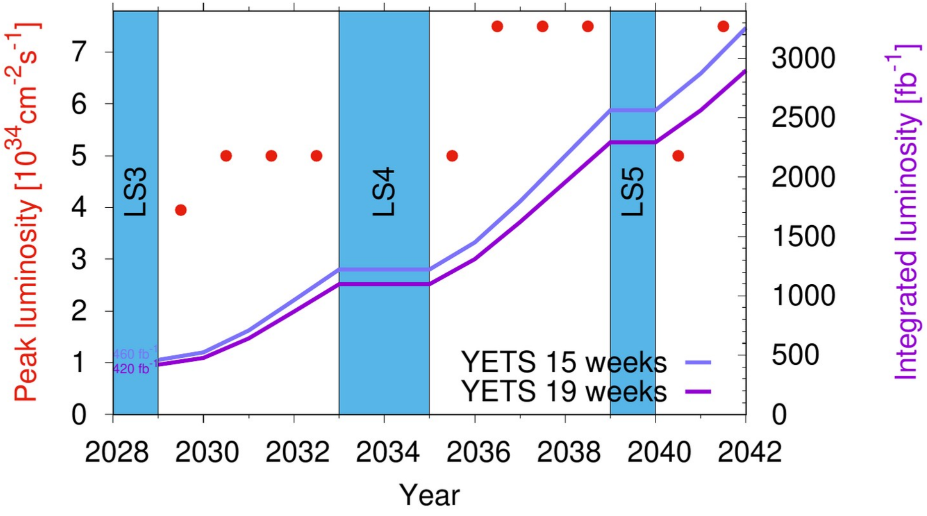
The High-Luminosity Large Hadron Collider (HL-LHC) is an **upgraded version of the LHC**

It will operate at a higher luminosity or, in other words, it will be able to produce more data

The HL-LHC will enter service in 2029, **increasing the volume of data** analysed by the experiments **by a factor of 10**

Increased Luminosity, and data production





https://lhc-commissioning.web.cern.ch/lhc-commissioning/schedule/HL-LHC-plots.htm

HL-LHC network requirements

ATLAS & CMS T0 to T1 per experiment

- 350PB RAW, taken and distributed during typical LHC uptime of 7M seconds (3 months)
  - 50GB/s or 400Gbps
- Another 100Gbps estimated for prompt reconstruction data tiers (AOD, other derived output)
- estiimated 1Tbps for CMS and ATLAS summed

ALICE & LHCb T0 Export

- 100 Gbps per experiment estimated from Run-3 rates

Minimal Model

- Sum (ATLAS,ALICE,CMS,LHCb)*2(for bursts)*2(overprovisioning) = **4.8Tbps expected HL-LHC bandwidth**

Flexible Model

- Assumes reading of data from above for reprocessing/reconstruction in 3 month
- Means doubling the Minimal Model: **9.6Tbps expected HL-LHC bandwidth**

Network requirements for HL-LHC

Tier1s:

- 1Tbps to the Tier0 (LHCOPN)
- 1 Tbps to the Tier2s (aggregated, LHCONE)

Tier2s

- 400 Gbps and more

Over provisioning main not always be an option. More efficient technology may be needed

How to get there: Data Challenges

2021: 10% of HL-LHC requirements - Done

2024: 25% of HL-LHC requirements - Done

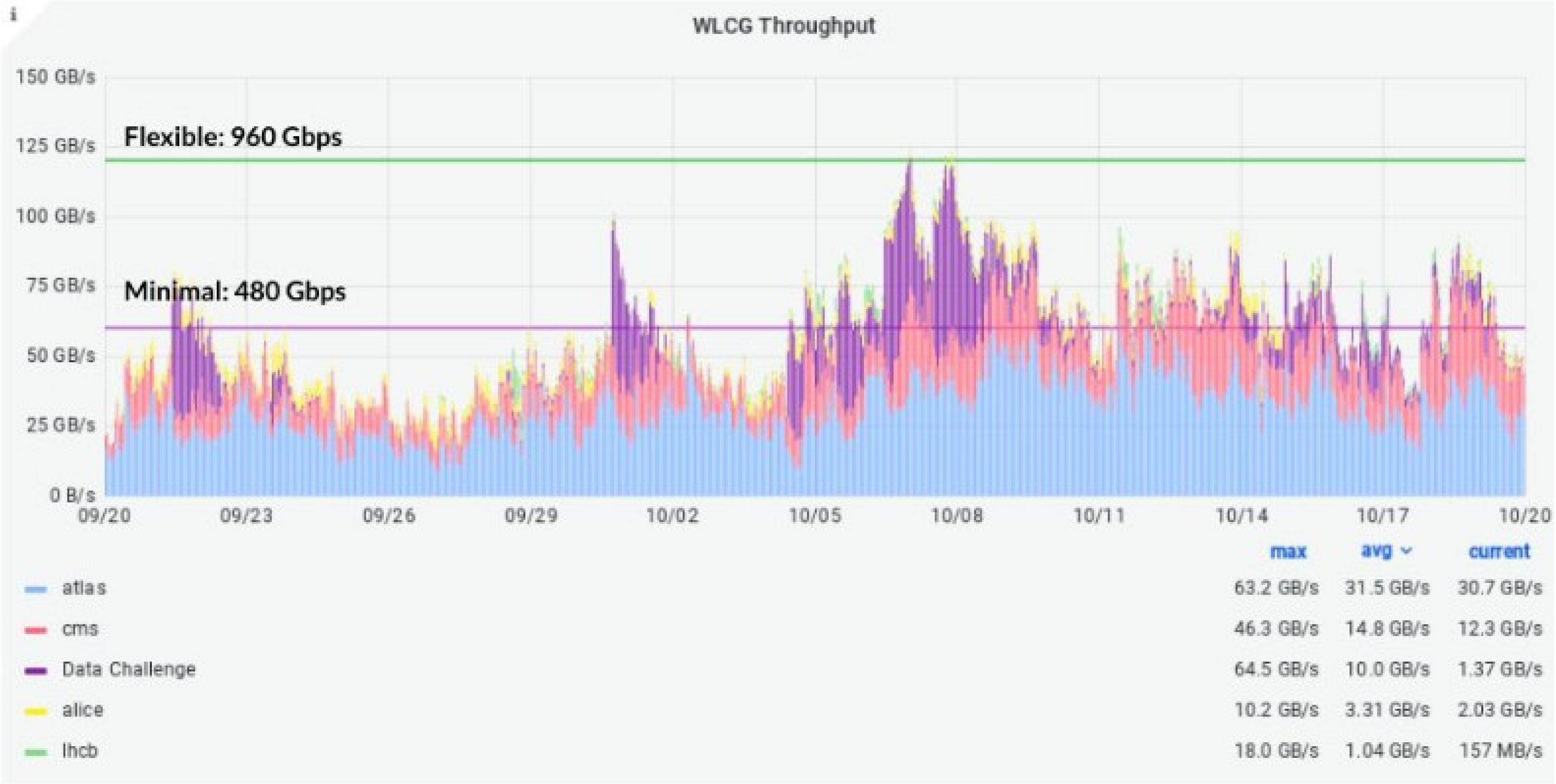
~2026: 50% of HL-LHC requirements

~2028: 100% of HL-LHC requirements

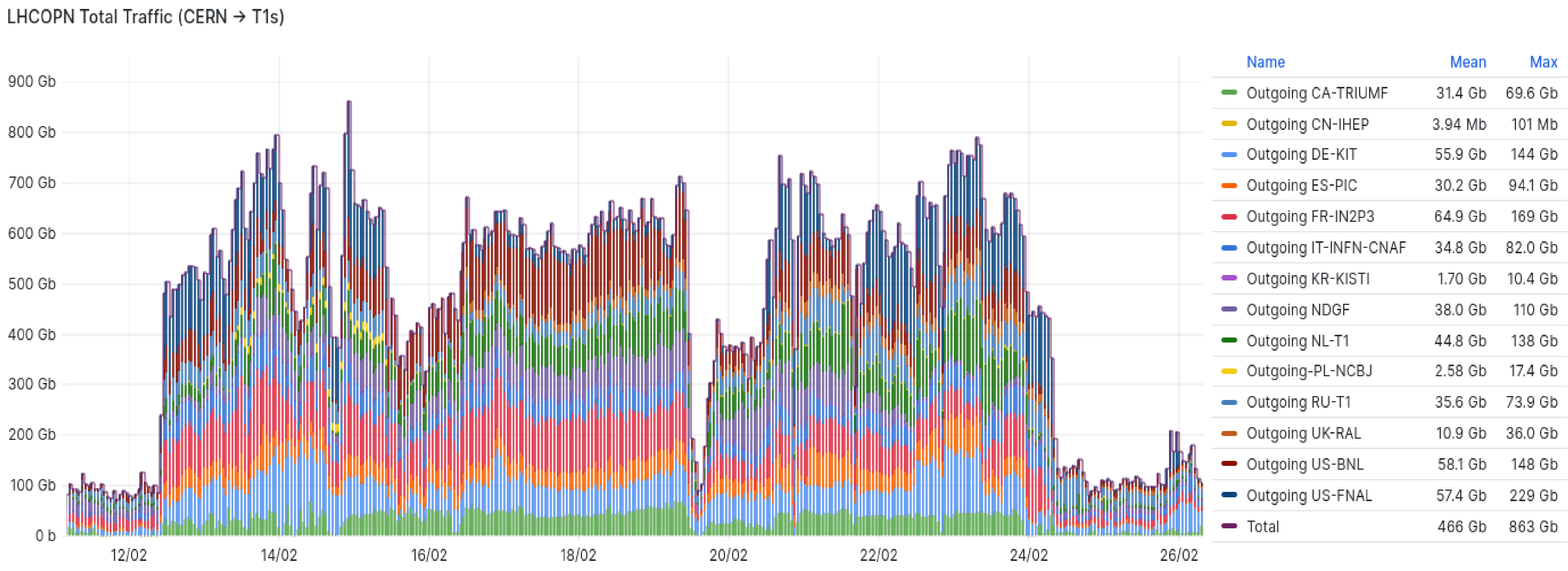
2029: start of HL-LHC (Run4)

DC21

Successfully reached the 10% minimal and flexible targets

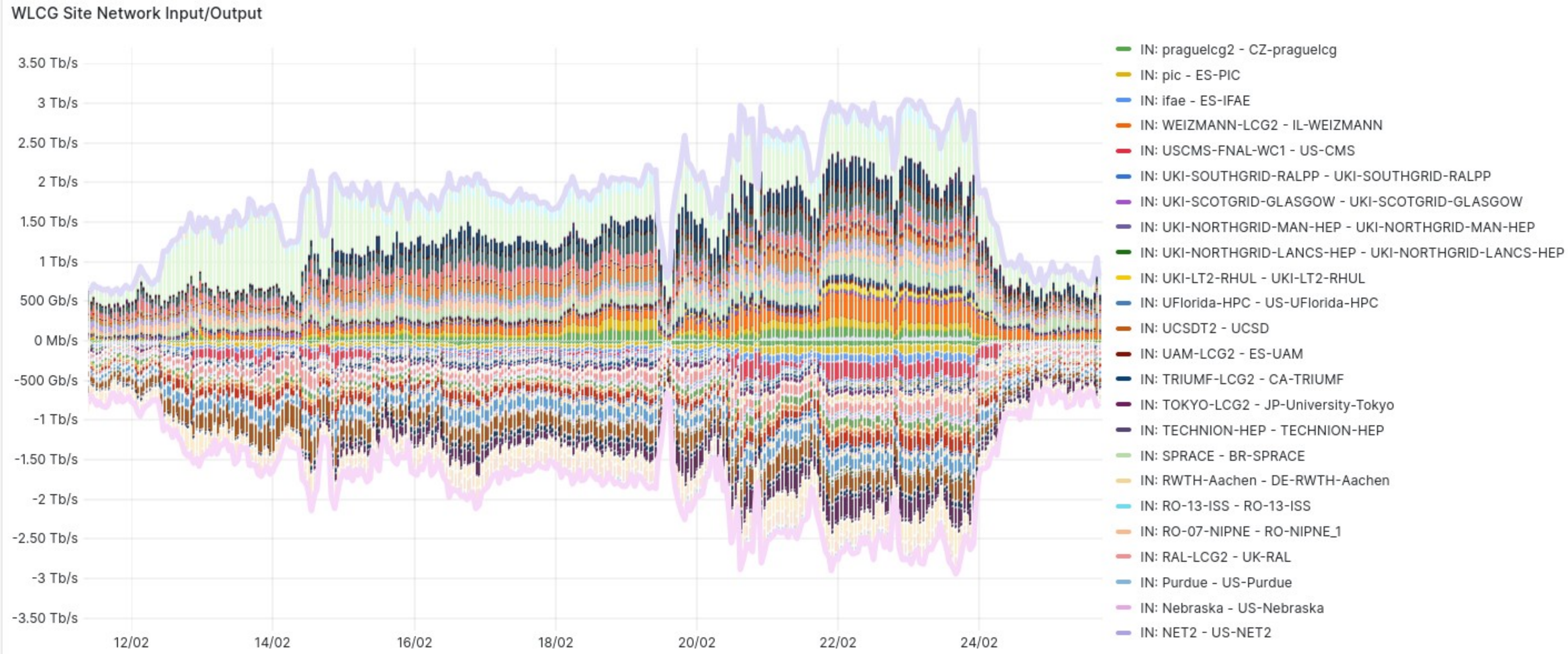


DC24: 800Gbps on LHCOPN



DC24: 3Tbps among WLCG sites

WLCG traffic exceeded 3Tbps



https://monit-grafana-open.cern.ch/d/MwuxgogIk/wlcg-site-network?from=1707638857216&orgId=16&to=1708880381142

WLCG guidelines

Message from Simone Campana, WLCG director:

In the next 10 years WLCG will be faced with two major network challenges:
- dealing with the HL-LHC data volumes and complexity
- the cohabitation with other experiments and sciences on the same infrastructure

The network community can play a leading role:
- modernize the network services, progressing with the ongoing R&D activities and bringing early prototypes in production
- engage with other experiments and sciences to drive the evolution of R&E networks

It’s not only HL-LHC

LHCONE is already used by other HEP collaborations

Belle II

Xe
XENON
Dark Matter Project

JUNO

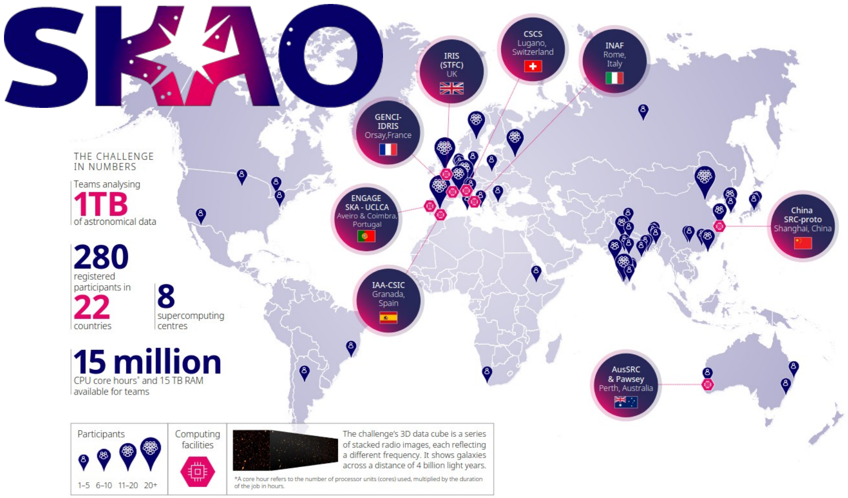
NOvA

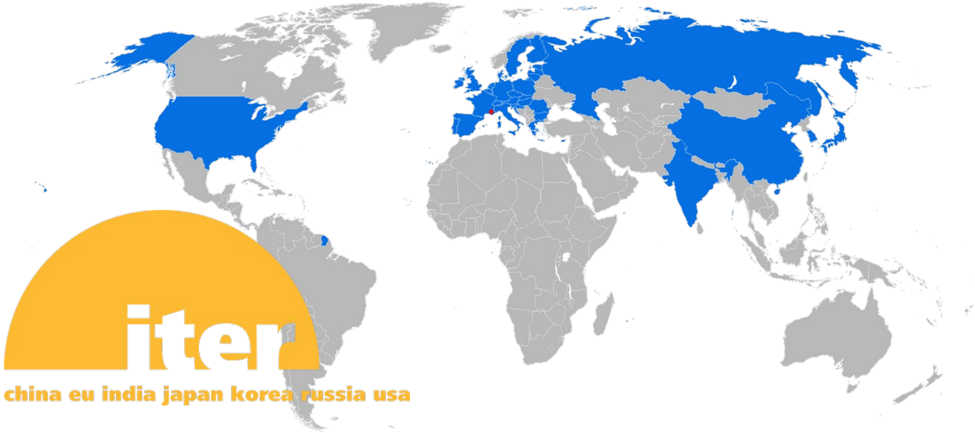
PIERRE AUGER OBSERVATORY

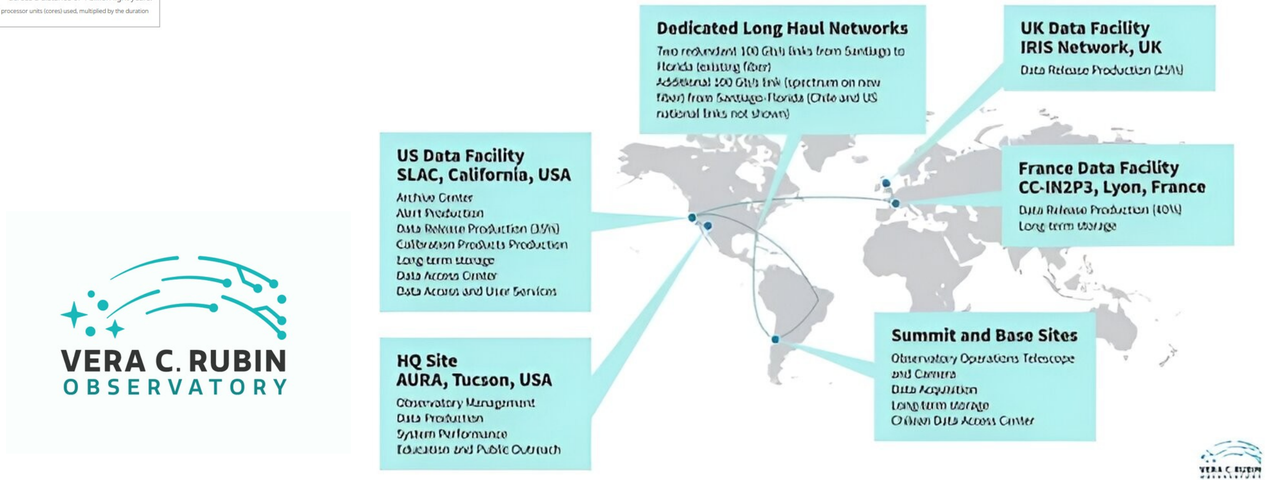
DUNE
DEEP UNDERGROUND
NEUTRINO EXPERIMENT

CERN

More Big Data sciences are coming on line







How are we getting ready?

Challenges

How to keep **high security** standards, while keeping very large data transfers at an affordable price?

Transoceanic bandwidth still subject to long cuts. Over-provisioning still expensive. How to reduce these bottlenecks?

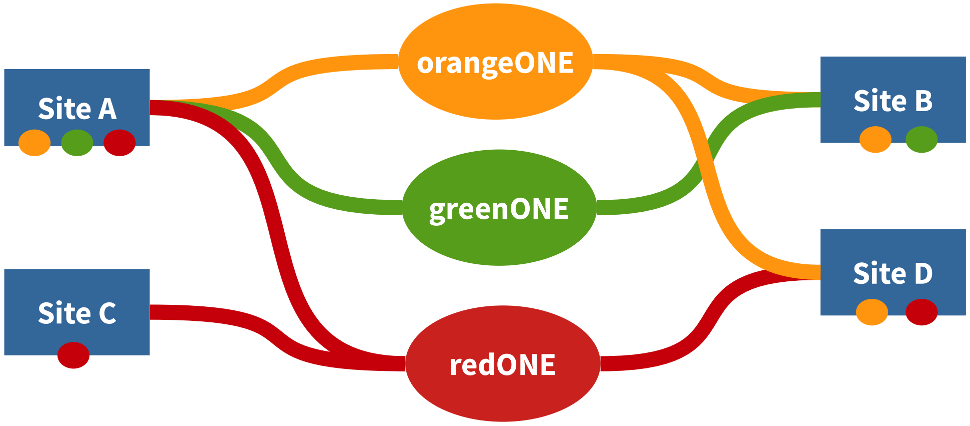
How to guarantee enough bandwidth for all sciences? Will we need some kind of **coordination**?

How to keep **sharing resources** in an increasingly divided world?

multiONE

LHCONE is already very large, it could become risky to include other large science projects.

MultiONE is a project which aims to keep the existing security for WLCG and any other Big Science project that may come

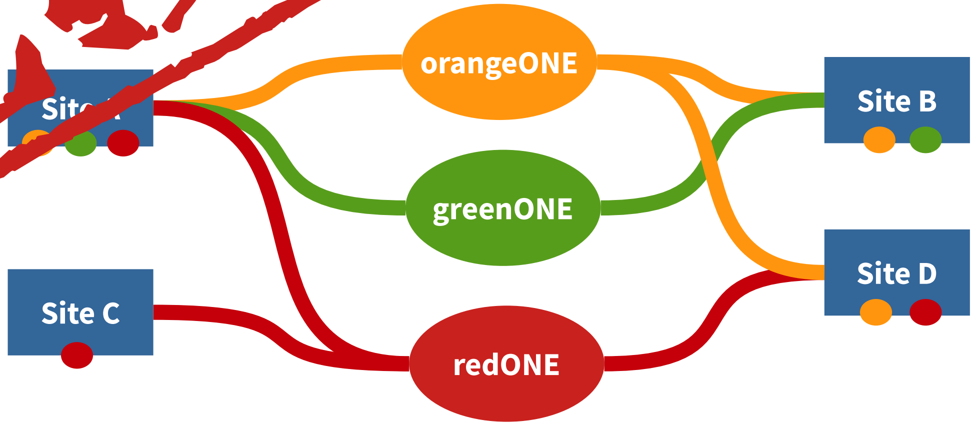


multiONE

LHCONE is already very large, it could become n[illegible]y to include other large science projects.

MultiONE is a project which aims to keep the existing security for WLCG and any other Big Science project that may come

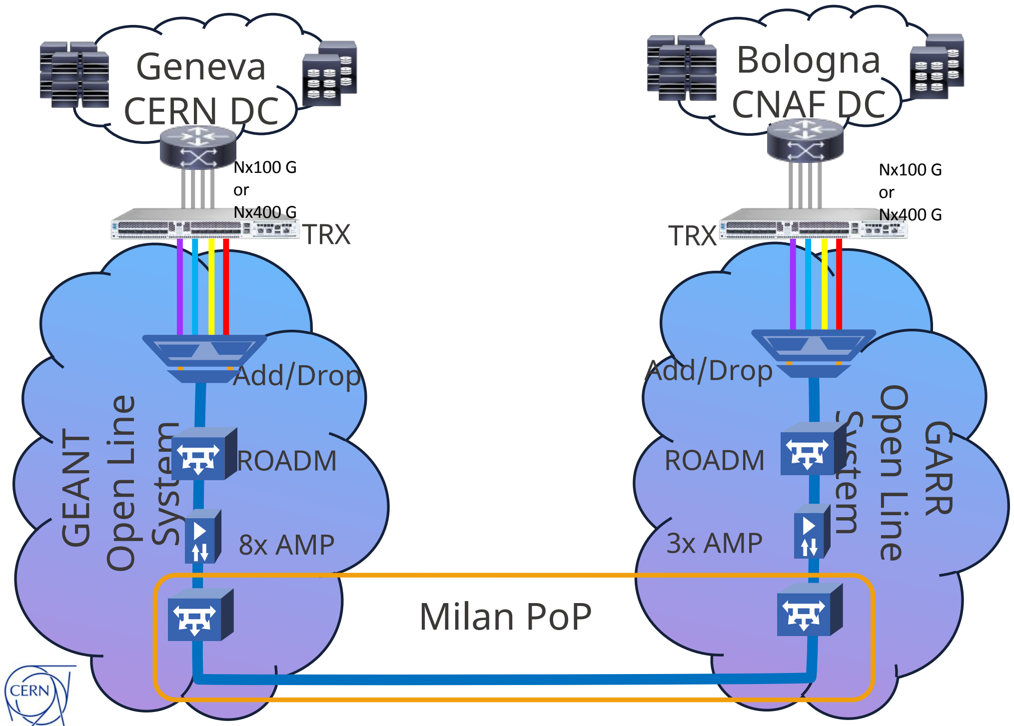
SECURITY

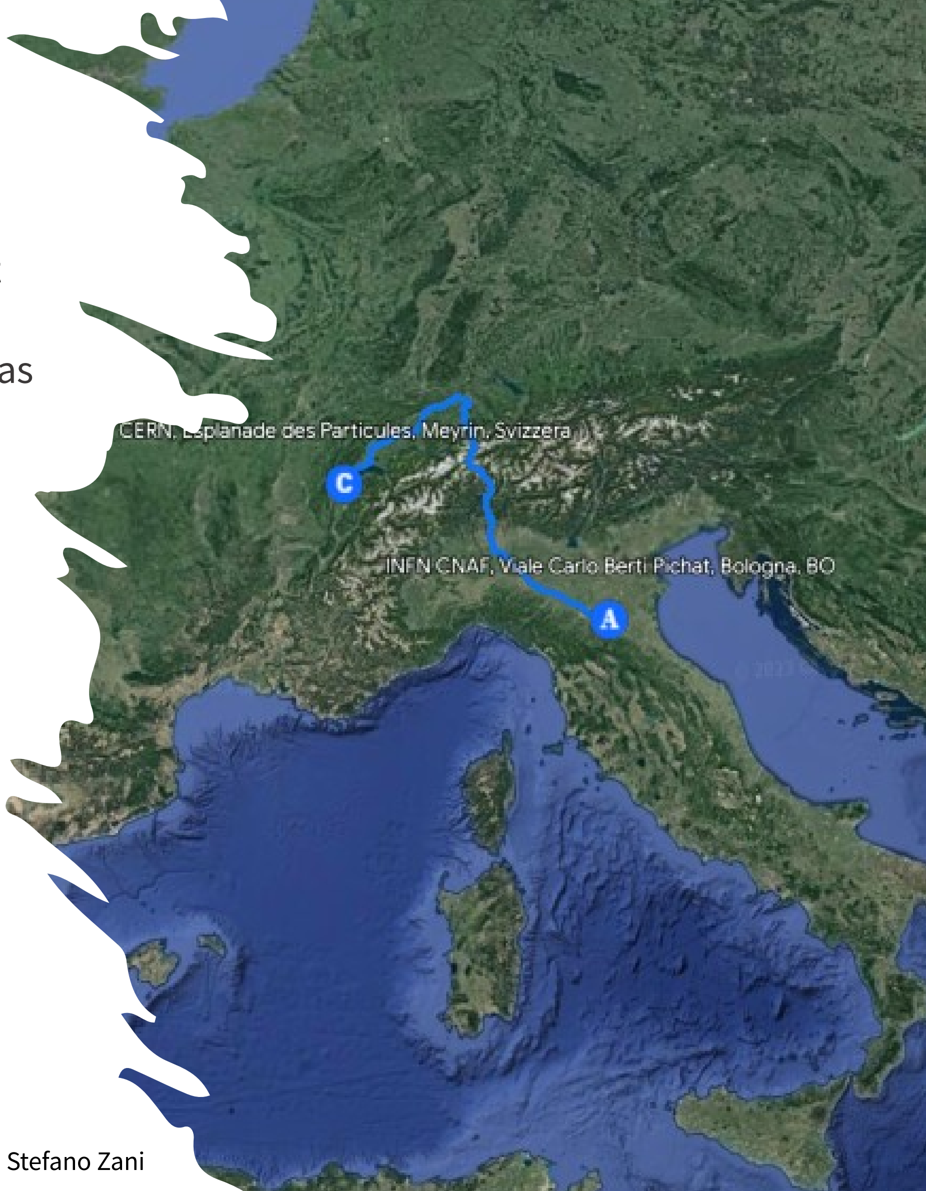


DCI on shared spectrum

Proposed in GEANT GN4-3 (WP7-T2) as a possible use case for experimenting the multi domain Spectrum Connection Service at about 1000 km of distance.

It is possible to reach 1.6 Tbps on this «Circuit» that could be used as up to 4x400Gb Ethernet or 16x100Gb Ethernet



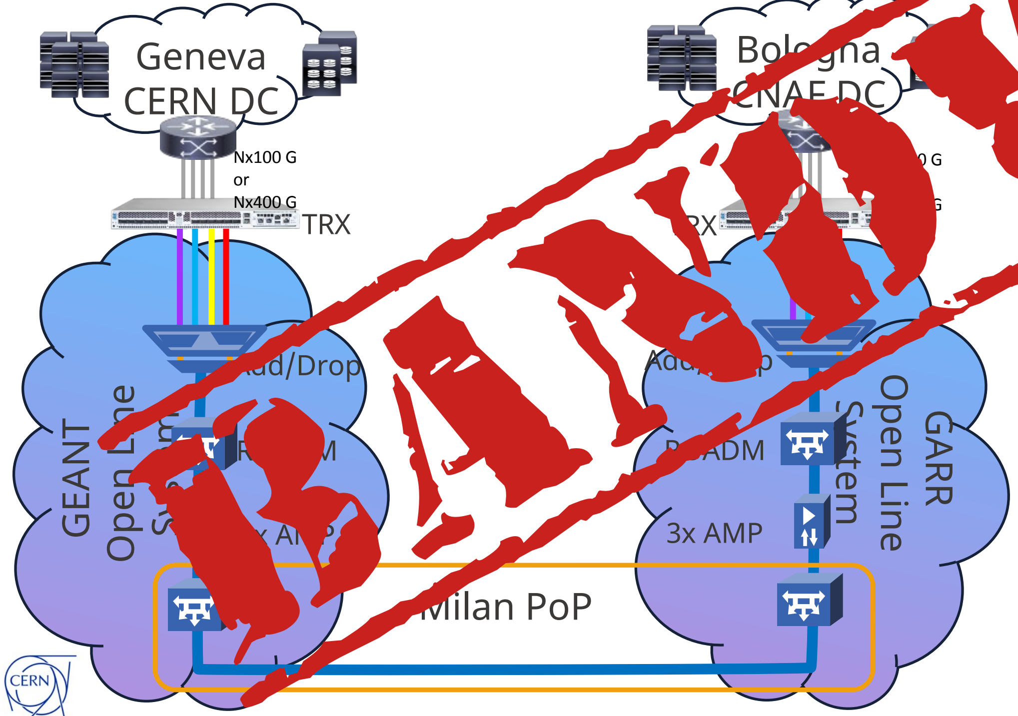


Credits: Stefano Zani

DCI on shared spectrum

Proposed in GEANT GN4-3 (WP7-T2) as a possible use case for experimenting the multi domain Spectrum Connection Service at about 1000 km of distance.

It is possible to reach 1.6 Tbps on this «Circuit» that could be used up to 4x400Gb Ethernet or 16x100Gb Ethernet



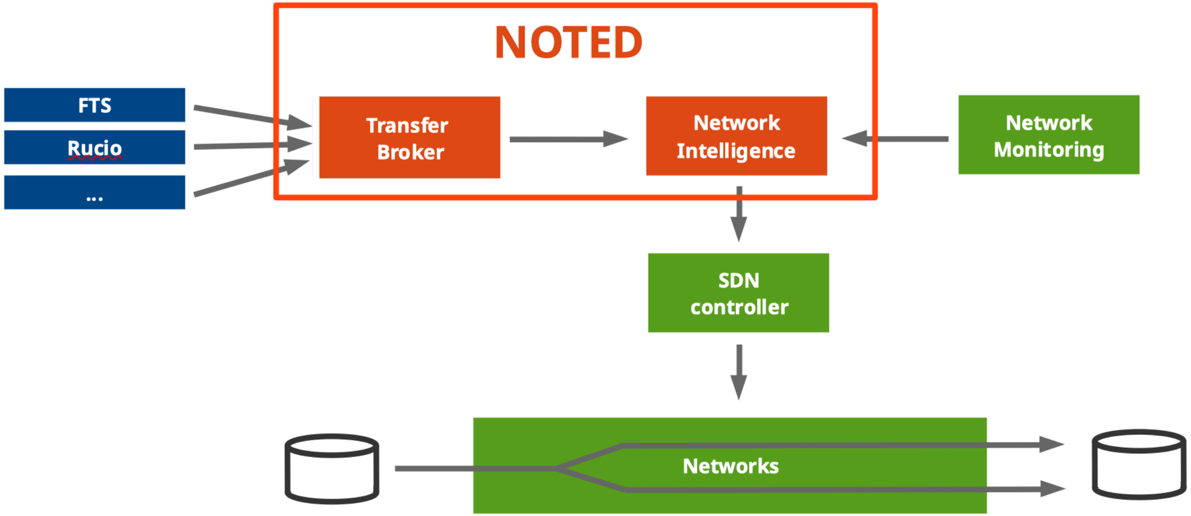
BANDWIDTH

Credits: Stefano Zani

NOTED SDN

NOTED is a framework that can detect large FTS data transfers and trigger network optimization actions to speed up the execution of the transfers

It can make a more efficient use of idle network connections

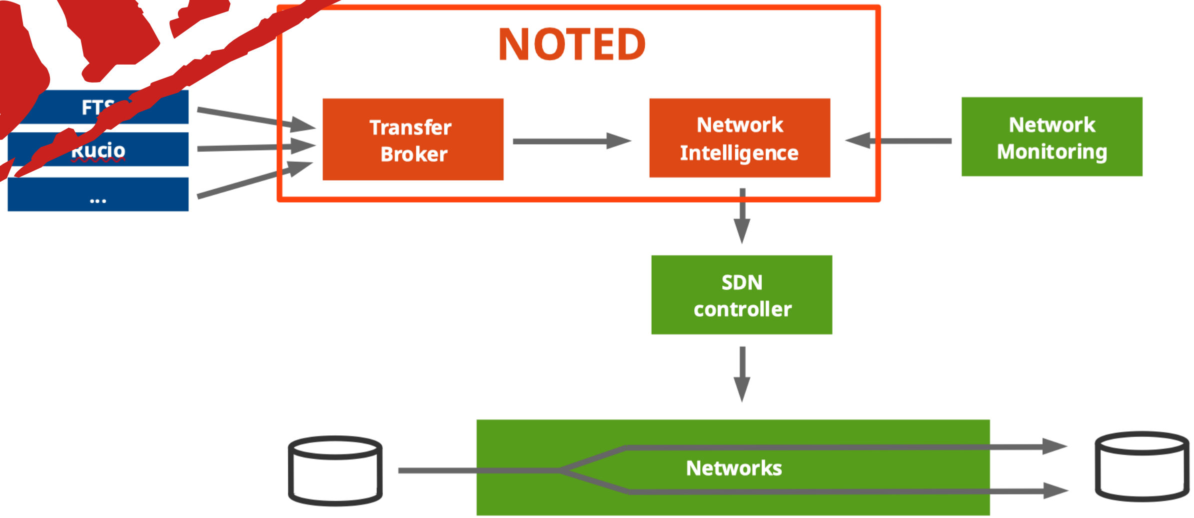


NOTED SDN

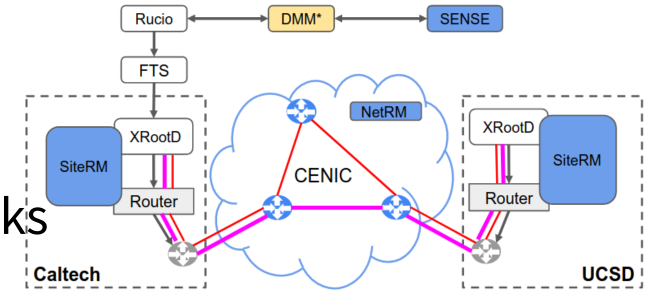
NOTED is a framework that can detect large FTS data transfers and trigger network optimization actions to speed up the execution of the transfers

It can make a more efficient use of idle network connections.

CONGESTION



Using SENSE to move CMS data in Rucio



Project led by UCSD and Caltech

The increased requirements of the HL-LHC requires to use any resource in the most efficient way, including networks

Objectives of the project:

#1 Make Rucio capable to schedule transfers on the network and prioritize them

#2 Predetermined transfer speed and quality of service (time to completion)

Demonstrated:

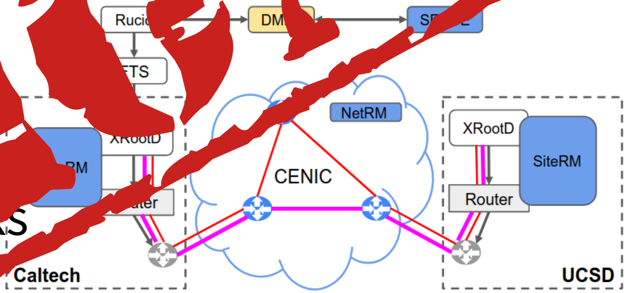
- SENSE can build VPNs between pairs of XrootD servers in charge of FTS transfers requested by Rucio

- QoS can be provisioned in the network to prioritize the traffic in the VPN

Using SENSE to move CMS data in Rucio

Project led by UCSD and Caltech

The increased requirements of the HL-LHC requires to use any resource in the most efficient way including networks



Objectives of the project:

#1 Make Rucio capable to schedule transfers on the network and prioritize them

#2 Predetermined transfer speed and quality of service (time to completion)

Demonstrated:

- SENSE can build VPNs between pairs of XrootD servers in charge of FTS transfers requested by Rucio

- QoS can be provisioned in the network to prioritize the traffic in the VPN

CONSISTENT

Possible use of data caches

Storage cache allows data sharing among users in the same region

- Reduce the redundant data transfers over the wide-area network
- Decrease data access latency
- Increase data access throughput
- Improve overall application performance

Pilot: Southern California Petabyte Scale Cache (SoCal Repo)

- Nodes at UCSD, Caltech, LBNL (RTT between 3 and 10ms)
- It could serve about 67.6% of files from its disk cache, while only 35.4% of bytes requested could be served from the cache
- During the period where fewer large files were requested (3/2022 – 5/2022), the network traffic was reduced by about 29TB per day

Possible use of data caches

Storage cache allows data sharing among users in the sam[illegible] regi[illegible]
- Reduce the redundant data transfers over the wide-area networ[illegible]
- Decrease data access latency
- Increase data access throughput
- Improve overall application performanc[illegible]

CONGESTION

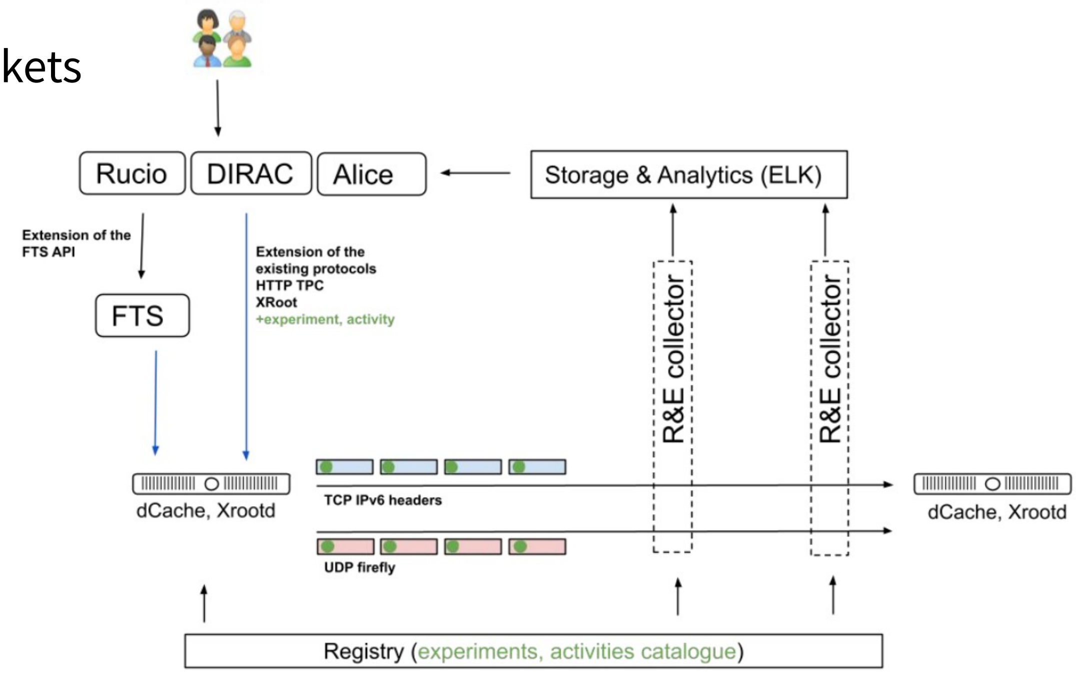
Pilot: Southern Californ[illegible]a Pe[illegible]abyte Sc[illegible] Ca[illegible]he (S[illegible]Cal Repo)
- Nodes at UCSD, Calte[illegible], LBN[illegible] (RT[illegible] netw[illegible] 3 a[illegible] 10m[illegible])
- It could serve a[illegible]out [illegible] f fi[illegible]es [illegible] its [illegible]sk cac[illegible]e, while only 35.4% of bytes requested could be served [illegible]om t[illegible] cache
- Duri[illegible]g the p[illegible]riod [illegible]re [illegible]wer [illegible]ge fil[illegible]s were requested (3/2022 – 5/2022), the network traffic was red[illegible]d [illegible] bou[illegible] [illegible]TB [illegible]r day

Packet Marking

Marking of data packets/flows with Experiments and Applications IDs for better accounting

Two options being investigated:

- Tag in the IPv6 flowlabel field
- Tag (and more) in UDP fireflies (UDP packets sent in parallel to each flow)



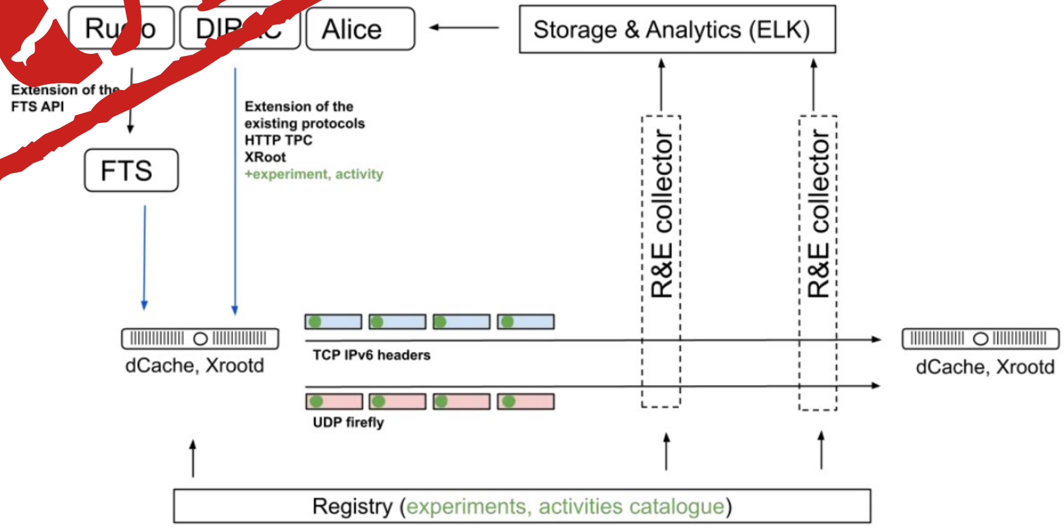
Packet Marking

Marking of data packets/flows with Experiments and Applications IDs for better accounting

Two options being investigated:

- Tag in the IPv6 flowlabel field
- Tag (and more) in UDP fireflies (UDP packets sent in parallel to each flow)

ACCOUNTING



Packet Pacing

A small amount of packet loss makes a huge difference in TCP performance, especially on long distance flows

TCP can send packets in burst. These burst can be a problem in case of:

- Shallow switch buffers
- Slower receivers
- Speed mismatch on the path

Goal of pacing is to limit the burst rate of a TCP flow

BBR TCP congestion protocol has built-in pacing (transmit based on a clock, not ACKs)

Packet Pacing

A small amount of packet loss makes a huge difference in TCP performance, especially on long distance flows

TCP can send packets in burst. These burst can be a problem in case of:

- Shallow switch buffers
- Slower receivers
- Speed mismatch on the path

Goal of pacing is to limit the burst rate of a TCP flow

BBR TCP congestion protocol has built-in pacing (transmit based on a clock, not ACKs)

CONGESTION

Conclusions

Conclusions

Networks more and more essential for big-data science projects. Demands will keep growing (out of control?)

Over-provisioning is simple, but may become too expensive (or maybe not)

Extended visibility and accounting is essential

What about application driven network automation?

Security at Tbps scale is one of the biggest concerns

Comments?

edoardo.martelli@cern.ch