LHC-OPN LHC-ONE
Note della riunione

https://indico.cern.ch/event/1234127/

FZU, Praga

18-19 Aprile 2023

LHC OPN Update

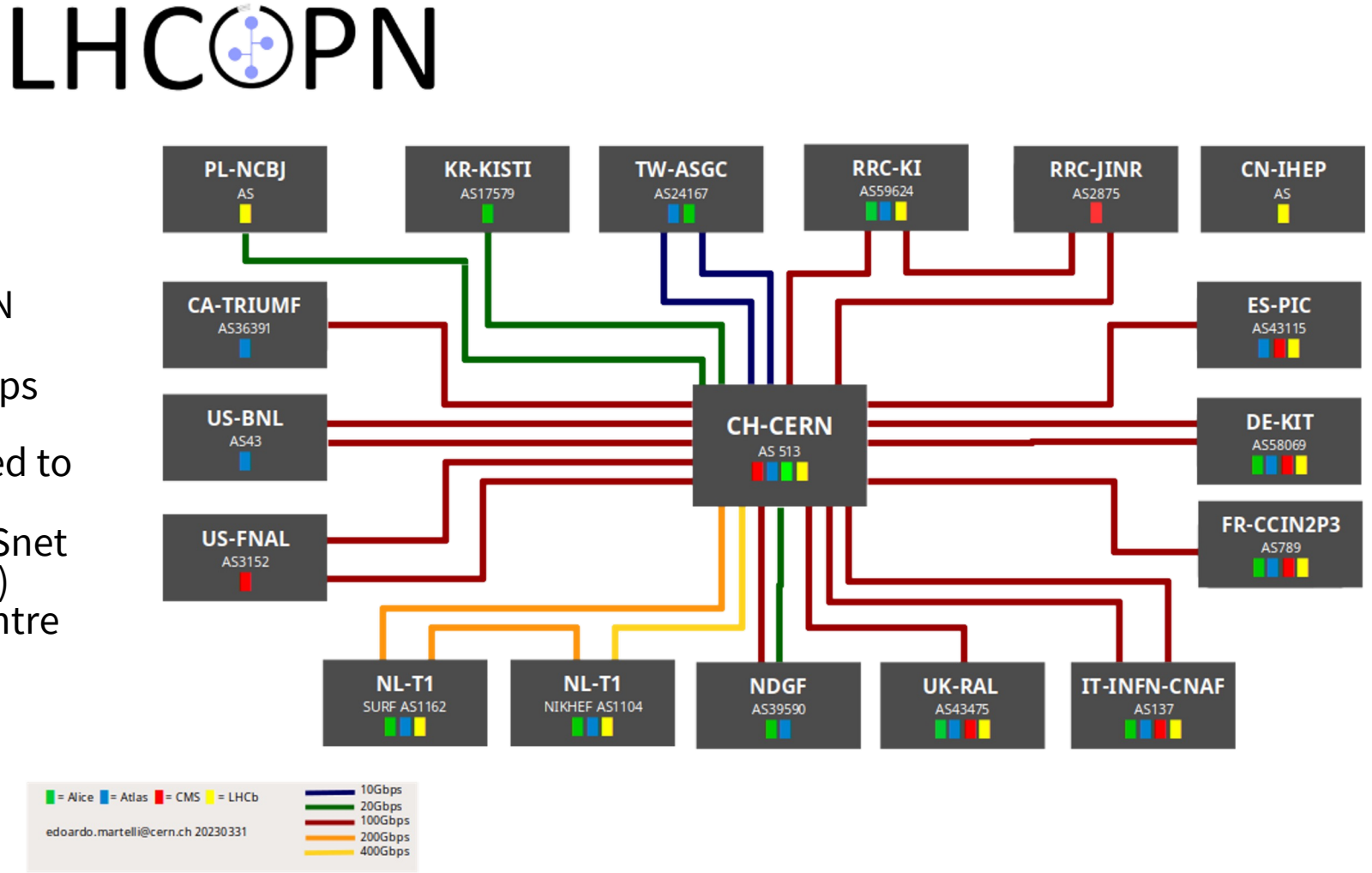
Nuovi T1 in arrivo

- Polonia
- Cina

TW-ASGC no longer Tier1. Effective from October 2023

CH-CERN:

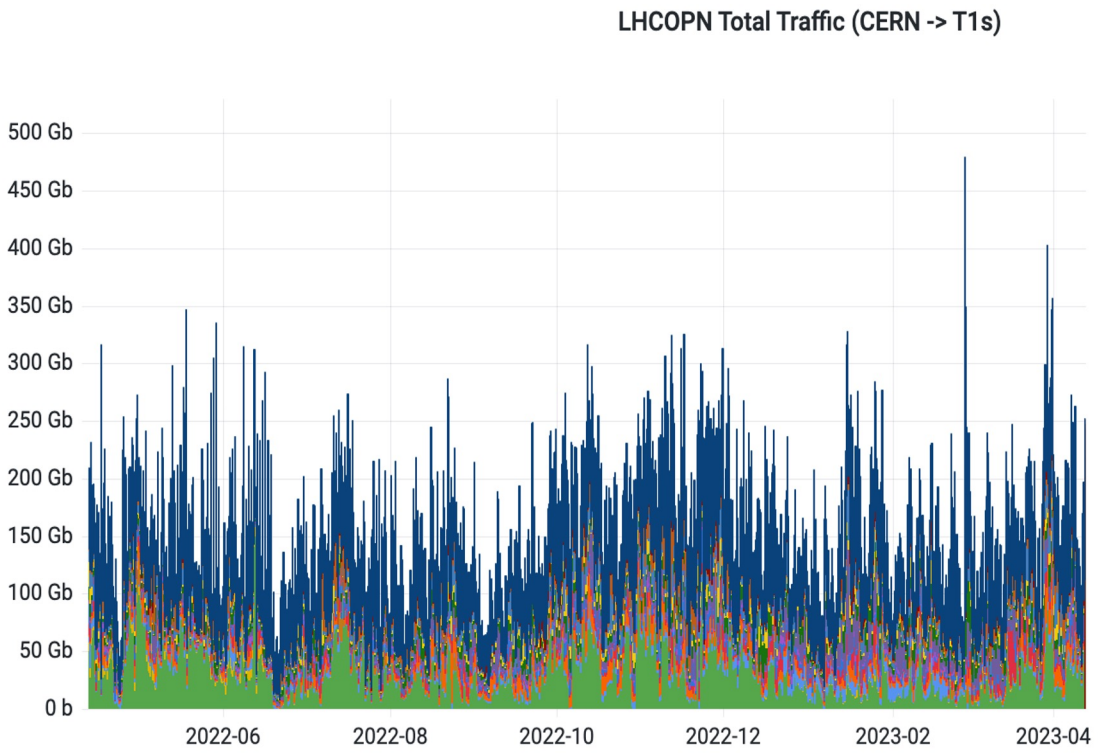
- Completed migration of second LHCOPN router from Juniper QFX-10002 to Juniper PTX-10001, to support 400Gbps connections
- LHCONE connections to GEANT upgraded to 2x 400Gbps in November
- Requested upgrade of connections to ESnet to 2x 400Gbps (LHCOPN and LHCONE)
- Construction of Prevessin Computer Centre is progressing well



LHC-OPN Traffic

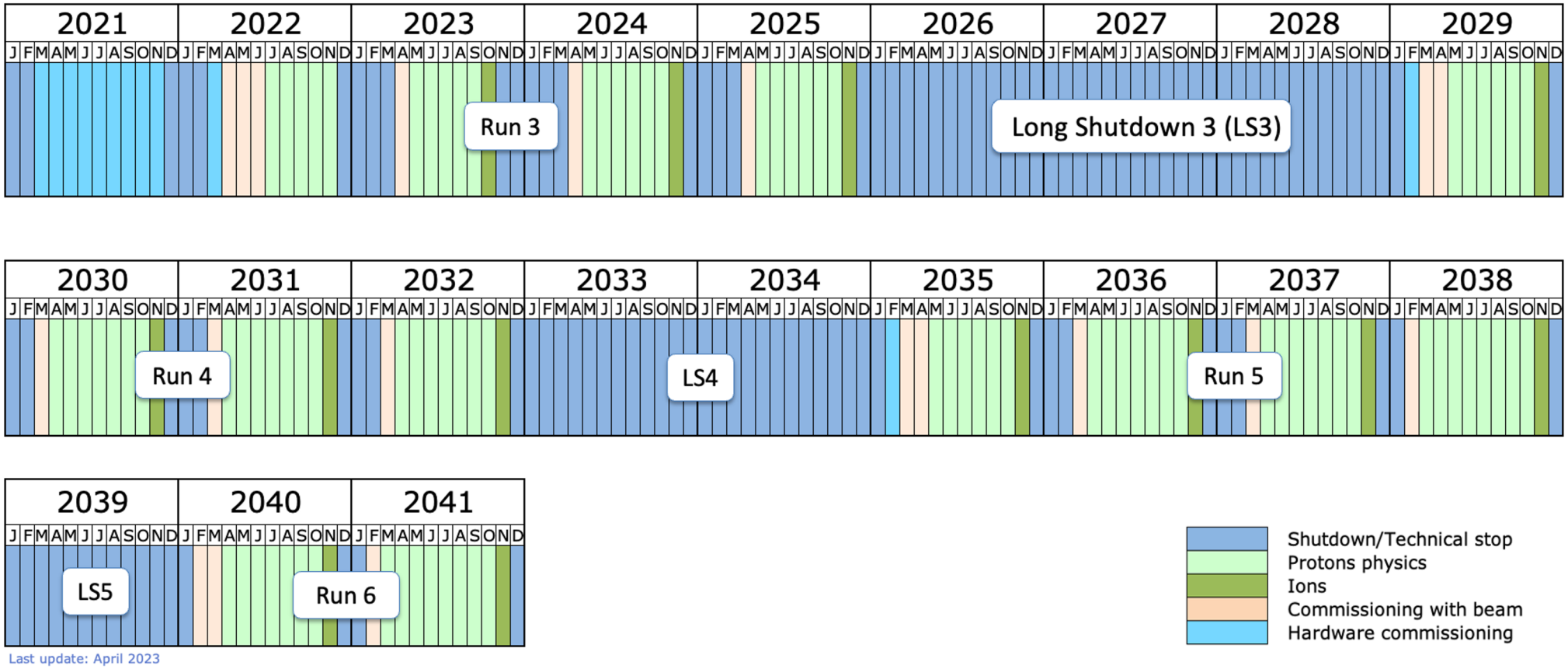
- Moved ~488 PB in the last 12 months
- +12% compared to previous year (433PB)
- Peak at ~479Gbps

Noi siamo il secondo T1 per traffico medio ed anche come valore di picco (127Gbps)



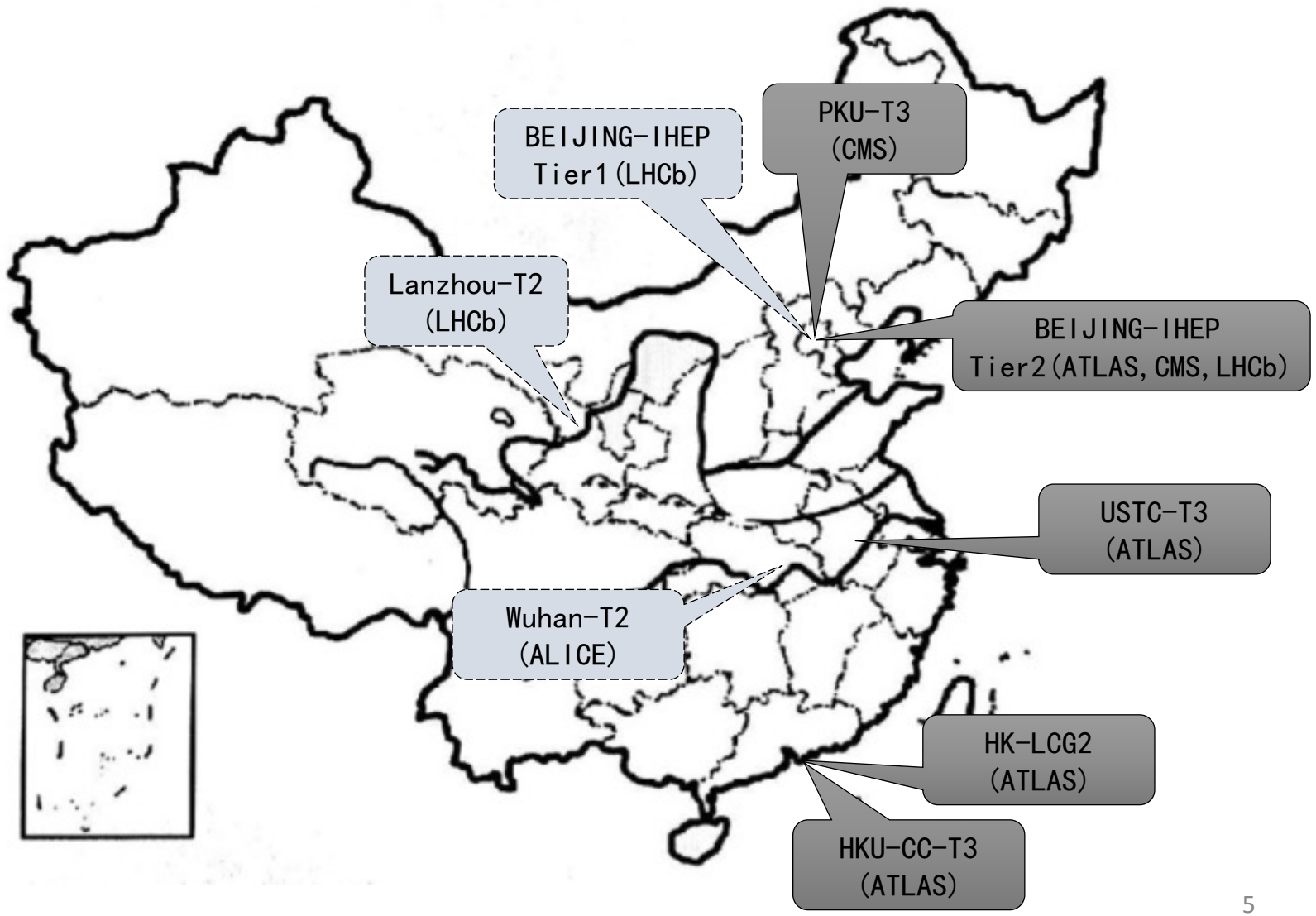
	Mean	Max
Outgoing DE-KIT	49.9 Gb	251 Gb
Outgoing KR-KISTI	872 Mb	11.4 Gb
Outgoing RU-T1	5.34 Gb	46.0 Gb
Outgoing FR-IN2P3	10.4 Gb	86.7 Gb
Outgoing NDGF	6.54 Gb	83.3 Gb
Outgoing NL-T1	6.46 Gb	94.4 Gb
Outgoing TW-ASGC	1.71 Gb	9.69 Gb
Outgoing IT-INFN-CNAF	11.5 Gb	127 Gb
Outgoing UK-RAL	7.28 Gb	38.8 Gb
Outgoing CA-TRIUMF	4.96 Gb	75.7 Gb
Outgoing US-BNL	10.1 Gb	87.7 Gb
Outgoing US-FNAL	5.76 Gb	82.3 Gb
Outgoing ES-PIC	2.91 Gb	54.3 Gb
Total	124 Gb	479 Gb

LHC Schduele



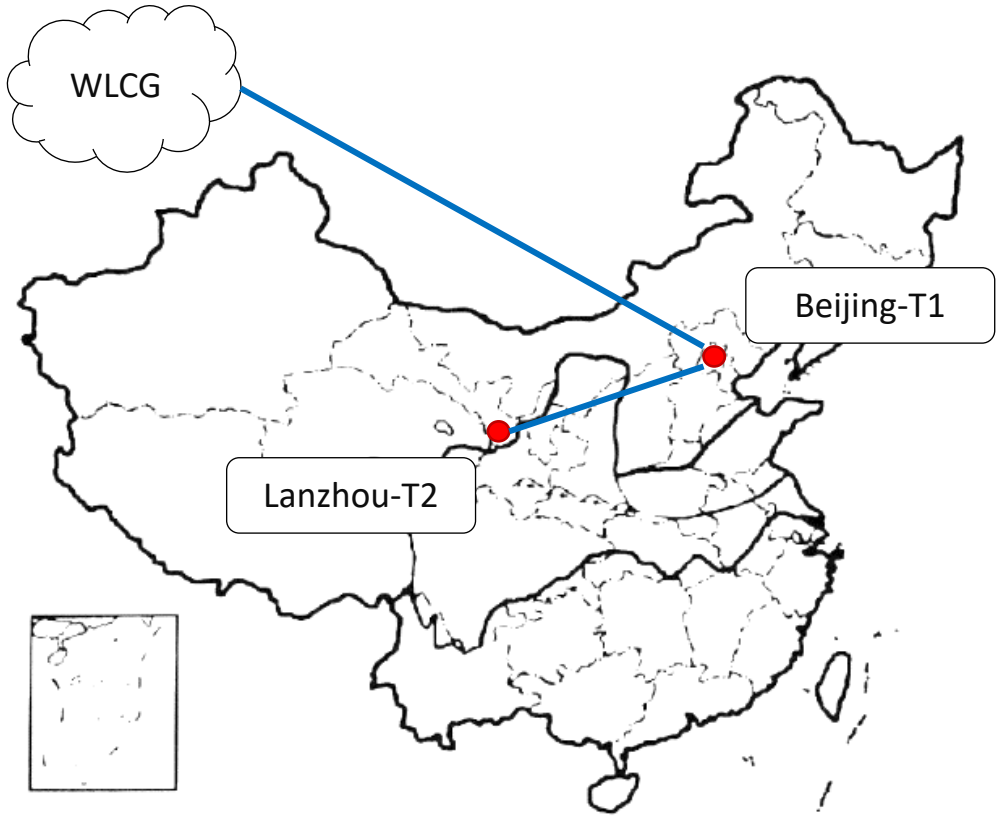
IHEP
Current Status of WLCG in China-Mainland

- Tier2 sites
  - BEIJING-IHEP
  - Atlas CMS and LHCb
- Tier3 sites
  - PKU-T3,USTC-T3
- Sites under developing
  - Tier1: IHEP-LHCb-T1
  - Tier2: Lanzhou-T2
  - Tier2: Wuhan-T2



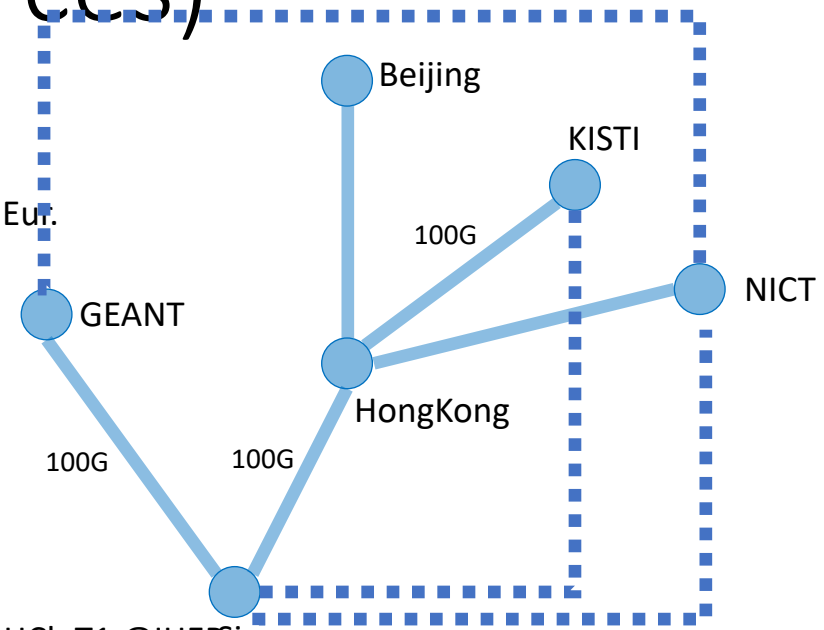
Proposed LHCb Tier1 Site @IHEP

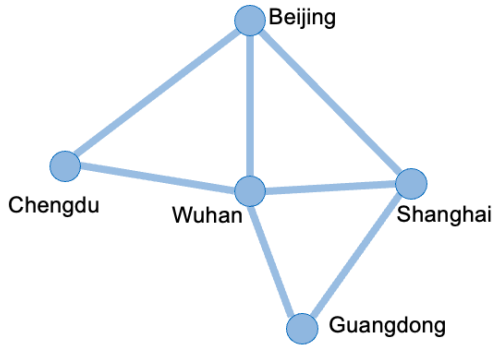
- The LHCb Tier-2 site at IHEP is proposed to be upgraded to a Tier-1 site.
- A new LHCb Tier-2 site will be built at Lanzhou University (LZU).
- Add all new Tier-1/2 sites into CN-IHEP Federation for WLCG (or changes to a new name?).
  - All the WLCG sites in China-mainland are supported by IHEPCC
  - CSTNet willing to be a member of the Federation

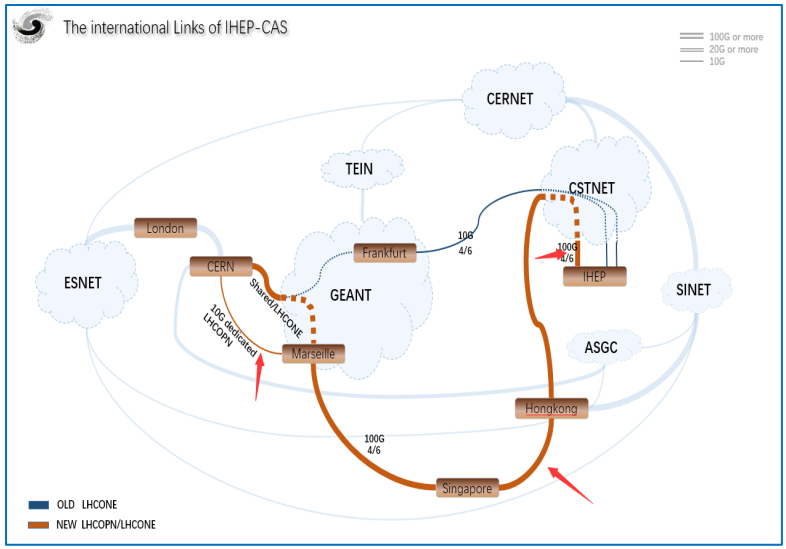


Resources (network resources)

- CSTNet is the internet service provider for IHEP
- International links
  - New connection will be launched to improve the bandwidth Between China and Eur.
  - CSTNet **– Beijing –Hongkong** – Singapore – Marseille - Geant
  - All the connections will be 100Gbps at the end of Apr.
- LHCONE
  - Old ： IHEP – CSTNET – GEANT(10G) – Frankfurt – CERN
  - New: IHEP – CSTNET – Singapore – Marseille – CERN
- LHCOPN
  - A new deliciated link will be launched between CERN and GEANT(Marseille) for LHCb T1 @IHEP
  - IHEP – CSTNET – Singapore – Marseille – CERN, hopefully be ready at the end of May.
- Domestic links
  - All the domestic connections will be upgraded from 10G to 100G
  - Ready at the end of 2023

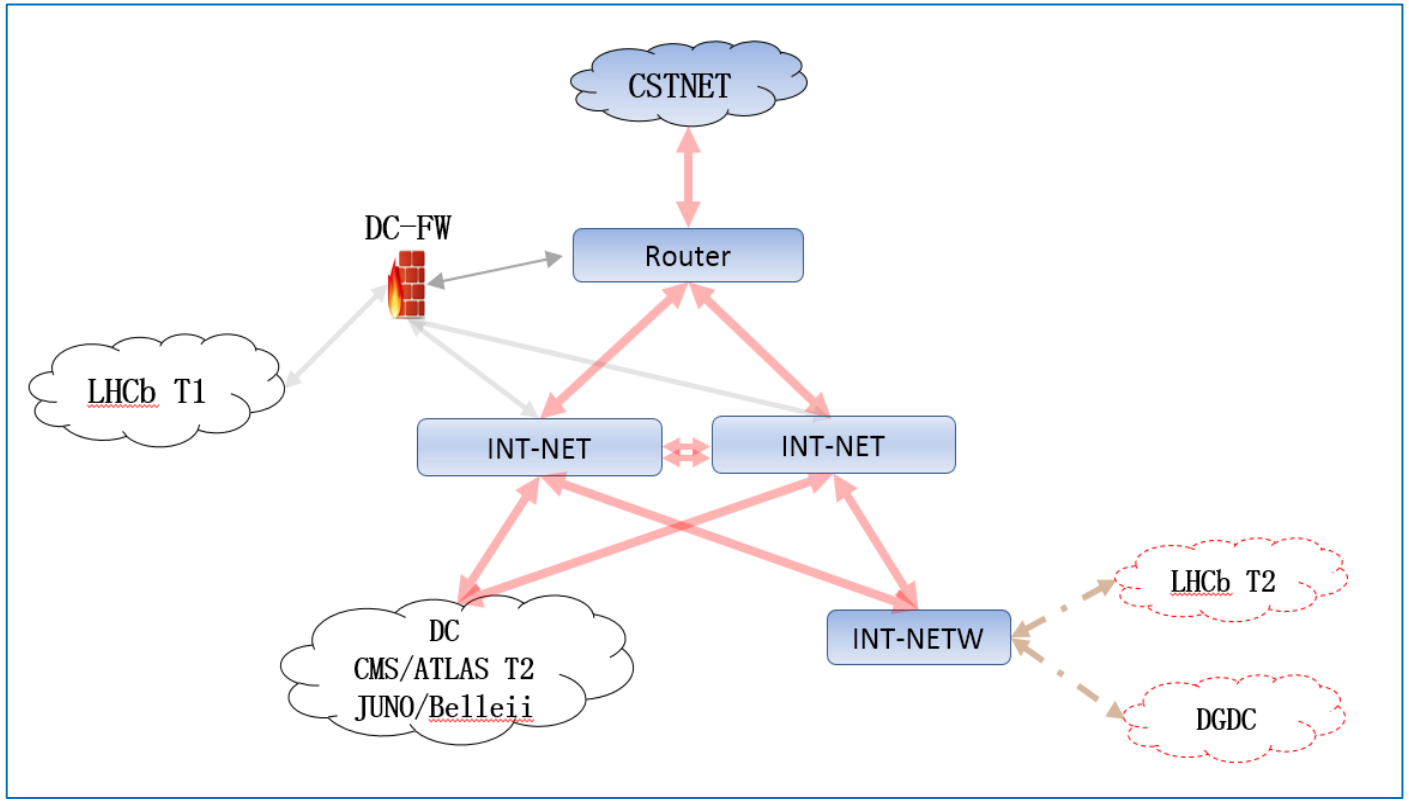






Resources (local network @IHEP)

- Current Status
  - IHEP to CSTNET
    - 100G, Dual Stack, Ready
  - Local Backbone
    - INT-NET to Router 2*10G -> 2*100G
    - DC to INT-NET 2*40G -> 2*100G
    - Ready at the end of May
  - Lanzhou LHCb Tier 2
    - LZU to IHEP 2Gb ready
  - DGDC
    - 10G ready



iHEP Summary

- LHCONE link will be upgraded to 100G at the end of Apr.
- LHCOPN link to CERN will be ready at the end of May
- IHEP LHCb Tier 1 site will be ready at the end of June
- New challenges not only for computing and storage, but also for network, to deploy and maintain the new T1 site

Proto Poland TIER1 NCBJ-CIS
LHCB ed in futuro anche CMS

- Non lontano da Varsavia
- 100G con PIONEER (Academic, Internet, GEANT)
  - 20Gb LHCONE VLAN

Next Steps

- 20Gb LHCOPN VLAN (On Going during 2023
- Support of iPv6

Resources 2024

7200-105000 kHS06 in 2024

Splitting between IPv4 ed IPv6 traffic

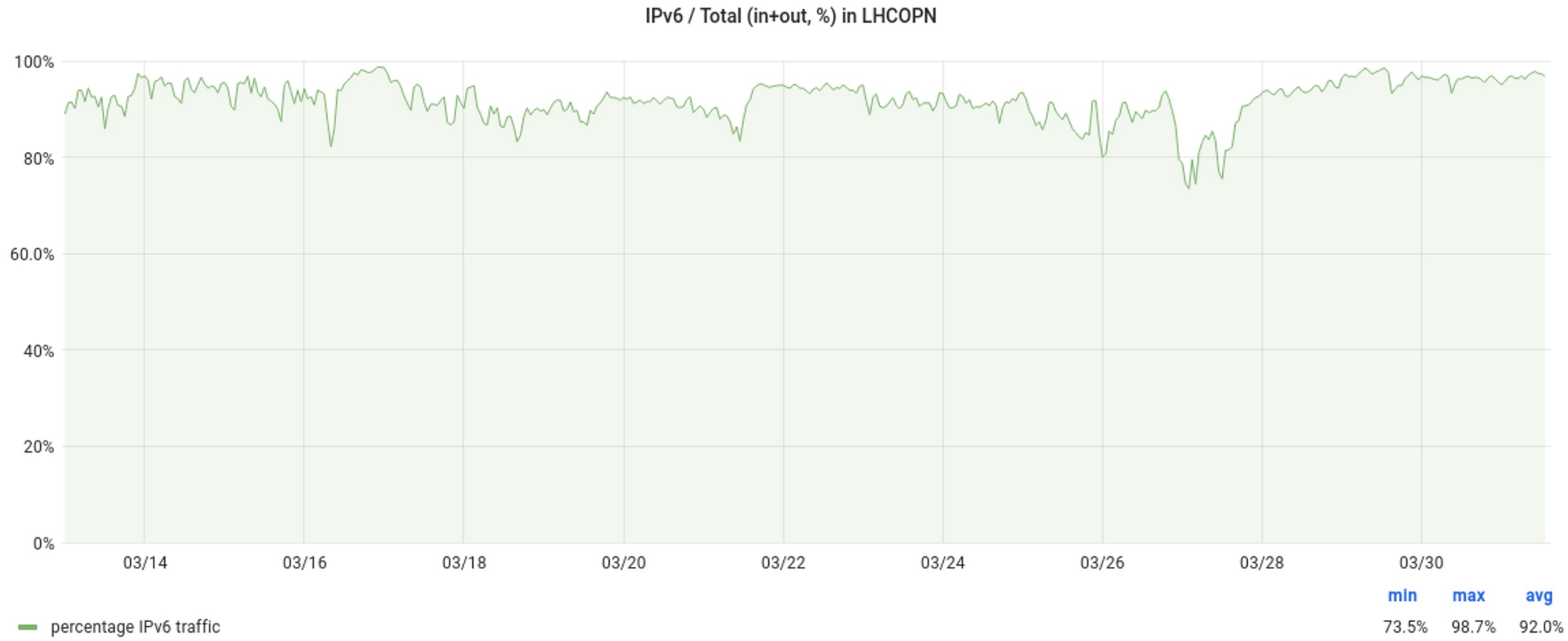
Forzati a splittare su due VLAN per problemi di contatori sui nuovi Juniper.

- On-going activity to separate IPv6 from IPv4 on LHCOPN links
- Prompted by unreliable sflow data on new CERN LHCOPN routers
- Implemented using two parallel VLANs
- Already done:
  CA-TRIUMF, DE-KIT, ES-PIC, FR-IN2P3, NDGF, NL-T1, RU-JINR, RU-KI, UK-RAL, US-BNL, US-FNAL

- Next: **IT-INFN-CNAF**, KR-KISTI (Da capire con Marletta di GARR se e come)

IPv6 may be already above 90%

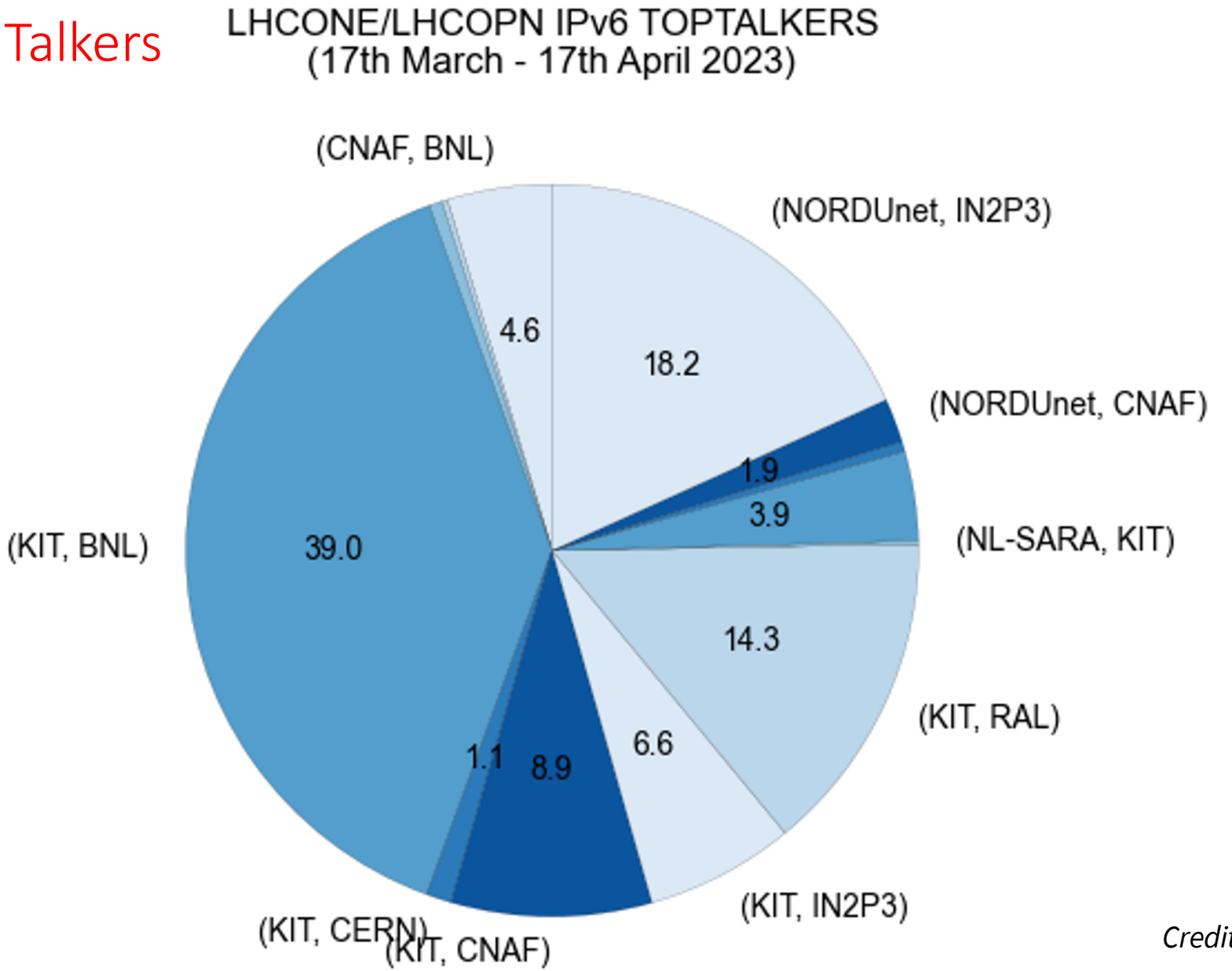
IT-INFN-CNAF and KR-KISTI not yet included



Ref: https://monit-grafana-open.cern.ch/d/cumEJJb4z/lhcopn-one-ipv6-vs-ipv4?orgId=16

Top talkers

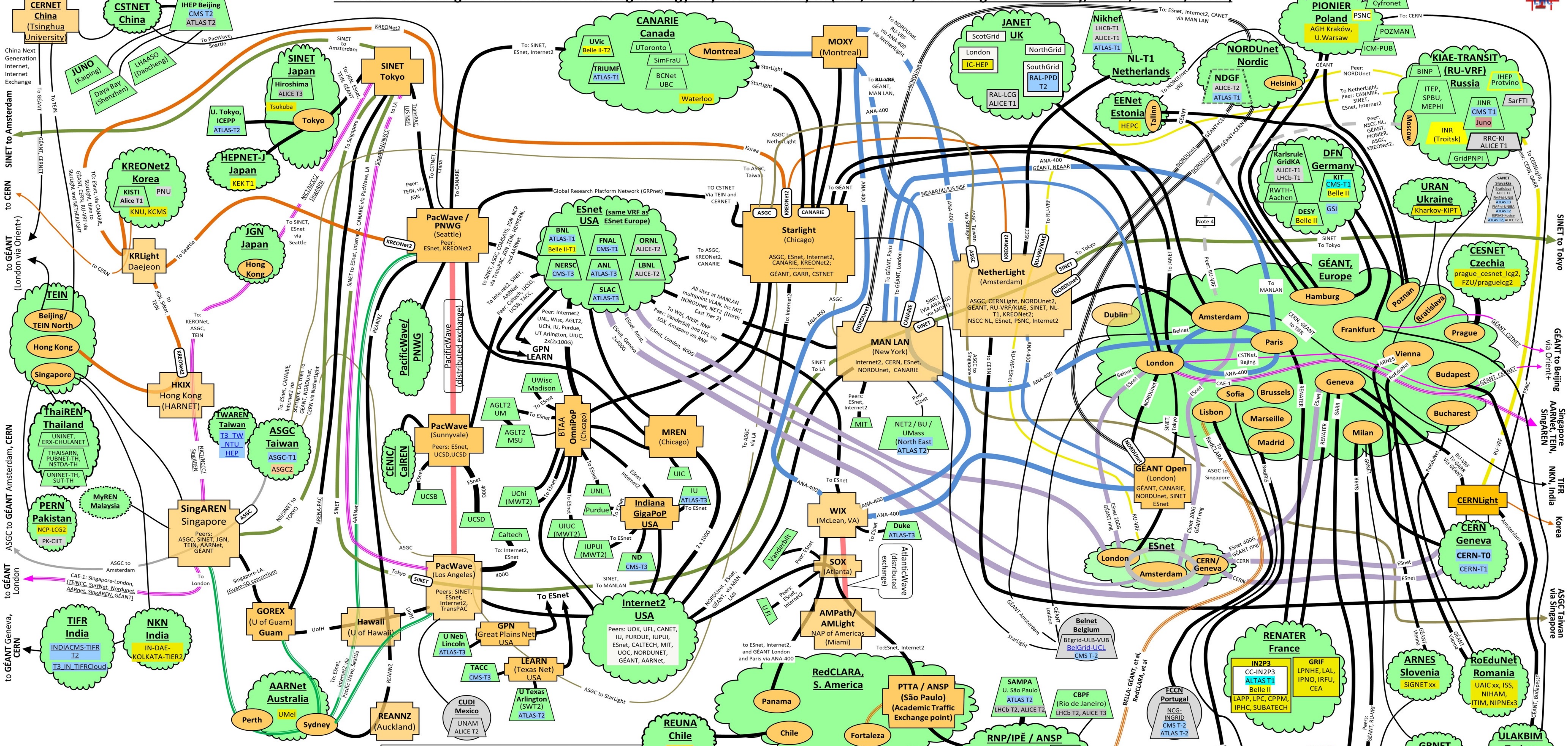
Siamo fra i Top Talkers



Credits: Carmen Misa (CERN)

LHC-ONE PART

LHCONE L3VPN: A global infrastructure for High Energy Physics data analysis (LHC, Belle II, Pierre Auger Observatory, NOvA, XENON, JUNO)



LHCONE Map Ver. 6.0, 2022-11-15 – WEJohnston, ESnet, wej@es.net

- GARR: LHCONE VRF domain/aggregator - A provider network.
- ANSP: Connector network – provides, e.g., an L2 path between VRFs.
- London: Provider network PoP router
- FCCN Portugal: WLCG sites that are not connected to LHCONE
- Exchange point
- SINET: NREN/site router at exchange point
- Communication links: <100G=1.5pt, 100G=4pt, 200G=5pt, 400G=6pt, 800G=7.5pt
- JGN – SingAREN/NSCC: Underlined link information indicates link provider, not use
- NDGF-T1: Double dash outline indicates distributed site
- KAUST Saudi Arabia DTN1: Future site

International infrastructure by provider/collaboration

- various
- AARNet
- GÉANT
- SINET, Japan, global ring
- ASGC, Taiwan
- ESnet transatlantic, USA
- NICT/NCCC/ SingAREN
- SINET
- NORDUnet
- KIAE, Russia
- KREONet2, Korea
- BELLA: GÉANT, et al, RedCLARA, et al
- ANA-300/400 - Various links provided by CANARIE, ESnet, GÉANT, Internet2, NORDUnet, SURFnet, SINET,IU/NSF

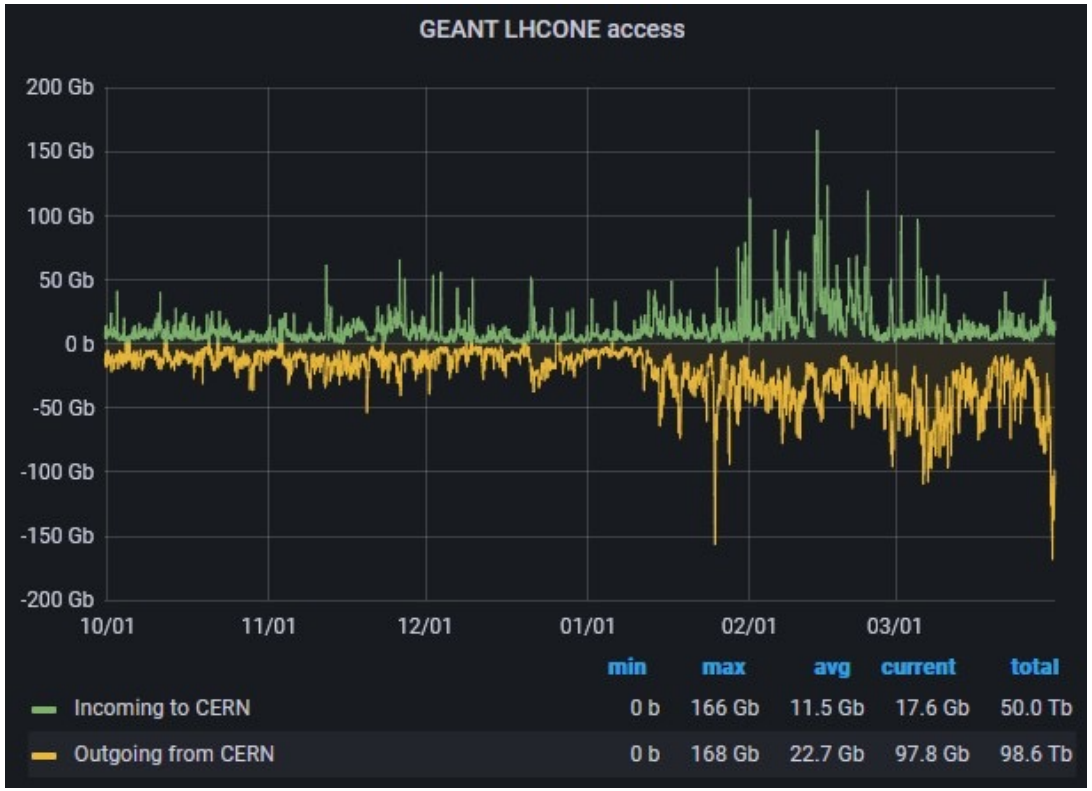
- LHCB-T1: LHC ALICE or LHCb site
- CNAF-T1: LHC Tier 1 ATLAS and CMS
- UChi: LHC Tier 2/3 ATLAS and CMS
- KEK: Belle II Tier 1/2
- JUNO: JUNO
- UNL: Sites that are standalone VRFs

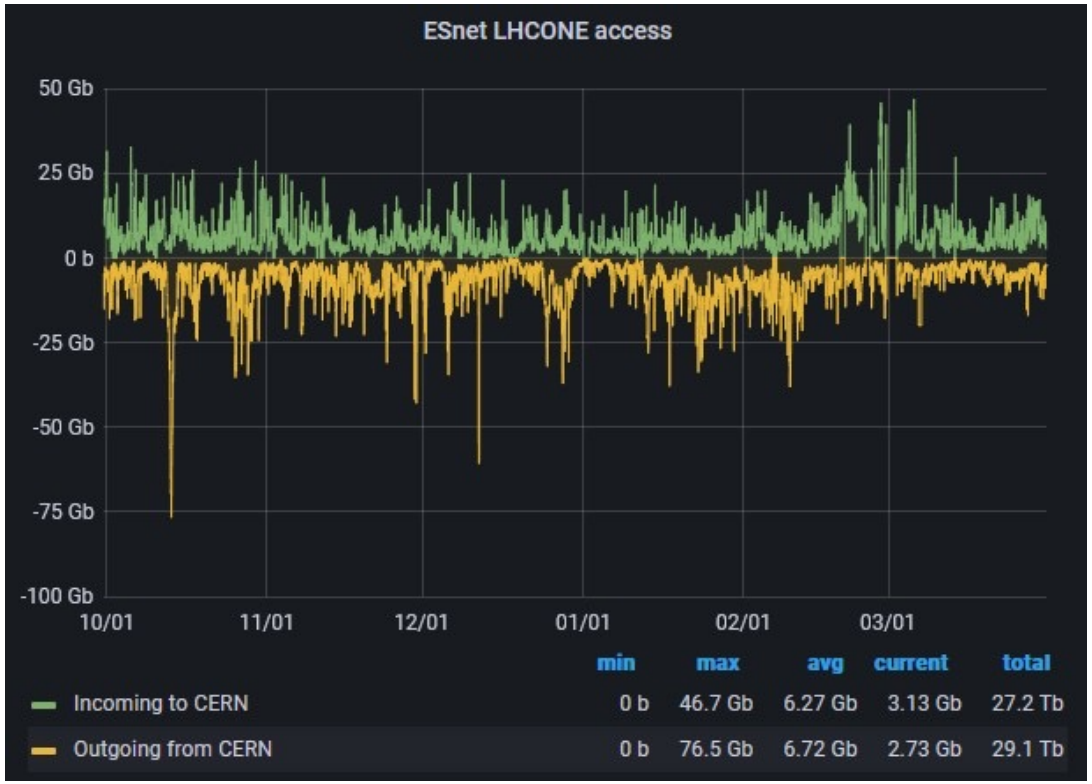
NOTES

1) ONLY links involved in LHCONE are shown
2) LHCOPN links are not shown on this diagram
3) For map explanation see "Interpreting the LHCONE Map" at https://www.dropbox.com/sh/padxfo58j0j1raz/AADsB5K8fISH9FfhCjA4eCtea?dl=0
4) GÉANT and CANARIE have shutdown the peering between their VRF and KIAE, as a result of the Ukraine war.

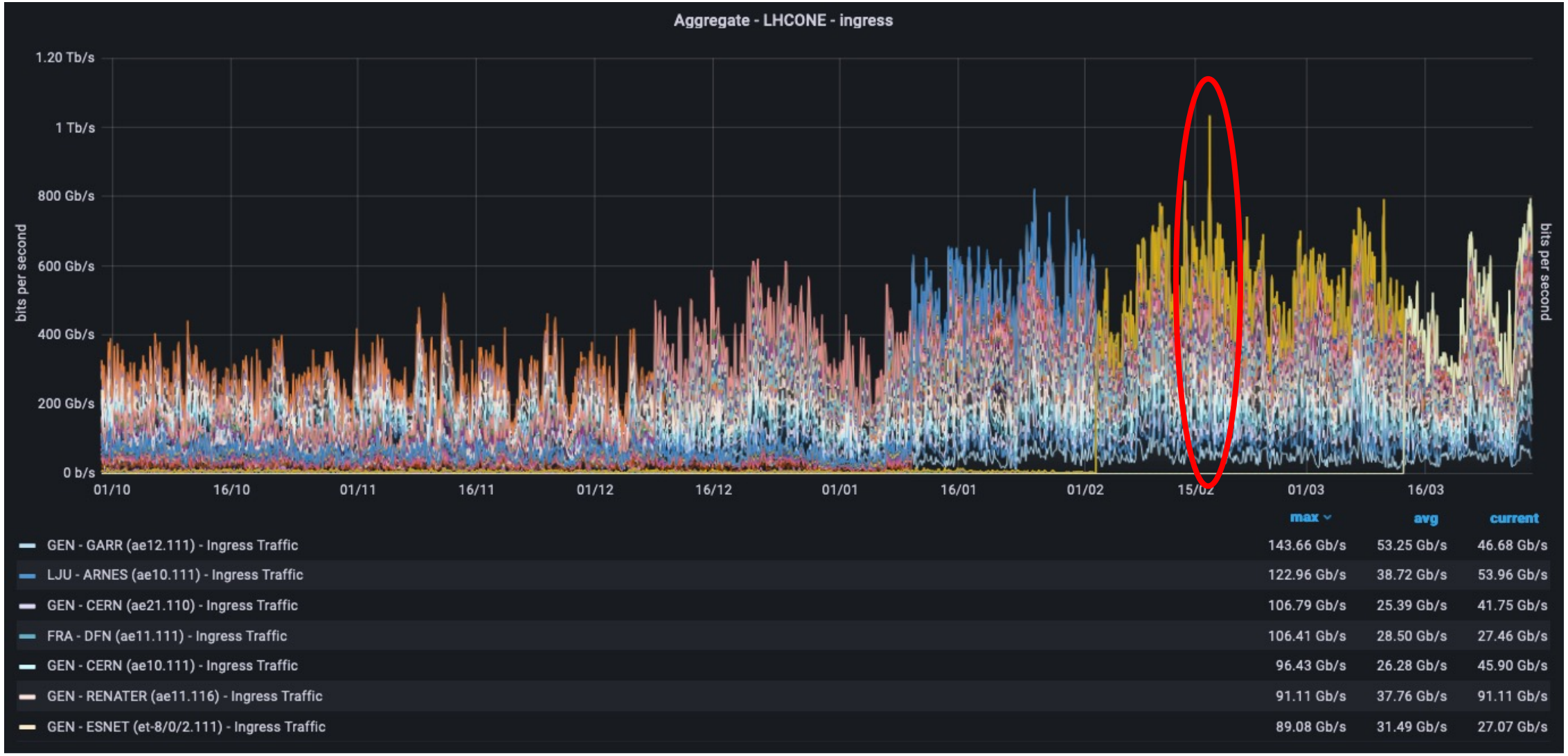
Novità e traffico LHCONE GEANT E ES-NET

- CERN upgraded LHCONE access to 2 x 400G
- RAL expanded the announced IP prefixes
- Lawrence Berkeley National Laboratory (ESnet)
- University of Massachusetts – Amherst (ESnet)

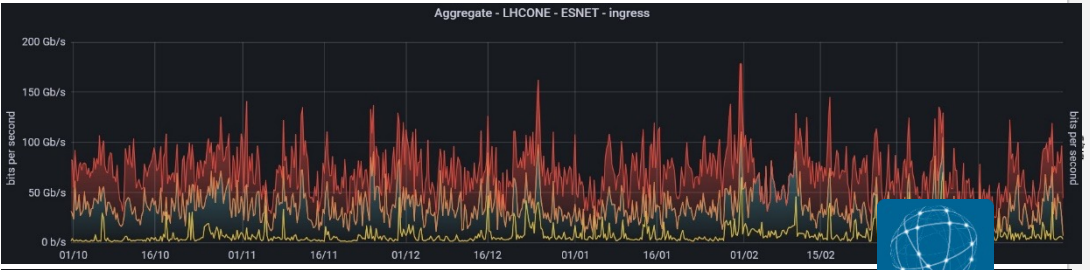


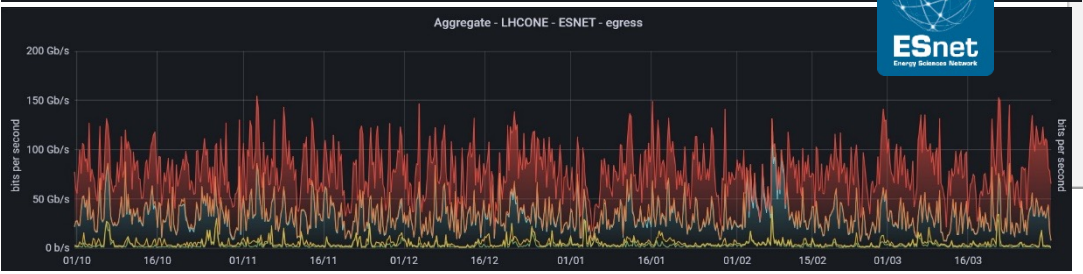


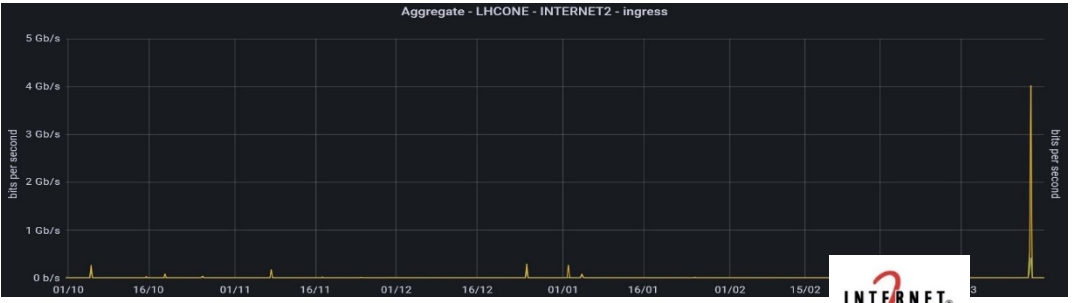
GÉANT overall

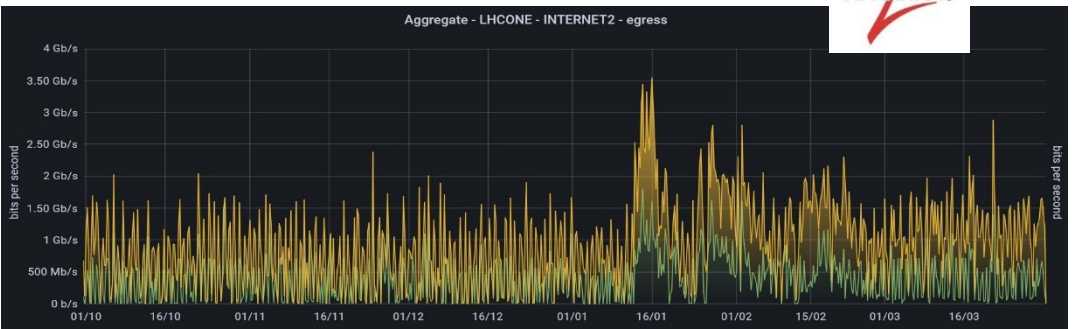


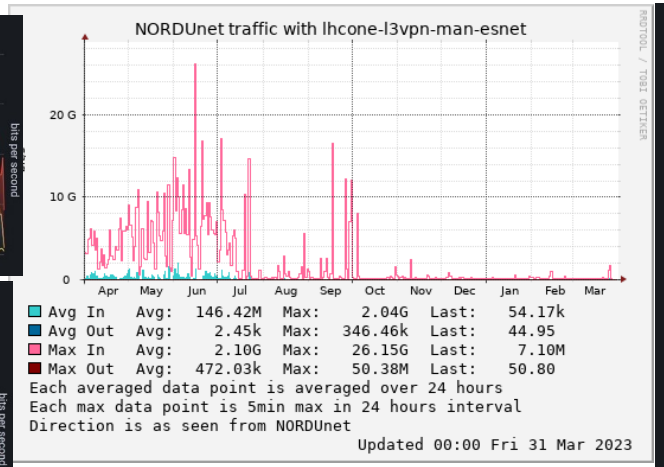
EU <-> North America

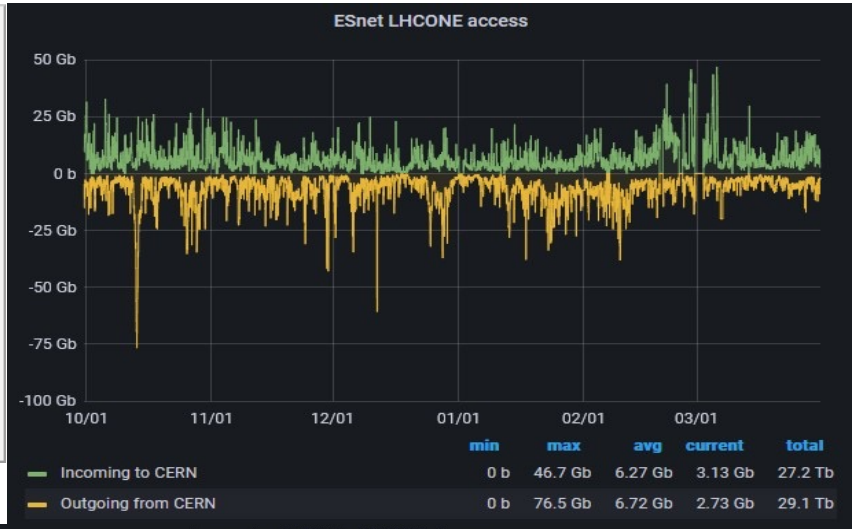


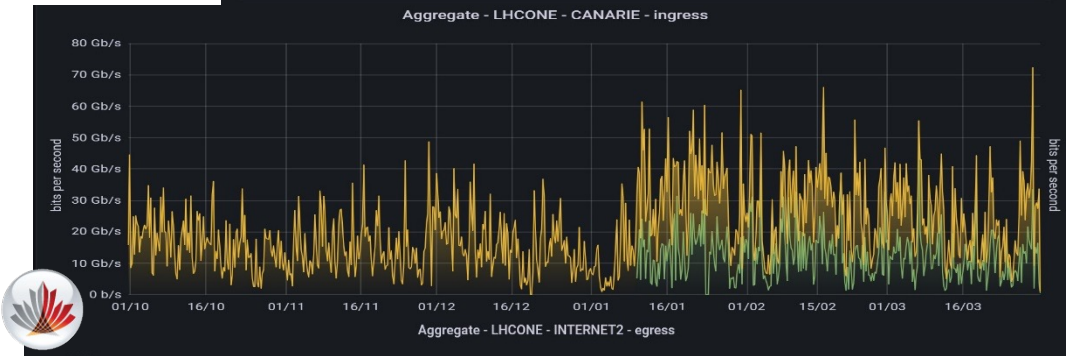


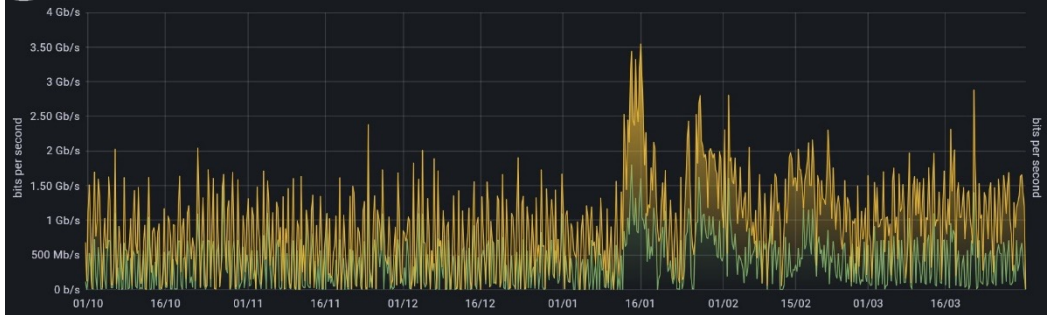












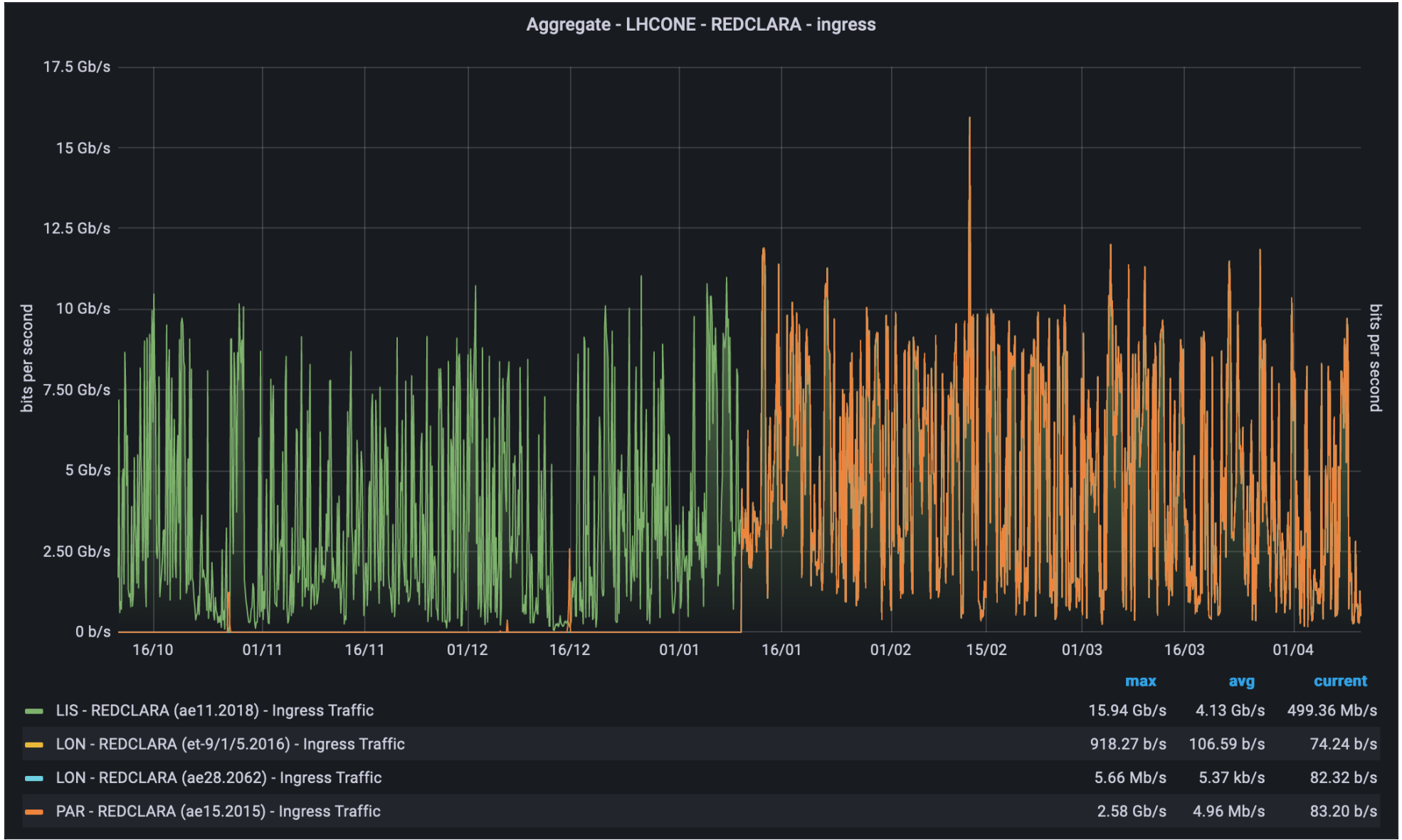
EU <-> Latin America

BELLA
Building the Europe Link to Latin America

RedCLARA
+ Red + Ciencia

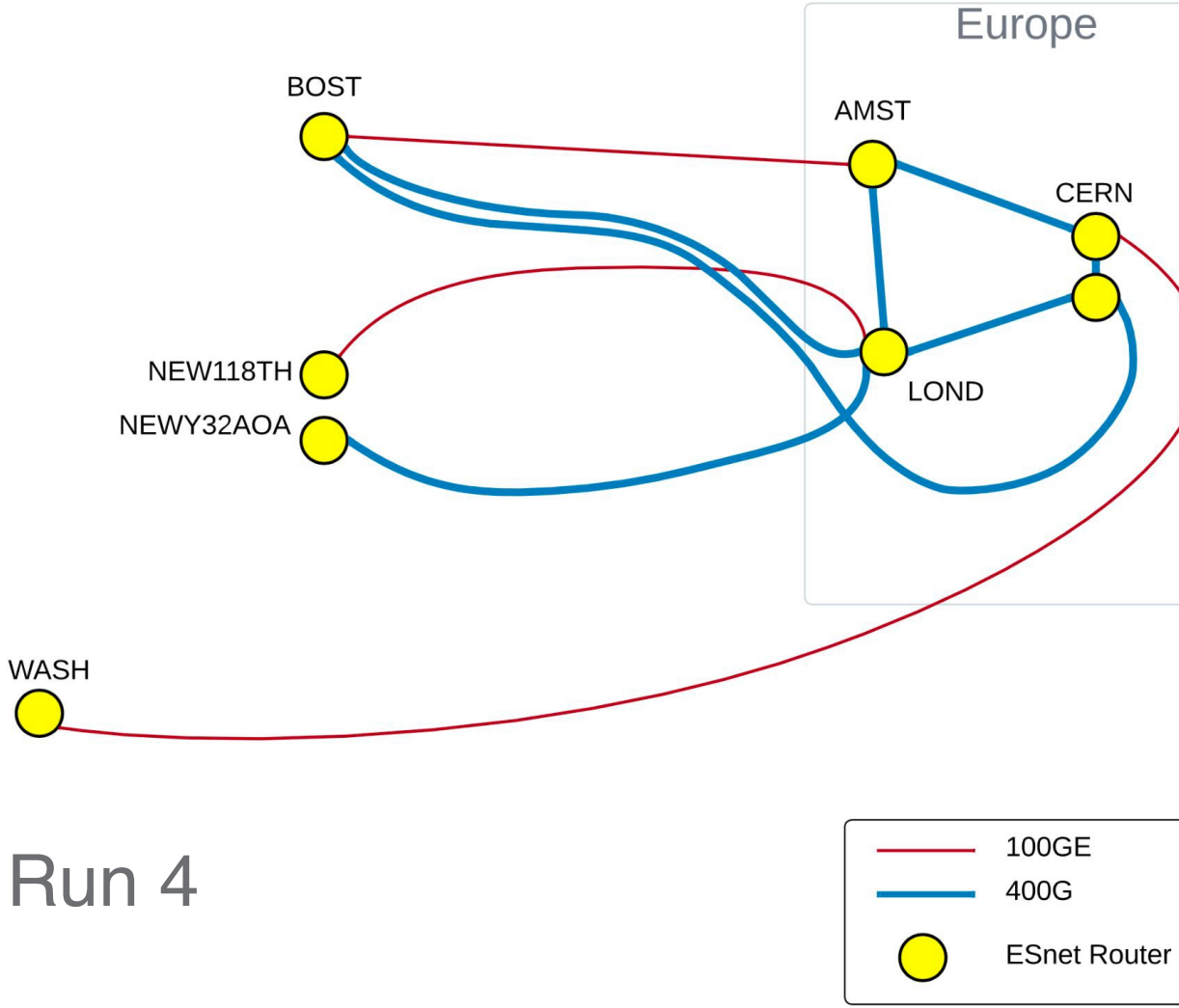
RNP

RedUNA



ESNET

- Now In Production:
  - 400G New York - London
- Currently underway:
  - 400G Boston - London (late fall)
  - 400G Boston - CERN (late fall)
  - 400G Europe Ring (late fall)
- Trans-Atlantic capacity targets
  - 1.5T in advance of DC24
- – ...
- 3.2T* in 2027, well in advance of Run 4



ESNET Cloud provider Peerings

ESnet6 built physical network into major commercial facilities

- via private fiber interconnects
  - 5x100G to Google (one more pending)
  - 6x100G to Oracle
- via fabrics
  - 5x100G to Microsoft
  - 5x100G to Amazon
- Private Cloud Interconnects to nearly any provider
- 5 locations (each 2x100G) to PacketFabric
  - OSCARS connectivity across Esnet
  - possibility for API-based provisioning end-to-end

ESNET DOE Site Connectivity

DOE Site Connectivity

- ESnet6 installed routers collocated at our sites
  - Most are connected to our optical system at 1.2Tbit + redundancy
- We are now ready to accommodate upgrades as sites are abl
- BNL - US ATLAS Tier 1
  - Current: 300G (2 x 100G + 1 x 100G)
  - Near Future: 800G (1 x 400G + 1 x 400G)
- FNAL - US CMS Tier 1
  - Current: 400G (2 x 100G + 2 x 100G)
  - Near Future: 800G (1 x 400G + 1 x 400G)
- NERSC
  - Current: 1T (2 x 400G + 2 x 100)

PerfSonar

- PerfSONAR 5 is OUT!
- Perfsonar will be used to help during next Data challenge

- http://my.es.net/
- http://www.es.net/
- http://fasterdata.es.net/

CRIC USE discussion

Come miglioreare l’uso di CRIC per mantenere informazioni aggiornate?

- https://apps.db.ripe.net/db-web-ui/query?searchtext=rs-lhcone
- https://wlcg-cric.cern.ch/core/networkroute/list/
- https://wlcg-cric.cern.ch/core/netsite/list/

Prossimo Meeting

- Hosted by University of Victoria (CA), Randal Sobie - Date: 16-20 of October 2023, co-located with HEPiX meeting Fall 2023 - Venue: University of Victoria Student Union Building Agenda: - Some presentations shared with HEPiX, during the Network Session - A full day for LHCOPN/ONE only (Wednesday or Thursday, to be agreed with HEPiX) - Informal dinner being planned for Thursday evening. No late afternoon flights, people should plan to leave the day after the meeting. Fee: - About 400CAD to attend both HEPiX and LHCOPN/ONE meetings - Free for LHCONE/OPN meeting only (thanks to CANARIE!)

Data Cahllenge 24 Planning

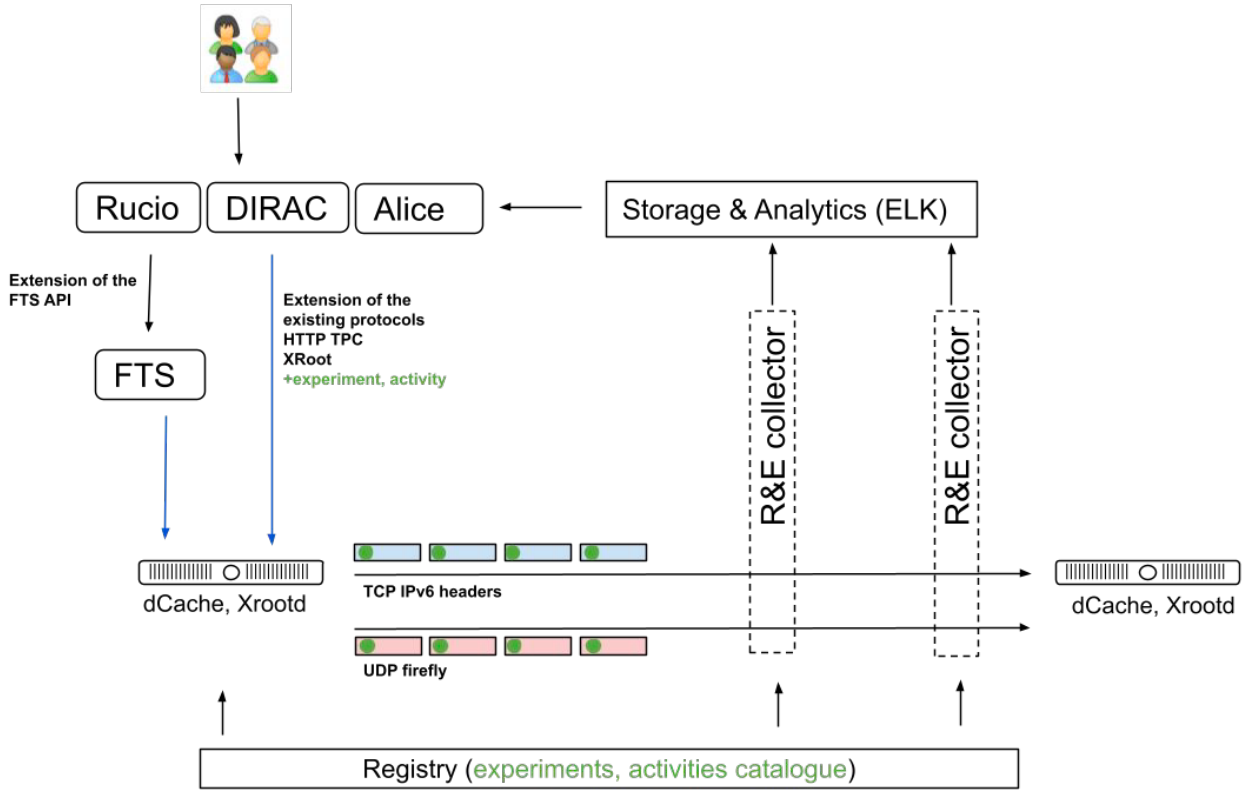
- Target Rate should be 25%
- Invitation to other sciences (SKA, JUNO, BELLE) to try the system to see if the load on the net could be critical in case o simultaneous data transfers.
- https://indico.cern.ch/event/1258343/

Shawn presentation:
https://docs.google.com/presentation/d/1ZZGafmF7imrqMBk8HO8AsCVkOdf5i4PG5rGelrZrjJU/edit#slide=id.g8036819354_0_7

  - Don’t want sites to prematurely spend money to reach numebers on the challenge
  - For High-Energy Physics (HEP), we have identified a need to better understand and optimize our network traffic to ensure we are using the network as effectively (for our science) as possible.
  - Scientific Network Tags (scitags) is an initiative promoting identification of the science domains and their high-level activities at the network level. (

SCITAGS

The scitags.org domain provides an API that can be consulted to get the standard values: https://api.scitags.org or https://www.scitags.org/api.json



SCITAGS

- The detailed technical specifications are maintained on a Google doc https://docs.google.com/document/d/1x9JsZ7iTj44Ta06IHdkwpv5Q2u4U2QGLWnUeN2Zf5ts/edit#heading=h.2msfykqhodwc
- ● The spec covers both Flow Labeling via UDP Fireflies and Packet Marking
- via the use of the IPv6 Flow Label.
- ○ Fireflies are UDP packets in Syslog format with a defined, versioned JSON schema.
- ■ Packets are intended to be sent to the same destination (port 10514) as the flow they
- are labeling and these packets are intended to be world readable.
- ■ Packets can also be sent to specific regional or global collectors.
- ■ Use of syslog format makes it easy to send to Logstash or similar receivers.
- ○ Packet marking is intended to use the 20 bit flow label field in IPv6 packets.
- ■ To meet the spirit of RFC6437, we use 5 of the bits for entropy, 6 for activity and 9 for
- owner/experiment.

SCITAGS STATUS

- Flow Marking (UDP firefly) implementation
  - Xrootd 5.4+ supports UDP fireflies
    - https://xrootd.slac.stanford.edu/doc/dev54/xrd_config.htm#_pmark
    - map2exp - can be used to map particular path to an experimen
    - map2act - can be used to map particular user/role to an activity
  - Flowd - prototype service
    - Issue fireflies from netstat for a given experiment (only for dedicated storages)
- Collectors
  - Initial prototype was developed by ESnet (available on scitags github)
  - ESnet and Jisc/Janet*
- Registry
  - Provides list of experiments and activities supported
  - Exposed via JSON at api.scitags.org
- Simplified deployment was tested during DC21
  - Flowd + ESnet collector + Registry
  - AGLT2, BNL, KIT, UNL and Caltech participated
  - Brunel, Glasgow and QMUL interested to help with further testing
- New flowd version will be ready to be deployed shortly (building packages)

For traffic pacing the group thinks to use Linux TC (Traffic Control)

https://man7.org/linux/man-pages/man8/tc.8.html

https://tldp.org/HOWTO/Traffic-Control-HOWTO/intro.html

Packet Pacing (Deep Buffer critcity)

Eli Dart Presentation
https://indico.cern.ch/event/1234127/contributions/5271809/attachments/2630790/4550132/20230419-dart-pacing-v2.pptx

- Different interface speed cause buffer issues and burst
- Burst cause problems.

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

- Reduce impact of burdts on buffers, receivers, etc..

- Basta un apparato low buffer per compromettere il trasferimento.
  - Un TOR Cheap può compromettere i trasferimenti
  - Medium or Deep buffer device costano moltissimo.

Problema di non facile suluzione...

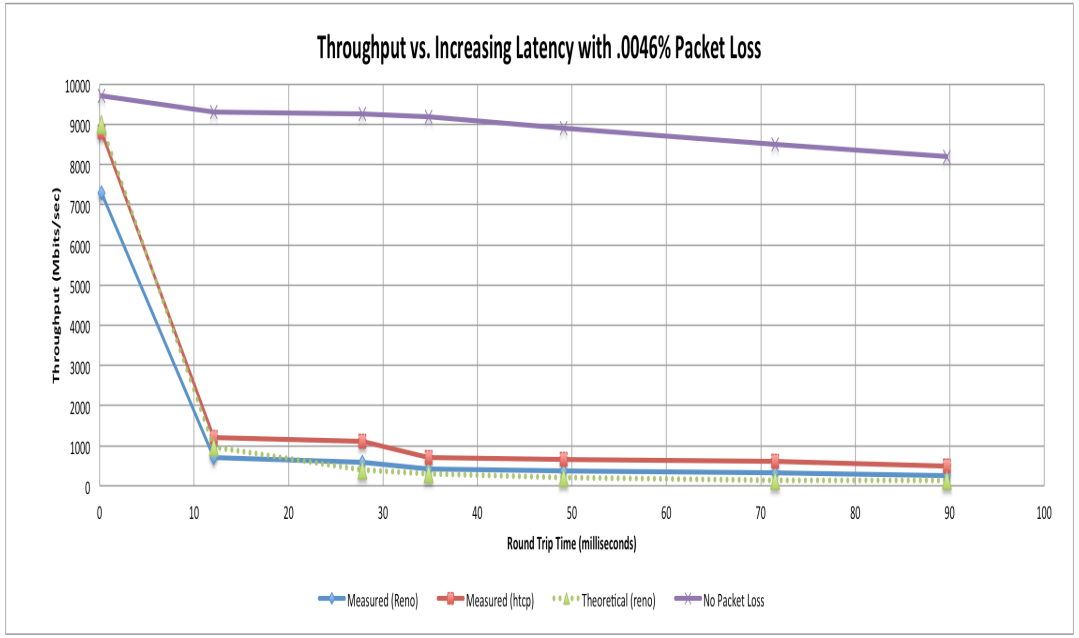
E’ utile capire quale è il PER FLOW DATARATE.

Framing and Context

- TCP has been and continues to be the workhorse protocol used by data transfer applications
  - Internet's Reliable byte-stream delivery protocol (in contrast to UDP)
  - Underlying mechanism used by HTTP, Globus, FTP, etc.
- TCP performance is badly impacted by even minor packet loss
- TCP's bursty behavior contributes to packet loss
  - This is why we deploy deep buffers
  - Deep buffers are expensive, and will go away in the future
- Is there something we can do about all this?

TCP is “bursty” – what does that mean?

- TCP sends data when two conditions are true
  - TCP has data to send, e.g. the application wrote to the socket
  - The receiver has advertised available window space
  - TCP sends data until one of these conditions is not true
- Most host interfaces can send at wire speed
  - This means that if TCP has data to send, and it hands the data off to the host NIC, the NIC will send packets at wire speed until done
  - 10G NICs send data at 10G, 100G NICs send data at 100G
- On average the rate may be lower, but the instantaneous rate is wire speed
  - A host that runs at 50% of wire speed on average might actually send at wire speed 50% of the time and sit idle 50% of the time



BBR TCP has built-in pacing (slide from Matt Mathis presentation, March 2020)

- BBR: new first principles for Congestion Control
  - BBR builds an explicit model of the network
    - Estimate max_BW and min_RTT
- The BBR core algorithm:
  - By default pace at a previously measured Max_BW
    - **Transmit based on a clock**, not ACKs
  - Vary the pacing rate to measure model parameters
    - increase to observe new max rates
    - decrease to observe the min RTT
    - gather other signals such as ECN (bbr2)
- BBR's "personality" is determined by the heuristics used to vary the rates and perform the measurements
  - These heuristics are completely unspecified by the core algorithm
  - Relatively easy to extend or adapt
  - Many different heuristics algorithms can work together

Ideas to Explore

- Can we come up with a simple pacing configuration for WLCG DTNs that improves performance?
- What is the mix of host and network interface speeds in WLCG, and how might that affect a global pacing configuration?
- What per-flow data rate do we need in WLCG? Is it different for different workflows? What does this mean for pacing config?
- BBRv2 will probably make manual pacing configurations obsolete, but BBRv2 is years away (Google has not yet merged it upstream, so it hasn’t even begun the path to production distro kernels)

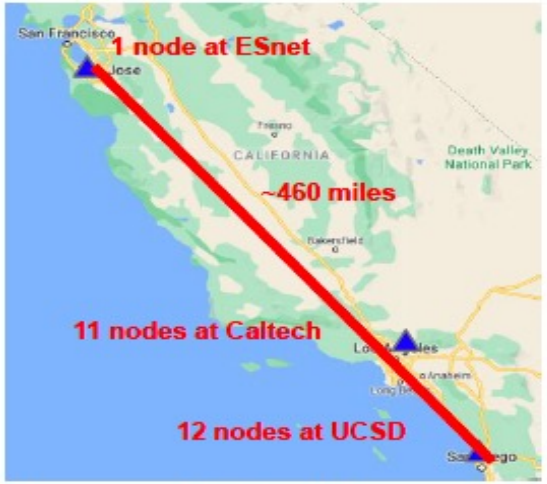
Es-Net **Network Caching**

Presentazione di Chin Guok (CTO di Esnet)

https://indico.cern.ch/event/1234127/contributions/5271810/attachments/2630824/4550194/ESnet%20In-Network%20Caching%20-%20LHCOPN-LHCONE%20Apr2023.pdf

Summary observations

- SoCal Repo could serve on average about 67.6% of files from its disk cache, while on average only 35.4% of bytes requested could be served from the cache
  - Because the large files are less likely to be reused
  - To avoid cache pollution from this particular usage pattern with large files, the operators have separated the two different types of files requests with different storage nodes.
- Over the whole period of observation, there is a five-month period where the large file requests are noticeably high, resulting in an average reduction of wide-area network traffic of about 12.3TB per day
- During the period where fewer large files were requested (3/2022 – 5/2022), the network traffic was reduced by about 29TB per day



Sunnyvale–San Diego
is the relevant distance scale

Discussion on mini challeges

Si discute la opportunità di fare mini challenges

https://docs.google.com/document/d/1o08dzU1MDSWxco4SJ1phU9b8_MqtZTyW2o4V_Sy4oHs/edit

Jumbo Frames discussion

Christopher Walker, Tim Chown

https://indico.cern.ch/event/1234127/contributions/5314839/attachments/2630823/4551821/LHCONE%20Jumbo%20frame%20discussion.pdf

The most frequent problems are related to end to end systems with jumbo talking on a path that doesn’t fully support junbos .

- In many cases problems are related to devices misconfigured in LAN or MAN close to one end.
- Other problems are related to ICMP filtering not allowing PMTU Discovery.

NRENs are supporting Junbo by decades. Sites has to decide to adopt or not Jumbo.

Overview

Jumbo

- Larger packets (MTU=9000, rather than 1500)
- Potential performance advantage of larger MTU
  - Higher link capacities
  - CPU clock speed not increasing
  - Larger frames intuitively make sense
- WLCG recommendation in 2018
  - MTU (“jumbo frames”) recommendation for LHCONE and LHCOPN (cern.ch)
  - Goal was to get NRENs to support jumbo frames
- Is the time right to try it out?

Network test data

- Iperf (Raul from Jisc)

Source	Destination	RTT	9000	1500
SURF (NL)	RNP (Brazil)	100ms	31 Gbit/s	20 Gbit/s
Jisc (London)	BNL (USA)	100ms	14 Gbit/s	6 Gbit/s

Source	Destination	RTT	9000	1500
London	Cambridge	3ms	37 Gbit/s	15.8 Gbit/s
London	AARnet	120ms	21 Gbit/s	3.4 Gbit/s

Thoughts?

- Do we want to have another push on this?
  - QMUL, RALPP already doing this
  - Data transfer tests desirable
- Is MTU=9000 agreed (at least at NREN level)?
  - Do we need to test this?
- What to advocate?
  - (e.g., tips like net.ipv4.tcp_mtu_probing=1)
- Next steps?

Next Meeting

Hosted by University of Victoria (CA), Randal Sobie

- Date: 16-20 of October 2023, co-located with HEPiX meeting Fall 2023
- Venue: University of Victoria Student Union Building

Agenda:

- Some presentations shared with HEPiX, during the Network Session
- A full day for LHCOPN/ONE only (Wednesday or Thursday, to be agreed with HEPiX)
- Informal dinner being planned for Thursday evening. No late afternoon flights, people should plan to leave the day after the meeting.

Fee:

- About 400CAD to attend both HEPiX and LHCOPN/ONE meetings - Free for LHCONE/OPN meeting only (thanks to CANARIE!)
- Wednesday afternoon for private LHCOPN/ONE meeting