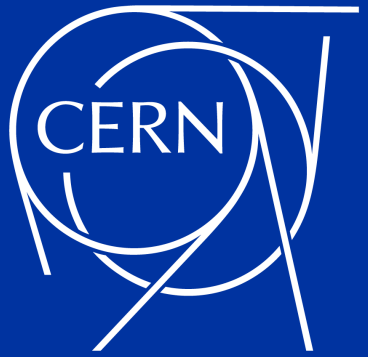


LHCOPN-LHCONE meeting #50 summary notes

GDB - 12 July 2023

edoardo.martelli@cern.ch

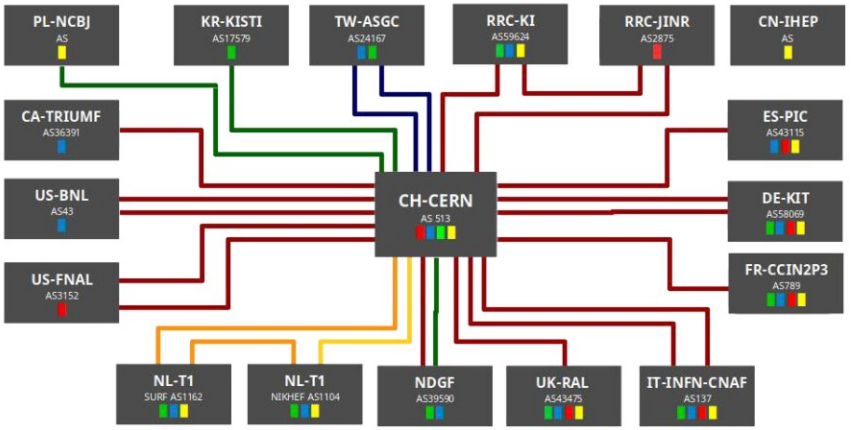
Venue

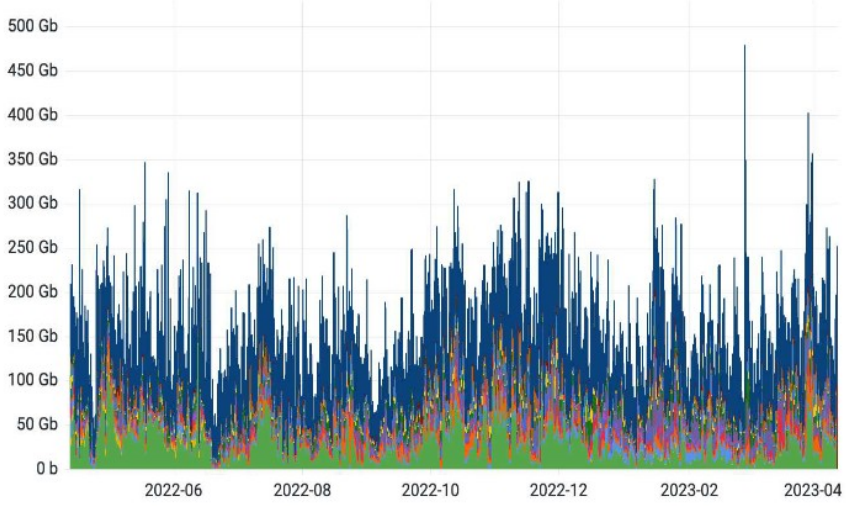
- 18-19 of April 2023
- Hosted by FZU in Prague (CZ)
  - Special Thanks to Jiri Chudoba!
- co-located with GEANT SIG-NGN meeting
- ~40 people in presence and ~30 connected remotely
- Agenda at https://indico.cern.ch/e/LHCOPNE50

FZU Institute of Physics of the Czech Academy of Sciences

LHCOPN - update

- 2.1Tbps of aggregated bandwidth to the Tier0
- Traffic stats: moved 488PB in the last 12 months. +12% compared to previous year (slower increase than previous years)
- ES-PIC upgraded to 100Gbps
- NL-T1 SurfSARA upgraded to 2x100G
- CA-TRIUMF second link upgraded to 100G
- TW-ASGC no longer Tier1. Effective from October 2023
- **NBCJ (PL) and IHEP (CN) in the process to become Tier1s**



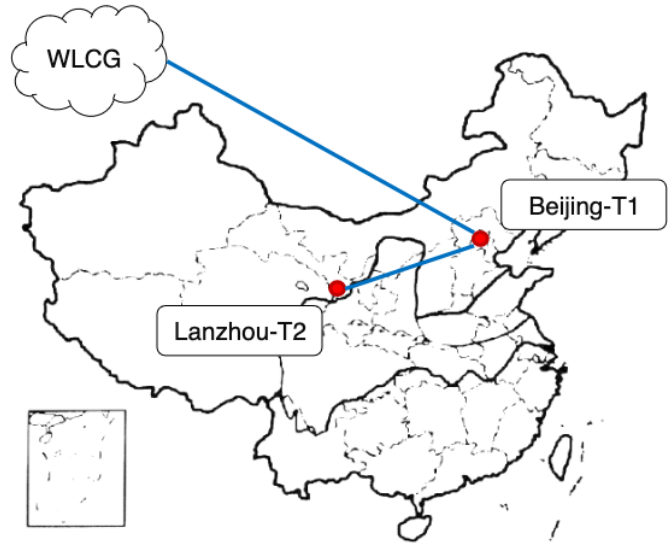


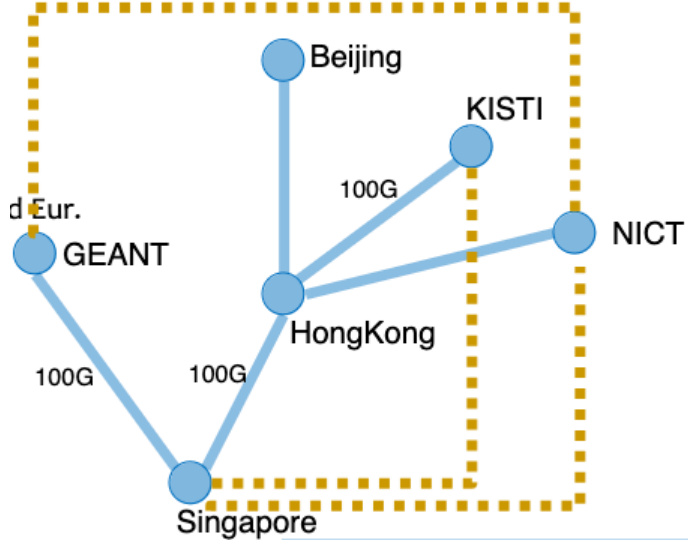
IHEP China: new LHCb Tier1

IHEP LHCb Tier-2 has started the procedure to become LHCb Tier-1

- CSTNet is the internet service provider for IHEP International links
- All domestic connections will be upgraded from 10G to 100G
- New international connections will be deployed to improve the bandwidth between China and Europe
- LHCONE link will be upgraded to 100G at the end of April [not ready yet]
- **LHCOPN: new link to CERN via GEANT (Marseille)**, planned for end of May [not ready yet]

https://indico.cern.ch/event/1234127/contributions/5231347/attachments/2630726/4550038/IHEP%20Tier1%20Report-20230418.pdf





NCBJ new Polish Tier1

NCBJ, National Centre for Nuclear Research in Warsaw has started the procedure to become a LHCb Tier1.

It hosts the Świerk Computing Centre (CIŚ)

- Computing: 1.4 PFLOPS, 36000 cores, 200 TB RAM
- Disk storage: 26 PB (Lustre, Isilon, Netapp, dCache)
- Tape storage: TSM4500, 16 PB (uncompressed)

Network resources:

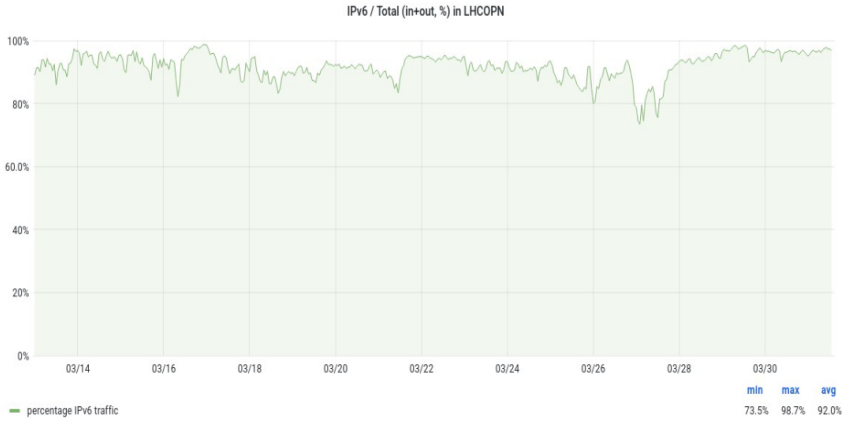
- 100 Gbps link to PIONIER (academic internet, GEANT)
- 20 Gbps dedicated VLAN to LHCONE
- Full speed achieved during 2022 Data Challenge
- **Additional 20 Gbps dedicated VLAN to LHCOPN** [completed]

IPv6 in LHCOPN

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 [done], KR-KISTI

Statistics show IPv6 usage above 90%

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



https://indico.cern.ch/event/1234127/contributions/5306049/attachments/2629065/4549722/LHCOPNE-20230418-50-LHCOPN-IPv6.pdf

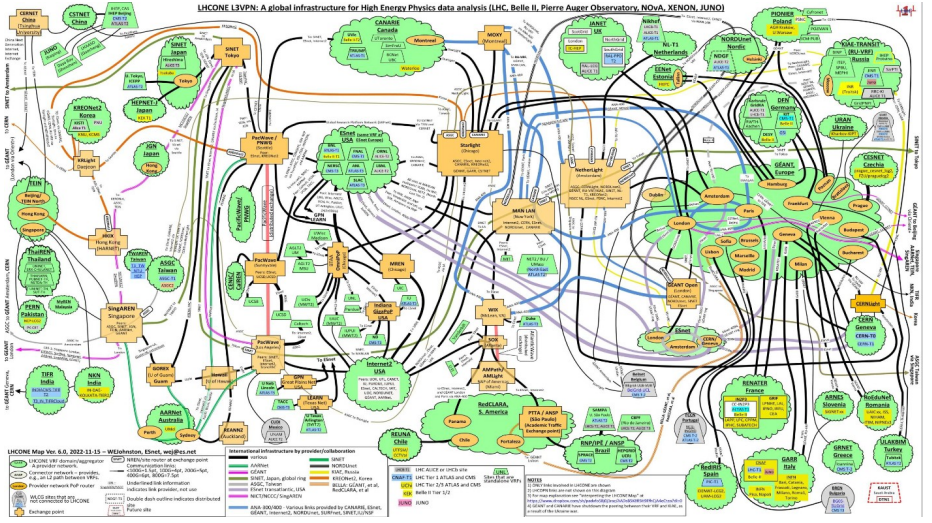
LHCONE L3VPN - update

News

- New sites:
  - Lawrence Berkeley National Laboratory (ESnet)
  - University of Massachusetts – Amherst (ESnet)
- CERN upgraded LHCONE GEANT access to 2x 400G
- New NORDUnet peering in Amsterdam
- KIFU (Hungarian NREN) joins LHCONE

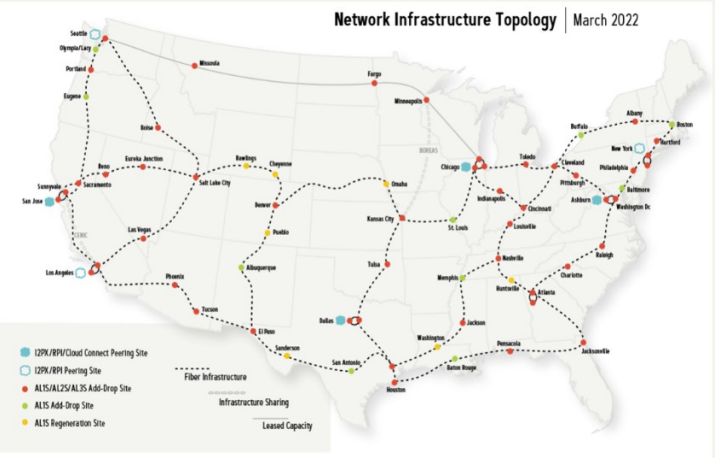
Traffic statistics:

- continue increase
- first peak above 1Tbps seen in GEANT



Internet2 update

- Internet2 Network: 400G native links, 18,700 miles of dark fibre
- New equipment: Ciena, Juniper, Cisco. Reduced carbon footprint of 66%
- **Transatlantic: 1x 400Gbps link shared with CANARIE**
- New Performance Assurance Service based on 47 perfSONAR nodes
- Implementation of Arroyo platform to support automation and orchestration of the new network and deliver new services to the community
- **Services for improved cloud connectivity:**
  - I2PX: peering at exchange points
  - I2CC: private connections over I2 infrastructure (up to 5Gbps)
  - I2RPI: private 10G interconnections
  - to Azure, AWS, Google, Oracle

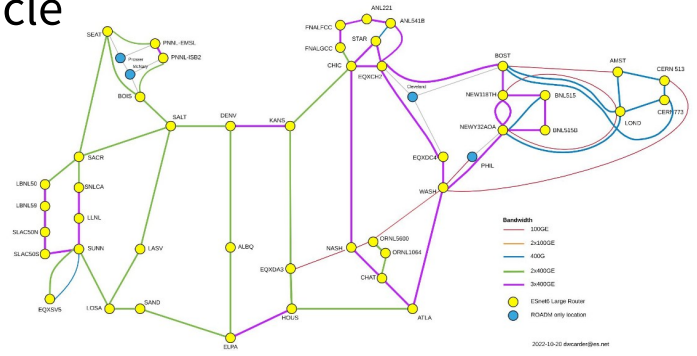


https://indico.cern.ch/event/1234127/contributions/5231144/attachments/2630831/4550203/I2-Network-Update-LHCONE-2023.pdf

ESnet update

Trans-Atlantic & EU ring upgrades:
- 400G New York – London (production)
- upcoming (late fall): 400G Boston – London, 400G Boston – CERN, 400G Europe Ring
- **Trans-Atlantic capacity targets: 1.5T in advance of DC24, 3.2T in 2027**

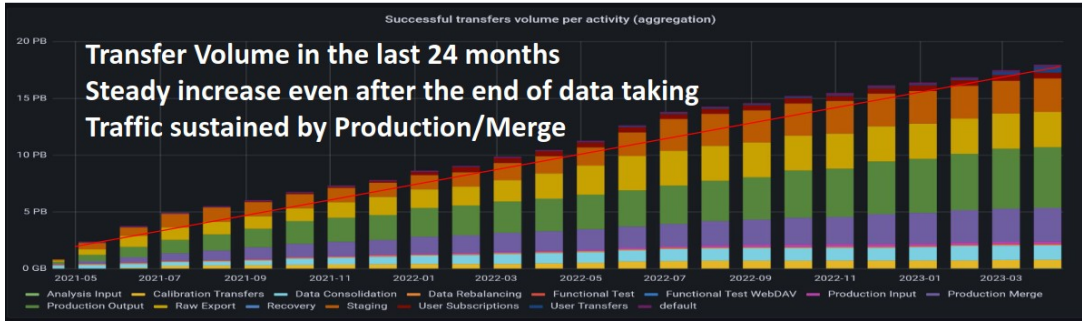
ESnet Cloud connectivity:
- private fibre interconnects: 5x100G to Google, 6x100G to Oracle
- via fabrics: 5x100G to Microsoft, 5x100G to Amazon
- **Private Cloud Interconnects to nearly any provider**



perfSONAR monitoring:
- deployment of new servers almost completed
- adding LHCONE connection to each server to increase LHCONE visibility

BelleII update

- Around 2.8PB of RAW Data Collected since 2019
- Currently in Long Shutdown for upgrade
- **Data taking will start again in the last quarter of 2023**
- Using Rucio for data distribution
- ~70% of storage reachable over IPv6
- On-going migration to DAVS for data transfers: 90% of transfers already on DAVS
- Token based authentication migration in progress
- **BelleII will participate to the WLCG DC24**. Main goal: Emulate data transfer conditions in a Belle II high-lumi scenario

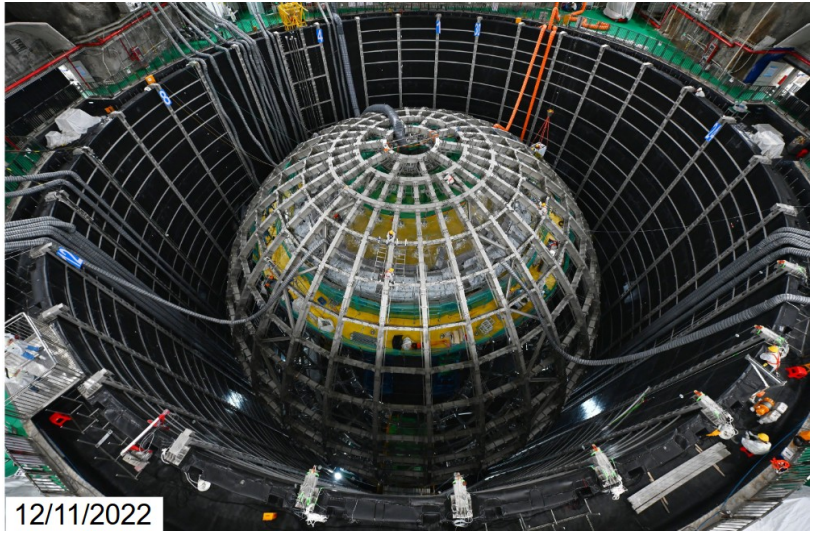


JUNO update

- Construction of JUNO detector is progressing well. **Data taking expected to start in 2024 H1**
- JUNO Distributed Infrastructure is working: usage and transfer increasing, **aiming to participate to WLCG DC24**

Networking:

- sites are connected to LHCONE, but Russians site are not reachable there. Alternative path works
- network challenge 2022/23 showed improvements compared to previous challenge



https://indico.cern.ch/event/1234127/contributions/5306050/attachments/2630988/4550508/JUNO%20update.pdf

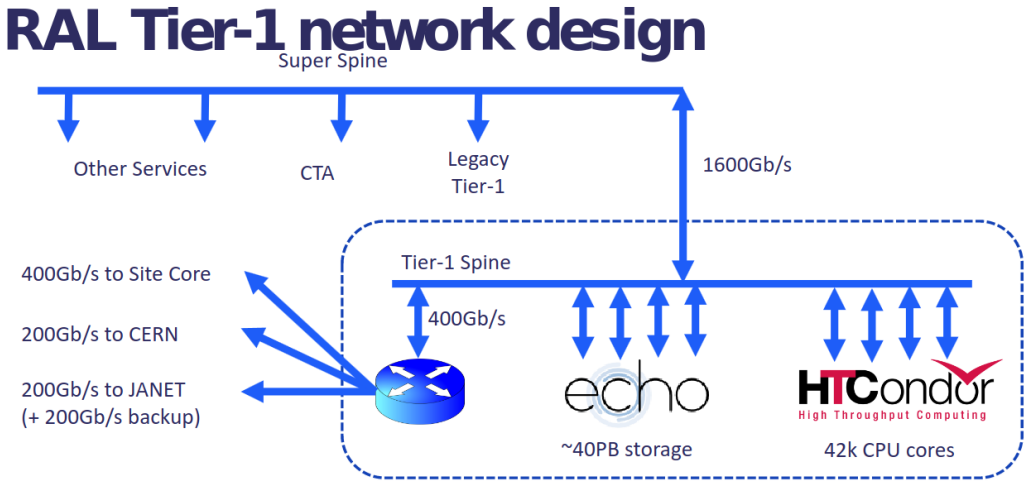
RAL update

RAL Tier1 has connected to LHCONE with storage, worker nodes and services

- 200Gbps link to Janet got saturated several time because of it

- Echo (disk) gateway will also be added to LHCONE

Second 100Gbps for LHCOPN soon in production [not yet]



https://indico.cern.ch/event/1234127/contributions/5341655/attachments/2630951/4550740/2023-04-18-RAL-T1-report.pdf

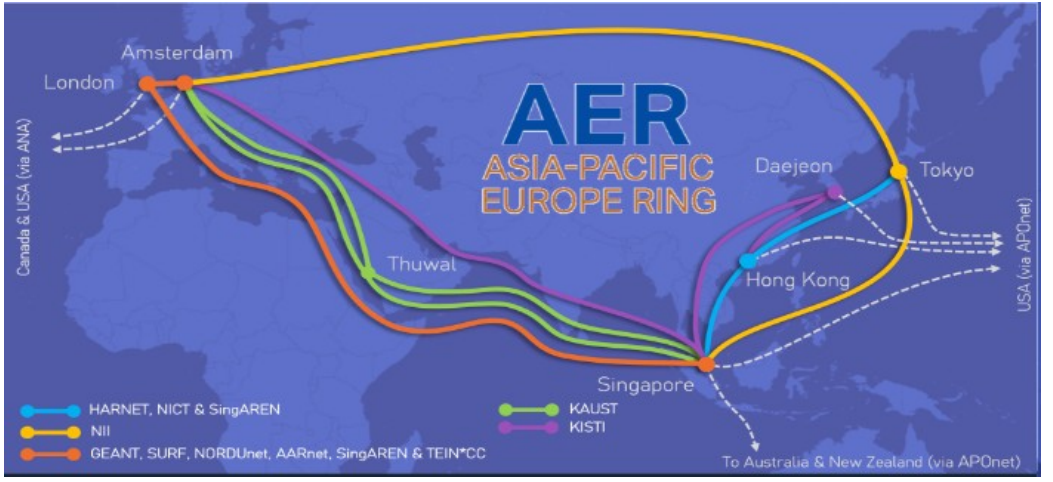
KREOnet update

Improved connectivity to Asia Pacific Oceania network (APOnet)

Link to Singapore-Amsterdam being upgraded to 100Gbps

AER (Asia-Pacific-Europe Ring) now has 500Gbps between Europe and Asia

Developed N-S-C, new orchestrator system to provide network,computing and monitoring services in KREOnet



https://indico.cern.ch/event/1234127/contributions/5347599/attachments/2631005/4550545/2023_4_18%20LHCOPN-LHCONE%20Meeting(bscho)v1.pdf

CRIC database status and development

Network information have been added to CRIC

- netsite: https://wlcg-cric.cern.ch/core/netsite/list/

- networkroute: https://wlcg-cric.cern.ch/core/networkroute/list/

Now **it's critical to keep it up to date**

Internet Routing Registry objects can be created to allow LHCONE Providers and Sites to build filters. Filters could be an incentive to update own records

CRIC devs will be asked to build user interfaces to query the database. E.g.

- is this address in a LHCONE or LHCOPN prefix? What site owns it?

- what are the ATLAS (/CMS/...) sites in LHCONE?

DC24 Planning

Full morning dedicated to DC24 planning:

- presented network activities that can be tested during DC24

- discussion followed

Research Network Technology WG - update

The RNTWG has made significant progress to identify network priority focus areas for the WLCG community. The current focus is on the network traffic visibility through the work on flow labeling and packet marking for DC24

Packet marking:

- standardized the “experiment” and “activity” fields used for both flow labelling and packet marking
- 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
- Flowd service: Flow and Packet Marking service developed in Python
- XRootD already provides SciTags implementation (from 5.0+)
- FTS/gfal2 needed to propagate SciTags to storages [discussions on-going]

Packet pacing:

- next work item to be tackled

scitags.org

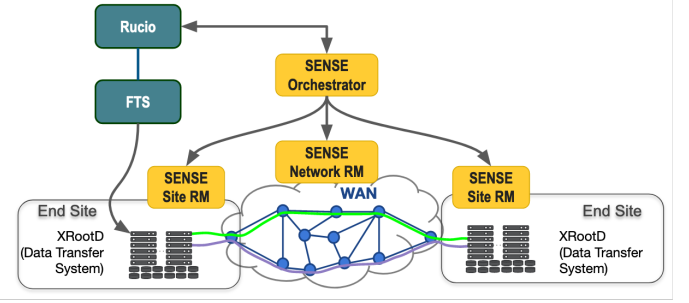
Packet pacing exploration

Overview of packet pacing. The activity is in the plan of the RNTWG and will start soon.

- 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 has built-in pacing (transmit based on a clock, not ACKs), but it is years away

https://indico.cern.ch/event/1234127/contributions/5271809/attachments/2630790/4550133/20230419-dart-pacing-v2.pdf

SENSE-Rucio project update



Project led by UCSD and Caltech

Functioning:

- Rucio identifies groups of data flows (IPv6 subnets) which are "high priority"
- SENSE takes flows from the site edge and "Traffic Engineers" paths across the WAN and End Sites
- Enables use of "multiple paths between sites" and provision of "deterministic" network resources to workflows

Components:

- Modified Rucio: which can prioritize certain transfers
- Data Movement Manager: react to Rucio's requests and translate them into network improvements
- E2E performance monitoring: evaluate the performance of the provided service

Pilots being tested at UCSD, Caltech, FNAL

https://indico.cern.ch/event/1234127/contributions/5233364/attachments/2631364/4551258/2023-04-19-sense-rucio-interoperation-LHCONE-50-mtg.pdf

perfSONAR strategy in support of DC24

perfSONAR 5 has just been issued:

- OS support: CentOS 7, Debian 10, Ubuntu 18 and Ubuntu 20. Alma 8,9, Debian 11 and Ubuntu 22 should follow soon
- ESmond (Postgresql+Cassandra) replaced by Elasticsearch

PerfSONAR will be used for DC24:

- Update and utilize perfSONAR to clean up links and fix problems before DC24
- Instrument and document site networks, for at least largest sites

A campaign to upgrade WLCG sites to perfSONAR 5 will be launched soon

Developing high-level services based on perfSONAR measurements that will help sites, experiments and R&Es receive targeted alarms/alerts on existing issues in the infrastructure

https://indico.cern.ch/event/1234127/contributions/5271808/attachments/2631309/4551135/perfSONAR%20Strategy%20in%20Support%20of%20DC24%20Preparation.pdf

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

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

DUNE participation to DC24

- Construction is in progress: caverns being drilled
- DUNE successfully utilizing WLCG resources
  - already ran successful data challenges

Working on:
- more complete Rucio integration
- developing new workflows and workflow management, including access to HPC
- integrate GPU software and hardware for processing

DUNE involvement in DC24:
- Simulate the archival of 25% of the raw data rate from the Far Detector
- Maintain continuous processing workload at distributed sites, utilize compute elements across the WLCG and OSG
- sites will need to opt-in to participate in order to not interfere with WLCG DC24 goals

Plans for DC24 and mini challenges

RENs stated their great interest and full support for WLCG DC24
DUNE, BelleII, JUNO, Pierre Augere will contribute

Items that can be tested during DC24:
- Packet and Flow marking
- SDN: NOTED, SENSE-Rucio
- Monitoring
- Cloud connectivity

Agreed to intensify the frequency of coordination meetings. DOMA general meetings will be used for reporting. The existing be-weekly LHCONE R&D call can be used for planning coordination [On-going]

Notes:
https://docs.google.com/document/d/1o08dzU1MDSWxco4SJ1phU9b8_MqtZTyW2o4V_Sy4oHs/edit?usp=sharing

DUNE use of LHCONE

DUNE officially requested to be allowed to use LHCONE, since most of the collaborating sites are already connected to LHCONE

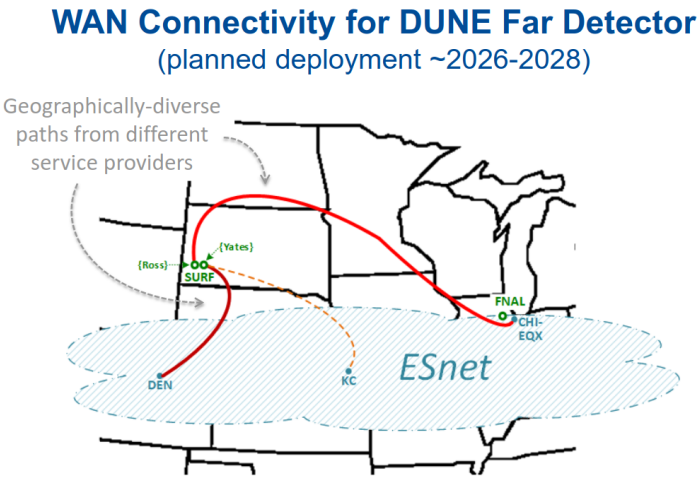
Computing and network requirements:

- Computing model: similar to LHC, but service-based
- Archival (raw data): 1st copy kept at FNAL, 2nd copy shared at other archival sites
- Computing sites: 35 sites distributed across 11 nations, nearly all are LHC sites
- DUNE projects its data volumes to be ~5-10% of a large LHC experiment

No objections in the room.

The procedure to include DUNE will be followed-up and the request presented to the WLCG management board for final endorsement

[request ready, discussion planned for September's MB]



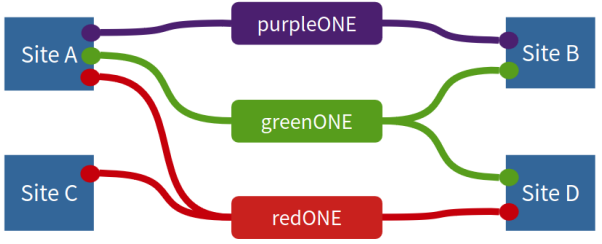
MultiONE: implementation proposal

Presented proposal about how to implement multiONE, multiple LHCONE-like VPNs for the different collaborations.

Goal: reduce exposure of sites to cyber-threats.

Main goal: be prepared on the day when a large(r) science domain ask to join LHCONE.

Proposal: MultiONE could be easily achieved if the routing into the correct VPN would be allowed to be loosely implemented.



Discussion followed:

- **Sites and RENs are not fond on increasing the network complexit**
- Sites recognize that the size of LHCONE is at its limit and an evaluation activity should be started. However, splitting the LHC experiments in different VPNs is not seen appropriate. New VPNs should be considered only for other large-science
- GEANT will elaborate a proposal and present it to the next meeting
- A coordination meeting involving LHC, SKAO, ITER will be organized for TNC24

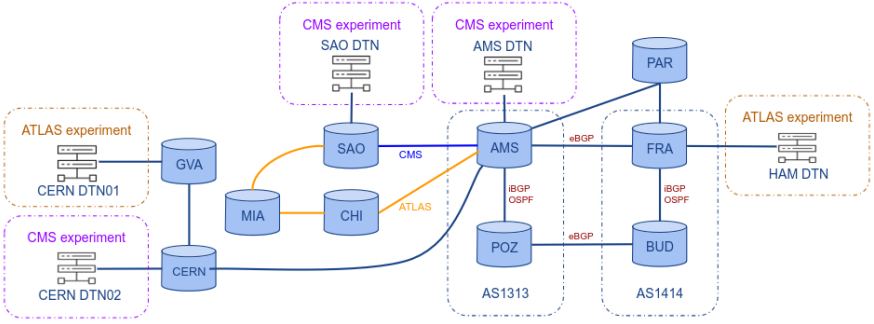
MultiONE demo implementation using P4 programmable switches

Presented prototype of multiONE implementation using P4 programmable switches

Implemented in the GEANT P4 Lab using Tofino switches running FreeRTR

- Emulated REN networks providing multiple VPNs

- Tofino switches at sites put the traffic in the correct VPN using the IPv6 Netflow value found in the packets



https://indico.cern.ch/event/1234127/contributions/5231153/attachments/2629625/4551233/CERN_LHCONE_Meeting50_P4lang_CarmenMisaMoreira.pdf

Use of Jumbo frames

Revamp of Jumbo frame deployment initiative
- Jumbo frames (large MTU, 9000 Bytes) can improve improve performance of data transfers
- The NREN networks already support Jumbo frames. LHCOPN and LHCONE too
- Transfer servers at LHCONE sites *should* support Jumbo frames

Discussion:
- Few WLCG sites are using Jumbo frames
- Others tried, but rolled back to normal frames because of unexpected operational issues
- Main concern are security filters that may break Path MTU discovery protocol
- Proposed to run survey to understand current use and interest in the community

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

Connectivity to Google cloud for ATLAS

ATLAS has an ongoing R&D project to demonstrate it can use Google Cloud resources to extend the current distributed computing infrastructure

So far the project has been successful. At this stage they need to calculate the Total Cost of Ownership (TCO).

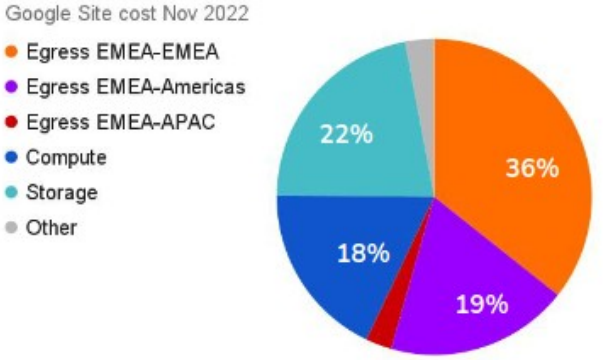
Cost of Network connectivity is variable and depends on the operating model of the computing elements and on the different connectivity options that Google can propose. In some scenarios the cost of the network reaches 60% of the TCO

ATLAS would like to test the Cloud Interconnect options with the help of a REN to understand how to reduce this cost

Discussion:

- GEANT is connected to AZURE and Oracle, but not Google
- ESnet and Internet2 are well connected to Google and can help

[on-going activity with ESnet to connect to Google in the US]



SC22 report and planning for SC23 demos

Two successful demonstrations were presented at SC22:

- NOTED with SENSE circuits between TRIUMF,CERN,KIT
- SciTags Packet and Flow marking at scale

Planning for SC23 (Denver)

- 1.2 Tbps StarLight facility to StarLight venue booth
- Network Research Exhibition (NRE) descriptions due June 2nd [submitted]
  - NRE: NOTED with the addition of FNAL
  - NRE: SciTags

https://indico.cern.ch/event/1234127/contributions/5231155/attachments/2631370/4551267/SC22%20Demonstrations%20Report%20And%20Plans%20For%20SC23%20Demonstrations.pdf

Tofino and P4: status and future

FreeRTR is the control plane developed by the RARE project

FreeRTR works on P4 programmable network processors and FPGAs

Intel has announce the discontinuation of the Tofino platform. No Tofino 3

FreeRTR will continue on the existing Tofino 1 and 2 platforms, while looking for alternatives (Broadcom and Marvel).

FABRIC and FAB update

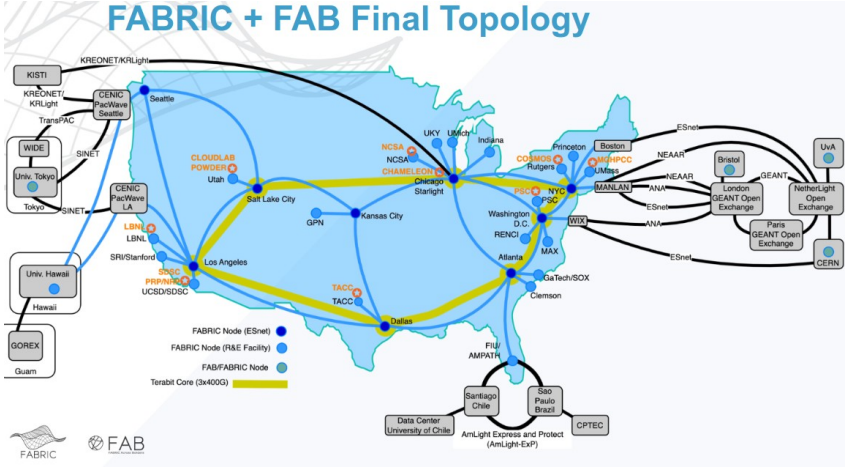
FABRIC is a distributed test-bed. Each node (hank) is composed by a set of advanced, programmable network and computing elements

FAB is the extension of FABRIC outside the US

- FABRIC: 27 sites deployed, connected at 100G and 400G via ESnet6

- FAB: CERN deployed, Bristol, Amsterdam [done] and Tokyo underway

The node at CERN will be connected to the CERN border routers with high speed access to the Tier-0 resources [done]



Conclusions

Summary

- LHCOPN: Two new Tier1s will soon connect. Almost 90% of the traffic is IPv6
- ESnet and Internet2 continue to upgrade their networks and are increasing the transatlantic capacity
- DC24: Network providers are eager to support it. BelleII, DUNE, JUNO and PierreAuger are willing to contribute. New network functionalities like NOTED and Flow marking will be tested
- MultiONE: not urgent, but activity should start in preparation of start of SKAO and ITER
- Jumbo frame: some interest in evaluating the possible benefits

Next meeting

University of Victoria (CA)
18-19 October 2023
Co-hosted with HEPiX Fall 2023

Agenda will be published here
https://indico.cern.ch/e/LHCOPNE51

References

Meeting agenda and presentations:
https://indico.cern.ch/e/lhcopne50

Questions?

edoardo.martelli@cern.ch