LHCOPN-LHCONE meeting #49 summary notes

GDB – 14 December 2022

edoardo.martelli@cern.ch

Venue

- 24-25 of October 2022
- Hosted at CERN, IT auditorium
- First in person meeting after the pandemic
- 35-40 people in presence and ~30 connected on Zoom
- Agenda at https://indico.cern.ch/e/LHCOPNE49

WLCG guidelines

Simone Campana, WLCG project leader, launched the meeting.

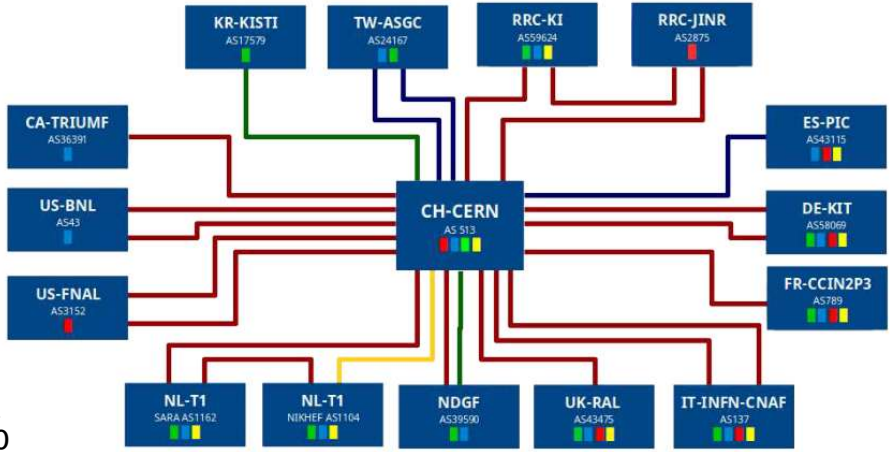
In the next 10 years WLCG will be faced with two major network challenges:
- dealing with the HL-LHC data volumes and complexity
- the cohabitation with other experiments and sciences on the same infrastructure

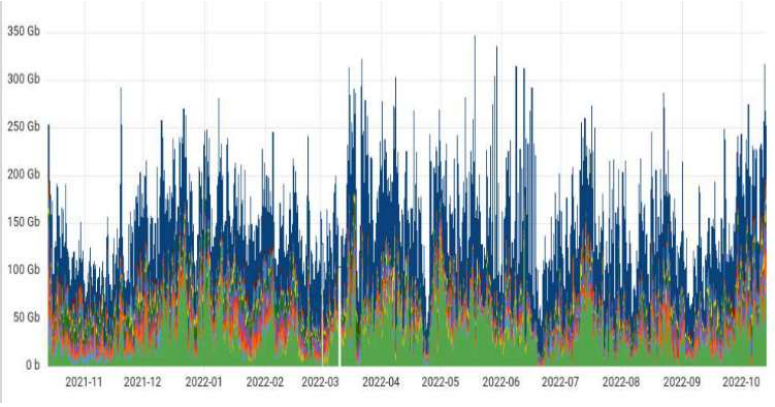
The network community can play a leading role:
- modernize the network services, progressing with the ongoing R&D activities and bringing early prototypes in production
- engage with other experiments and sciences to drive the evolution of R&E networks

https://indico.cern.ch/event/1146558/contributions/4893865/attachments/2533543/4360510/LHCONE-2022.pdf

LHCOPN - update

- Most of the link now 100Gbps. NIKHEF first site connected with 400Gbps link
- 1.9Tbps of aggregated bandwidth to the Tier0
- Traffic stats: moved 457PB in the last 12 months. +34% compared to previous year
- PIC's 100Gbps still waiting for GEANT upgrade [now fully operational]
- IT-INFN-CNAF now load balancing on their two 100Gbps links
- Construction of new CERN data-centre (PCC) progressing well. Procurement of servers starting soon. Ready to use in Q3 2023





NLT1 update

- SURFsara to be renamed SURF Internal Services
- SURF is working on increasing the LHCOPN and LHCONE capacity and redundancy for NIKHEF and SURFsara
- **LHCOPN connectivity will be re-engineered to add alternative back-up paths.** The Autonomous Systems of SURFsara and NIKHEF will appear behind SURF (Network) AS 1103
- Upgrade of equipment and links to 400Gbps is on-going
- SURF is implementing its own LHCONE VRF for the Dutch sites

https://indico.cern.ch/event/1146558/contributions/4907736/attachments/2533840/4360359/20221024%20-%20LHCOPN%20SURF%20update.pdf

LHCONE L3VPN - update

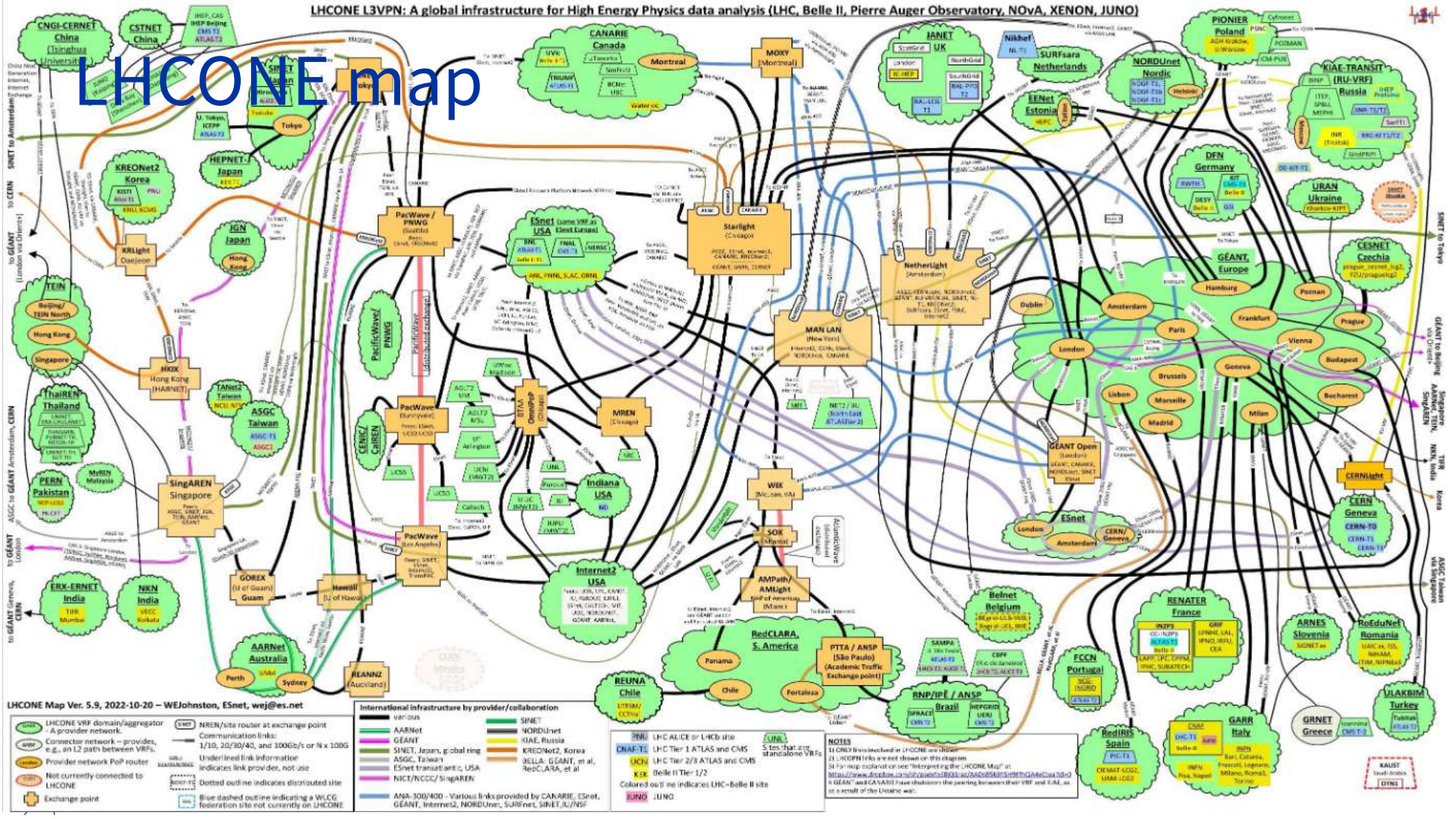
News

- SURF LHCONE instance being implemented
- GARR is upgrading the Italian LHCONE backbone to 400Gbps
- DFN has now a 200Gbps connection to GEANT LHCONE
- CERN is going to connect at 2x 400Gbps to GEANT LHCONE

Traffic statistics:

- Slovenia traffic has grown considerably, due to its HPC intensively used by WLCG
- LHCONE traffic has increased 30-40% in all regions
- IPv6 now dominating everywhere

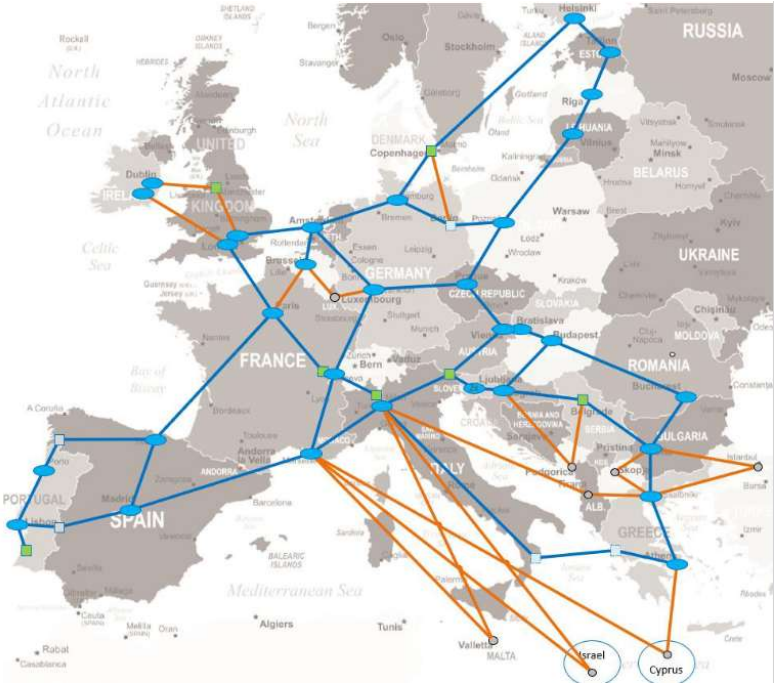
https://indico.cern.ch/event/1146558/contributions/5022652/attachments/2533838/4360354/2022-10-24_ECapone_LHCONE_L3VPN.pdf


LHCONE L3VPN: A global infrastructure for High Energy Physics data analysis (LHC, Belle II, Pierre Auger Observatory, NOvA, XENON, JUNO)
LHCONE map
LHCONE Map Ver. 5.9, 2022-10-20 – WEJohnston, ESnet, wej@es.net
International infrastructure by provider/collaboration
various
AARNet
GÉANT
SINET, Japan, global ring
ASGC, Taiwan
ESnet transatlantic, USA
NICT/NCCC/ SingAREN
ANA-300/400 - Various links provided by CANARIE, ESnet, GÉANT, Internet2, NORDUnet, SURFnet, SINET,R//NSF
SINET
NORDUnet
KIAE, Russia
KREONet2, Korea
RELLA: GÉANT, et al, RedCLARA, et al
LHC ALICE or LHCb site
LHC Tier 1 ATLAS and CMS
LHC Tier 2/3 ATLAS and CMS
Belle II Tier 1/2
Colored outline indicates LHC+Belle II site
JUNO
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
4) GÉANT and CANARIE have shutdown the peering between their VRF and ESNET as a result of the Ukraine war.

GEANT update

Developing its next generation network:

- i**ncreasing dark-fibre footprint**
- reducing power consumption ~20%
- reducing cost ~40%
- 15+ years fibre leases
- partially disaggregated Optical Network
- offering wavelength to NRENs
- tender for new IP/MPLS routers starting now



https://indico.cern.ch/event/1146558/contributions/4935053/attachments/2533862/4360500/G%C3%89ANT%20Network%20Update.pdf

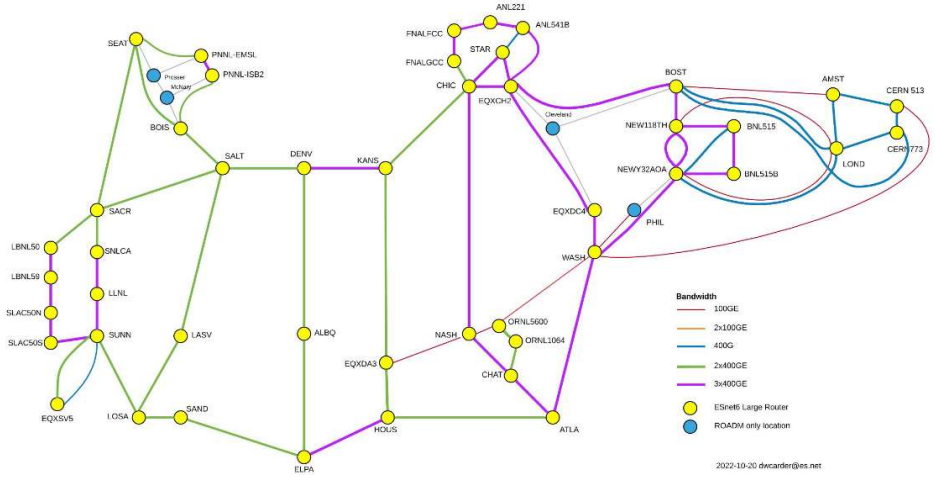
ESnet update

ESnet6 has been officially launched

Upgrades of EU links to 400Gbps is on-going

Trans-Atlantic capacity targets:
– 500G now
– 1.5T in Q3 2023
– 3.2T in 2027, well in advance of Run 4

US sites:
- BNL: now 300Gbps, soon 800Gbps
- FNAL: now 400Gbps, soon 800Gbps
- Tier2s will be upgraded to 400Gbps by 2027



https://indico.cern.ch/event/1146558/contributions/4950620/attachments/2533903/4361099/LHCONE%20Meeting%20-%20ESnet%20Update%202022-10-24%20version%202.pdf

LHCONE monitoring - update

- perfSONAR 5 beta is out and being tested. Still some stability issues due to the scale of the deployment and number of tests [post meeting note: v5 has been delayed to 2023]
- perfSONAR 5 will use Elasticsearch and Grafana
- Infrastructure: the message bus is being phased out and data will be sent directly to Elasticsearch
- 100G mesh: data is now shown correctly , but results are not great. Work in progress
- Sites need to plan to update hardware as well as keeping the perfSONAR software updated. Especially needed for DC24
- pS-Dash: Implemented AS traceroute, tool that hides the noise caused by load-balancing
- Total IN/OUT bandwidth: sites asked to provide URL with json of total in/out network counters. URLs stored in CRIC

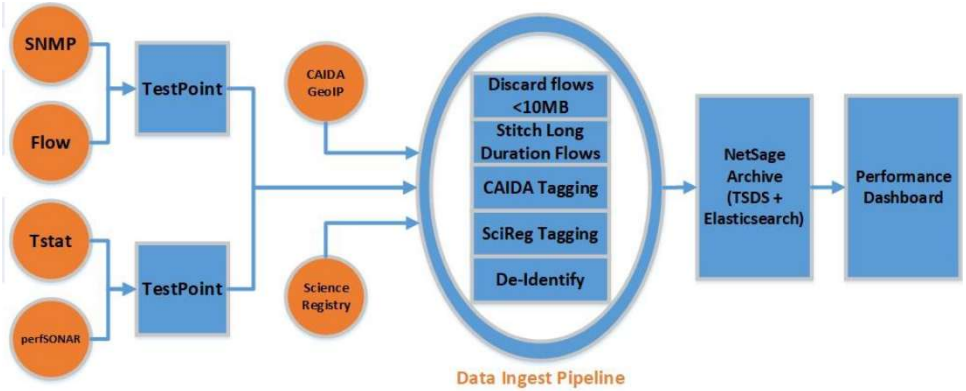
https://indico.cern.ch/event/1146558/contributions/5022661/attachments/2533607/4359869/LHCONE_LHCOPN%20Monitoring%20Update%20Fall%202022.pdf

NetSage update

NetSage is a tool to make network monitoring stats easily accessible and understandable
- It can ingest data from many sources: SNMP counters, net/s/flow, perfSONAR and others
- It is committed to privacy, fully GDPR compliant

News:
- Dev Team: several people has left, work is on-hold while re-organizing
- Work on ingest pipeline not yet completed. Moving to KAFKA

Pilot for LHCONE and LHCOPN
- https://lhc.netsage.global/
- Already showing data from CERN
  LHCOPN/ONE border routers



Other collaborations and sciences

Updates from Juno and BelleII, already members of LHCONE

Invited talks of ITER and SKA, major science projects which in the future may compete with WLCG on network utilization

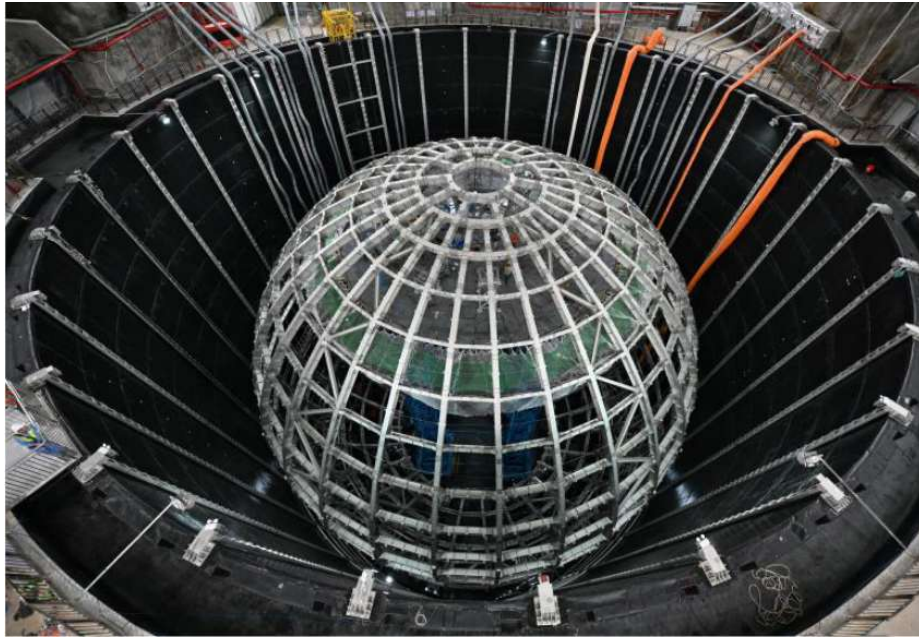
DUNE was also discussed: its sites are already members of LHCONE and it may formalize the use of LHCONE by signing the AUP

MultiONE ideas should be revamped

There's a growing need for major Science collaborations to coordinate their requirements to allow an organic grow of the R&E networks

JUNO update

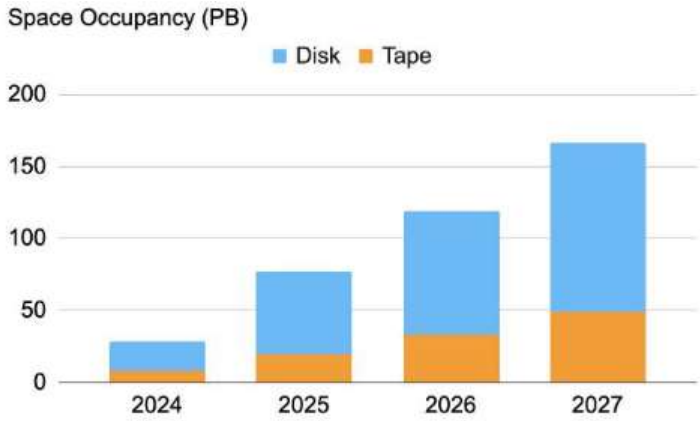
- Data generated at IHEP (CN), then sent to INFN-CNAF (IT) and from there to CCIN2P3 (FR) and JINR (RU)
- Expected volumes: 2PB/y raw data, 200TB/y reconstructed data
- Construction of detector is progressing well
- Data challenge:
  - mostly positive, but some sites below minimum requirement
  - The stop of LHCONE for Russia has reduced the transfer speed between IHEP and JINR



https://indico.cern.ch/event/1146558/contributions/5072290/attachments/2533749/4360156/JUNO%20update.pdf

BelleII update

- More than 2PB of RAW Data Collected so far, since 2019
- Estimated size of collected data-sets to grow O(10) PB per year
- Currently in Long Shutdown for upgrade
- **Data taking will start again in the last quarter of 2023**
- Using Rucio for data distribution
- On-going migrating to DAVS protocols for data transfers



https://indico.cern.ch/event/1146558/contributions/5050215/attachments/2534147/4360905/BelleII%20update%20-%20LHCONE_LHCOPN-Oct-2022-1.pdf

ITER

ITER Scientific Data and Computing Centre:

– The SDCC is under construction and expected to finish in 2023. Operation is scheduled for early 2024.

- Total scientific data rate is expected around 30-50+ GB/sec, total archive capacity 90-2200 TB/day. Data is expected to be in the Exabyte scale around 2035

On-going projects:

– ITER global connectivity via Marseille. 200-400 Gbit initial capacity, scalable to 3-6 Tbit/s

– Data storage design of complete data chain via PoCs, ongoing

– Cloud HPC burst capacity (AWS, Azure, Google Cloud - done) and Cloud Storage Test for long term archive and data distribution

– Data challenge tests planned with Renater, GEANT, ESNET etc. In 2023, a separated archive must be provided >50 km from the primary storage

https://indico.cern.ch/event/1146558/contributions/5011650/attachments/2534215/4361031/SDCC%20CERN%202022%20Presentation%20.pdf

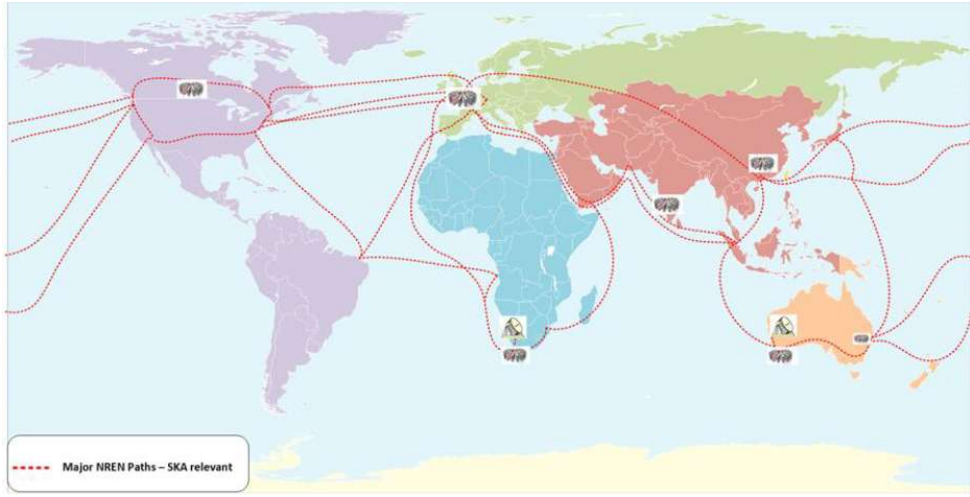
SKAnet

Observatories (data sources) in South Africa and Australia

Custodial and analyses in Europe, North-America and Asia

- Data stored: 300PB/year
- Data flows from observatories: 2x 100Gbps

GEANT is helping SKA to develop its network architecture. Proposed a VPN design similar to LHCONE



Needs to coordinate with WLCG for use of International backbones in the coming years

https://indico.cern.ch/event/1146558/contributions/5108504/attachments/2534190/4367594/SKAnet_Oct22_v2.pdf

LHC experiments

The LHCONE meetings at CERN give the opportunity to meet the computer coordinators of the major LHC experiments

ATLAS, ALICE, CMS and LHCb presented the latest updates on current and future use of the computing infrastructure

ATLAS

Excellent performance of the distributed computing infrastructure; Grid, HPCs, Cloud

- The WAN is mostly used for asynchronous Third Party Copies (TPC) with FTS and Rucio to place data where computing is available. FTS peaked at 480Gbps
- **Storage-less sites are a strategic evolution of the infrastructure**. Funding agencies are evaluating the possibility to consolidate storage in fewer sites
- **For storage-less sites the WAN should be at the level of the LAN**
- HPC is also important. There are HPC centres available in Africa and Asia, but the network to reach them is not very good. It would be very useful to improve connectivity there
- Caches have given good results
- Monitoring is necessary to understand and modulate the use of the network

https://indico.cern.ch/event/1146558/contributions/4917779/attachments/2534301/4361201/2022%20LHCOPN_ONE%20-%20ATLAS%20%26%20Networking.pdf

ALICE

ALICE is happy with LHCOPN/ONE and in general with the network performance

- The computing model favours local data access
- WAN used only for file replication and in case of issues with local storage
- WAN utilization is~ 4% of LAN

Run 3 model will continue using the same principles

- File transfers volume will continue at the current level
- T0 to T1s data transfer of Pb-Pb data: higher LHCOPN use for 2-3 months/year. More data produced, but no significant increase of pressure on LHC networking

ALICE would like to better understand the topology and capacity of the network to better place the data

LHCb

LHCb will increase network usage by an order of magnitude in Run3 and beyond

- Dominated by real data coming from the detector
- Fast and reliable network is at the basis of our successful computing operations and ultimately of the physics productivity of LHCb
- LHCb favours LAN over WAN
- when running on a Tier2, LHCb favours the national network before going abroad.

The new detector output has increased the throughput of 30x, however the connectivity requirements for Tier1 are well below the network in place

CMS

Networking is a strategic concern in an experiment like CMS with a flexible computing model.

Improving monitoring is key to improving usage:

- Making the XRootD monitoring more complete
- Packet marking for in-depth understanding of usage of both scheduled and unscheduled traffic

CMS encourages and participates in R&D and other improvements, including SDNs: NOTED, Rucio and SENSE integration

WLCG Data Challenge goals are a good baseline for HL-LHC needs

The storage landscape is becoming more heterogeneous and so the network has to evolve with it.

https://indico.cern.ch/event/1146558/contributions/5063563/attachments/2534544/4361711/221025-CMS-LHCOPN-LHCONE.pdf

WLCG Data Challenges

WLCG Data Challenge 2021

- Achieved expectations (10% of HL-LHC)
- Network not saturated, but somehow stressed at exchange points

Next data challenge (aka DC24)

- Originally planned for 2023, but most likely delayed to Spring 2024
- Comments on the target of 30% of HL-LHC requirements: **30% is 3x increase and sites may not have enough hardware; on the other hand, DC21 could have achieved more than 10%**
- Network providers have already planned upgrades. Network capacity should be enough for DC24
- The community needs to plan what is necessary to be tested. Packet marking and SDN should be part of it

https://indico.cern.ch/event/1146558/contributions/4908525/attachments/2534611/4361807/The%20path%20to%20the%20next%20WLCG%20Data%20Challenge.pdf

DC24 Planning

The LHCONE meeting was followed by a dedicated meeting for the preparation of the next WLCG Data Challenge in 2024

Important points discussed:

- This period before the challenge should be used to pre-test individual components with mini-challenges and milestones
- The duration of DC24 should be extended to allow to understand possible problems and also to allow to test different scenarios
- New network functionalities like packet marking and SDN projects should be part of the challenge
- Sites should not buy hardware just to meet DC24 requirements (the hardware may be obsolete by the time of Run4). Goal of 30% may be reviewed

Notes from the meeting here

The matter was also discussed at the WLCG workshop. Shawn's presentation is here

US-ATLAS/CMS workshop on HPC and Cloud

Mandate to conduct a blueprint process about the usage of Cloud/HPC resources in U.S. ATLAS and U.S. CMS collaborations

Main topics: Workflows, Integration, Total cost of operation, R&D and development,Benchmarking and Accounting, recommendations to ATLAS and CMS

ESnet evaluated the connectivity part:

- Sites and HPC centres connected by ESnet are already undergoing upgrades to 400Gbps
- ESnet backbone is already connected to major cloud providers at high speed
- Challenges:
  - flat traffic pattern not usual for commercial providers
  - providers may serve diffidently the various regions
  - LHC traffic may get mixed with other

https://indico.cern.ch/event/1146558/contributions/5086476/attachments/2534784/4362174/ESnet%20Report%20from%20the%20USATLAS-USCMS%20HPC_Cloud%20Blueprint%20workshop.pdf
https://indico.cern.ch/event/1146558/contributions/5086482/attachments/2534857/4362359/Joint%20USATLAS_USCMS%20HPC_Cloud%20workshop%20Summary.pdf

LHCONE R&D

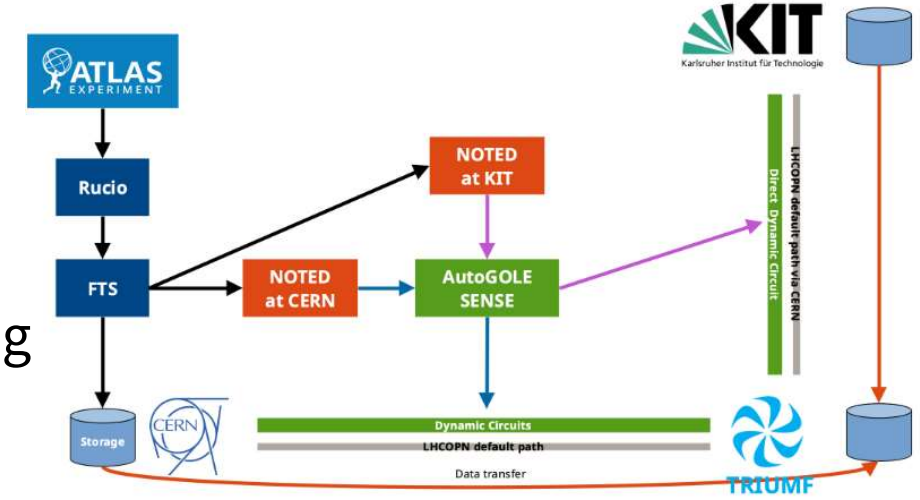
Large R&D session

Many projects presented status updates and future plans

NOTED and its SC22 demo

NOTED 2.0 is ready and being tested with production transfers
- The python code is now available on PiP for download.
- Evaluating possible integration with FTS to tune the FTS Optimizer when a new link is provisioned.

A demo for SC22 is in preparation.
It will be similar to the SC21 demo, but running the new Python code and with two separated instances, each one taking care of one of the paths TRIUMF-CERN and TRIUMF-KIT [Demo successful]



Research Network Technology WG - update

The RNTWG has made significant progress on network traffic visibility through the work on IPv6 flowlabel tagging and Firefly flow marking

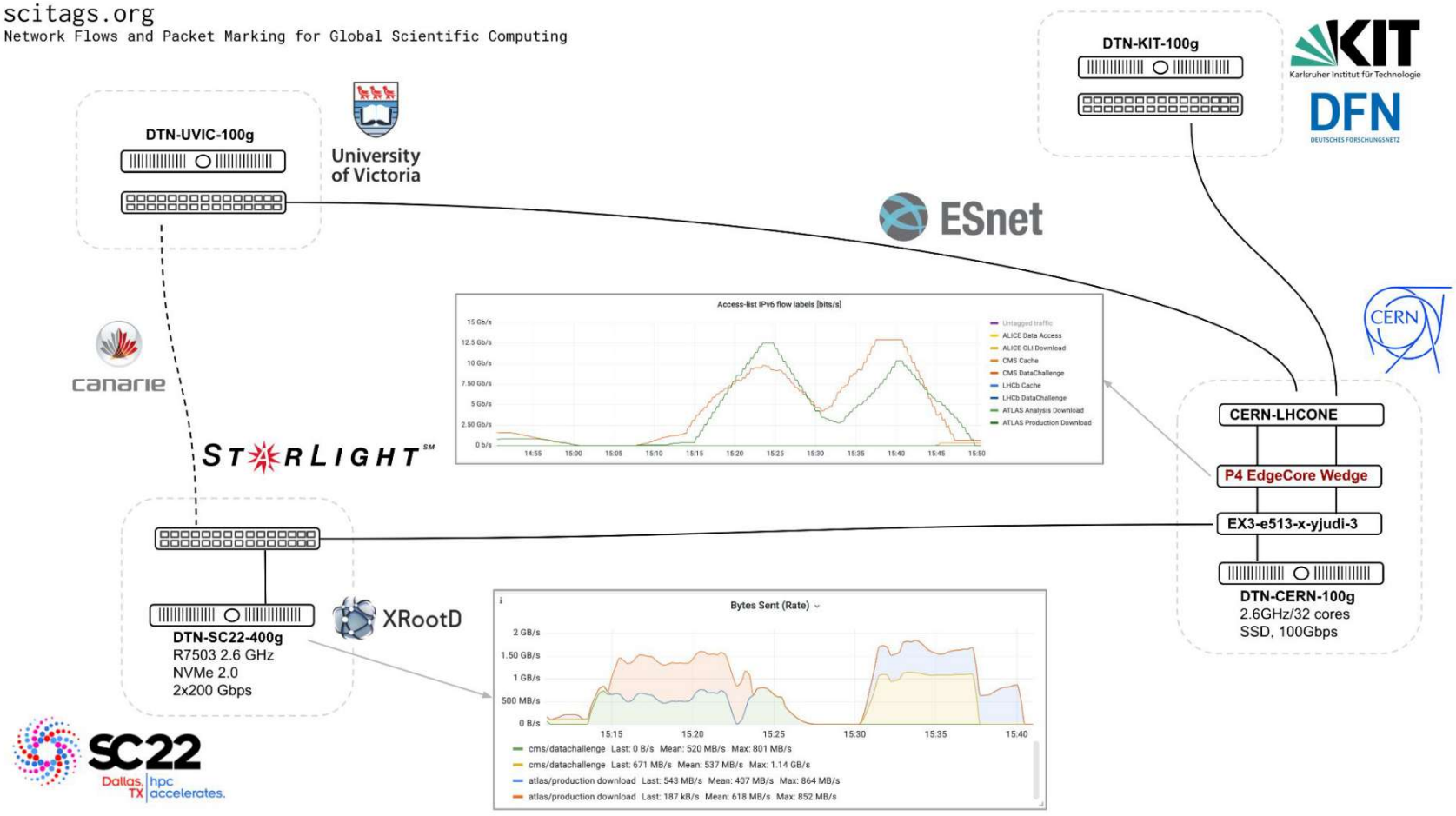
Flowd development: flow and packet marking service developed in Python
- Supports plugins to get to know which connections have to be marked
- Backends are used to make the marking and other tasks:
  - eBPF backend for flowlabel tagging
  - Prometeus backend to expose marked connections

scitags.org

Presented demo that will be shown to SC22 [also this demo was successful]

https://indico.cern.ch/event/1146558/contributions/5011791/attachments/2534747/4362099/LHCONE_LHCOPN%20Research%20Networking%20Technical%20WG%20Update%20%2349.pdf

Packet tagging demo at SC22



Use of SONIC and FreeRTR

Overview of Network OSes for open switches

Sonic: Open and multivendor, focused on data-centre, support traditional and programmable switches. Based on Linux

RARE/FreeRTR: Supported by the Research and Education community. Provides most of the Internet protocols. Runs on programmable switches, FPGAs and Linux DPDK

Global P4 lab: Network of SONIC and FreeRTR switches provided by the R&E community

https://indico.cern.ch/event/1146558/contributions/5011798/attachments/2534916/4362473/2022-10-25%20-%20LHCONE49%20-%20Use%20of%20SONIC%20and%20FreeRTR%20on%20programmable%20switches%20-%20Marcos%20Schwarz.pdf

PolkA

PolKA: Polynomial Key-based Architecture for Source Routing

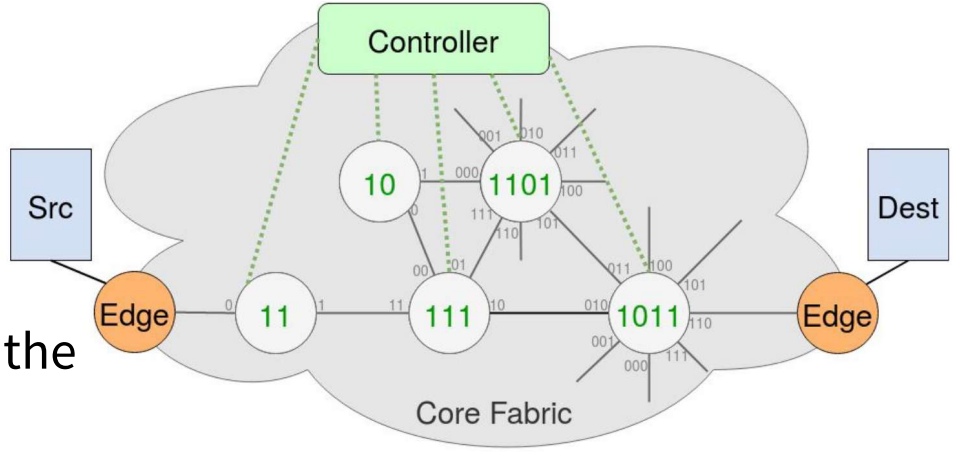
Source Routing approach that meets these requirements:

open source, interoperable, topology agnostic, multipath routing, fixed length header, implementable on programmable switches

Based on mathematics:

- Polynomial Residue Number System (RNS)
- Chinese Remainder Theorem (CRT)
- Packet forwarding based on an arithmetic operation: remainder of division

Being implemented in FreeRTR and tested in the GEANT P4 Lab



https://indico.cern.ch/event/1146558/contributions/5011804/attachments/2535013/4362653/LHC-ONE%20R%26D%20meeting%20PolKA-SR.pdf

AutoGOLE and SENSE - an update

AutoGOLE: Infrastructure which provides “end-to-end” network services in a fully automated manner

Open-source software framework based on:

- Network Service Interface (NSI): multidomain network provisioning
- SENSE: end-system provisioning and realtime integration with network services

Persistent Infrastructure, somewhere in between production and a testbed

AutoGOLE, NSI and SENSE work together to provide the mechanisms for complete end-to-end services that include network and attached End Systems DTNs

Circuit provisioning functionality being used by NOTED and Scitags demo for SC22

https://indico.cern.ch/event/1146558/contributions/5011801/attachments/2535212/4363029/2022-10-25-LHCOPN-LHCONE-AutoGOLE-SENSE.pdf

Transport Control Networks and FTS

TOP is an initial effort to design an application-layer, efficient, controllable transport scheduler system for data-intensive networks.

The transport scheduling: decides

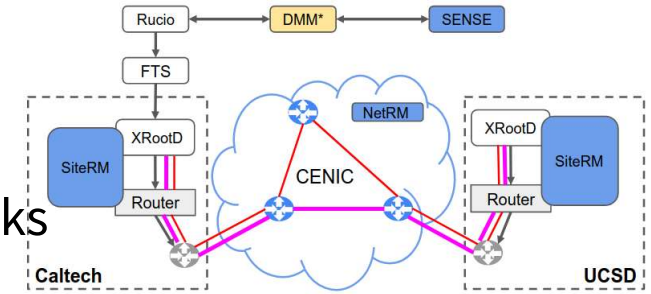
- when to start the transport of a transfer request,
- with how much transport resource
- Transport resource broadly defined: networking transport, storage transport, memory buffer used for transport

Benefits of well-designed transport scheduling

- Efficiency: not overwhelm transport resources, not leave resources unused
- Multiplex sharing: multiple users coexist and share the resources with control, extract statistical multiplexing gain
- Application-level objectives: schedule according to application needs

https://indico.cern.ch/event/1146558/contributions/5022826/attachments/2534888/4362418/2022-10-24-alto.pdf

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

Wish to participate to DC24

HEPiX IPv6 working group meeting

The in person meeting was co-hosted at CERN after the LHCONE/OPN meeting

Currently focusing on understanding why some traffic between dual-stack machines still use IPv4.

Lot of progress done recently thanks to the analyses work done on the logs of FTS and XRootD

The work of the WG in getting WLCG sites and applications ready for IPv6 has paved the way for an easier implementation of new techniques that can leverage special IPv6 functionalities , like the flowlabel tagging

Agenda at https://indico.cern.ch/event/1185115/

Conclusions

Summary

- LHCOPN and LHCONE traffic keeps increasing
- ESnet6 has been launched, GEANT upgrade is progressing well. Ready for Run3 and preparing for DC24
- LHC experiments satisfied with the performance of the network so far. They wishes more visibility on the behaviour of the network and the possibility to interact with it
- The packet and flow marking activities are progressing well and should be part of DC24
- The next WLCG Data challenge will most likely be delayed to 2024
  Stakeholders has to define all the components that have to be tested
- Coordination with ITER, SKA and other major science projects should be established before they start sending data

Next meeting (TBC)

Venue: Amsterdam TBC (Catania no longer an option)
Date: April 2023, exact days being discussed with SURF
In person meeting

Agenda will be published here
https://indico.cern.ch/e/LHCOPNE50

References

Meeting agenda and presentations:
https://indico.cern.ch/e/lhcopne49

Questions?

edoardo.martelli@cern.ch