LHCOPN, LHCONE and evolution in research network

ATCF8 TIFR Mumbai - 2nd September 2024

edoardo.martelli@cern.ch

Content

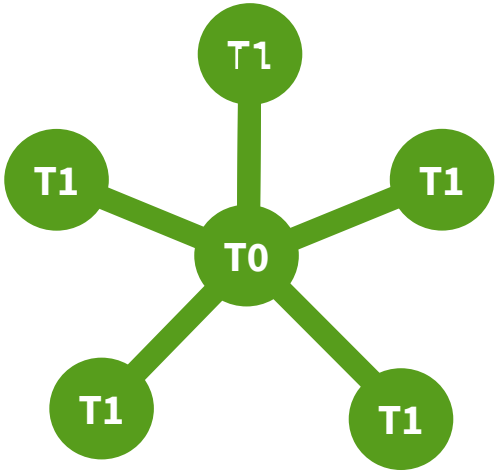
LHCOPN, LHCONE, MultiONE, DC24, research projects and transfer efficiency

LHCOPN

LHCOPN

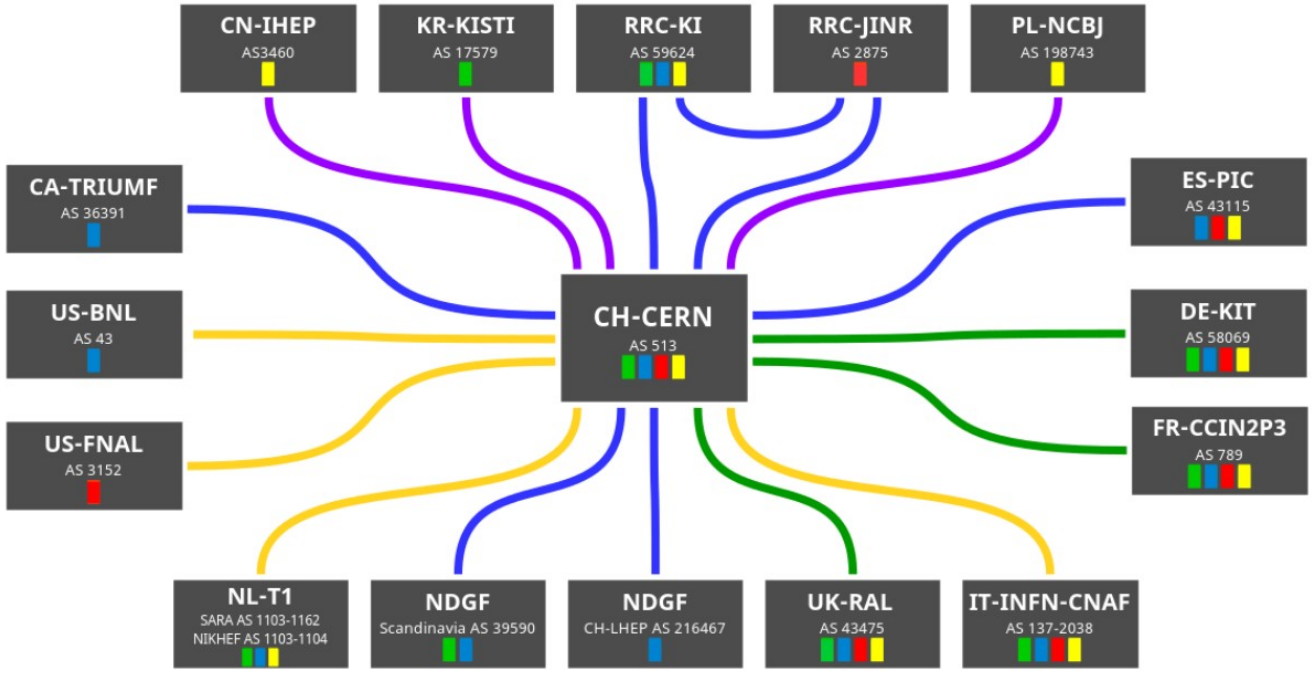
Private network connecting Tier0 and Tier1s

Secure:
- Dedicated to LHC data transfers
- Only declared IP prefixes can exchange traffic
- Can connect directly to Science-DMZ,
  bypass perimeter firewalls



Advanced routing:
- BGP communities for traffic engineering

LHCOPN



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

Numbers

- 17 sites for 15 Tier1s + 1 Tier0
- 14 countries in 3 continents
- 2.86 Tbps to the Tier0
- CN-IHEP, PL-NCBJ and NDFG-LHEP last additions
- TW-ASGC has left

Latest news

CH-CERN:

- The Prevessin Data Centre (PDC) is ready and in production
- Two rooms equipped with batch nodes

CERN Prevessin Data Centre

NLT1:

- SURF has tested a 800Gbps link on a single wavelength CERN-Amsterdam
- Dutch Tier1s will be connected with 2x 400Gbps (currently 400Gbps)

LHCOPN latest news

CH-LHEP (NDGF):

- Activated 100Gbps primary link. Provided by SWITCH

CN-IHEP

- Activated primary (via Marseille) and backup (via London) links, 20Gbps each.
  Provided by GEANT and CERnet

FR-IN2P3

- Activated second 100Gbps link and configured in load-balancing with existing one.
  Provided by RENATER

LHCOPN latest news

IT-INFN-CNAF:

- 4x100Gbps over DCI connection activated and used during DC24, now in production
- Legacy 2x 100Gbps will be kept as backup
- New CNAF data-centre being equipped and expected in production in October 2024.
  It will be connected with multiple 400Gbps from the DCI system

US-FNAL:

- Now with 400Gbps capacity for LHCOPN

US-BNL:

- Now with 400Gbps capacity for LHCOPN

LHCOPN latest news

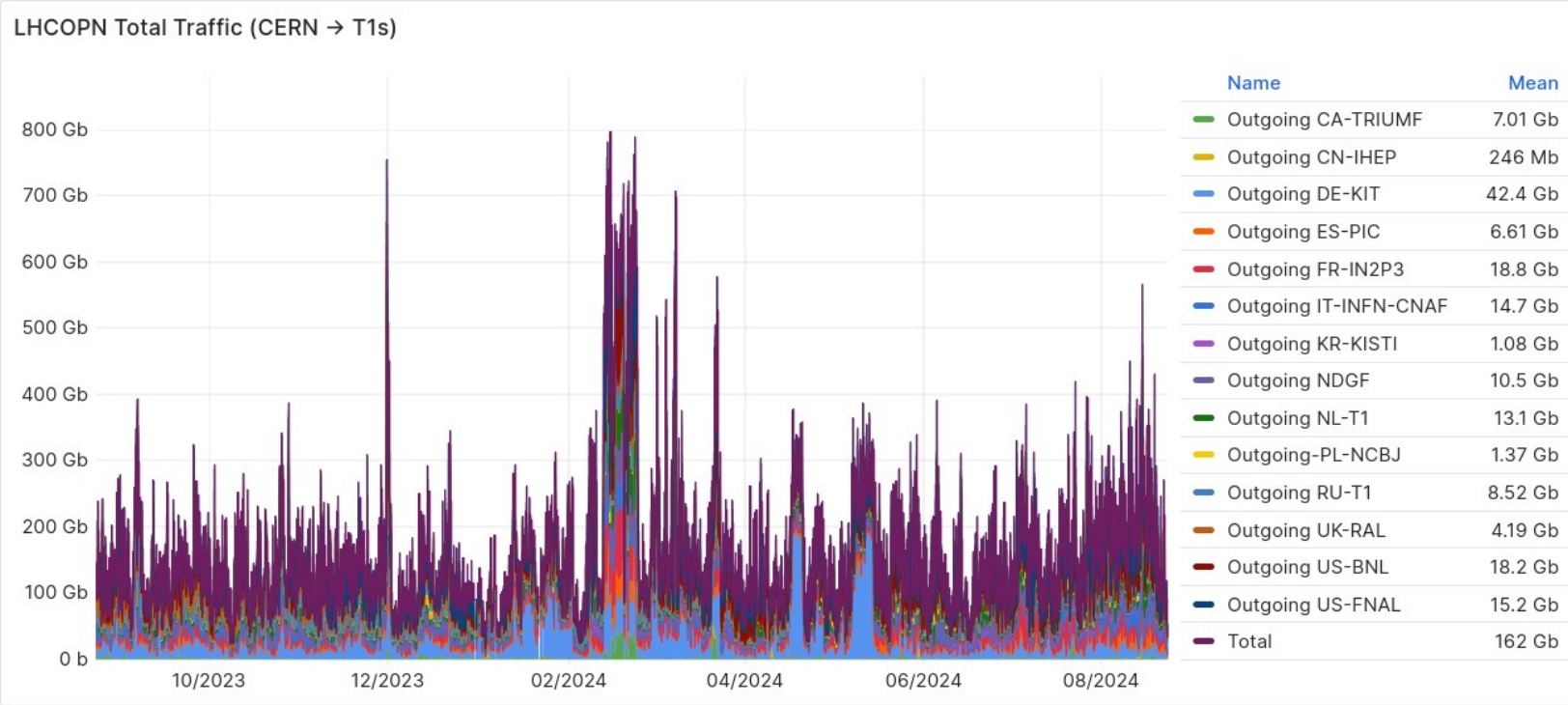
New Serbian CMS Tier1:

- The Vinča Institute of Nuclear Sciences in Belgrade is working to become a CMS Tier1. They have a datacentre in Kragujevac, south of Belgrad.
- Developed project document and designated the project leader. Work in progress

TW-ASGC:

- Phase-out completed. All LHCOPN links and peerings removed. ASGC stays as a Tier2

LHCOPN Traffic – last 12 months



Numbers:

Moved ~638 PB in the last 12 months

+18% compared to previous year (540PB)

Peak at ~800Gbps during DC24

Ref: https://monit-grafana-open.cern.ch/d/HreVOyc7z/all-lhcopn-traffic?orgId=16

LHCONE

LHCONE L3VPN service

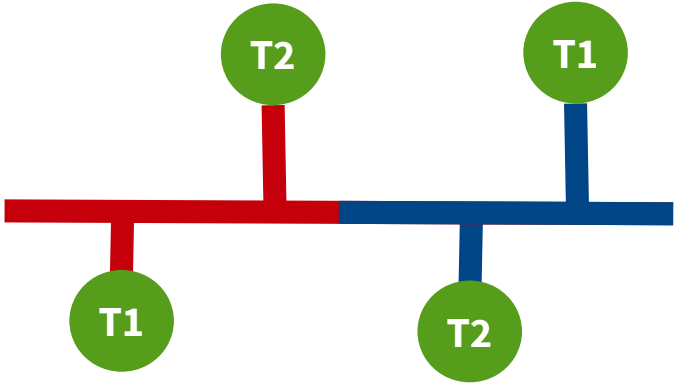
Private network connecting Tier1s and Tier2s

Secure:

- Dedicated to LHC data transfers
- Only declared IP prefixes can exchange traffic
- Can connect directly to Science-DMZ, bypass perimeter firewalls

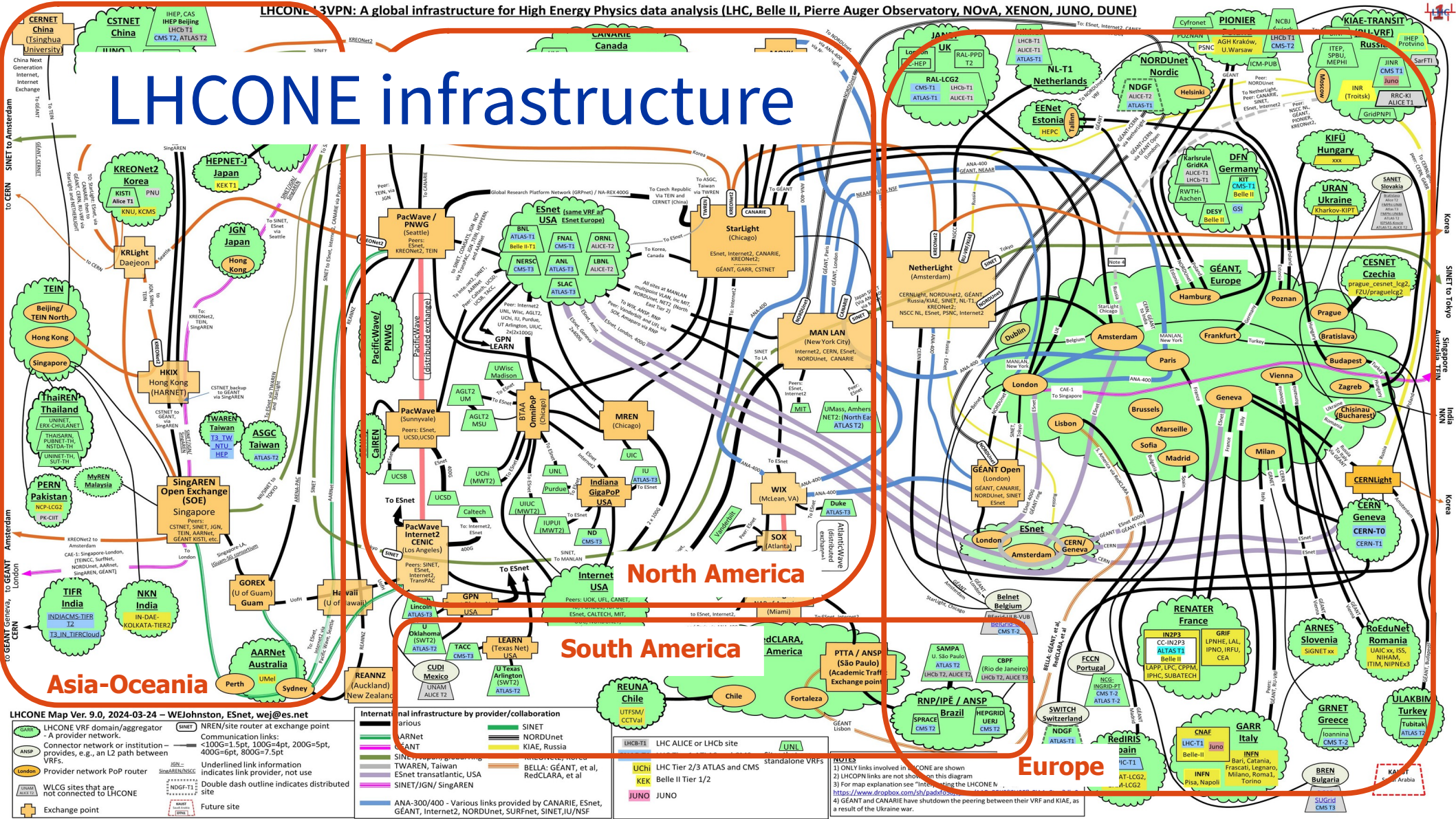
Advanced routing:

- Multi domain L3 VPN
- BGP communities for traffic engineering



Open to other HEP collaborations

LHCONE infrastructure



LHCONE L3VPN – latest news

WLCG Data Challenge 2024 (DC24) impact well visible in all the LHCONE networks

New LHCONE network providers:

- SWITCH (Switzerland) for University of Bern-LHEP
- FCCN (Portugal) for NCG-INGRID-PT

Capacity upgrades:

- GARR (Italy) access to GEANT upgraded to 2 x 300G (MIL, MAR)
- RENATER (France) access to GEANT upgraded to 400G in Geneva, 300G in Paris (about to become 400G)

LHCONE status

- VRFs: 30 national and international Research Networks

- Connected sites: ~110 in Europe, North and South America,Asia, Australia

- Trans-Atlantic connectivity provided by ESnet, GEANT, Internet2, RedCLARA, NORDUnet, CANARIE and SURF

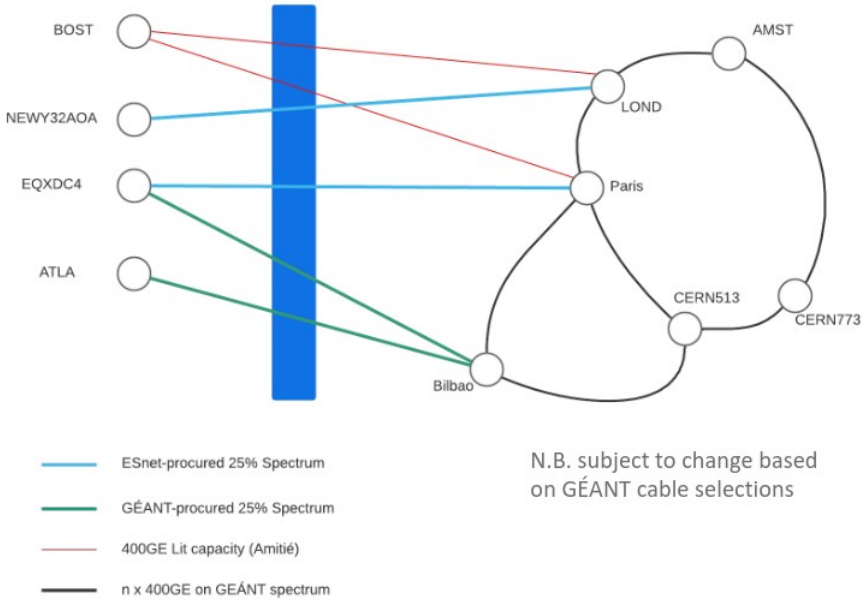
- Trans-Pacific connectivity provided by KREOnet, SINET, TransPAC

- Interconnections at Open Exchange Points including NetherLight, StarLight, MANLAN, WIX, CERNlight, Hong Kong, Singapore and others

ESnet update

Trans-Atlantic upgrades

- Now In Production:
  - 400G New York - London
  - 400G Boston - CERN
  - 400G Boston - London
- Trans-Atlantic capacity targets:
  - 3.2T in 2027, in advance of Run 4



US - Europe Connectivity Plans

- Collaborating with GEANT to share spectrum on subsea
- Two additional EU PoPs: Paris (firm), Bilbao? (tbd)
- n x 400G EU rings in partnership with GÉANT

US-EU traffic engineering

- Most of the US<->EU LHCONE traffic use ESnet transatlantic links

perfSONAR monitoring update

Updates to perfSONAR and OSG/WLCG network measurement platform
- perfSONAR 5.1 is out with new features; it requires sites to update OS.
- Plan to adapt the network measurement platform to benefit from changes in 5.1

Ongoing efforts in network analytics and ML methods for WLCG data
- Focus on pre-processing (gaps, predictive models) and anomaly detection
- Opportunity to collaborate on models and data sets

Monthly meetings with site network teams:
- Discuss how sites are deploying, managing and planning for WLCG networking requirements

perfSONAR

Network information in CRIC

CRIC (Computing Resources Information Catalogue) is the database used by WLCG to document the available resources. It is used also to store network information related to LHCOPN and LHCONE

Easily accessible

- Netsite: https://wlcg-cric.cern.ch/core/netsite/list/ (login required)
- NetworkRoute: https://wlcg-cric.cern.ch/core/networkroute/list/ (login required)
- Json view: https://wlcg-cric.cern.ch/api/core/rcsite/query/?json (no login)

MultiONE with BGP communities

Trust of LHCONE

The major benefit of LHCONE is the trust in the connected sites: it allows the **LHCONE fat links to bypass slow and expensive security inspection**

Due to the inclusions of other collaborations (BelleII, DUNE…), the increasingly growing number of connected sites may reduce the trust

The MultiONE project aims to reduce the exposure of the sites, so to increase the trust in LHCONE

Agreed MultiONE implementation

Don’t add any additional VPN

Each prefix announced to LHCONE is tagged with BGP communities* that identify the collaborations served by the site

The tagging is done by the sites. Or by the connecting Network Provider, if a site is unable to do it

Later on, sites can decide to drop prefixes of collaborations they are not working with

** BGP is the routing protocol used in LHCONE. BGP communities are numeric tags that can be added to the network prefixes announced to the BGP peers*

Benefits

- Simple and commonly used technique, no additional VPNs to configure
- Tags are useful to document the use of the network and to double check what is declared in CRIC
- Reduced exposures of sites when filtering will be implemented
- Tagging and Filtering can be implemented progressively
- No changes at sites when a new site connects to LHCONE

Limitations

Not 100% secure:

- Any sites will still be able to send packets to sites that tag and filter. However TCP connections should not work.

- A malicious sites can tag its prefixes with all the existing tags and get the prefix accepted. This could be mitigated if Network Providers validate the tagging

Implementation

1 - Reach out all the LHCONE sites and request to implement the tagging, while reviewing their prefix declarations in CRIC (on-going)

2 - Monitor the progress of the tagging in the LHCONE routing tables

Milestone: all prefixes tagged by LHCONE meeting #54 (Spring 2025)

3 - Implement filtering at (some) WLCG sites during year 1 of LHC LS3 (2027) and in preparation for the next Data Challenge

Documentation

https://twiki.cern.ch/twiki/bin/view/LHCONE/MultiOneBGPcommunities

Community format:

- Standard BGP community
- **Format: 61139:ExpID**
- The ExpIDs are defined by the SciTags initiative

Collaboration	BGP Community (AS:ExpId)
ALICE	61339:5
ATLAS	61339:2
BelleII	61339:6
CMS	61339:3
DUNE	61339:8
ILC	61339:10
JUNO	61339:12
LHCb	61339:4
NOvA	61339:13
Pierre Auger Observatory	61339:11
XENON	61339:14
Operational	
perfSONAR servers	61339:60001
LHCONE backbone	61339:60002
Demo/Prototype/Lab	61339:60003

What your site should do

1) Identify the experiments served by your site
2) Tag your LHCONE IP prefixes with the correspondent BGP communities

E.g.
- Site ABC participate to ATLAS, CMS and BelleII

- ABC's prefixes announced to LHCONE must be tagged with the BGP communities 61339:2, 61339:3, 61339:3

Status update

Date: 20240823
Number of tagged prefixes: 41
Total number of LHCONE prefixes: 722
Number of ASes originating tagged prefixes: 9

List of tagging ASes:
3,16, 160, 513, 2505, 3152, 35296, 43115, 58069

Network R&D during WLCG Data Challenges 2024

Data Challenges for HL-LHC

WLCG has planned for a series of data challenges to prepare for HL-LHC data taking

- Demonstrate readiness for the expected HL-LHC data rates with:
  - Increasing volume/rates
  - Increase complexity (e.g. additional technology)
- A data challenge roughly every two years

2021: 10% of HL-LHC requirements *(480Gbps minimal – 960Gbps flexible)*
2024: 25% of HL-LHC requirements *(1.2Tbps minimal – 2.4Tbps flexible)*
2026/7: 50% of HL-LHC requirements *(date and % to be confirmed)*
2028/9: 100% of HL-LHC requirements *(date and % to be confirmed)*
2030: start of HL-LHC (Run4) *(4.8Tbps minimal – 9.6Tbps flexible)*

HL-LHC network requirements

ATLAS & CMS T0 to T1 per experiment

- 350PB raw data per year; average of 50GB/s or 400Gbps during LHC running time
- Another 100Gbps estimated for prompt reconstruction data tiers (AOD, other derived output)
- estimated 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(safety-margin) =

  4.8Tbps expected HL-LHC bandwidth

Flexible Model

- Experiments may need to reprocess and reconstruct the collected data during the year
- This requires doubling the bandwidth of the Minimal model:

  9.6Tbps expected HL-LHC bandwidth

Overall network requirements for HL-LHC

Each Major Tier1s:

1 Tbps to the Tier0 (LHCOPN)

1 Tbps to the Tier2s (aggregated, LHCONE)

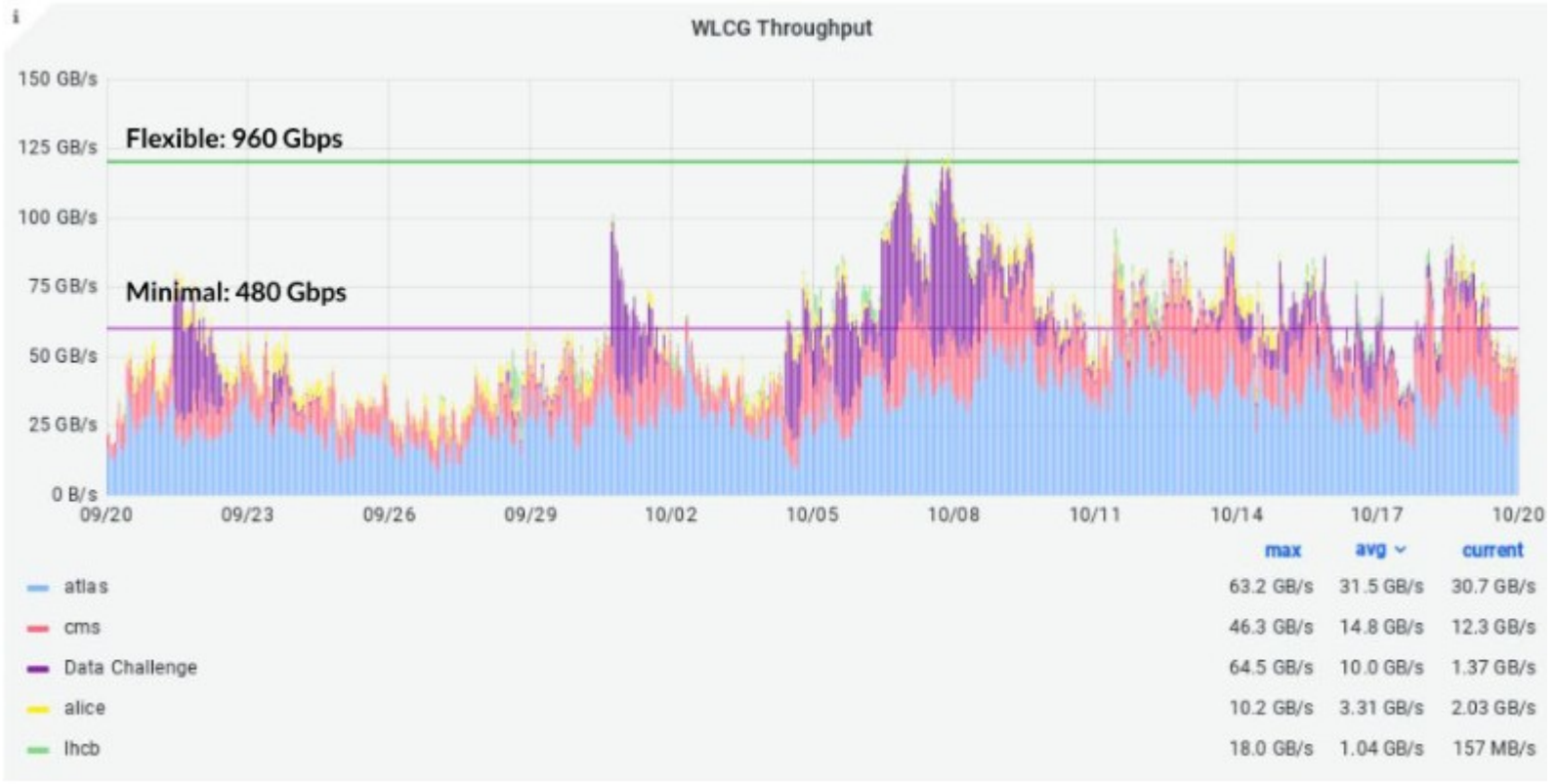
Each Major Tier2s:

>400 Gbps (LHCONE)

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

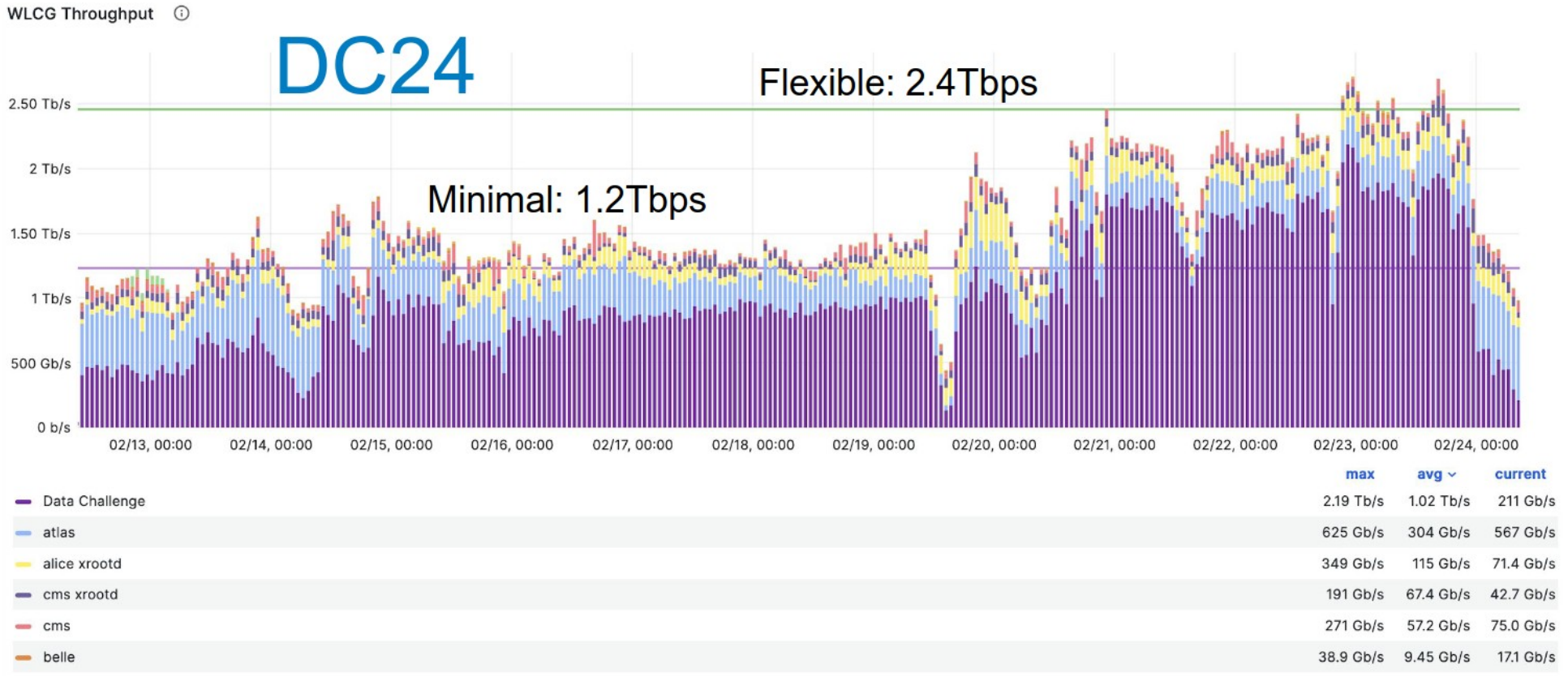
DC21

Mostly production transfers. Flexible model reached for a short time



DC24

Achieved full throughput of minimal model. Flexible model achieved only for short time.



DC24 results: applications

- Ran 12-23 February 2024
- Real data moved disk-to-disk using production applications: ad-hoc transfers on top of production traffic
- Applications pushed to uncharted levels showed unexpected limitations
- Using tokens for authorization added instabilities, but testing at such scale it was a necessary step to take

DC24 results: Networking

- **Global Research & Educations (REN) networks demonstrated more than sufficient capacity and reliability during DC24 and were NOT a bottleneck for any of the experiments.** Unexpected undersea cable cuts properly backed-up by alternative connectivity

- Some sites did identify local network bottlenecks or non-optimal architectures

- Various network technologies (NOTED, SENSE, BBR, perfSONAR, SciTags, Spectrum sharing…) were successfully tested during DC24 and showed promising results, now working to put them into production.

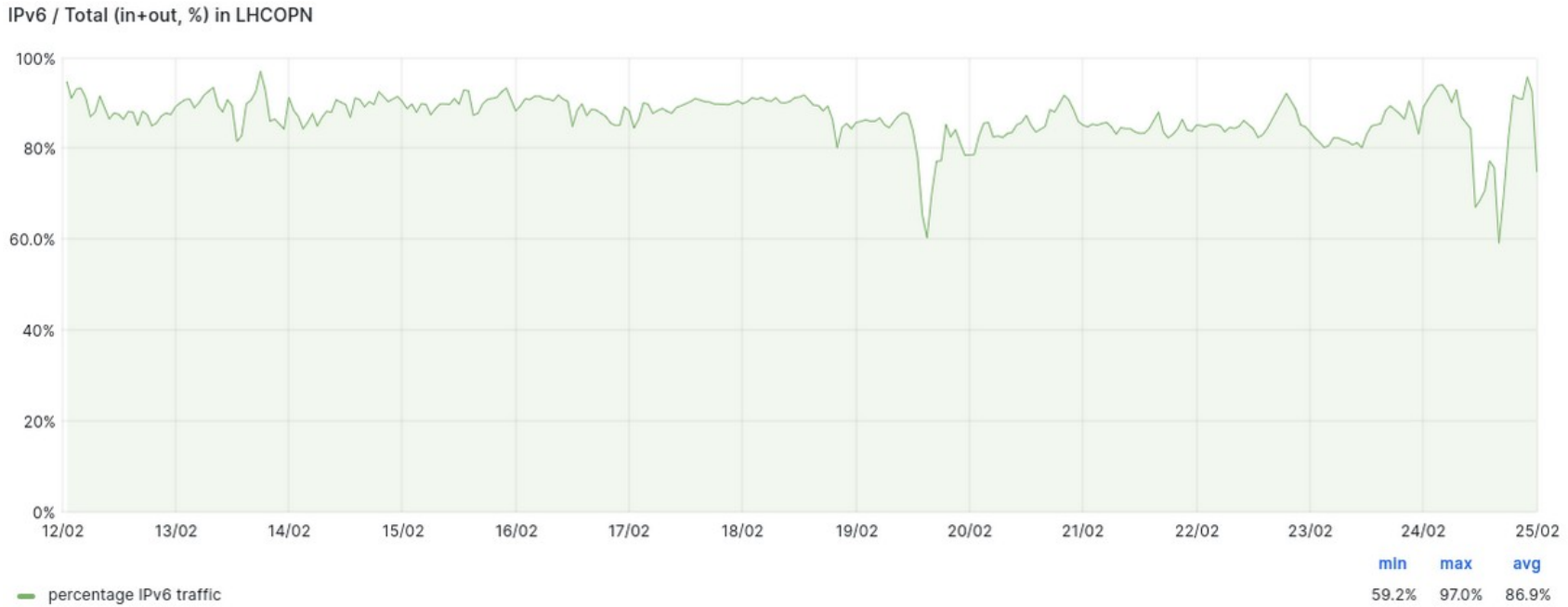
- Efforts to better monitor networks has become part of WLCG operational toolkits

- Regular mini-challenges will be run to track progress and prepare for DC26/7

- Storage infrastructure and middleware are being improved and the networks need to keep pace. Bandwidth upgrades may be needed for DC26/7

DC24 Network measurements: IPv6

- IPv6 Traffic in LHCOPN: 86.9% of the total
- Some sites testing IPv6 because of IPv4 scarcity

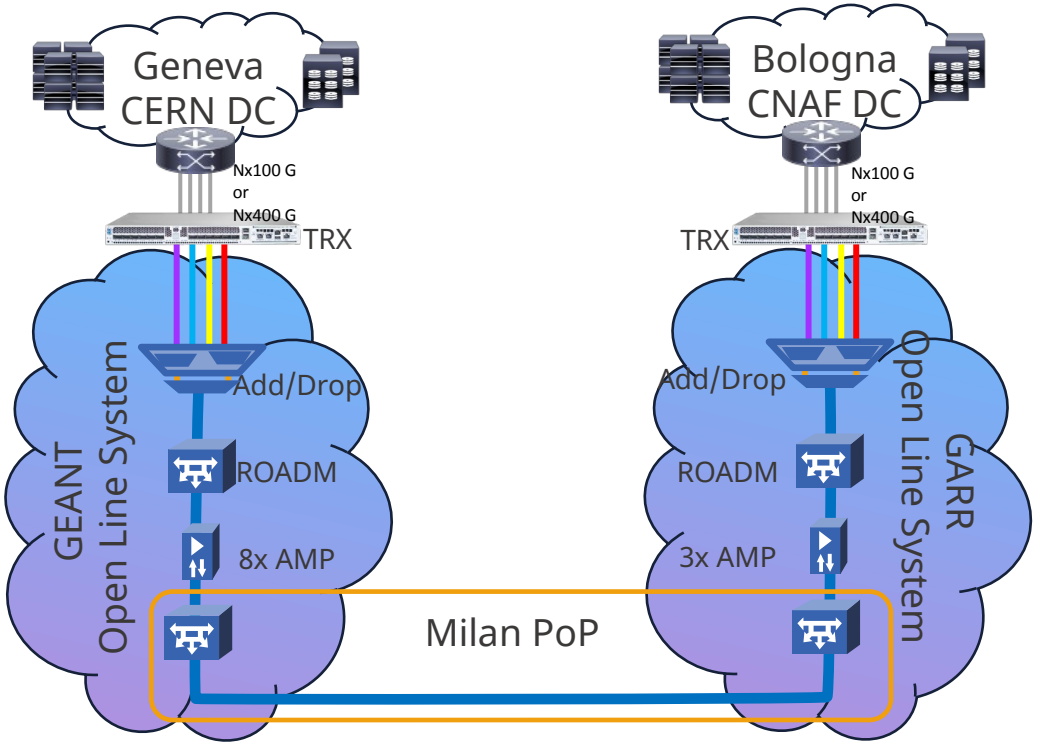


DC 24 Network R&D: CNAF-CERN DCI

- spectrum sharing over GEANT and GARR dark fibres

- 4x100Gbps links between CERN and CNAF used for DC24 and now in production

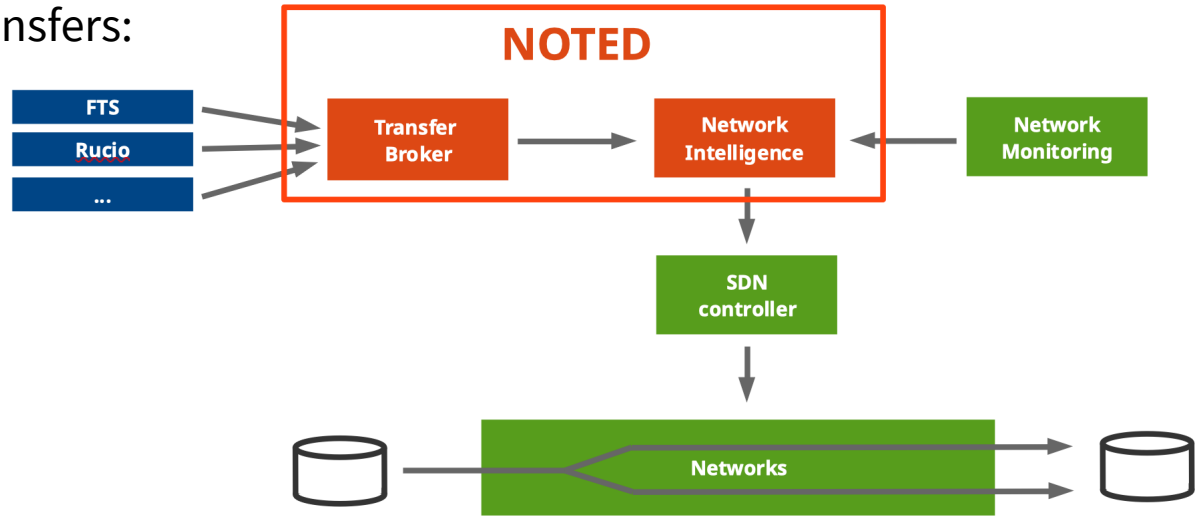
- cost effective technique to get >1Tbps LHCOPN connections already today



DC24 Network R&D: NOTED SDN

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

Already used with production data transfers:
- During SC22 and SC23
- New version with triggers from Network Monitoring tested during DC24
- Aiming to production-ready version for DC26



NOTED: DC24 test description

NOTED ran during the whole DC24

On the first 10 days it was not taking any action, just triggering a warning when additional bandwidth was needed

On the last 3 days, NOTED was taking real actions for TRIUMF, PIC and KIT

- TRIUMF: load-balancing traffic over the primary and backup links
- PIC and KIT: load-balancing over the LHCOPN and LHCONE links

NOTED DC24 test: conclusions

Useful exercise to fix some NOTED driven router configuration issues (e.g. load-balancing routing when a site had multiple LHCOPN links, full BGP config on both the LHCOPN routers…)

Also useful to implement additional checks on FTS

Overall happy of the results: Large Transfers detection worked properly in most of the cases

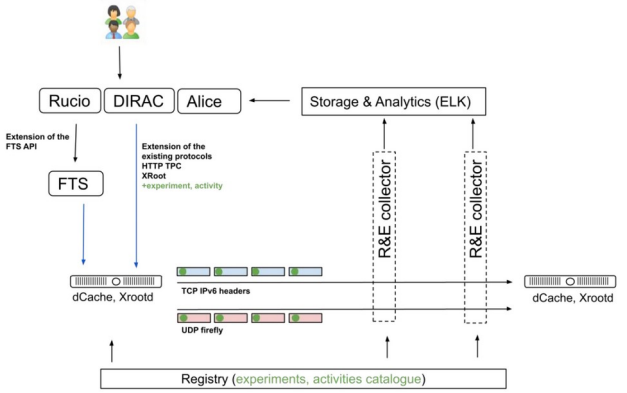
Though no congestion to demonstrate the added value

DC24 Network R&D: SciTags

Science Tags: marking of data packets and flows with Experiment and Application IDs for better network accounting

Two options being implemented:

- Tag in the IPv6 flowlabel field (proposed IETF draft: draft-cc-v6ops-wlcg-flow-label-marking)
- Tags (and more infor) in UDP fireflies (UDP packets sent in parallel to each flow)



SciTags at DC24: results

SciTags was largely tested during DC24:

- 80% of EOS CMS (production), UNL production storage
- Flow labeling functionality (fireflies)

Results:

- Confirmed the capability to propagate Scitags all the way to the storages (for both ATLAS and CMS)
- Sending fireflies (from XRootd, EOS storages)
- Collection and visualisation at ESnet collector worked properly

SciTags Implementation status

Propagation:

- Rucio supports Scitags from 32.4.0
- FTS/gfal2 support Scitags from 3.2.10/2.21.0

Storages:

- XRootD provides Scitags implementation (from 5.0+)
- EOS provides Scitags support from 5.2.19+
- Working on a project for production rollout at CERN (for WLCG)
- dCache prototype exists, roadmap for release pending
- Also working with StoRM and Pelican

Packet Pacing

A small amount of packet loss makes a huge difference in TCP performance, especially on long distance flows. A proper pacing of the packets travelling on the network can prevent losses

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 TCP flows

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

BBRv1 DC24 test

During the last 3 days, 40 EOS servers at CERN used by ATLAS and CMS were flipping their TCP congestion protocol every 2 hours (every 6 hours from Thursday on)

Results:

- No evidence of gain nor loss using BBRv1. It shows advantages in congested lines, but there wasn't much congestion in the networks

Data transfers performance

Slow data transfers

ESnet discussed some findings they observed in the packet by packet analyses of the DC24 traffic, using the data produce by the ESnet High Touch service.

- **Almost no sign of retransmission**. Which indicates no congestion nor problematic links. And makes the need for Jumbo frames and BBR and other congestion avoidance measures less stringent

- **The average flow travelled at less than 1Gbps**. There is room for improvement

Use of Jumbo frames

Benefits of jumbo frames are evident on long distance transfers, less on the short distance

Operational issues are also evident, but they can be mitigated by sharing deployment experiences

Agreed to put effort on testing at larger scale. CERN (historically reluctant) has agreed to push the testing on some production servers

Jumbo frames survey

Results of Jumbo Frame survey among WLCG/LHCONE sites

53 answers received during ~1 month
- 26 sites have Jumbo in the storage elements (50%)
- 19 sites of the 26 have Jumbo also on worker nodes (40%)
- 4 sites of the 26 didn’t have any problem with implementing Jumbo (20%)
- 10 sites of the 20 without Jumbo don’t want to implement Jumbo (~20% of total)

BBRv3 preliminary results

Results of BBRv3 testing run by ESnet:

- BBRv3 is still not part of the mainline Linux Kernel

- BBR helps on some paths
  – only minor improvements over CUBIC on clean paths with large buffered devices
  – need to find paths will small buffered devices or congestion to really see improvements

- Detailed analysis of pacing behaviour is difficult: more work needed

Jumbo and BBR: next steps

Agreed to continue the work on BBR and Jumbo frames.

BBRv3 requires more testing in congested situation

Jumbo frames may improve the efficiency of the file transfer by relieving the storage servers' CPU and allow higher throughput for FTS.

CERN will put more effort in setting up a pilot to have a group of EOS servers running with Jumbo frames and run tests with other sites

Conclusions

Summary

- LHCOPN: upgraded several links for DC24; three new Tier1s connected
- LHCONE: continues to grow
- MultiONE: implementation with BGP communities started
- WLCG DC24: positive results from the network, R&D projects could be used for future challenges
- Data transfer efficiency can be improved

Questions?

edoardo.martelli@cern.ch