tnc23
DIGITAL GENERATIONS
TIRANA, ALBANIA | 5-9 JUNE 2023

Networking at the WLCG: R&D activities

Carmen MISA MOREIRA

Edoardo MARTELLI

CERN IT Department CS Group

6th June 2023

GÉANT

Co-funded by the European Union

Outline

LHC, WLCG, networking

WLCG Network R&D

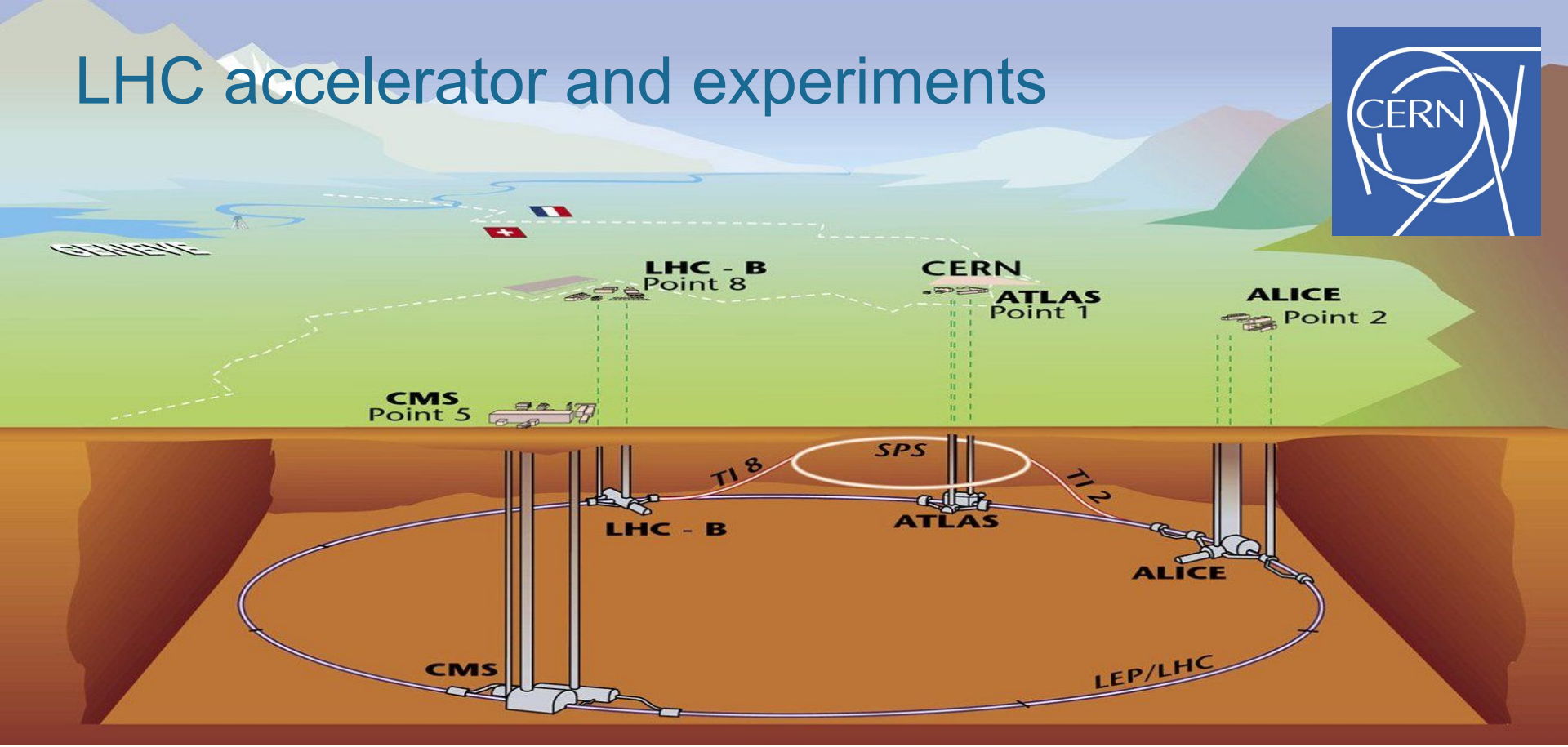
$P4_{flow}$

MultiONE

NOTED

LHC, WLCG, networking

LHC accelerator and experiments



LHC data flow

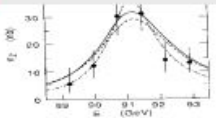
Filter and first selection

CERN Data Centre - Tier0

Create sub-samples

World-Wide Analysis

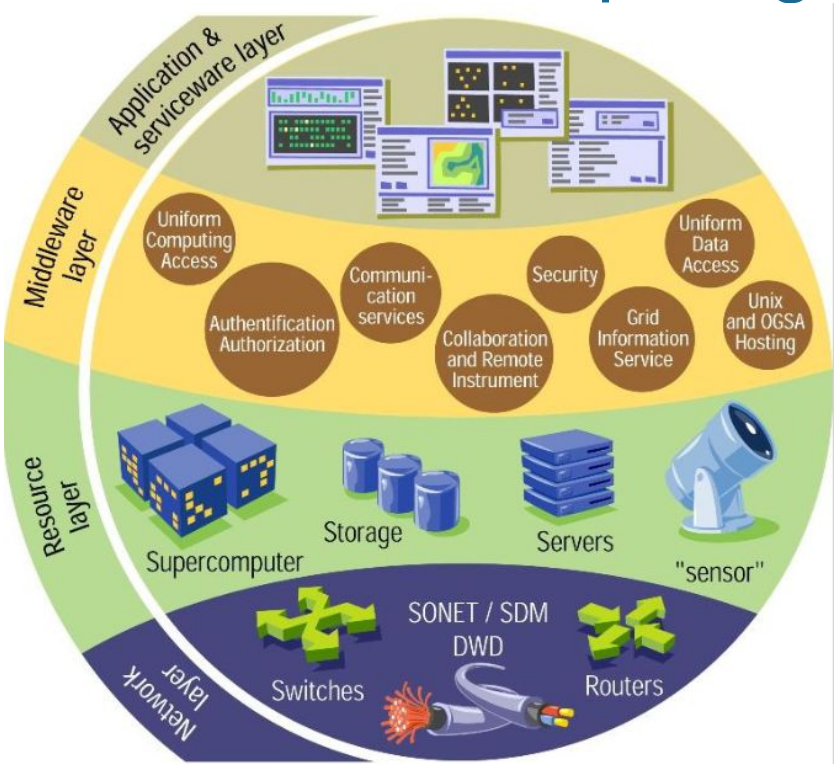
Physics
Explanation of nature



Store on disk and tape

Export copies

WLCG: Worldwide LHC Computing Grid

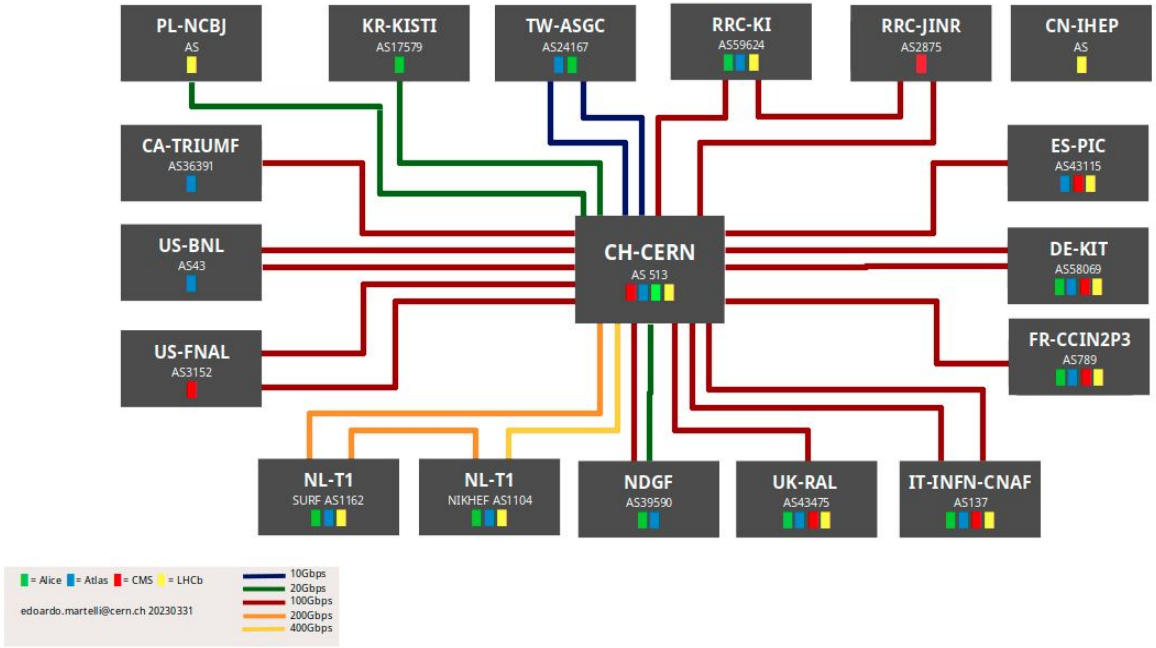


LHCOPN (LHC Optical Private Network)

LHCOPN

Tier0 to Tier1s network

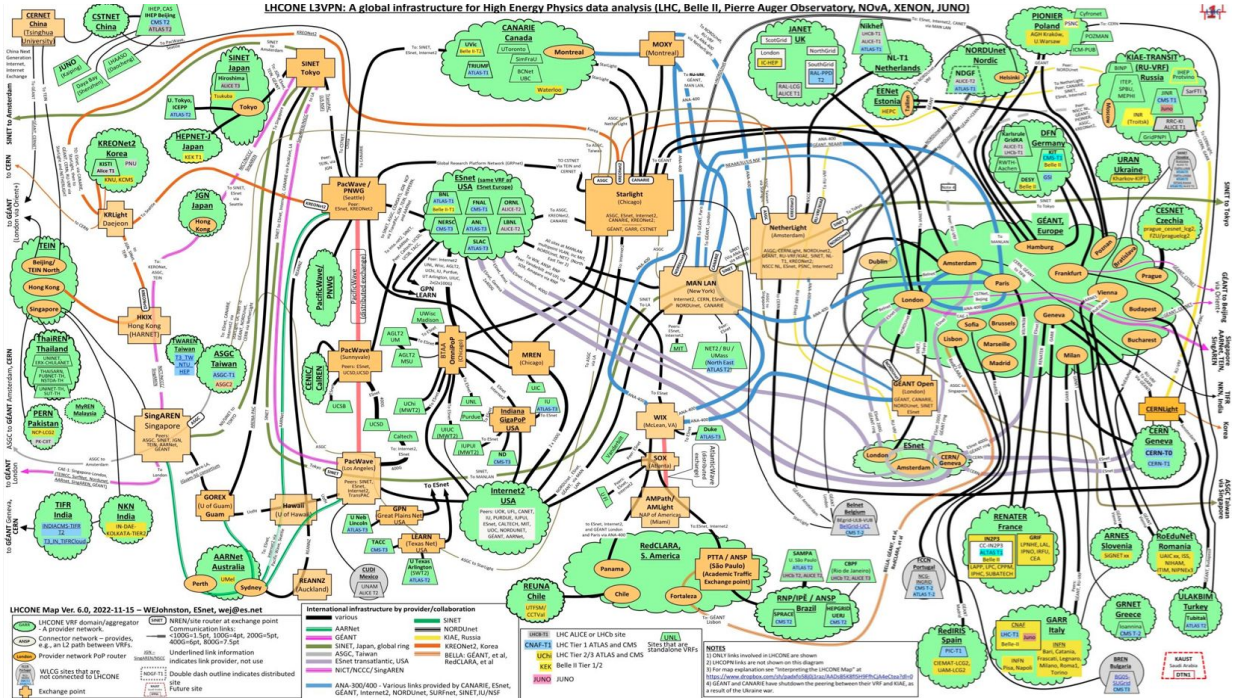
- Dedicated 10/100/400Gbps links
- 16 Tier1s + 1 Tier0
- 12 countries in 3 continents
- 2.1 Tbps to the Tier0



LHCONE (LHC Open Network Environment)

Tier1s and Tier2s network

- Overlay VPN on RENs networks
- 31 REN Network Providers
- 117 Connected Sites
- 5 Continents



LHCOPN, LHCONE and Science DMZs

Science DMZ is a local network at a science laboratory that connects high speed data transfer servers

LHCOPN and LHCONE are Virtual Private Networks (VPN) that interconnect Science DMZs

LHCOPN and LHCONE are trusted network that can bypass expensive perimeters firewalls

WLCG Network R&D

Challenges

- Expensive links may run idle for long time
- Some links get congested, while parallel ones run low
- Increasing number of LHCONE connected sites exposes internal resources to more risks
- Scarce visibility and understanding of the L7 origin of the network traffic

Motivation

- Make effective use of bandwidth, especially on the expensive transoceanic links, reduce idle periods
  → NOTED
- Increase visibility and understanding of network utilization
  → Scitags and $P4_{flow}$
- Reduce exposure of internal resources
  → MultiONE and $P4_{flow}$

R&D projects

- NOTED
  - Can improve data transfers when the network is the bottleneck
- Scitags
  - Enables tracking and correlation of WLCG transfers within Research and Education Networks
  - Makes it possible to monitor network flows at socket level
- $P4_{flow}$
  - Easy prototyping of advanced data plane services
  - Scitags accounting
- MultiONE
  - Aiming to increase security and control
  - Could take advantage of the packet marking initiative

P4flow

Programmable switches for accounting scitags-based IPv6 packets

Programming Protocol-independent Packet Processors: P4 language

Language for programming the data plane of network devices.

- ❏ Define how packets are processed.
- ❏ P4 program structure: header types, parser/deparser, match-action tables, user-defined metadata and intrinsic metadata.

Domain-specific language designed to be implementable on a large variety of targets

- ❏ Programmable network interface cards, FPGAs, software switches and hardware ASICs.

EdgeCore Wedge100BF-32QS

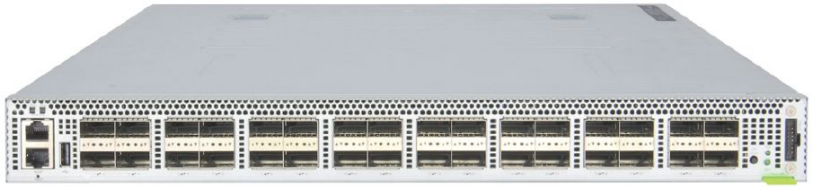
100GbE Data Center Switch
- ❏ Bare-Metal Hardware
- ❏ L2/L3 Switching
- ❏ 32xQSFP28 Ports

Data-Plane Programmability
- ❏ Intel Tofino Switch Silicon
- ❏ Barefoot Networks

Quad-Pipe Programmable Packet Processing Pipeline
- ❏ 6.4 Tbps Total Bandwidth

CPU: Intelx86 Xeon 2.0GHz
- ❏ 8-core/48GB/2TB SSD

Intel Tofino P4-programmable Ethernet Switch ASIC


EdgeCore Wedge100BF-32QS

Network Operating System

RARE/FreeRtr

- Controls the data plane by managing entries in routing tables
- Free and open source router operating system
- Export forwarding tables to DPDK or hardware switches
  - via OpenFlow or P4lang
- No global routing table
  - Every routed interface must be in a virtual routing table

SciTags

Scientific Network Tags (scitags) is an initiative promoting identification of the science domains and their high-level activities at the network level.

- Enable tracking and correlation of our transfers with Research and Education Network Providers (R&Es) network flow monitoring
- Experiments can better understand how their network flows perform along the path
- Sites can get visibility into how different network flows perform

Packet and flow marking specification

Flow label field of IPv6 header: 20 bits

- 5 entropy bits to match RFC 6436
- 9 bits to define the science domain
- 6 bits to define the application/type of traffic

Scitags → Scientific network tags initiative [1]

Bits 12 - 13 Entropy	Bits 14 - 22 Science Domain	Bit 23 Entropy	Bits 24 - 29 Application	Bits 30 - 31 Entropy

Astro/HEP Science Domains:

Reserved	0
Default	65536
ATLAS	32768
CMS	98304
LHCb	16384
ALICE	81920
Bellell	49152
SKA	114688
LSST	73728
DUNE	8192

Application:

Reserved	0
Default	1
perfSONAR	2
Cache	3
DataChallege	4
AnalysisDownload	9
DataAccess	10
CLIDownload	13
ProductionDownload	19

First approach at IP level (layer 3)

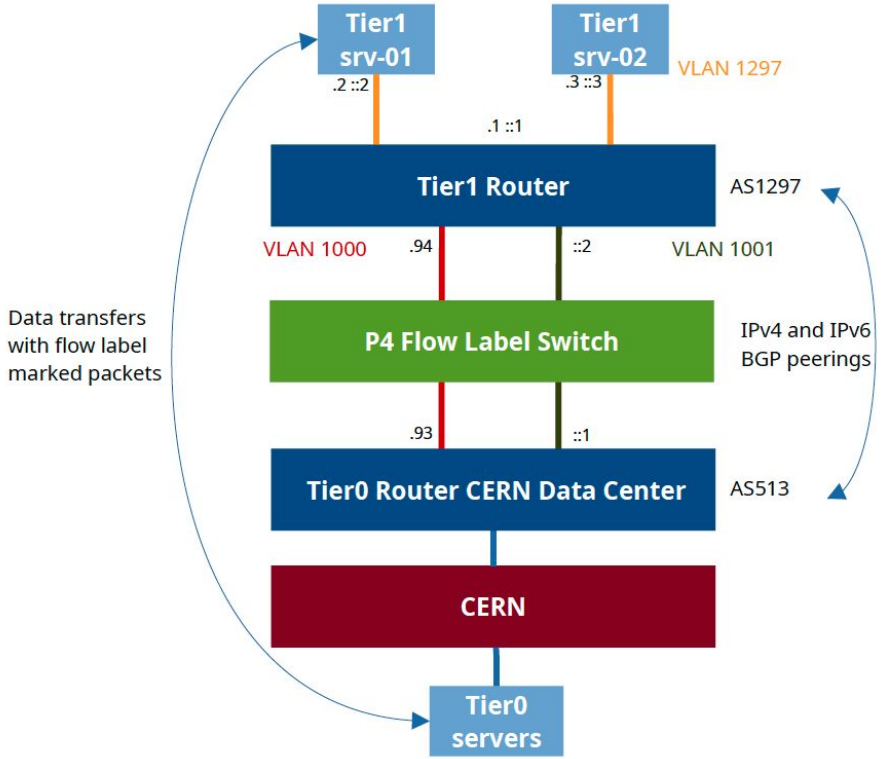
Network configuration:

- Virtual Routing Forwarding
- Policy-based routing based on flow label field value
  - Flow label 10 → VLAN 40
  - Flow label 20 → VLAN 41

Second approach at Ethernet level (layer 2)

Network configuration:

- Emulates a Tier 1/0 link
- Tier1/0 routers
  - IPv4/IPv6 BGP peerings
- Tier0 router
  - LHCOPN production border router
- Pure layer 2 bridges
  - VLAN 1000: IPv4 traffic
  - VLAN 1001: IPv6 traffic
- Tier0 servers
  - OpenStack product servers



P4 switch network configuration

Access-list:

```
access-list acl_all_ipv6_flowlabels
     # Match <Experiment> and <ANY Application>
     sequence 10 permit all any all any all flow 131076 & 261884
     sequence 11 permit all any all any all flow 65540  & 261884
     sequence 12 permit all any all any all flow 196612 & 261884
     sequence 13 permit all any all any all flow 32772  & 261884
     # Match <Experiment> and <perfSONAR Application>
     sequence 20 permit all any all any all flow 131072 & 261632
     sequence 21 permit all any all any all flow 65536  & 261632
     sequence 22 permit all any all any all flow 196608 & 261632
     sequence 23 permit all any all any all flow 32768  & 261632
     # Permit the rest of the traffic
     sequence 30 permit all any all any all
     exit
```

ATLAS <any>
CMS <any>
LHCb <any>
ALICE <any>

ATLAS <perfSONAR>
CMS <perfSONAR>
LHCb <perfSONAR>
ALICE <perfSONAR>

Pure layer 2 bridges:

```
interface sdn1.1000
     description [VLAN ID=1000]
     bridge-group 1
     no shutdown
     no log-link-change
     exit

interface sdn1.1001
     description [VLAN ID=1001]
     bridge-group 2
     bridge-filter ipv6in acl_all_ipv6_flowlabels
     no shutdown
     no log-link-change
     exit
```

VLAN 1000 belongs to bridge 1

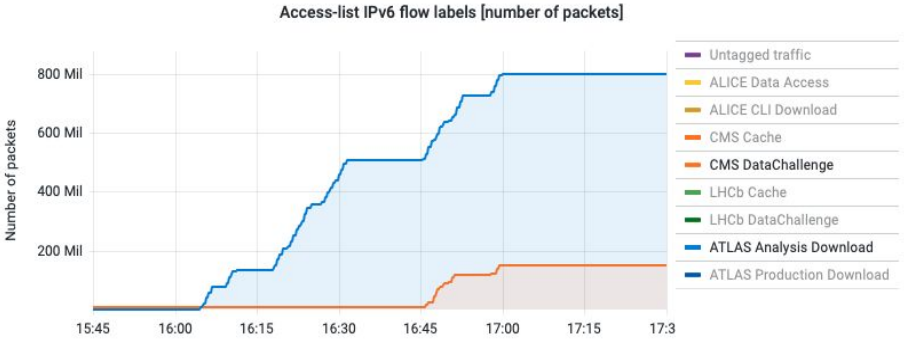
VLAN 1001 belongs to bridge 2
Filter IPv6 traffic at the input based on the access-list sentences

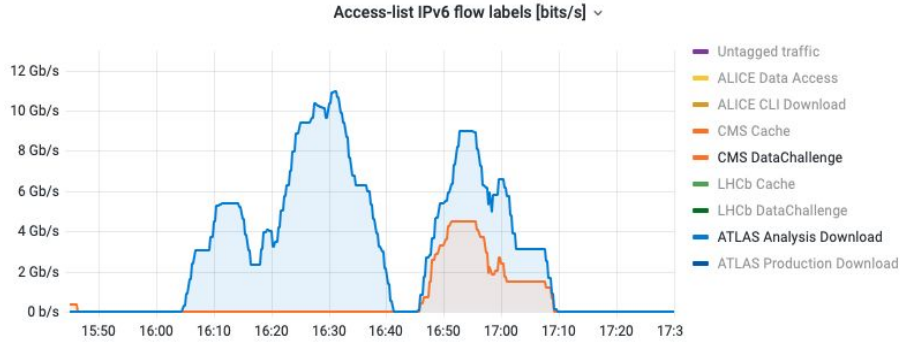
Statistics

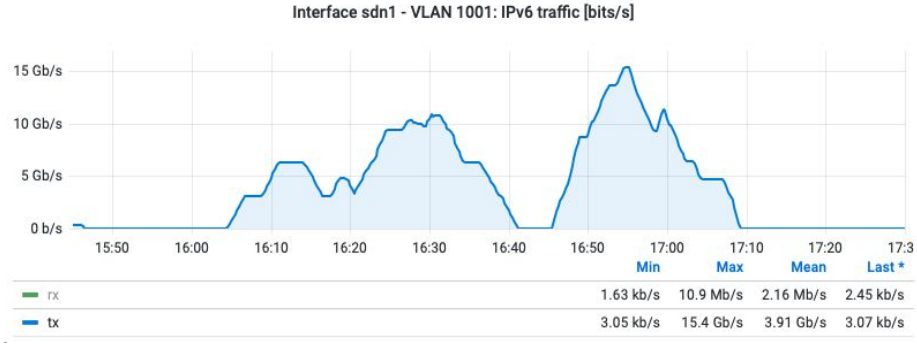
The counters of the access-list are exported to Prometheus DB and Grafana:

```
E513-E-YECWH-1#show access-list acl_all_ipv6_flowlabels
seq  txb  txp  rxb              rxp          last      timout    cfg
10   0+0  0+0  0+12374638771    0+8743031    00:03:02  00:00:00  permit all any all any all flow 131076&261884   ATLAS <any>
11   0+0  0+0  0+37019728635    0+24984028   00:02:30  00:00:00  permit all any all any all flow 65540&261884    CMS   <any>
12   0+0  0+0  0+23940164205    0+15797973   00:02:00  00:00:00  permit all any all any all flow 196612&261884   LHCb  <any>
13   0+0  0+0  0+18150017192    0+12017039   00:02:00  00:00:00  permit all any all any all flow 32772&261884    ALICE <any>

20   0+0  0+0  0+30346726207    0+20005622   00:01:29  00:00:00  permit all any all any all flow 131072&261632   ATLAS <perfSONAR>
21   0+0  0+0  0+25281078379    0+16663278   00:01:29  00:00:00  permit all any all any all flow 65536&261632    CMS   <perfSONAR>
22   0+0  0+0  0+28556351375    0+19008806   00:00:58  00:00:00  permit all any all any all flow 196608&261632   LHCb  <perfSONAR>
23   0+0  0+0  0+37078713993    0+25770785   00:00:26  00:00:00  permit all any all any all flow 32768&261632    ALICE <perfSONAR>
30   0+0  0+0  0+2715536713     0+1802921    00:00:26  00:00:00  permit all any all any all
```

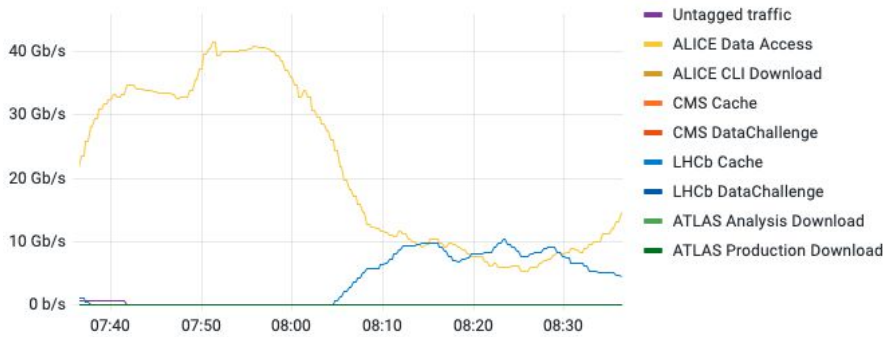


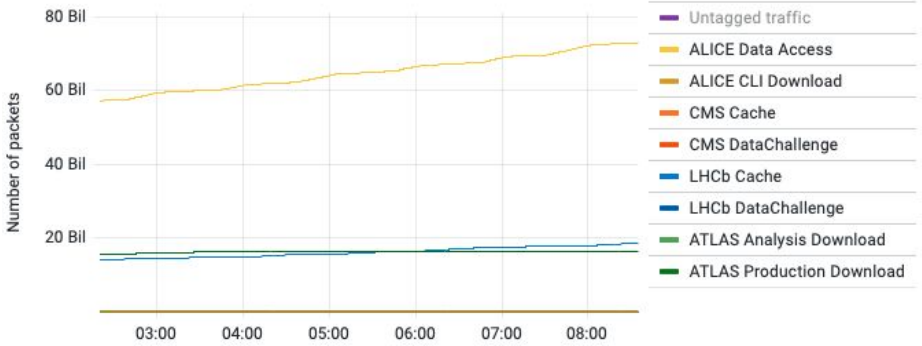


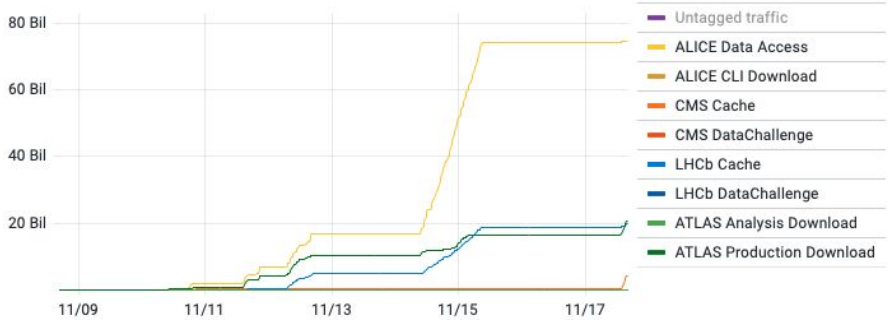


P4$_{flow}$ demo for SC22

- We demonstrated the accounting of tagged packets is feasible.







MultiONE

Programmable switches for routing scitags-based IPv6 packets

LHCONE success

- LHCONE has proved to be very useful to interconnect transfer servers to their clients at high speeds
- Over the years, other High Energy Physics collaborations joined it (BelleII, NOVAE, Pierre Auger observatory, XENON, JUNO, DUNE)

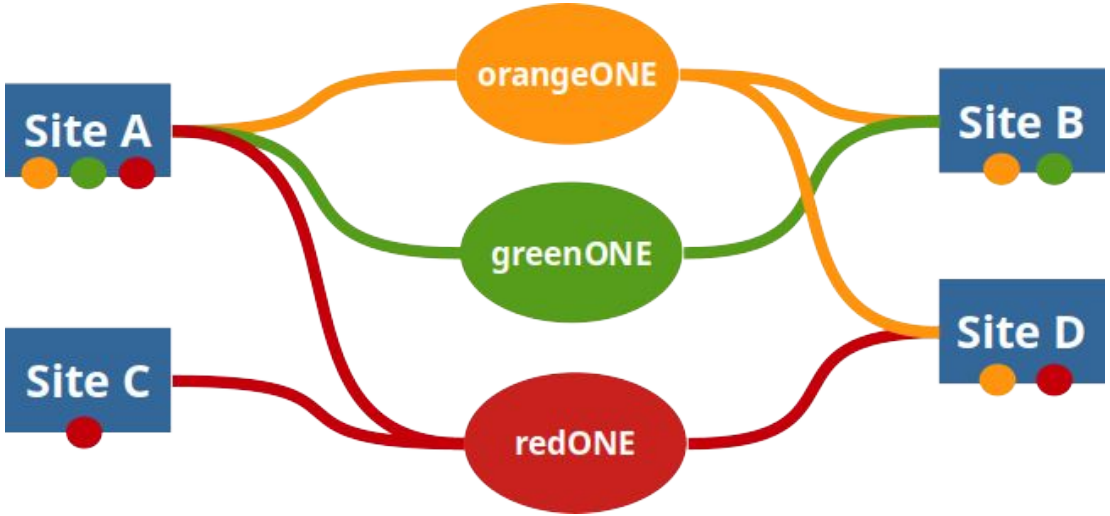
Problems adding Collaborations

- The more sites join LHCONE, the less trustable it becomes
- The more the traffic volume grows in a single domain, the more difficult for RENs is to shape the load in their networks
- Funding agencies prefers to have a clear distinction of who is using the resources they fund

From LHCONE to MultiONE

A solution would be to implement a VPN for each Collaboration:

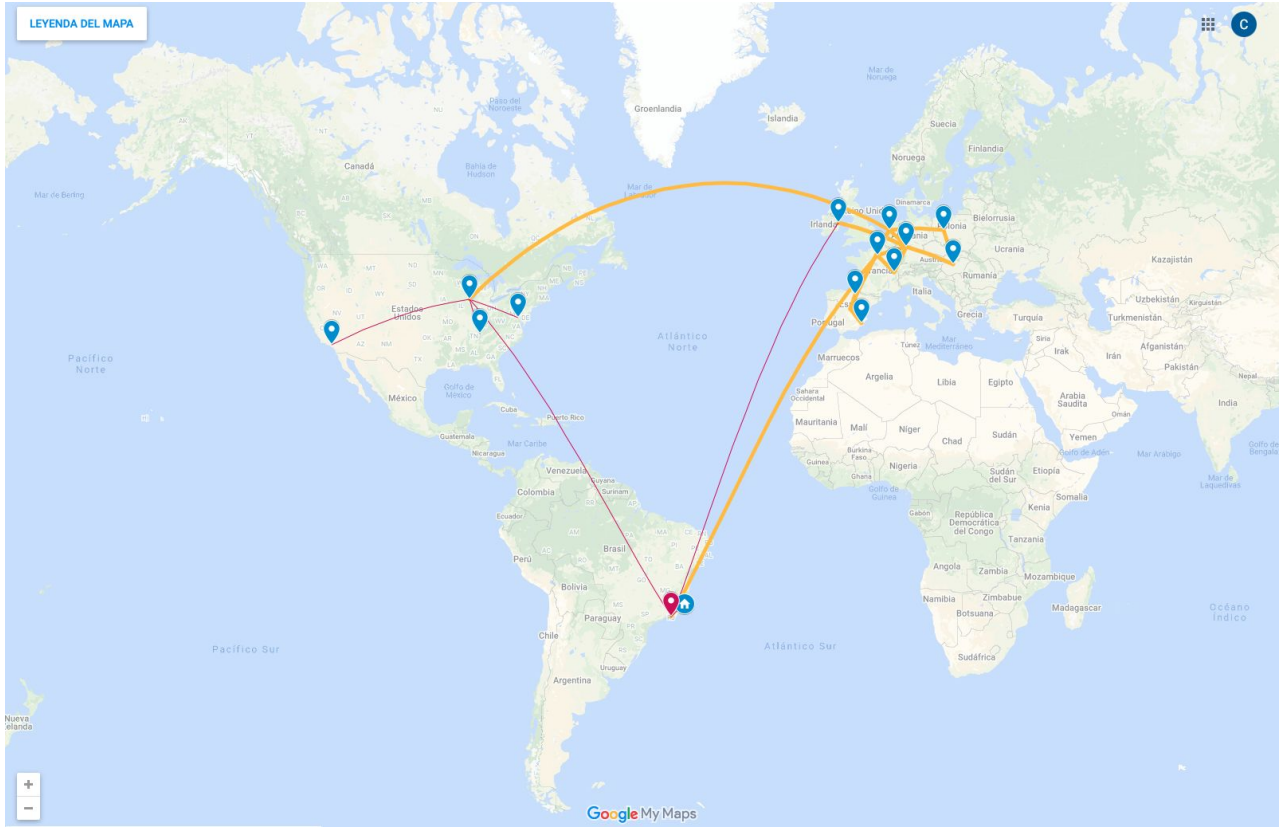
- Each site joins only the VPNs it is collaborating with, to reduce the exposure of their data-centre
- Flow label tagging could be used to policy route the traffic into the correct VPN



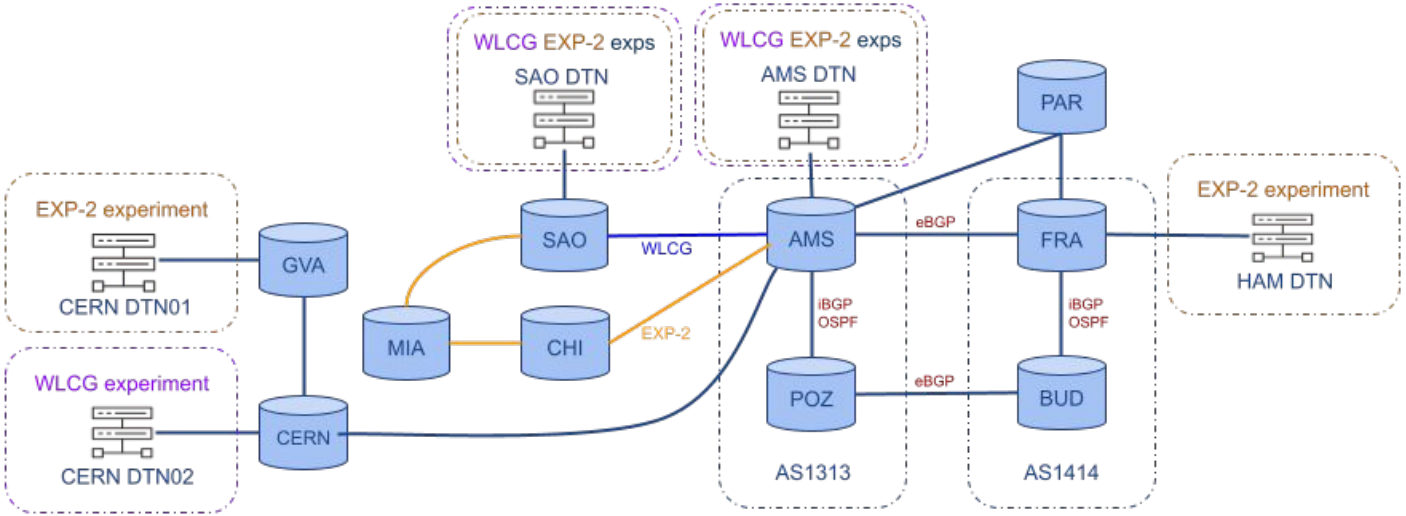
MultiONE prototype with Tofino switches

- Tofino switches can match the flow label field the IPv6 header
- A MultiONE Proof of Concept using flow label policy-based routing has been implemented in the GEANT P4lab

GÉANT P4Lab



MultiONE testbed in GP4Lab



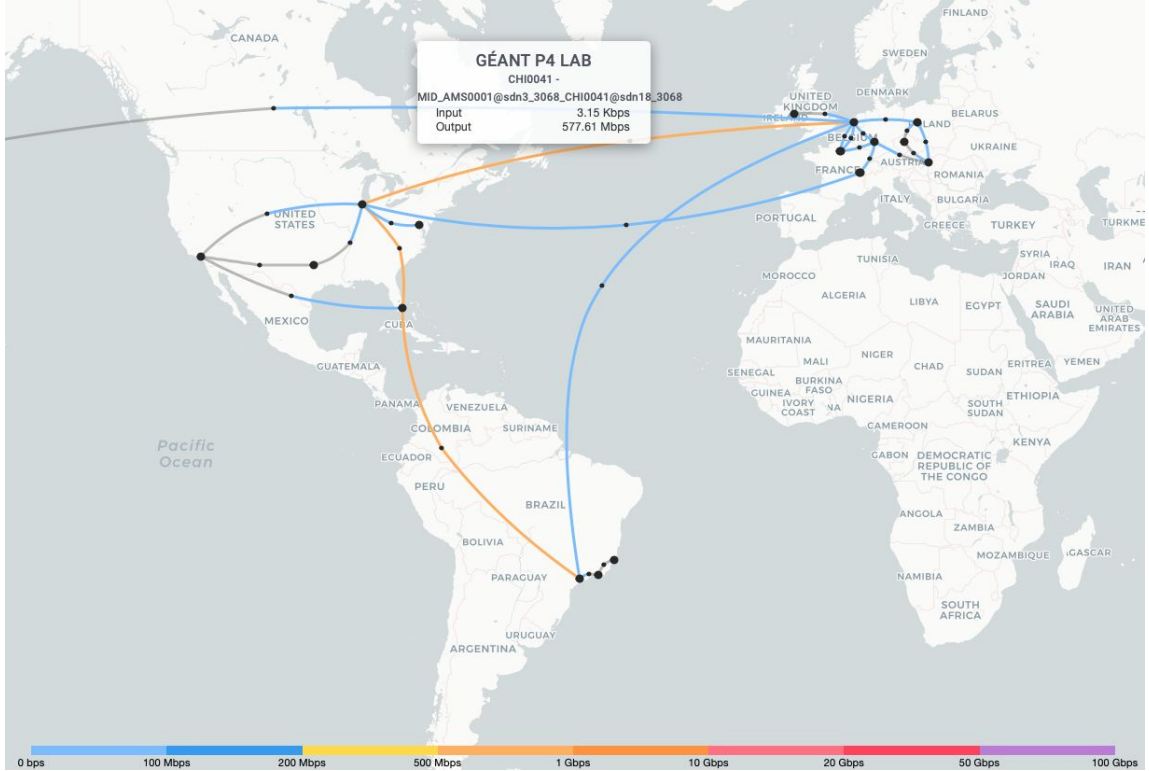
SAO P4 switch routes the traffic with PBR rules based on an access-list.

- WLCG traffic routing: SAO DTN → SAO → AMS → AMS DTN
- EXP-2 traffic routing: SAO DTN → SAO → MIA → CHI → AMS → FRA → FRA DTN

CERN DTNs generates tagged traffic to AMS DTN and HAM DTN.

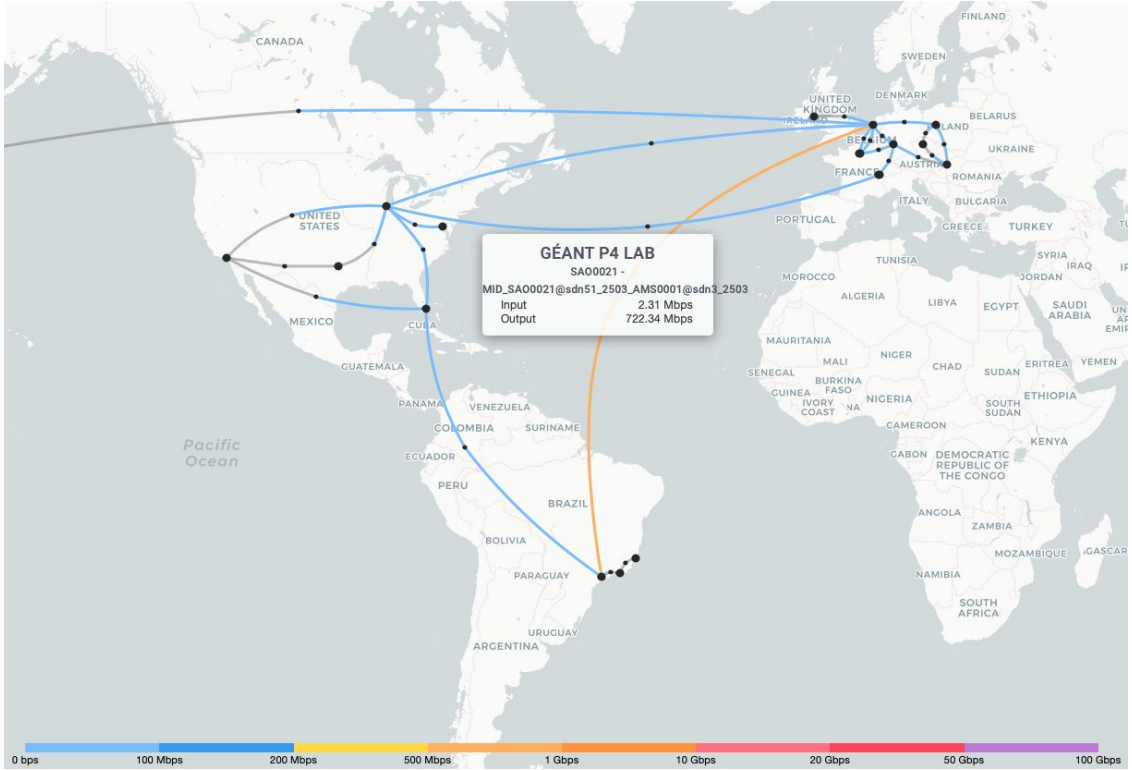
- The traffic is routed in the squared topology to WLCG or EXP-2 VPN so that LHCONE sites can only access other sites belonging to the same experiment and organization.

MultiONE testbed in GP4Lab



EXP-2 traffic routing from São Paulo to Amsterdam via Chicago and Miami.

MultiONE testbed in GP4Lab

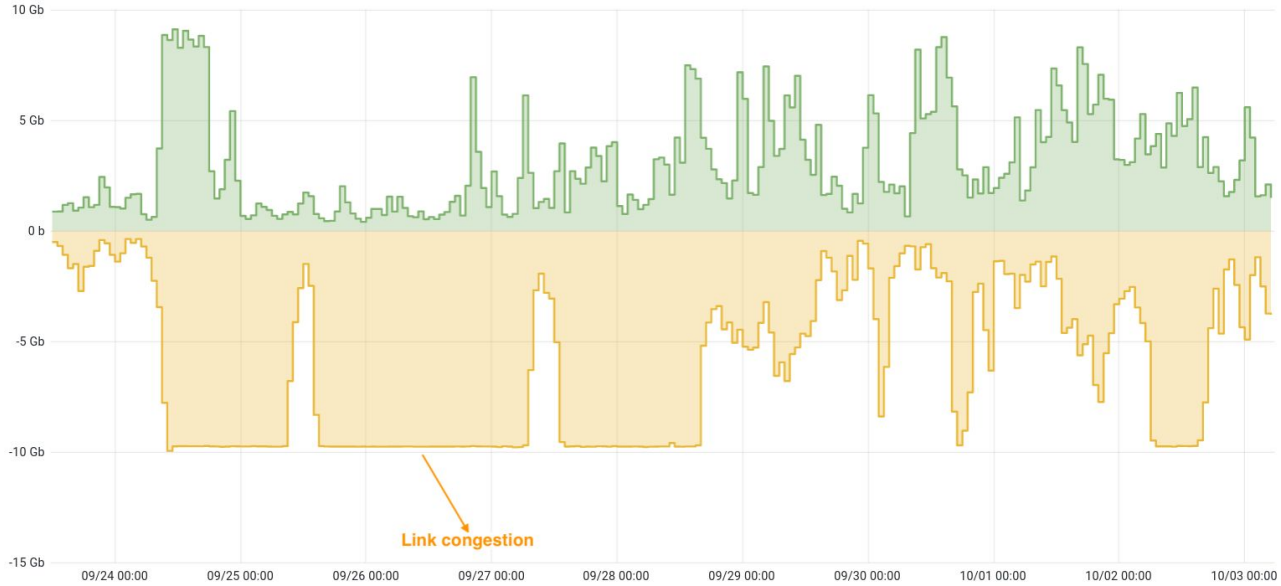


WLCG traffic routing from São Paulo directly to Amsterdam.

NOTED

Network Optimized Transfer of Experimental Data

Problematic

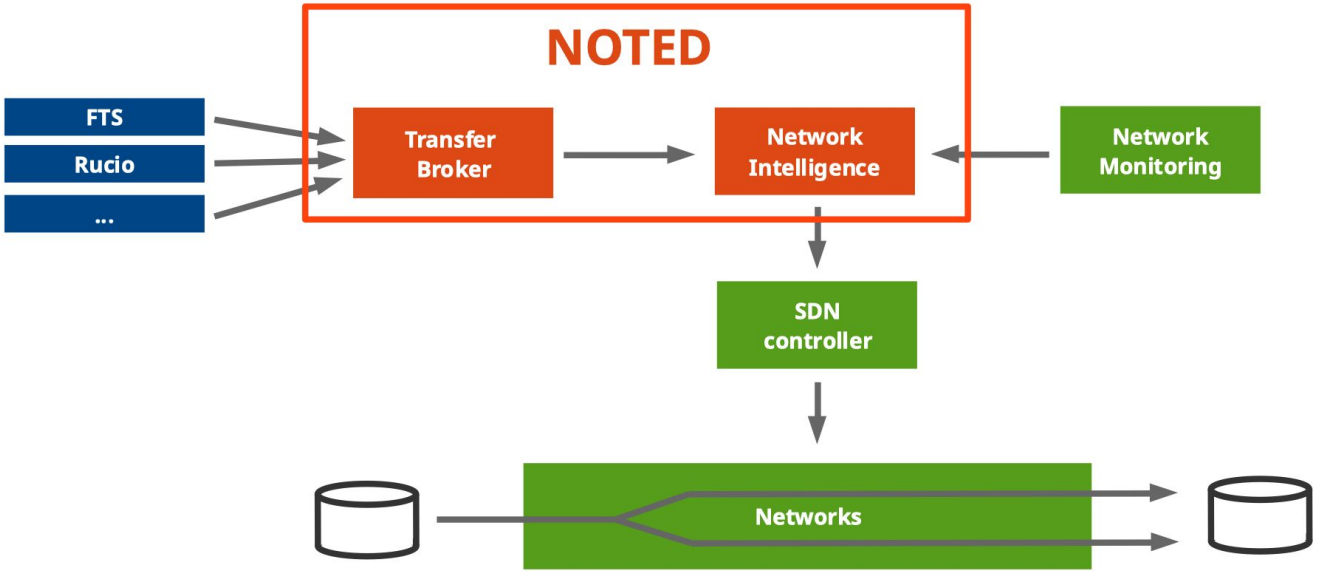


- Large data transfers can saturate network links while alternative paths may be left idle

Architecture

NOTED (Network Optimized Transfer of Experimental Data)

- An intelligent network controller to improve the throughput of large data transfers in FTS (File Transfer Services) by handling dynamic circuits.



Elements

FTS (File Transfer Service):

- Analyse data transfers to estimate if any action can be applied to optimise the network utilization → get on-going and queued transfers.

CRIC (Computing Resource Information Catalog):

- Use the CRIC database to get an overview and knowledge of the network topology → get IPv4/IPv6 addresses, endpoints, rcsite and federation.

Interaction with FTS

query monit_prod_fts_raw_queue* → ~ 50 lines per job

- {source se, dest se}: source and destination endpoints involved in the transfer.
- {throughput, filesize avg}: throughput [bytes/s] and filesize [bytes] of the transfer.
- {active count, success rate}: number of TCP parallel windows and successful rate of the transfer.
- {submitted count, connections}: number of transfers in the queue and maximum number of transfers that can be held.


```
"_source": {
    "data": {
      "source_se": "davs://grid-se.physik.uni-wuppertal.de",
      "dest_se": "davs://webdav.mwt2.org",
      "timestamp": 1662470909066,
      "throughput": 180269,
      "throughput_ema": 51234.889998671875,
      "duration_avg": 1,
      "filesize_avg": 581514.1612903225,
      "filesize_stddev": 581514.1612903225,
      "success_rate": 100,
      "retry_count": 0,
      "active_count": 0,
      "submitted_count": 25229,
      "connections": 200,
      "rationale": "Good link efficiency",
      "endpnt": "bnl"
    },
    "metadata": {
      "hostname": "monit-amqsource-ee2e71080d.cern.ch",
      "partition": "10",
      "type_prefix": "raw",
      "kafka_timestamp": 1662470912200,
      "topic": "fts_raw_queue_state",
      "producer": "fts",
      "_id": "d00e3711-9ba0-60e9-b4c9-36ac801d6ef2",
      "type": "queue_state",
      "timestamp": 1662470910441
    }
}
```

Dataset structure and workflow

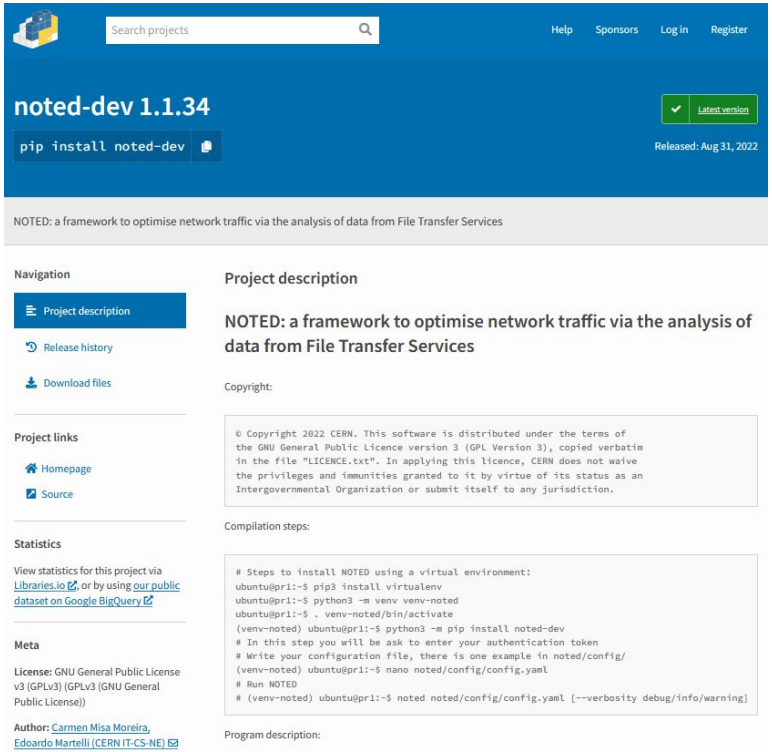
Configuration given by the user → a list of {src rcsite, dst rcsite} pairs.

1. Enrich NOTED with the topology of the network:
   - ❏ Query CRIC database → get the endpoints (α_i, β_i) that could be involved in the transfers for the given {src rcsite, dst rcsite} pairs.
2. Analyse on-going and upcoming data transfers:
   - ❏ Query FTS recursively → get the on-going transfers for each set of endpoints (α_i, β_i).
     Network utilization = $\sum \varphi_{\text{on-going transfers}}(\alpha_i, \beta_i)_{\text{involved}}$
3. Network decision:
   - ❏ When NOTED detects that the link is going to be congested → provides a dynamic circuit via Sense/AutoGOLE.

Source endpoint	Destination endpoint	Data [GB]	Throughput [Gb/s]	Parallel transfers	Queued transfers
davs://ccdavatlas.in2p3.fr	davs://webdav.echo.stfc.ac.uk	139.3726	54.0827	453	28557
srm://dcsrm.usatlas.bnl.gov	davs://dcgftp.usatlas.bnl.gov	121.9655	53.6442	422	28538
davs://dav.ndgf.org	davs://dcgftp.usatlas.bnl.gov	202.7864	82.0855	862	57880
davs://atlaswebdav-kit.gridka.de	davs://eosatlas.cern.ch	205.3606	82.0725	888	57790
srm://dcsrm.usatlas.bnl.gov	davs://dcgftp.usatlas.bnl.gov	193.5176	58.8136	530	26294
davs://f-dpm000.grid.sinica.edu.tw	davs://webdav.lcg.triumf.ca	210.2710	51.0323	567	26314
davs://ccdavatlas.in2p3.fr	davs://webdav.echo.stfc.ac.uk	332.0009	81.7908	905	50152
srm://dcsrm.usatlas.bnl.gov	davs://dcgftp.usatlas.bnl.gov	326.5855	80.1554	903	50028

Status of software

Available in https://pypi.org/project/noted-dev/



Common steps:

```
# Create a virtual environment:
$ pip3 install virtualenv
$ python3 -m venv venv-noted
$ . venv-noted/bin/activate
```

Ubuntu installation:

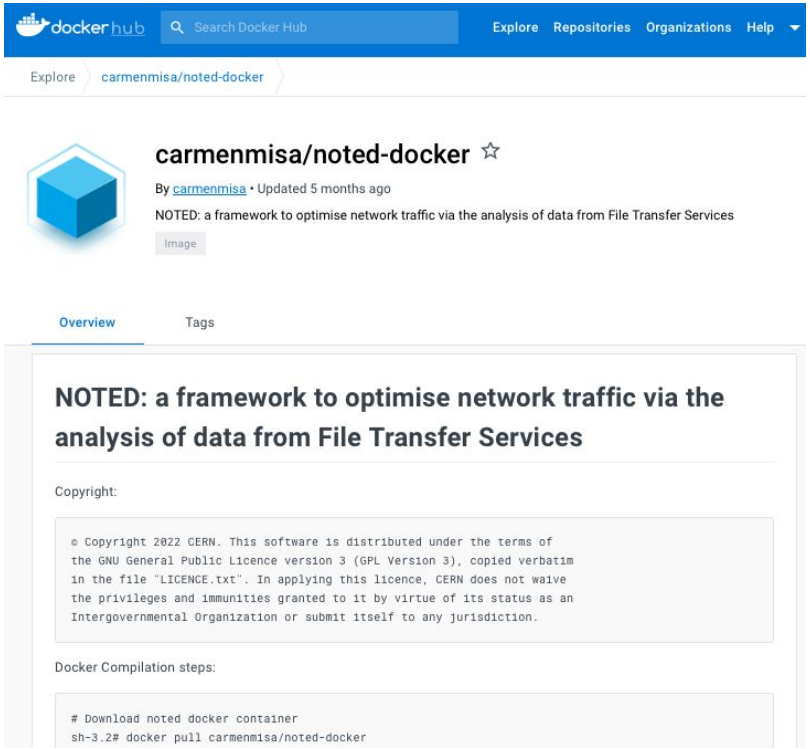
```
# Install noted-dev
(venv-noted) $ python3 -m pip install noted-dev
# Write your configuration file
(venv-noted) $ nano noted/config/config.yaml
# Run NOTED
(venv-noted) $ noted noted/config/config.yaml
```

CentOS installation:

```
# Download noted-dev.tar.gz
(venv-noted) $ wget url pypi repo tar gz
# Install noted-dev
(venv-noted) $ tar -xf noted-dev-1.1.62.tar.gz
(venv-noted) $ pip install noted-dev-1.1.62/
# Run NOTED
(venv-noted) $ noted noted/config/config.yaml
```

Status of software

Available in https://hub.docker.com/r/carmenmisa/noted-docker



Installation:

```
# Download noted docker container
$ docker pull carmenmisa/noted-docker
# Run docker container
$ docker run --detach --entrypoint /sbin/init
--network="host" --privileged --name
noted\_controller carmenmisa/noted-docker
# Copy your configuration file into the container
$ docker cp src/noted/config/config-example.yaml
noted\_controller:/app/noted/config
# Run commands in the container from outside
$ docker exec noted\_controller noted -h
$ docker exec noted\_controller
/app/src/noted/scripts/setup.sh mail
# Run NOTED
$ docker exec noted\_controller noted
config/config-example.yaml &
```

Configuration file

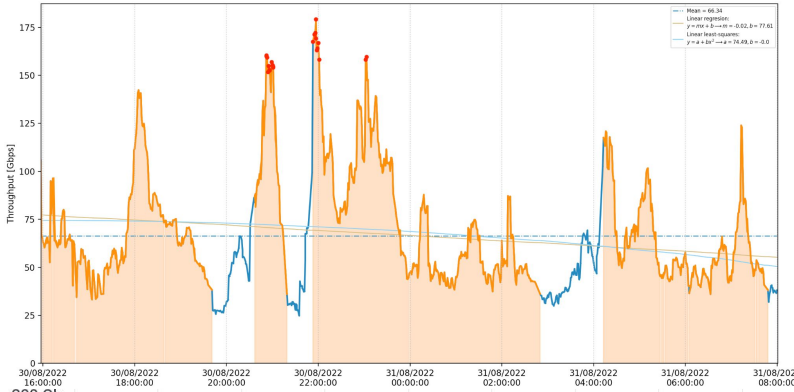
- Usage: `$ noted [-h] [-v VERBOSITY] config file`

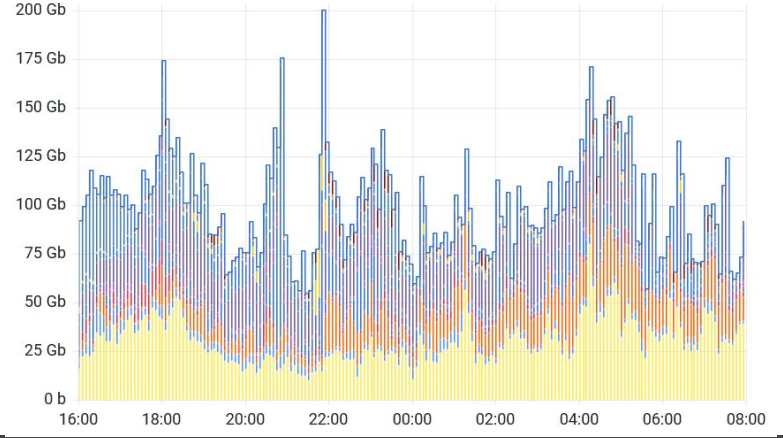
```
noted -h
positional arguments:
      config file the name of the configuration file [config-example.yaml]
optional arguments:
      -h, --help show this help message and exit
      -v VERBOSITY, --verbosity VERBOSITY defines logging level [debug, info, warning]
```

- Example of config.yaml:

```
src rcsite: ['rc site 1', 'rc site 2', 'rc site 3', 'rc site 4']# Source RC Sites
dst rcsite: ['rc site 1', 'rc site 2', 'rc site 3', 'rc site 4']# Destination RC Sites
events to wait until notification: 5 # Events to wait until email notification
max throughput threshold link: 80# If throughput > max throughput -> START
min throughput threshold link: 20# If throughput < min throughput -> STOP
unidirectional link: False # If False both TX and RX paths will be monitoring
number of dynamic circuits: 2 # Number of dynamic circuits
sense uuid: 'sense uuid 1'# Sense-o UUID dynamic circuit
sense vlan: 'vlan description 1'# VLAN description
sense uuid 2: 'sense uuid 2' # Sense-o UUID dynamic circuit
sense vlan 2: 'vlan description 2' # VLAN description
from email address: 'email 1' # From email address
to email address: 'email 1, email 2' # To email address
subject email: 'subject' # Subject of the email
message email: "message" # Custom message
auth token: auth token # Authenticantion token
```

Transfers of WLCG sites in LHCONE (31st of August 2022)

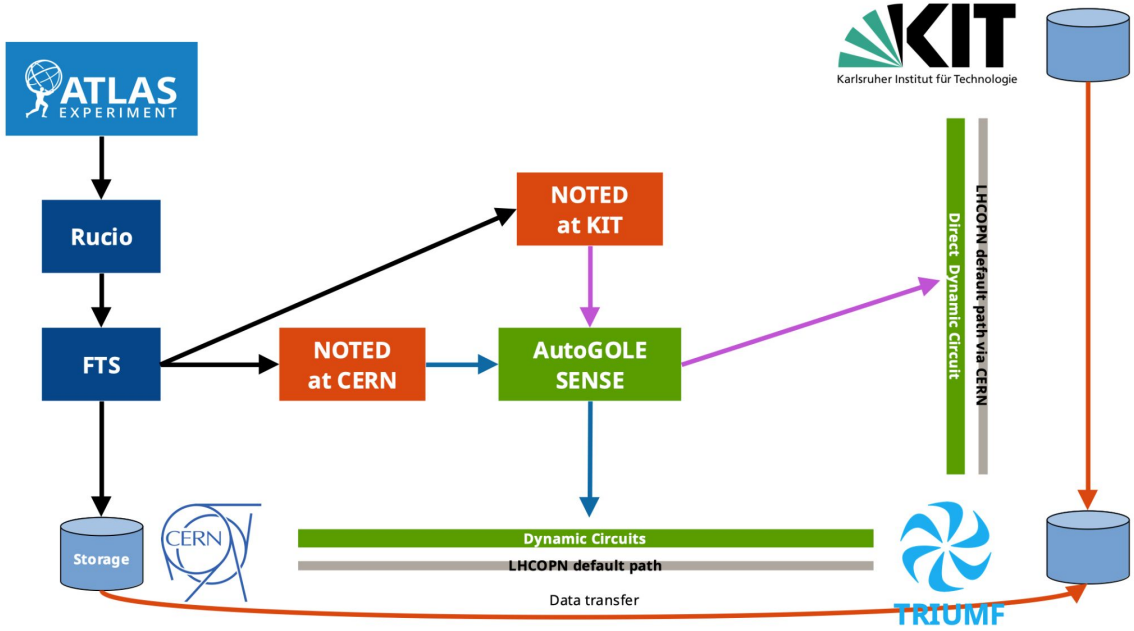




- If throughput > 80 GB/s → NOTED provides a dynamic circuit. When throughput < 40 GB/s → NOTED cancels the dynamic circuit and the traffic is routed back to the default path.
- Observations of NOTED about the network utilization correspond with the reported ones in Grafana by LHCONE/LHCOPN production routers.

Therefore, by inspecting FTS data transfers it is possible to get an understanding of the network usage and improve its performance by executing an action in the topology of the network.

NOTED demo for SC22



1. NOTED looks in FTS for large data transfers.
2. When it detects a large data transfer → request a dynamic circuit by using the SENSE provisioning system.
3. LHCOPN routers at CERN will route the data transfers over the new dynamic circuit.
4. When the large data transfer is completed → release the dynamic circuit, the traffic is routed back to the LHCOPN production link.

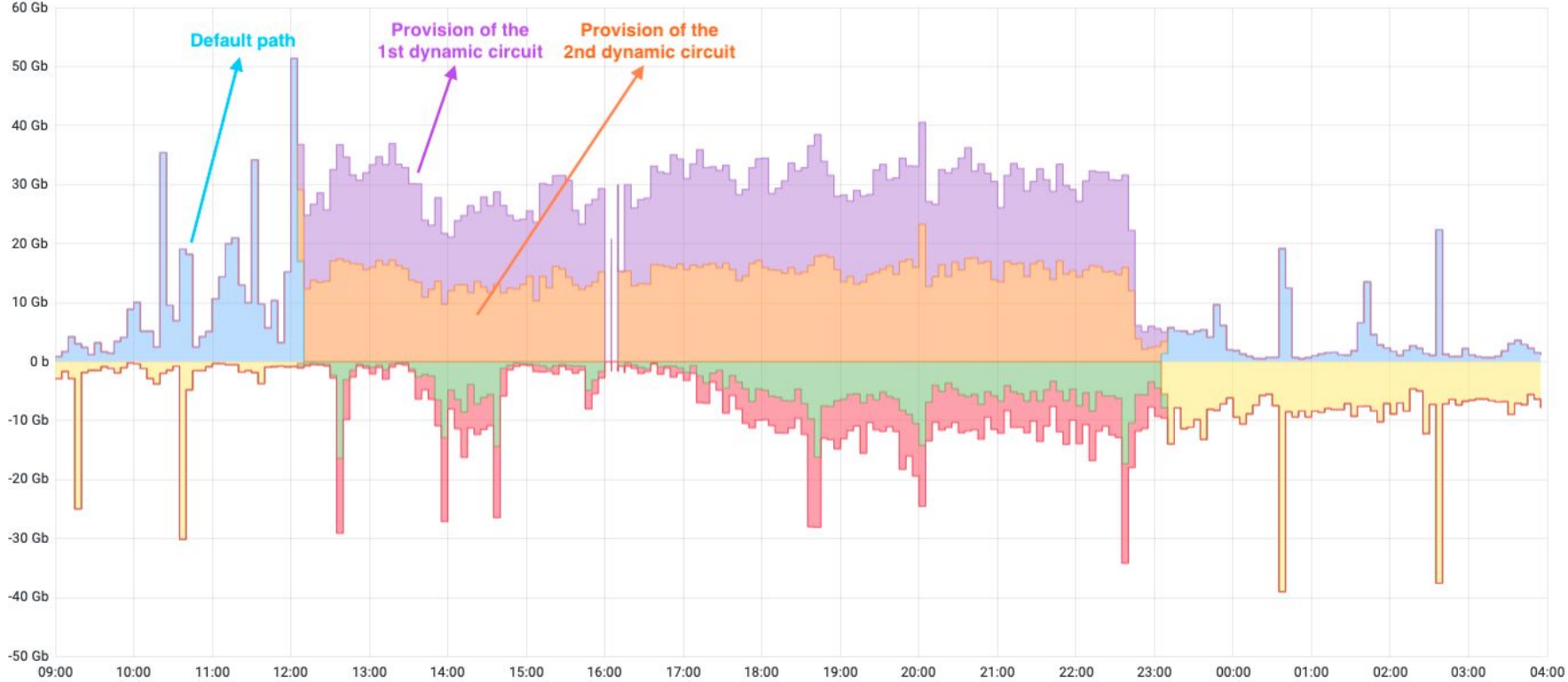
NOTED demo for SC22

Components:

- ❏ NOTED controller and FTS at CERN
- ❏ NOTED controller at KIT
- ❏ Data storage at CERN, TRIUMF, KIT
- ❏ AutoGOLE/SENSE circuits between CERN-TRIUMF and KIT-TRIUMF SENSE circuits are provided by ESnet, CANARIE, STARLIGHT, SURF

Participants:

NOTED demo for SC22



Conclusions

NOTED:

- NOTED can reduce the duration of large data transfers and improve the efficient use of network resources. It has been demonstrated with production FTS transfers at SC22.

$P4_{flow}$:

- The IPv6 flow label accounting and forwarding can be implemented both at layer 3 and layer 2. It was demonstrated at SC22.

MultiONE:

- By using the GP4Lab we demonstrated that MultiONE can be implemented by using PBR rules based on an access-list with the flow label definitions on the clients to control the access to each VPN.

tnc23
DIGITAL GENERATIONS
TIRANA, ALBANIA | 5-9 JUNE 2023

Thank you

Any questions?

Carmen MISA MOREIRA
Edoardo MARTELLI

GÉANT
Co-funded by the European Union