NOTED: a congestion driven network controller

tnc24
RENDEZVOUS À RENNES
Rennes, France | 10-14 JUNE 2024

Carmen MISA MOREIRA

CERN IT Department CS Group

12th June 2024

Co-funded by the European Union

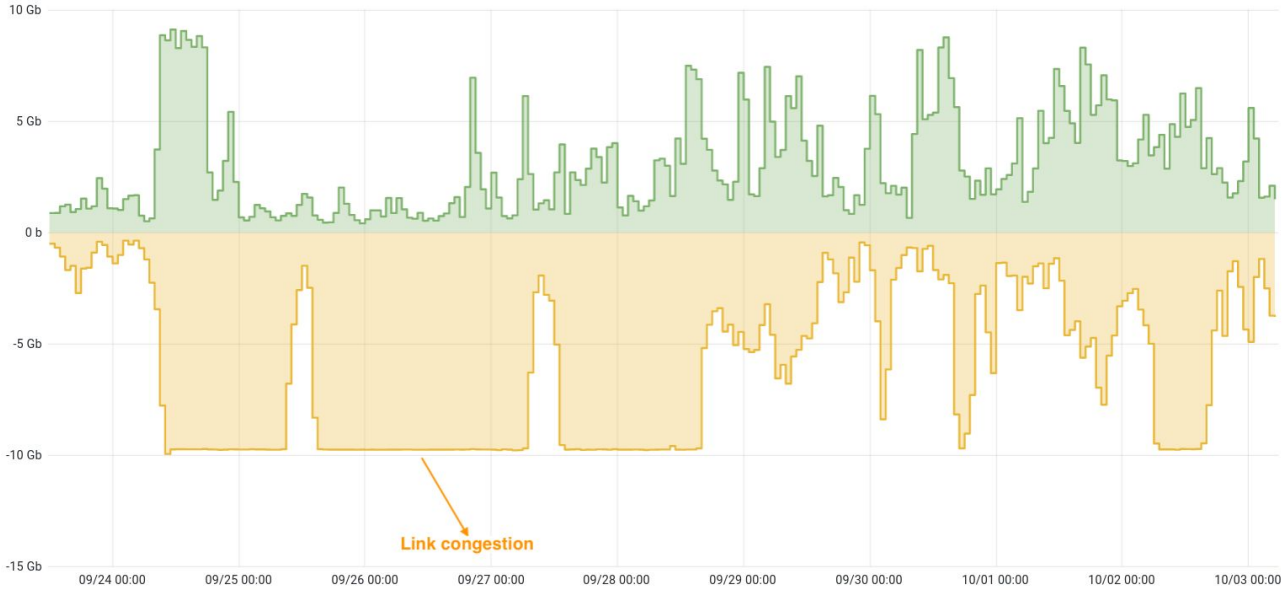
GÉANT

Outline

- Introduction
  - Motivation
  - Architecture
  - Elements
- Modes of operation
  - Configuration file
  - Flowchart
- States of execution
- NOTED in MONIT Grafana
  - Database parameters
- Package distribution
- NOTED demonstrations at SC22, SC23 and WLCG DC24
- Conclusions and future work
- Publications

Introduction

Motivation

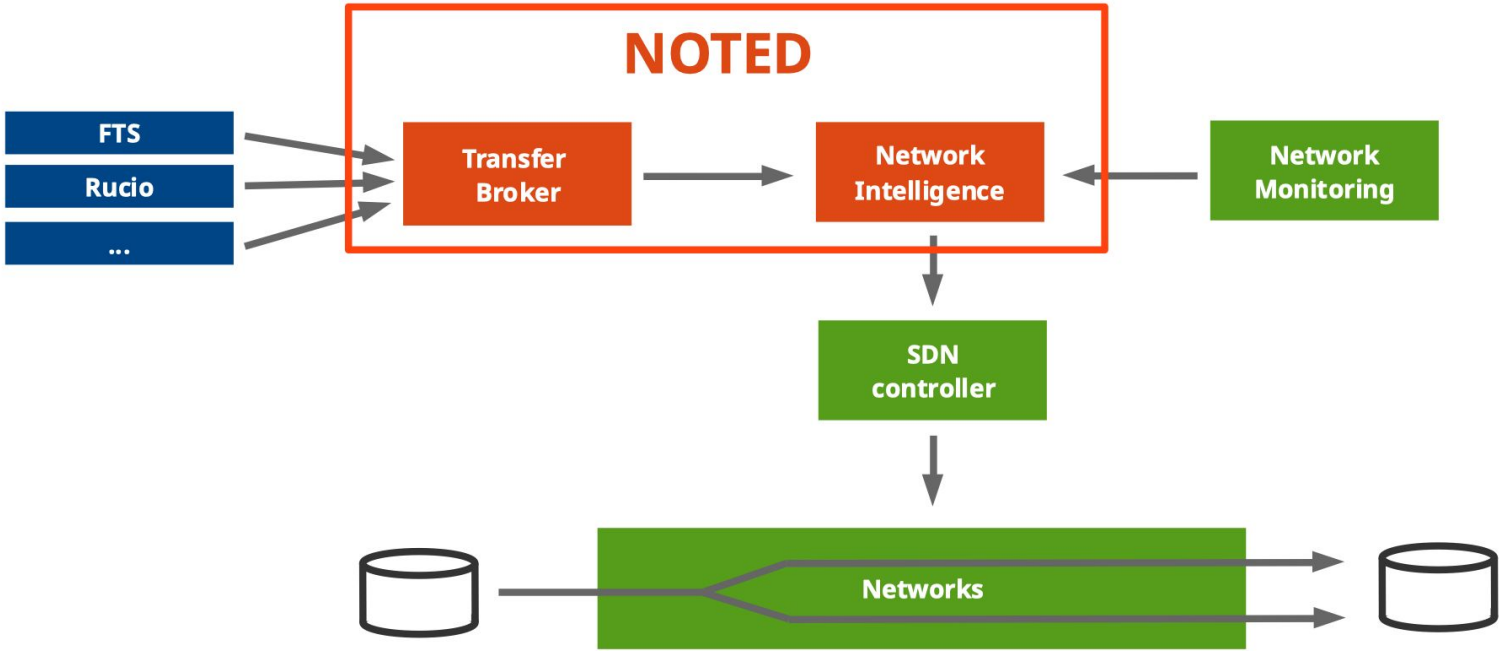


- Large data transfers can saturate network links while alternative paths may be left idle

Architecture

NOTED (Network Optimized Transfer of Experimental Data)

- An intelligent network controller to improve the throughput of large data transfers in FTS (File Transfer Service)



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 file size [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
    }
  }
```

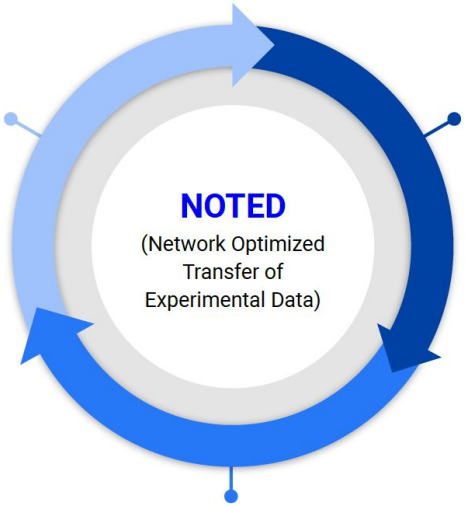
Interaction with CRIC

query rcsite*


```
"FZK-LCG2": {
    "country": "Germany",
    "description": "Tier 1",
    "federations": [ "DE-KIT" ],
    "infourl": "http://www.gridka.de",
    "latitude": 49.099049,
    "longitude": 8.432665,
    "name": "FZK-LCG2",
    "netroutes": {
      "FZK-LCG2-LHCOPNE": {
        "lhcone_bandwidth_limit": 200,
        "lhcone_collaborations": [
          "WLCG",
          "BelleII",
          "PierreAugerObservatory",
          "XENON"
        ],
        "networks": {
          "ipv4": [
            "157.180.228.0/22",
            "157.180.232.0/22",
            "192.108.45.0/24",
            "192.108.46.0/23",
            "192.108.68.0/24"
          ],
          "ipv6": [
            "2a00:139c::/45"
          ]
        }
      }
    },
    "rc_tier_level": 1,
    "services": [
      {
        "arch": "",
        "endpoint": "cloud-htcondor-ce-1-kit.gridka.de",
        "flavour": "HTCONDOR-CE",
        "state": "ACTIVE",
        "status": "production",
        "type": "CE",
      },
      {
        "arch": "",
        "endpoint": "grid-ce-1-rwth.gridka.de",
        "flavour": "HTCONDOR-CE",
        "state": "ACTIVE",
        "status": "production",
        "type": "CE",
      },
      {
        "arch": "",
        "endpoint": "perfsonar-de-kit.gridka.de",
        "flavour": "Bandwidth",
        "state": "ACTIVE",
        "status": "production",
        "type": "PerfSonar",
      },
    ],
    "sites": [
      {
        "name": "FZK",
        "tier_level": 1,
        "vo_name": "alice"
      },
      {
        "name": "FZK-LCG2",
        "tier_level": 1,
        "vo_name": "atlas"
      },
      {
        "name": "LCG.GRIDKA.de",
        "tier_level": 1,
        "vo_name": "lhcb"
      },
      {
        "name": "T1_DE_KIT",
        "tier_level": 1,
        "vo_name": "cms"
      }
    ],
    "state": "ACTIVE",
    "status": "production",
},
```

Modes of operation

Modes of operation



CUSTOM

NOTED is working based on the parameters written in a config.yaml file by the network administrator to monitor FTS data transfers

LHCOPN

When CERN NMS raises an alarm on an interface in one of the LHCOPN border routers, NOTED identifies the Tier 1 and starts to monitor FTS data transfers → automatically!

LHCONE

When CERN NMS raises an alarm on an interface in one of the LHCONE border routers, NOTED identifies the Tier 2, Tier 3 and starts to monitor FTS data transfers → automatically!

- Much more complex for LHCONE since a single path is shared by multiple sites ~ 100

Configuration file (CUSTOM version)

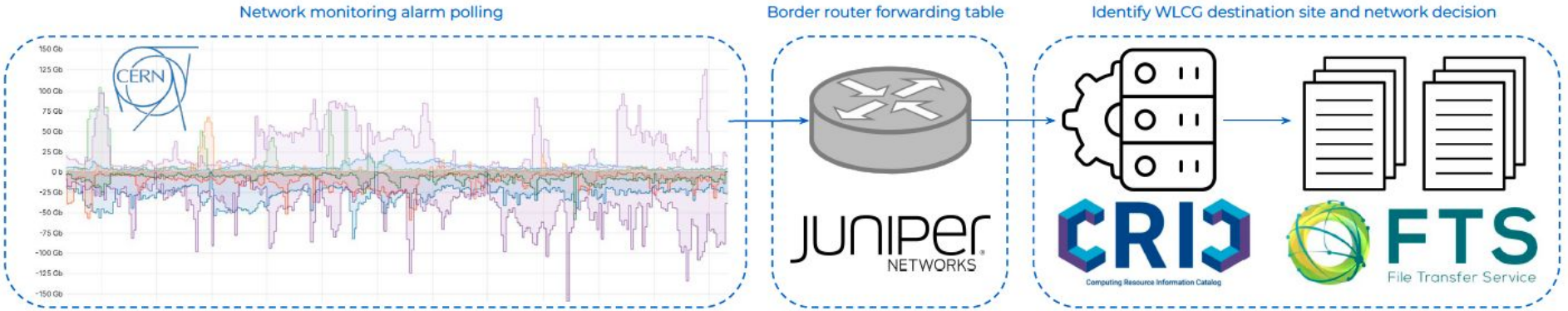
- Usage: `$ noted [-h] [-v VERBOSITY] config_file`

```
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 # Authentication token
```

Flowchart (LHCOPN, LHCONE version)



Network monitoring alarm polling (LHCOPN, LHCONE version)

- Poll the alarms IN/OUT LOAD THRESHOLD EXCEEDED generated by Spectrum, the CERN Network Monitoring System.

Alarms / Spectrum Alarms ☆ 

Instance: Production | Entity name: Enter variable value | Cause ID: Enter variable value | Exclude secstring: ---

IT/CS Alarm History

Severity	Occ	Entity name	Type	Class	Alarm name	Ack	Start at	Cleared at	Duration
MINOR	1	l513-e-rjup1-1_irb.2126	Gen_IF_Port	Port	OUT LOAD THRESH...	No	2023-10-06 08:59:23		
MINOR	1	l513-e-rjup1-1_irb.3530	Gen_IF_Port	Port	OUT LOAD THRESH...	No	2023-10-06 08:43:05		
MINOR	1	l513-v-rjuxl-12_xe-1_0_11	Gen_IF_Port	Port	OUT LOAD THRESH...	No	2023-10-06 08:21:10		
MINOR	1	l513-e-rjup1-1_irb.3530	Gen_IF_Port	Port	OUT LOAD THRESH...	No	2023-10-06 08:08:03	2023-10-06 08:22:50	00:14:47
MINOR	1	l513-e-rjup1-1_irb.3530	Gen_IF_Port	Port	IN LOAD THRESHO...	No	2023-10-06 07:28:02	2023-10-06 07:32:47	00:04:45
MINOR	1	l513-e-rjup1-1_irb.3530	Gen_IF_Port	Port	IN LOAD THRESHO...	No	2023-10-06 06:58:02	2023-10-06 07:07:47	00:09:45
MINOR	1	l513-v-rjuxl-12_xe-1_0_12	Gen_IF_Port	Port	OUT LOAD THRESH...	No	2023-10-06 06:46:00	2023-10-06 07:36:14	00:50:14
MINOR	1	l513-e-rjup1-1_irb.2126	Gen_IF_Port	Port	OUT LOAD THRESH...	No	2023-10-06 06:34:23	2023-10-06 08:44:08	02:09:45
MINOR	1	l513-v-rjuxl-12_xe-1_0_13	Gen_IF_Port	Port	OUT LOAD THRESH...	No	2023-10-06 06:15:58	2023-10-06 07:55:57	01:39:59
MINOR	1	l513-e-rjup1-1_irb.3530	Gen_IF_Port	Port	OUT LOAD THRESH...	No	2023-10-06 05:53:02	2023-10-06 06:27:47	00:34:45

Border router forwarding table (LHCOPN, LHCONE version)

Identify the prefixes routed via the alarmed interface:

- Find the IP of the next hop:

```
BORDER-ROUTER>  show interfaces irb.3530 terse

Interface               Admin Link Proto    Local                 Remote
irb.3530                up    up   inet     172.24.18.9/30
                                    inet6    2001:1458:302:38::1/64
```

- Find the routed prefixes:

```
BORDER-ROUTER>  show route next-hop 2001:1458:302:38::2

2a00:139c::/45     *[BGP/170] 2d 23:16:51, MED 10, localpref 100
                     AS path: 58069 I, validation-state: unverified
                   > to 2001:1458:302:38::2 via irb.3530
```

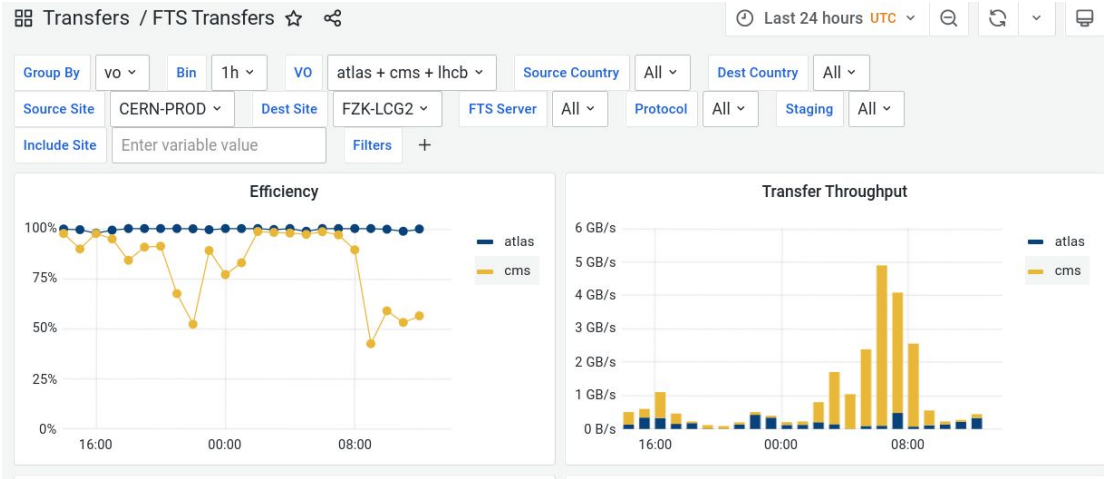
Identify WLCG destination site (LHCOPN, LHCONE version)

- Lookup routed prefixes in CRIC to identify the destination site
- Look for FTS transfers and make a network decision if it is causing congestion

NetworkRoute: FZK-LCG2-LHCOPNE

Network Site	DE-KIT
ASN	58069
More specific	False
Monitoring URL	*Not set*
Networks	

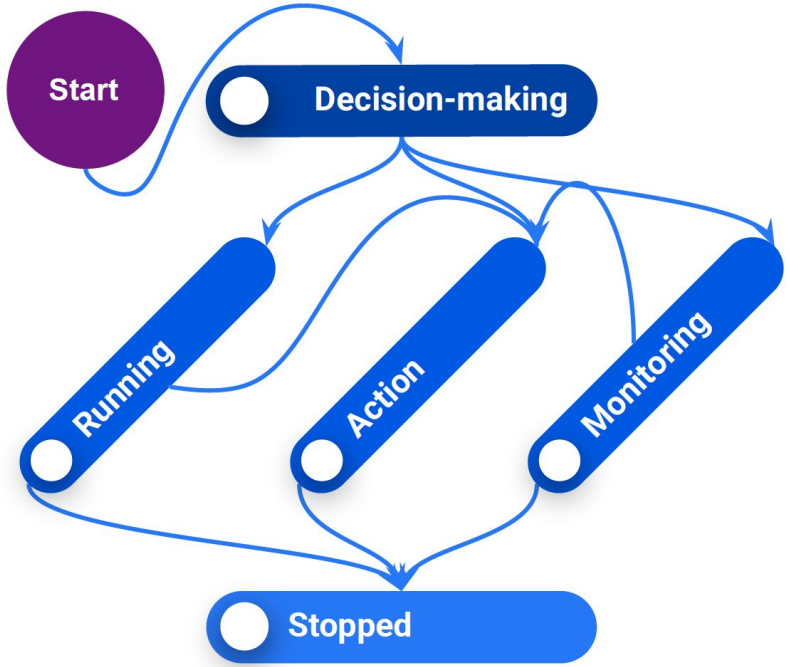
- 157.180.228.0/22
- 157.180.232.0/22
- 192.108.45.0/24
- 192.108.46.0/23
- 192.108.68.0/24
- 2a00:139c::/45



States of execution

States of execution

- Decision-making: NOTED is making the network decision to potentially execute an action or not.
- Running: NOTED is running but there are no transfers in FTS so NOTED is waiting and running until the link-saturation alarm is cleared.
- Monitoring: NOTED is running and there are on-going FTS transfers, but they are below the defined bandwidth threshold that we establish.
- Action: NOTED is running and has triggered an SDN action to provide more bandwidth.
- Stopped: NOTED has stopped because there are no transfers in FTS and the link-saturation alarm has cleared.



NOTED in MONIT Grafana

Database parameters

Alarm parameters

Alarm ID (int):
an unique identifier assigned by the network controller incrementally

Alarm name (str):
by convention is defined as the direction of the alarm from source to destination

Alarm start (timestamp):
timestamp that defines the start of the alarm generated by CERN network monitoring system

Alarm end (timestamp):
timestamp that defined the end of the alarm generated by CERN network monitoring system

Router interface (str):
describes the router and interface that is suffering network congestion

NOTED parameters

NOTED start (timestamp):
timestamp that defines the start of the network controller, i.e. NOTED is monitoring the link

NOTED end (timestamp):
timestamp that defined the end of the network controller, i.e. NOTED stops monitoring the link

NOTED version (str):
defines the version of NOTED that is being running, i.e. {custom, LHCONE, LHCOPN}

NOTED status (str):
defines the status of NOTED, i.e. {action, decision-making, running, monitoring, stopped}

NOTED description (str):
brief description about network usage and actions taken by the network controller

Max FTS throughput (float):
maximum throughput achieved during the large data transfer in FTS

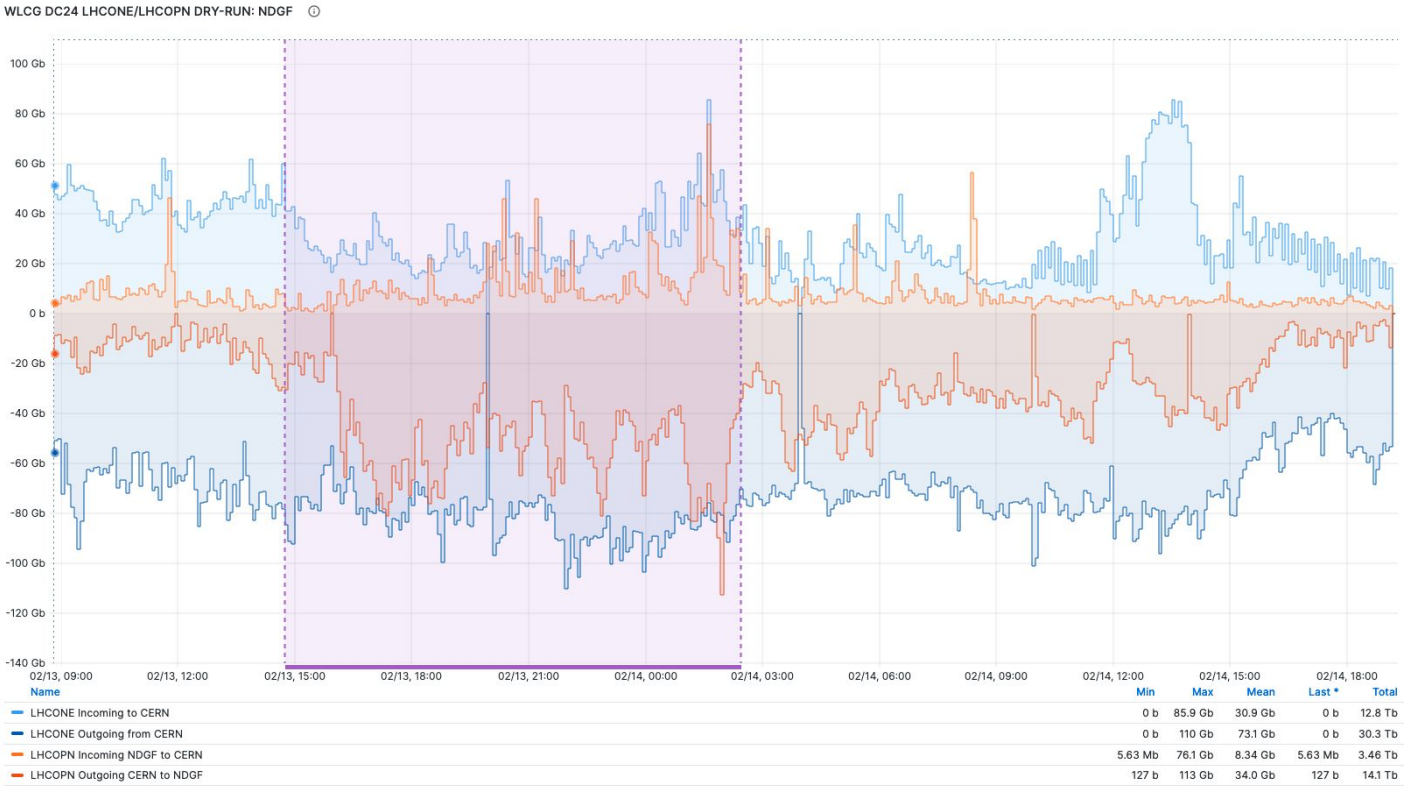
SDN parameters

SDN status (str):
defines the action taken by the network controller, i.e. {provided, released, not provided}

SDN start (timestamp):
timestamp that defines the start of the action in the network, i.e. NOTED is adding capacity

SDN end (timestamp):
timestamp that defined the end of the action in the network, i.e. NOTED stops adding the link

MONIT Grafana



- Network congestion

MONIT Grafana

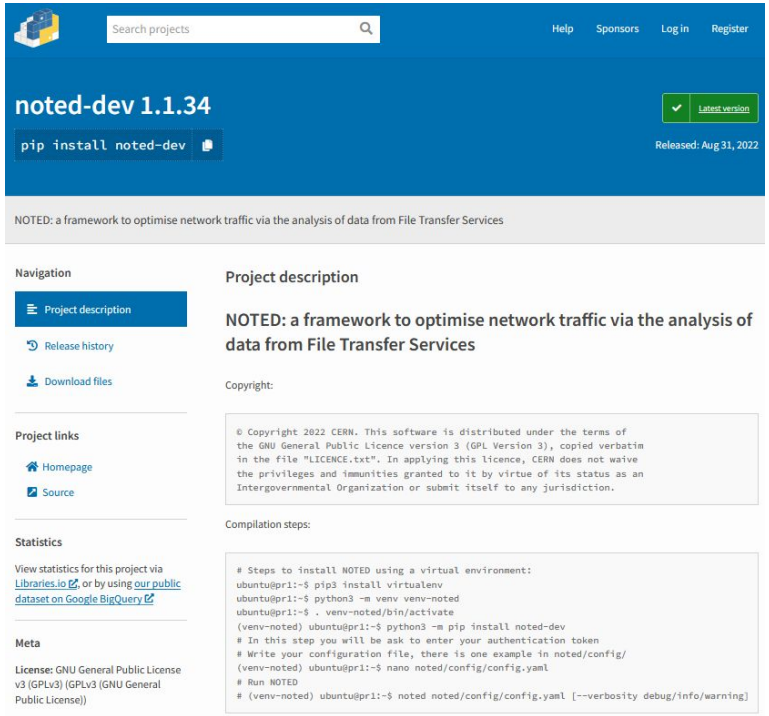
NOTED Alarms

ID	Alarm name	Version	NOTED status	NOTED action	SDN status	Max FTS Throughput [Gb/s]	Interface
184	CH-CERN to CA-TRIUMF	CUSTOM	Running	Spectrum generated an alarm: NOTED is inspecting FTS.	Not provided	0	
187	DE-KIT to CA-TRIUMF	CUSTOM	Action	On-going SDN. FTS throughput [Gb/s]: 5.56	Provided	9.94	
211	CH-CERN to FR-CCIN2P3	CUSTOM	Monitoring	No transfers found in FTS. NOTED is still running until Spectrum clears the alarm.	Not provided		
219	DE-KIT to CA-TRIUMF	CUSTOM	Stopped	The large data transfer is finished.	Released	22.3	
73	ES-ATLAS-T2 to CH-CERN	LHCONE	Decision-making	An action on the link may be required: number of events: 1. Throughput [Gb/s]: 4.12	Not provided		I513-e-rjup1-1_irb.111
83	FR-CCIN2P3 to CH-CERN	LHCONE	Action	On-going SDN. FTS throughput [Gb/s]: 4.94	Provided	7.52	I513-e-rjup1-1_irb.111
84	RO-LCG to CH-CERN	LHCONE	Stopped	The large data transfer is finished.	Released	10.3	I513-e-rjup1-1_irb.111
85	ES-PIC to CH-CERN	LHCONE	Action	On-going SDN. FTS throughput [Gb/s]: 5.94	Provided	12.6	I513-e-rjup1-1_irb.111
107	FR-GRIF to CH-CERN	LHCONE	Monitoring	No transfers found in FTS. NOTED is still running until Spectrum clears the alarm.	Not provided		I513-e-rjup1-1_irb.111
108	IT-INFN-T2 to CH-CERN	LHCONE	Stopped	The large data transfer is finished.	Released	27.9	I513-e-rjup1-1_irb.111
116	UK-SouthGrid to CH-CERN	LHCONE	Running	Spectrum generated an alarm: NOTED is inspecting FTS.	Not provided		I513-e-rjup1-1_irb.111
29	AU-ATLAS to CH-CERN	LHCOPN	Stopped	The large data transfer is finished.	Released	8.79	I513-e-rjup1-1_irb.3530
30	CH-CERN to CA-TRIUMF	LHCOPN	Action	On-going SDN. FTS throughput [Gb/s]: 7.45	Provided	31.5	I513-e-rjup1-1_irb.2126
31	CH-CERN to DE-KIT	LHCOPN	Stopped	The large data transfer is finished.	Released	17.7	I513-e-rjup1-1_irb.3530
32	CH-CERN to DE-KIT	LHCOPN	Monitoring	No transfers found in FTS. NOTED is still running until Spectrum clears the alarm.	Not provided	0	I513-e-rjup1-1_irb.3530
36	NL-T1 to CH-CERN	LHCOPN	Decision-making	An action on the link may be required: number of events: 1. Throughput [Gb/s]: 6.48	Not provided		I513-e-rjup1-1_irb.3530
37	DE-KIT to CH-CERN	LHCOPN	Running	Spectrum generated an alarm: NOTED is inspecting FTS.	Not provided		I513-e-rjup1-1_irb.3530

Package distribution

Package distribution and installation

- Available in PyPI: 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:

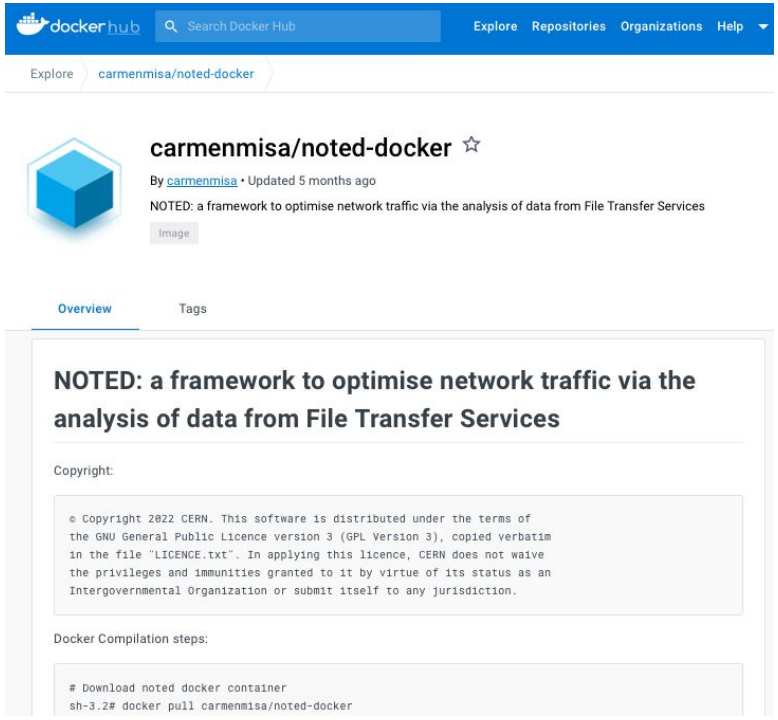
```
# Install noted-dev
(venv-noted) $ python3 -m pip install noted-dev
# Write your configuration file
(venv-noted) $ nano noted/config/config.yaml
# Run NOTED
(venv-noted) $ noted noted/config/config.yaml
```

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
```

Package distribution and installation

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



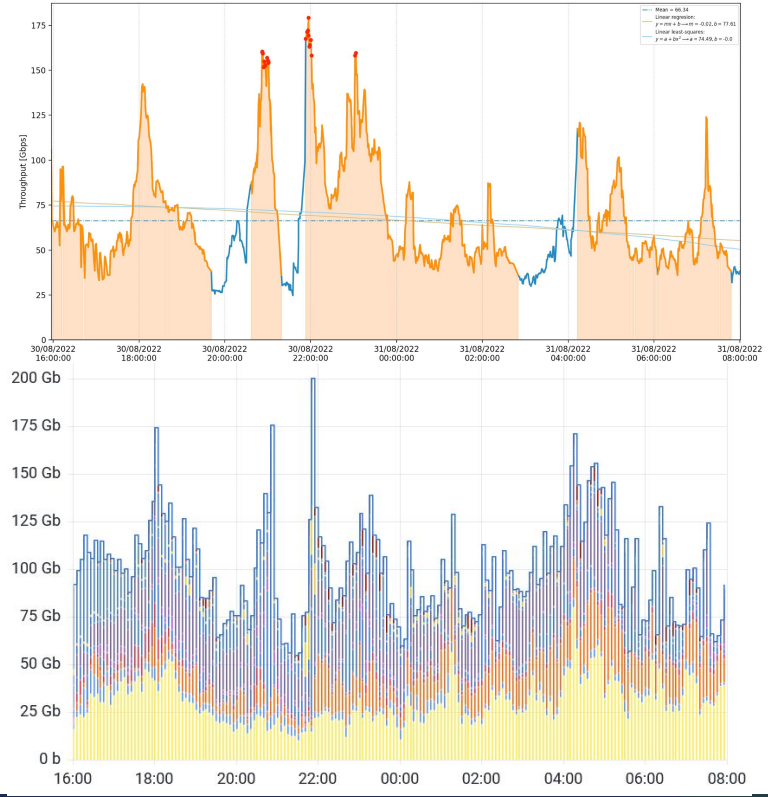
Installation:

```
# 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 &
```

NOTED demonstrations at SC22, SC23

(Provision of dynamic circuits by using SENSE as an SDN provider)

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

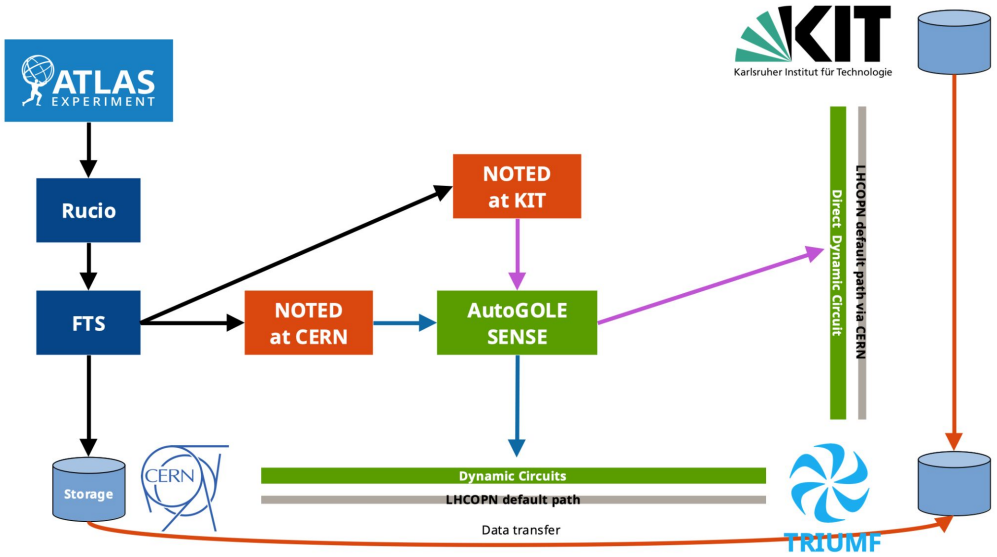


- If throughput > 80 Gb/s → NOTED provides a dynamic circuit. When throughput < 40 Gb/s → NOTED cancels the dynamic circuit and the traffic is routed back to the default path.

- Observations of NOTED about the network utilisation correspond with the reported ones in Grafana by LHCONE/LHCOPN production routers.

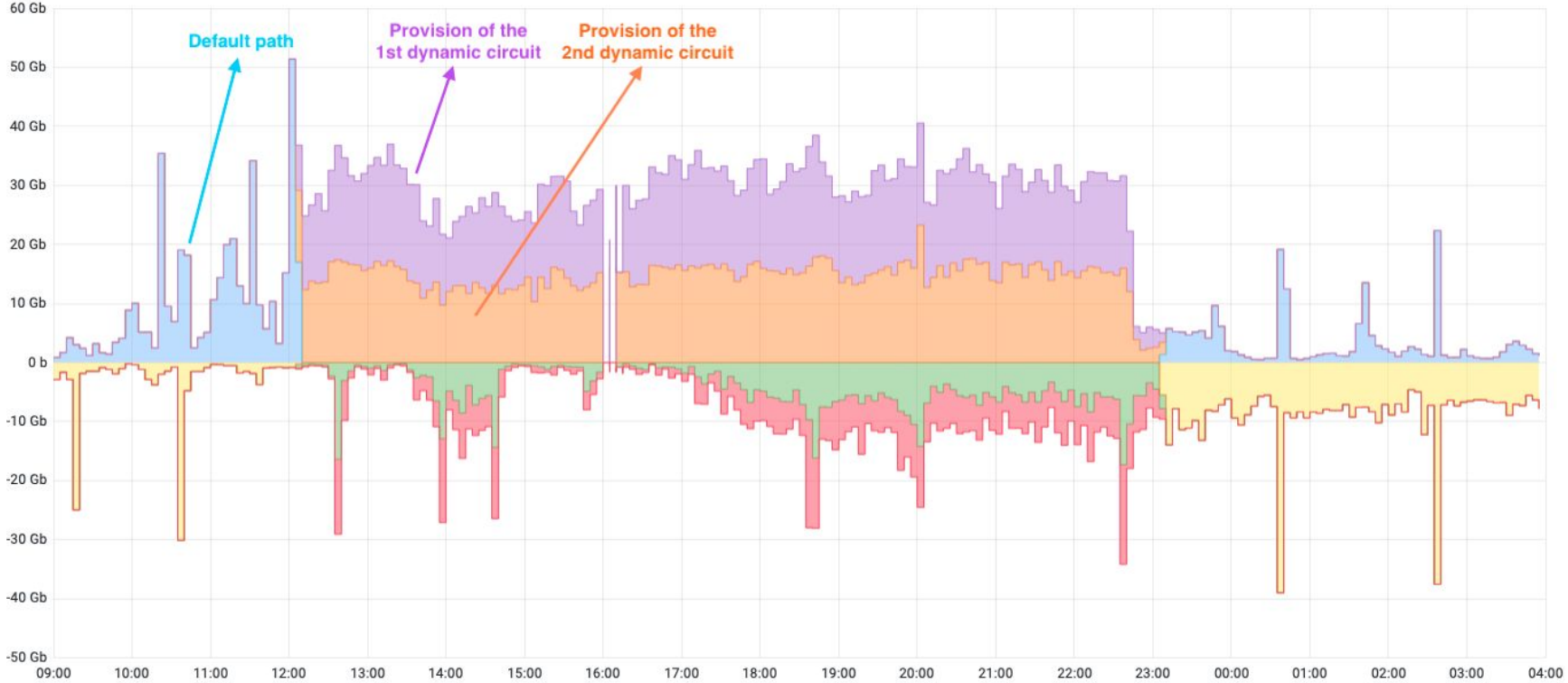
Therefore, by inspecting FTS data transfers it is possible to get an understanding of the network usage and improve its performance by executing an action in the topology of the network.

NOTED demo at Super Computing 2022 (CUSTOM version)

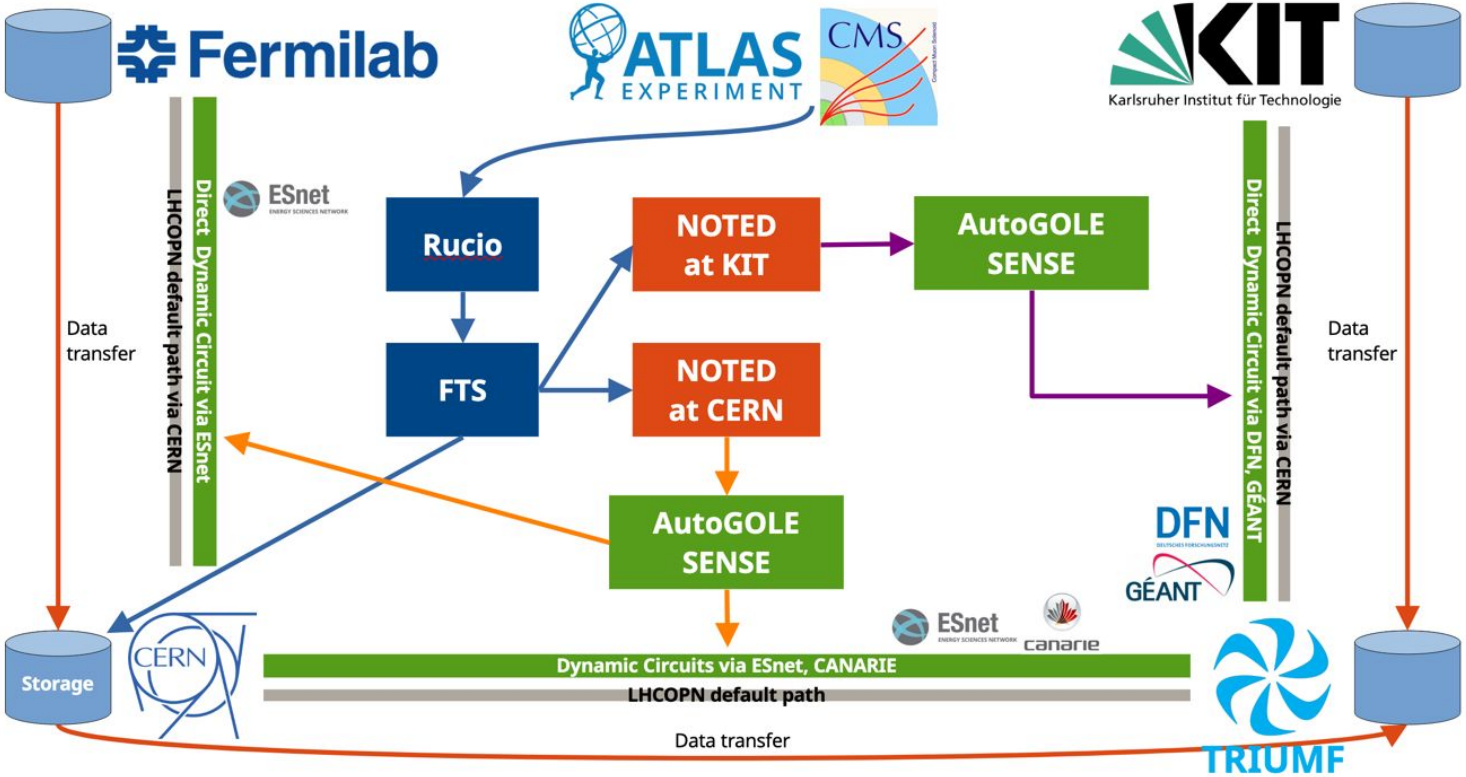


1. NOTED looks in FTS for large data transfers.
2. When it detects a large data transfer → request a dynamic circuit by using the SENSE/AutoGOLE provisioning system.
3. LHCOPN border routers at CERN will route the data transfers over the new dynamic circuit.
4. When the large data transfer is completed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

NOTED demo at Super Computing 2022 (CUSTOM version)

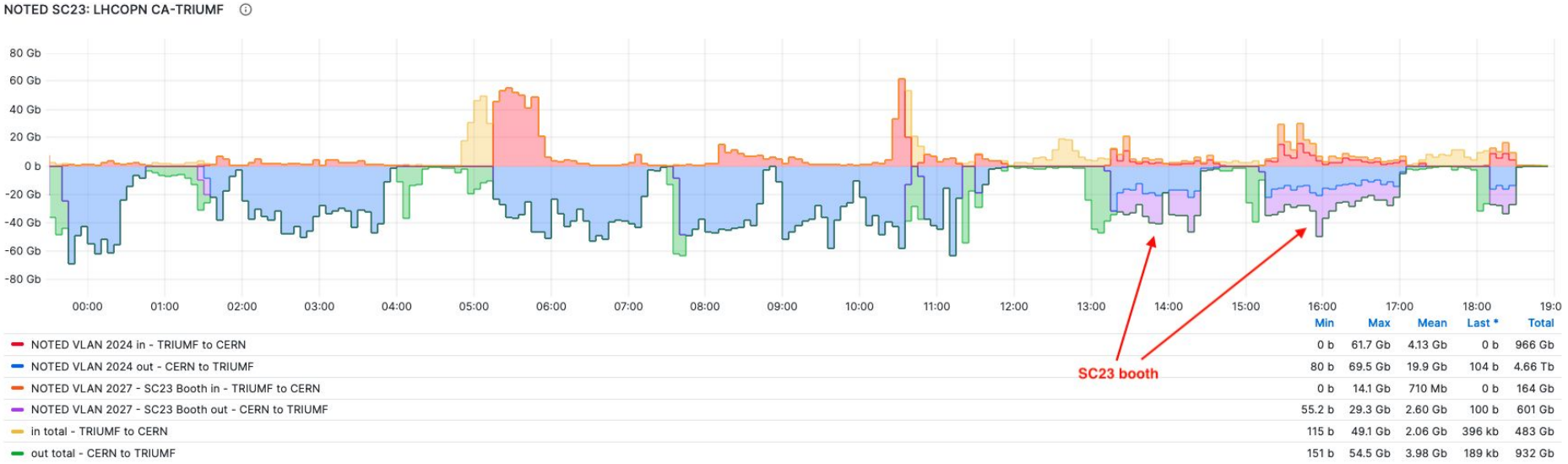


NOTED demo at Super Computing 2023 (LHCOPN, LHCONE and CUSTOM version)



NOTED demo at Super Computing 2023 (LHCOPN, LHCONE and CUSTOM version)

- Results of 14th November 2023.
- Data transfers between CH-CERN → CA-TRIUMF through SC23 booth.

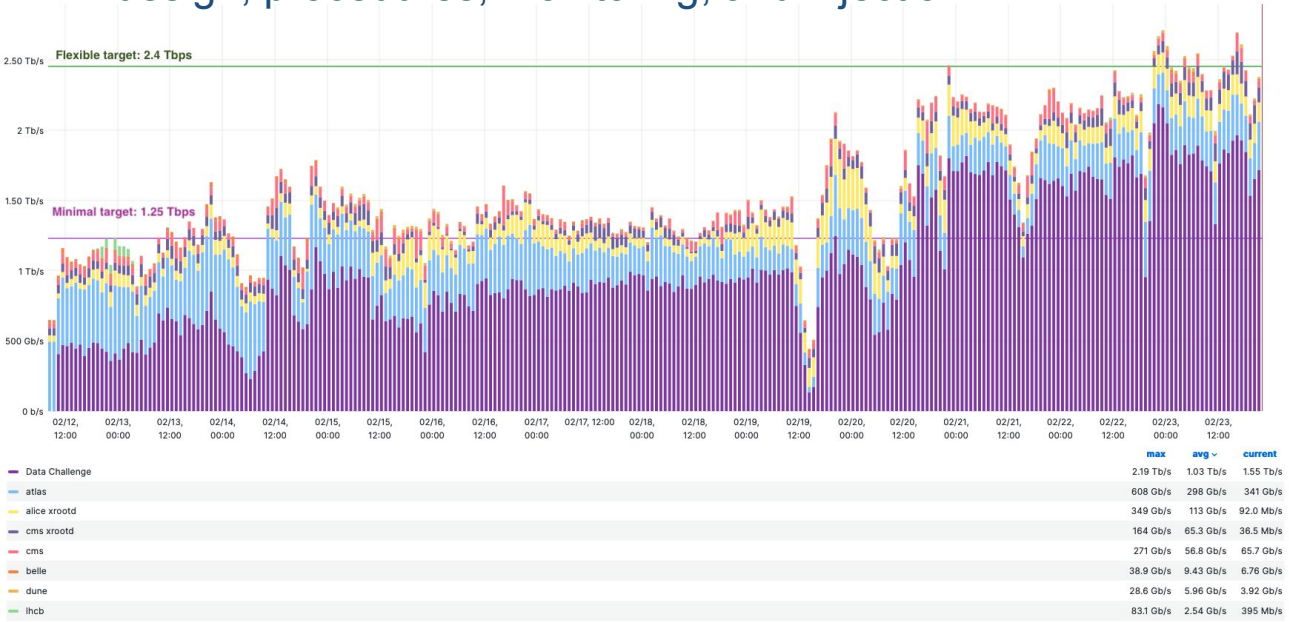


NOTED demonstrations at WLCG DC24

(Load balance between LHCONE and LHCOPN networks)

WLCG Data Challenge 2024

- Objective: demonstrate readiness for HL-LHC data rates by 2029.
- Lots of efforts on coordinating the data challenges across multiple experiments in terms of design, procedures, monitoring, and injection.

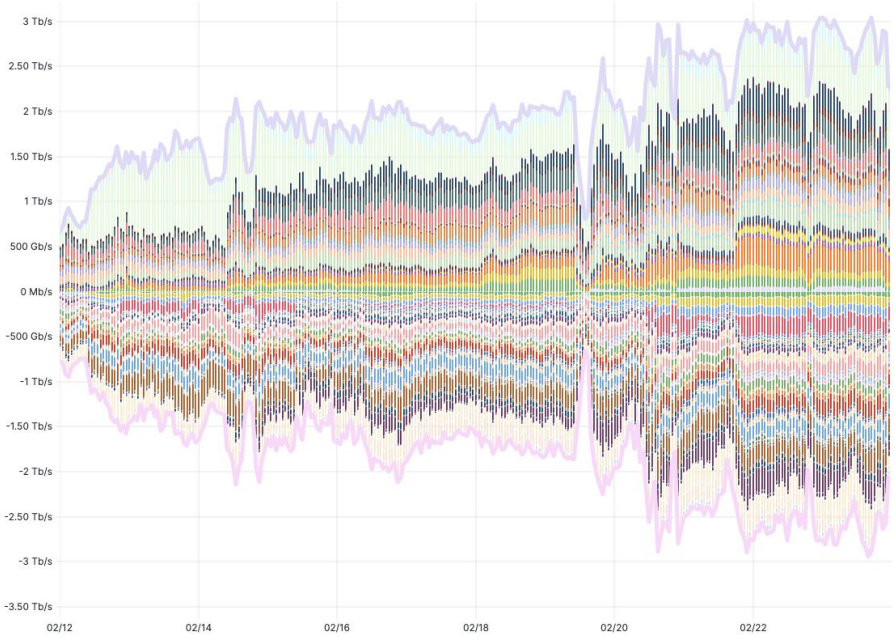


	max	avg	current
Data Challenge	2.19 Tb/s	1.03 Tb/s	1.55 Tb/s
atlas	608 Gb/s	298 Gb/s	341 Gb/s
alice xrootd	349 Gb/s	113 Gb/s	92.0 Mb/s
cms xrootd	164 Gb/s	65.3 Gb/s	36.5 Mb/s
cms	271 Gb/s	56.8 Gb/s	65.7 Gb/s
belle	38.9 Gb/s	9.43 Gb/s	6.76 Gb/s
dune	28.6 Gb/s	5.96 Gb/s	3.92 Gb/s
lhcb	83.1 Gb/s	2.54 Gb/s	395 Mb/s

- Target goal of WLCG Data Challenge 2024: 25% rate of HL-LHC
  - Peak at 3 Tbps
  - Reached 2.5 Tbps for ~9 hours
  - Flexible model: full mesh

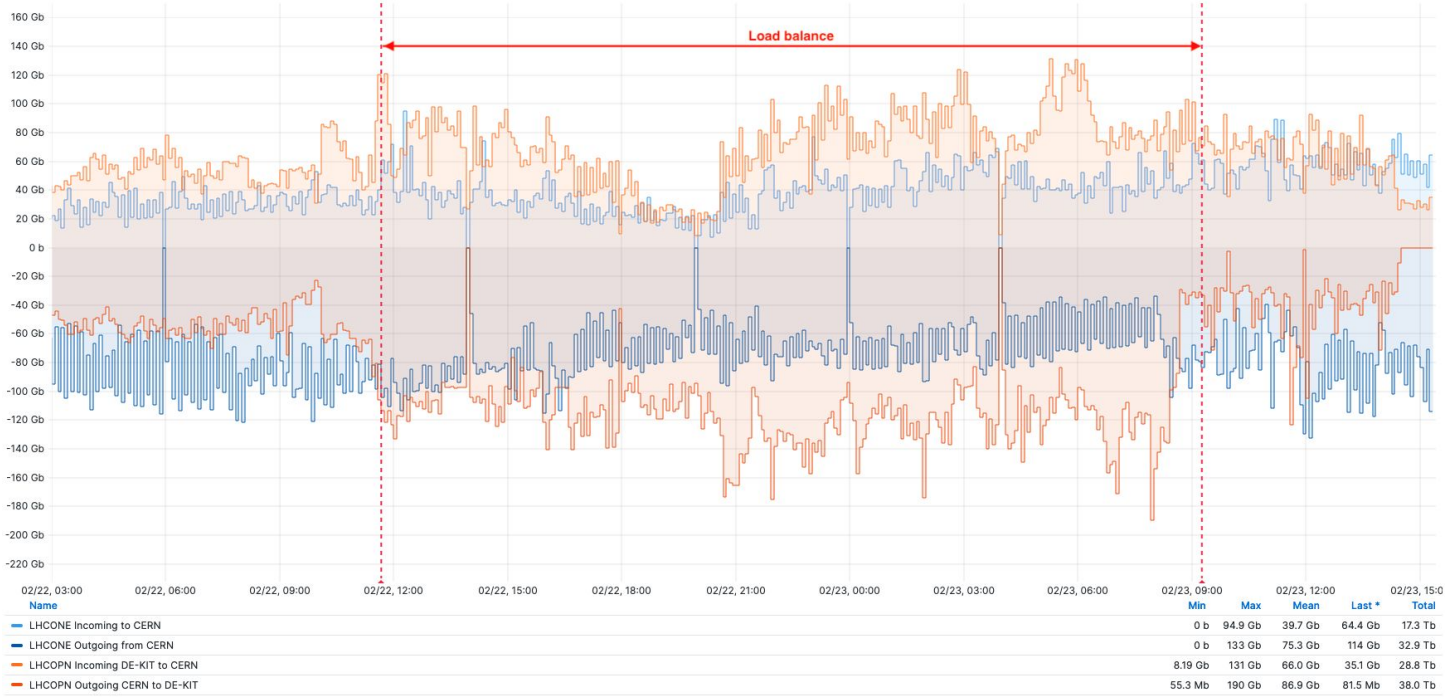
WLCG Data Challenge 2024

- The backbone network exhibit great network performance. The bottlenecks were mostly due to storage configuration and storage hardware limitations.



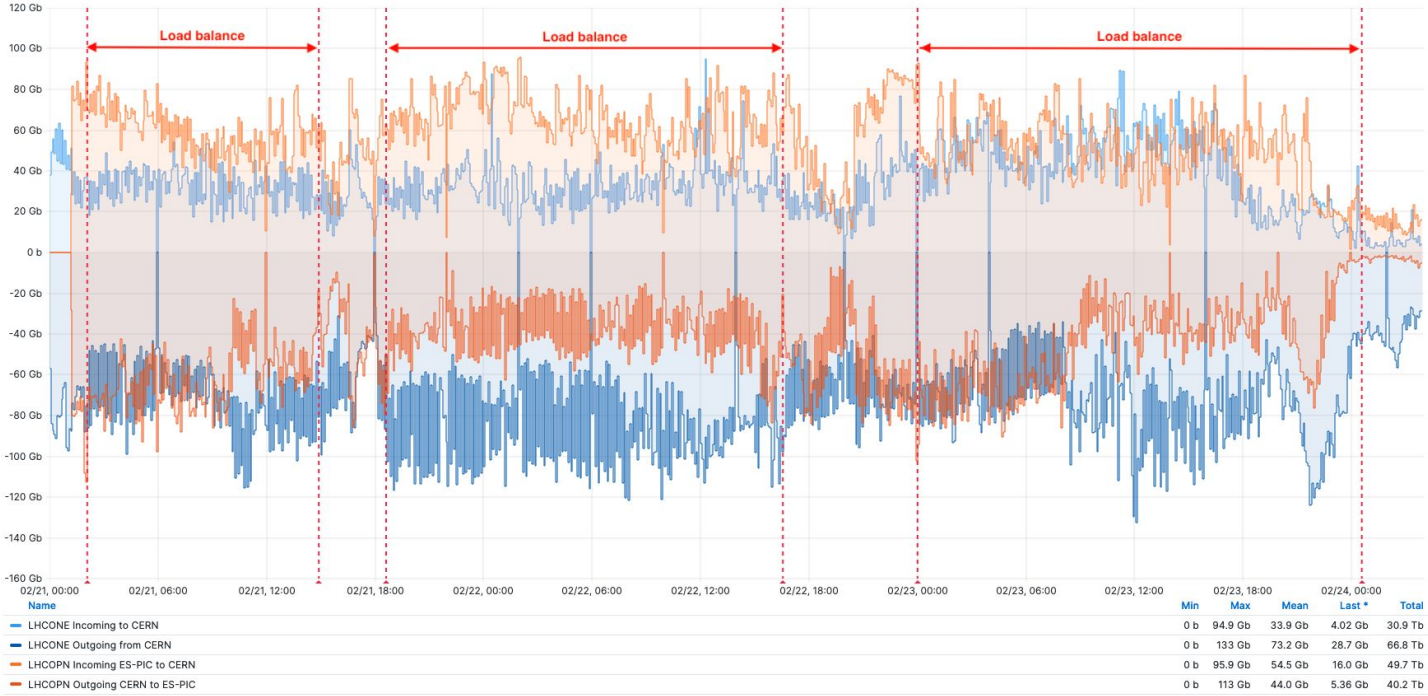
- Useful exercise to find bottlenecks within sites
- Stress tests impacted on the network sites and overloaded storage endpoints
- Test scalability and push services to extreme rates above their normal operation
  - FTS ran 2x of its normal transfer rate

NOTED demo at DC24 (LHCOPN, LHCONE version)



- DE-KIT load balancing between LHCOPN and LHCONE (from 22nd to 23rd of February 2024)

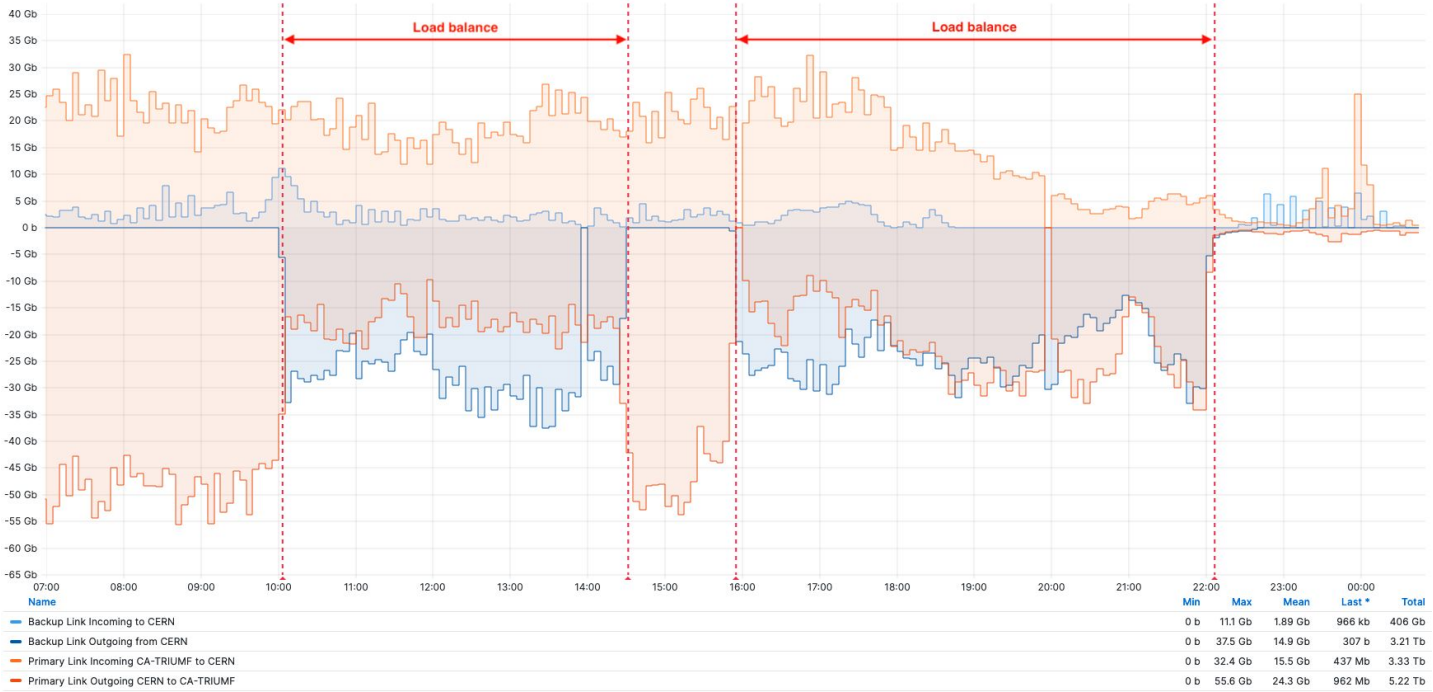
NOTED demo at DC24 (LHCOPN, LHCONE version)



Name	Min	Max	Mean	Last *	Total
LHCONE Incoming to CERN	0 b	94.9 Gb	33.9 Gb	4.02 Gb	30.9 Tb
LHCONE Outgoing from CERN	0 b	133 Gb	73.2 Gb	28.7 Gb	66.8 Tb
LHCOPN Incoming ES-PIC to CERN	0 b	95.9 Gb	54.5 Gb	16.0 Gb	49.7 Tb
LHCOPN Outgoing CERN to ES-PIC	0 b	113 Gb	44.0 Gb	5.36 Gb	40.2 Tb

- ES-PIC load balancing between LHCOPN and LHCONE (from 21st to 23rd of February 2024)

NOTED demo at DC24 (LHCOPN, LHCONE version)



Name	Min	Max	Mean	Last *	Total
Backup Link Incoming to CERN	0 b	11.1 Gb	1.89 Gb	966 kb	406 Gb
Backup Link Outgoing from CERN	0 b	37.5 Gb	14.9 Gb	307 b	3.21 Tb
Primary Link Incoming CA-TRIUMF to CERN	0 b	32.4 Gb	15.5 Gb	437 Mb	3.33 Tb
Primary Link Outgoing CERN to CA-TRIUMF	0 b	55.6 Gb	24.3 Gb	962 Mb	5.22 Tb

- CA-TRIUMF load balancing between LHCOPN and its backup link (from 21st to 23rd of February 2024)

Conclusions and future work

Conclusions:

- NOTED can reduce duration of large data transfers and improve the efficient use of network resources. It has been demonstrated with production FTS transfers.
- NOTED makes decisions by watching and understanding the behaviour of transfer services. Transfer applications don’t need any modification to work with NOTED.

Future work:

- Improve decision-making as much as possible, predict the duration and traffic forecasting by using machine learning.

Publications

- C. Busse-Grawitz, E. Martelli, M. Lassnig, A. Manzi, O. Keeble and T. Cass, *The NOTED software tool-set improves efficient network utilization for rucio data transfers via FTS*, CHEP 2020.
- J. Waczynska, E. Martelli, E. Karavakis and T. Cass, *NOTED: a framework to optimise network traffic via the analysis of data from file transfer services*, CHEP 2021.
- J. Waczynska, E. Martelli, S. Vallecorsa, E. Karavakis and T. Cass, *Convolutional LSTM models to estimate network traffic*, CHEP 2021.
- C. Misa-Moreira, E. Martelli and T. Cass, *NOTED: an intelligent network controller to improve the throughput of large data transfers in file transfer services by handling dynamic circuits*, CHEP 2023.
- C. Misa-Moreira and E. Martelli, *NOTED: a congestion driven network controller*, ISGC 2024.

Thank you

Any questions?

carmen.misa.moreira@cern.ch

tnc24
RENDEZVOUS À RENNES
Rennes, France | 10-14 JUNE 2024

Co-funded by the European Union

GÉANT