
CERN

NOTED: A Congestion Driven Network Controller

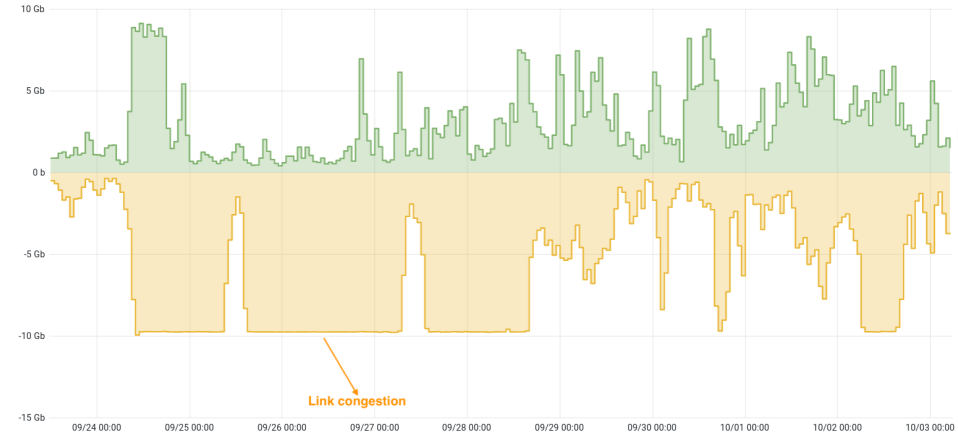
CERN
IT Department CS Group

International Symposium on Grids and Clouds (ISGC24)
24th - 29th March 2024

Carmen Misa Moreira
Edoardo Martelli

Outline

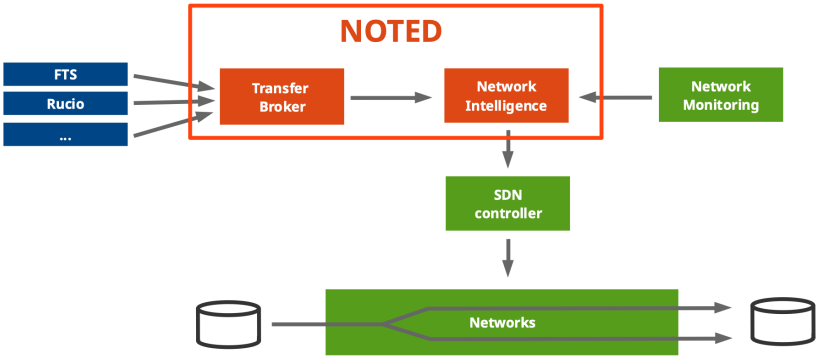
Motivation



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

Architecture

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 of the network topology → get IPv4/IPv6 addresses, endpoints, rcsite and federation.

FTS
File Transfer Service

elasticsearch

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
    }
  }
```

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",
},
```

Dataset structure and workflow

Configuration given by the network administrator → a list of {src_rcsite, dst_rcsite} pairs.

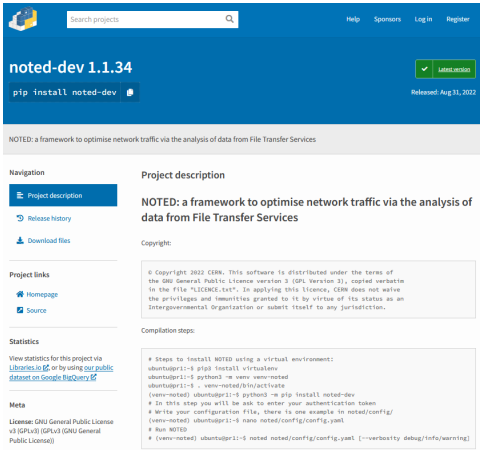
1. Enrich NOTED with the topology of the network:
   - Query CRIC database → get the endpoints (α_i, β_i) that could be involved in the transfers for the given {src_rcsite, dst_rcsite} pairs.
2. Analyse on-going and upcoming data transfers:
   - Query FTS every minute → get the on-going transfers for each set of endpoints (α_i, β_i). *Netwok utilization* $= \sum_{i=0}^{N} \varphi(\alpha_i, \beta_i)_{involved}$
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

Package distribution and installation

PyPI package

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:

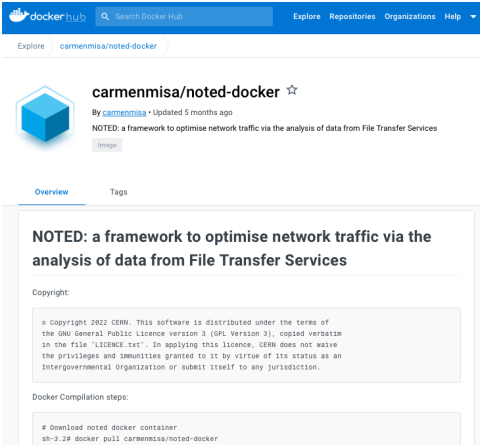
```
# Install noted-dev
(venv-noted) $ python3 -m pip install noted-dev
# Write your configuration file
(venv-noted) $ nano noted/config/config.yaml
# Run NOTED
(venv-noted) $ noted noted/config/config.yaml
```

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
```

Docker container

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



Installation:

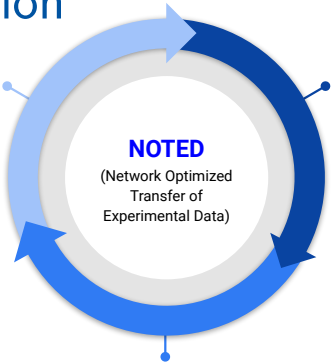
```
# Download noted docker container:
$ docker pull carmenmisa/noted-docker

# 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 &
```

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
```

Network monitoring alarm polling (LHCOPN, LHCONE version)

- Poll the alarms IN/OUT LOAD THRESHOLD EXCEEDED generated by the CERN NMS

Alarms / Spectrum Alarms ☆ 

Last

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	I513-e-rjup1-1_irb.2126	Gen_IF_Port	Port	OUT LOAD THRESH...	No	2023-10-06 08:59:23		
MINOR	1	I513-e-rjup1-1_irb.3530	Gen_IF_Port	Port	OUT LOAD THRESH...	No	2023-10-06 08:43:05		
MINOR	1	I513-v-rjuxl-12_xe-1_0_11	Gen_IF_Port	Port	OUT LOAD THRESH...	No	2023-10-06 08:21:10		
MINOR	1	I513-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	I513-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	I513-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	I513-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	I513-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	I513-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	I513-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 versions)

Identify the prefixes routed via the alarmed interface

- Find the IP of the next hop:

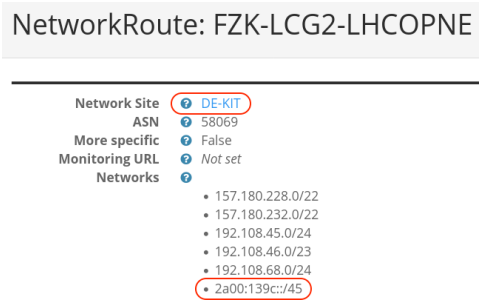
```
BORDER-ROUTER>  show interfaces irb.3530 terse

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:

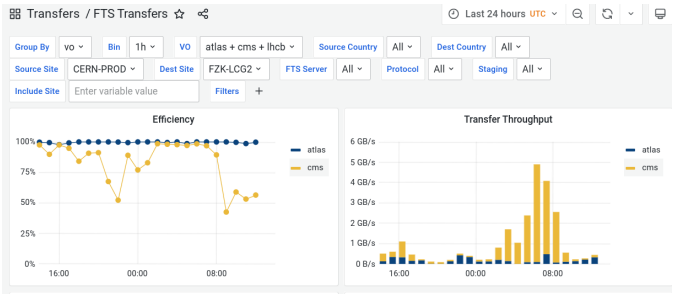
```
BORDER-ROUTER>  show route next-hop 2001:1458:302:38::2

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 versions)

- Lookup routed prefixes in CRIC to identify the destination site:



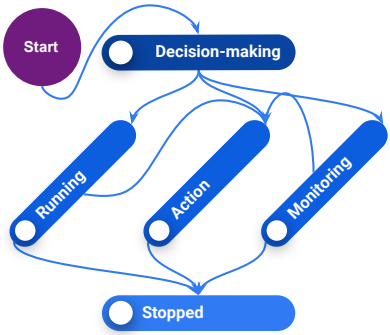
- Look for FTS transfers and make a network decision if it is causing congestion:



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.



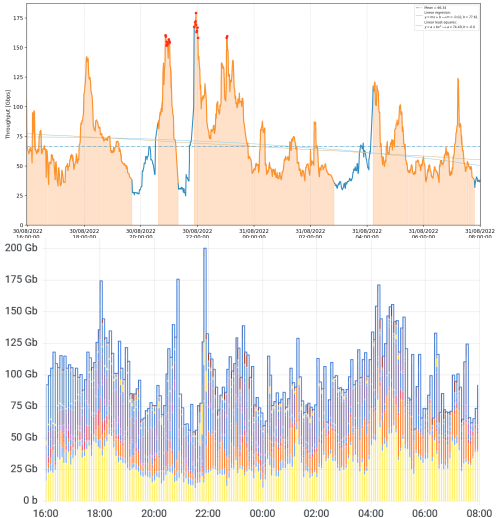
NOTED alarms in MONIT Grafana [Link to the dashboard]

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

NOTED demonstrations

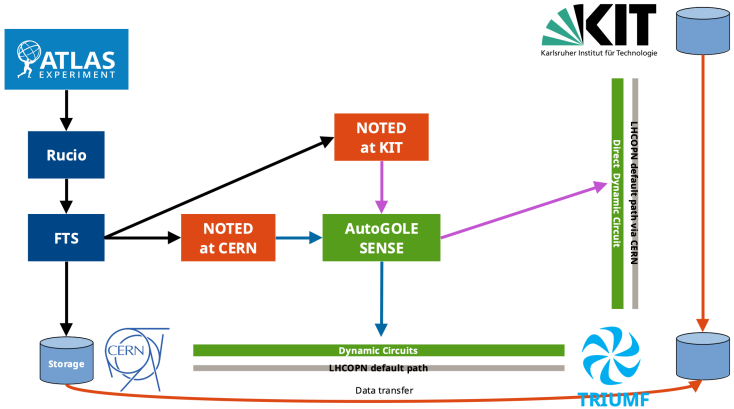
Transfers of WLCG sites in LHCONE (31st of August 2022)



- If throughput > 80 GB/s $\rightarrow$ NOTED provides a dynamic circuit. When throughput < 40 GB/s $\rightarrow$ 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.

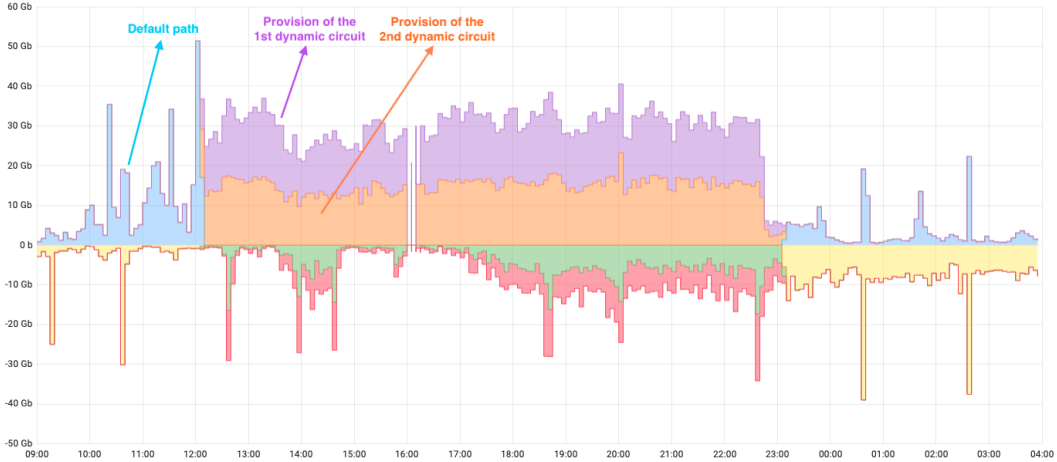
Therefore, by inspecting FTS data transfers it is possible to get an understanding of the network usage and improve its performance by executing an action in the topology of the network.

NOTED demo at SC22 (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 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 SC22 (CUSTOM version)



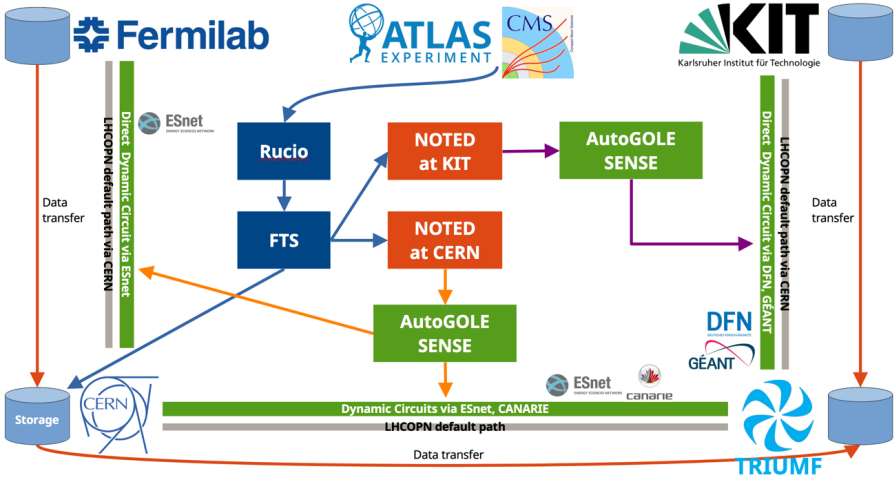
SC22 participants

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.

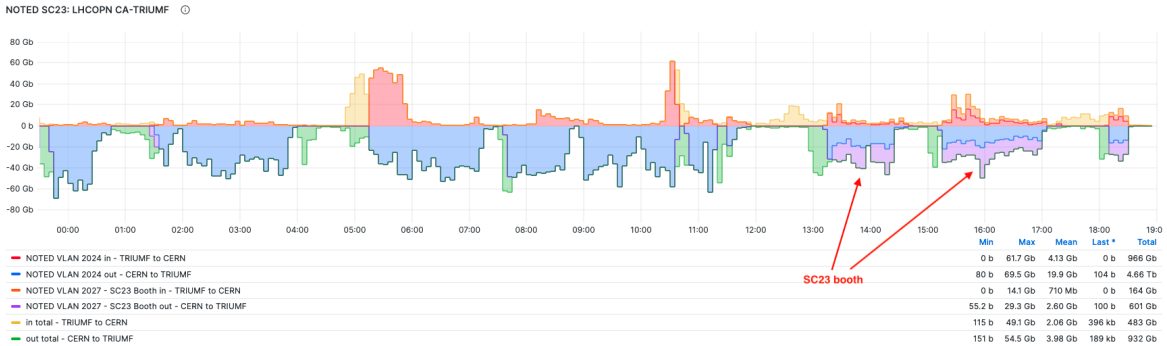
Participants:

NOTED demo at SC23 (LHCOPN, LHCONE and custom versions)



NOTED demo at SC23 (LHCOPN, LHCONE and custom versions)

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



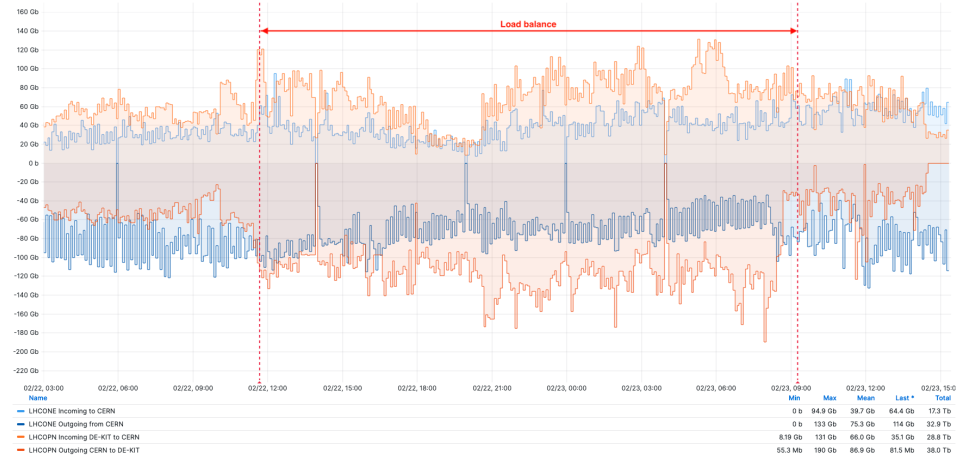
SC23 participants

Components:

- 3x NOTED controllers and FTS at CERN.
  - 2x custom version for TRIUMF and Fermilab.
  - 1x LHCOPN/LHCONE version.
- 1x NOTED custom controller at KIT.
- Data storage at CERN, TRIUMF, KIT and Fermilab.
- AutoGOLE/SENSE circuits between CERN-TRIUMF, CERN-Fermilab and KIT-TRIUMF.
  - SENSE circuits are provided by ESnet, CANARIE, DFN and GÉANT.

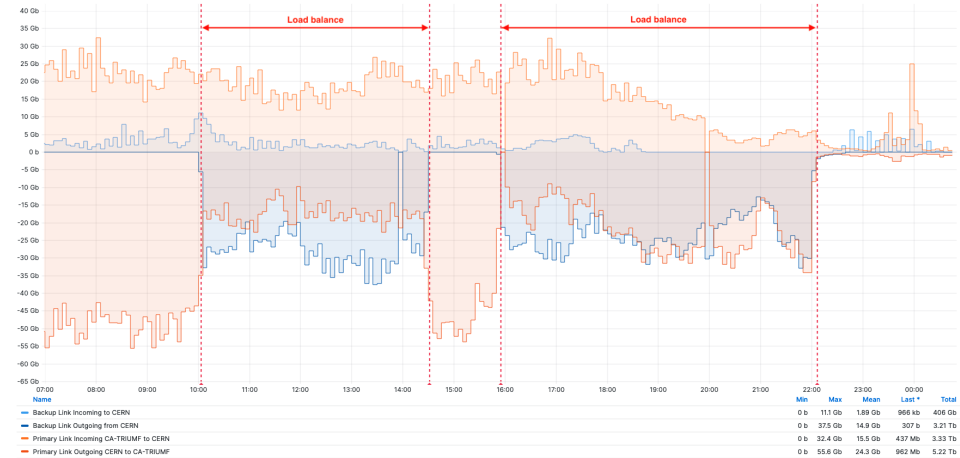
Participants:

NOTED demo at DC24 (LHCOPN, LHCONE versions)



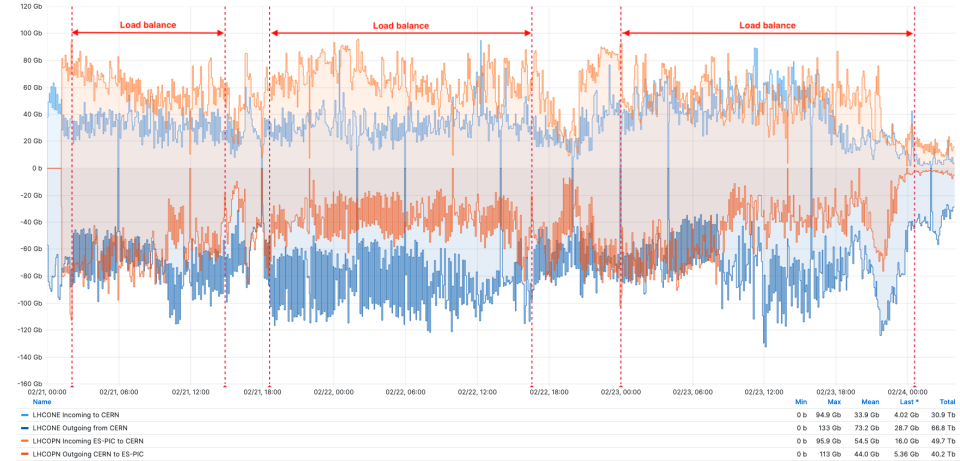
- DE-KIT load balancing between LHCOPN and LHCONE (from 22nd to 23rd of February 2024)

NOTED demo at DC24 (LHCOPN, LHCONE versions)



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

NOTED demo at DC24 (LHCOPN, LHCONE versions)



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

DC24 participants

- Monitoring of LHCONE and LHCOPN links at CERN.
- For CA-TRIUMF: load balance with their backup link.
- For ES-PIC and DE-KIT: load balance between LHCOPN and LHCONE.
- Dry-run mode for the rest of Tier 1's.

Conclusions and future work

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.

Thanks for your attention!


CERN