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
Architecture
   Elements
   Interaction with FTS and CRIC
Package distribution and installation
   PyPI package
   Docker container
Modes of operation
   Network monitoring and alarm polling
   Border router forwarding table
   Identify WLCG destination site
States of execution
NOTED demonstrations
   Transfers of WLCG sites in LHCONE
   NOTED demo at SC22
   NOTED demo at SC23
   NOTED demo at DC24
Conclusions and future work
Motivation

- Large data transfers can saturate network links while alternative paths may be left idle.
Architecture
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.
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.
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"]}
    }
  },
  "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" }
  ],
  "rc_tier_level": 1,
  "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"
},

28th March 2024 - ISGC24 NOTED: Network Optimised Transfer of Experimental Data
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 \((\alpha_i, \beta_i)\) 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 \((\alpha_i, \beta_i)\).
   
   \[
   \text{Network utilization} = \sum_{i=0}^{N} \varphi(\alpha_i, \beta_i)_{involved}
   \]

3. Network decision: when NOTED detects that the link is going to be congested → provides a dynamic circuit via Sense/AutoGOLE.

<table>
<thead>
<tr>
<th>Source endpoint</th>
<th>Destination endpoint</th>
<th>Data [GB]</th>
<th>Throughput [Gb/s]</th>
<th>Parallel transfers</th>
<th>Queued transfers</th>
</tr>
</thead>
<tbody>
<tr>
<td>davs://ccdavatlas.in2p3.fr</td>
<td>davs://webdav.echo.stfc.ac.uk</td>
<td>139.3726</td>
<td>54.0827</td>
<td>453</td>
<td>28557</td>
</tr>
<tr>
<td>davs://dav.ndgf.org</td>
<td>davs://dcgftp.usatlas.bnl.gov</td>
<td>202.7864</td>
<td>82.0855</td>
<td>862</td>
<td>57880</td>
</tr>
<tr>
<td>davs://atlaswebdav-kit.gridka.de</td>
<td>davs://eosatlas.cern.ch</td>
<td>205.3606</td>
<td>82.0725</td>
<td>888</td>
<td>57790</td>
</tr>
<tr>
<td>davs://f-dpm000.grid.sinica.edu.tw</td>
<td>davs://webdav.lcg.triumf.ca</td>
<td>210.2710</td>
<td>51.0323</td>
<td>567</td>
<td>26314</td>
</tr>
<tr>
<td>davs://ccdavatlas.in2p3.fr</td>
<td>davs://webdav.echo.stfc.ac.uk</td>
<td>332.0009</td>
<td>81.7908</td>
<td>905</td>
<td>50152</td>
</tr>
</tbody>
</table>

28th March 2024 - ISGC24
NOTED: Network Optimised Transfer of Experimental Data
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
NOTED (Network Optimized Transfer of Experimental Data)

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

<table>
<thead>
<tr>
<th>Severity</th>
<th>Occ</th>
<th>Entity name</th>
<th>Type</th>
<th>Class</th>
<th>Alarm name</th>
<th>Ack</th>
<th>Start at</th>
<th>Cleared at</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>MINOR</td>
<td>1</td>
<td>1513-e-rjup1-1_irb.2126</td>
<td>Gen_IF_Port</td>
<td>Port</td>
<td>OUT LOAD THRESH...</td>
<td>No</td>
<td>2023-10-06 08:59:23</td>
<td>2023-10-06 08:59:23</td>
<td>00:00:00</td>
</tr>
<tr>
<td>MINOR</td>
<td>1</td>
<td>1513-e-rjup1-1_irb.3530</td>
<td>Gen_IF_Port</td>
<td>Port</td>
<td>OUT LOAD THRESH...</td>
<td>No</td>
<td>2023-10-06 08:43:05</td>
<td>2023-10-06 08:43:05</td>
<td>00:00:00</td>
</tr>
<tr>
<td>MINOR</td>
<td>1</td>
<td>1513-e-rjup1-1_irb.3530</td>
<td>Gen_IF_Port</td>
<td>Port</td>
<td>OUT LOAD THRESH...</td>
<td>No</td>
<td>2023-10-06 08:21:10</td>
<td>2023-10-06 08:21:10</td>
<td>00:00:00</td>
</tr>
<tr>
<td>MINOR</td>
<td>1</td>
<td>1513-e-rjup1-1_irb.3530</td>
<td>Gen_IF_Port</td>
<td>Port</td>
<td>OUT LOAD THRESH...</td>
<td>No</td>
<td>2023-10-06 08:08:03</td>
<td>2023-10-06 08:08:03</td>
<td>00:00:00</td>
</tr>
<tr>
<td>MINOR</td>
<td>1</td>
<td>1513-e-rjup1-1_irb.3530</td>
<td>Gen_IF_Port</td>
<td>Port</td>
<td>OUT LOAD THRESH...</td>
<td>No</td>
<td>2023-10-06 07:32:47</td>
<td>2023-10-06 07:32:47</td>
<td>00:00:45</td>
</tr>
<tr>
<td>MINOR</td>
<td>1</td>
<td>1513-e-rjup1-1_irb.3530</td>
<td>Gen_IF_Port</td>
<td>Port</td>
<td>IN LOAD THRESH...</td>
<td>No</td>
<td>2023-10-06 07:07:47</td>
<td>2023-10-06 07:07:47</td>
<td>00:00:45</td>
</tr>
<tr>
<td>MINOR</td>
<td>1</td>
<td>1513-e-rjup1-1_irb.3530</td>
<td>Gen_IF_Port</td>
<td>Port</td>
<td>OUT LOAD THRESH...</td>
<td>No</td>
<td>2023-10-06 06:58:02</td>
<td>2023-10-06 06:58:02</td>
<td>00:00:45</td>
</tr>
<tr>
<td>MINOR</td>
<td>1</td>
<td>1513-e-rjup1-1_irb.3530</td>
<td>Gen_IF_Port</td>
<td>Port</td>
<td>OUT LOAD THRESH...</td>
<td>No</td>
<td>2023-10-06 06:46:00</td>
<td>2023-10-06 06:46:00</td>
<td>00:00:45</td>
</tr>
<tr>
<td>MINOR</td>
<td>1</td>
<td>1513-e-rjup1-1_irb.2126</td>
<td>Gen_IF_Port</td>
<td>Port</td>
<td>OUT LOAD THRESH...</td>
<td>No</td>
<td>2023-10-06 06:34:23</td>
<td>2023-10-06 06:34:23</td>
<td>00:00:45</td>
</tr>
<tr>
<td>MINOR</td>
<td>1</td>
<td>1513-e-rjup1-1_irb.13</td>
<td>Gen_IF_Port</td>
<td>Port</td>
<td>OUT LOAD THRESH...</td>
<td>No</td>
<td>2023-10-06 06:15:58</td>
<td>2023-10-06 06:15:58</td>
<td>00:00:45</td>
</tr>
<tr>
<td>MINOR</td>
<td>1</td>
<td>1513-e-rjup1-1_irb.3530</td>
<td>Gen_IF_Port</td>
<td>Port</td>
<td>OUT LOAD THRESH...</td>
<td>No</td>
<td>2023-10-06 05:53:02</td>
<td>2023-10-06 05:53:02</td>
<td>00:00:45</td>
</tr>
</tbody>
</table>
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:
  - NetworkRoute: FZK-LCG2-LHCOPNE
  - DE-KIT
  - 58069
  - False
  - Not set
  - 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

- 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]

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

NOTED SC23: LHCOPN CA-TRIUMF

![Graph showing data transfers between CH-CERN and 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:
DE-KIT load balancing between LHCOPN and LHCONE (from 22\textsuperscript{nd} to 23\textsuperscript{rd} 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 21\textsuperscript{st} to 23\textsuperscript{rd} 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!