PROCEEDINGS OF SCIENCE

NOTED: A Congestion Driven Network Controller

Carmen MISA MOREIRA[a,*] **and Edoardo MARTELLI**[a]

[a] *CERN - Conseil Européen pour la Recherche Nucléaire,*
Esplanade des Particules 1, 1211 Meyrin, Geneva, Switzerland, IT department CS group

E-mail: carmen.misa.moreira@cern.ch, edoardo.martelli@cern.ch

Abstract: NOTED is an intelligent network controller that aims to improve the throughput of large data transfers in FTS (File Transfers Services), which is the service used to exchange data transfers between WLCG sites, to better exploit the available network resources. For a defined set of source and destination endpoints, NOTED retrieves data from FTS to get the on-going data traffic and uses the CRIC (Computing Resource Information Catalog) database to get comprehensive understanding about the network topology. This feature has shown successful results during the SC22 and SC23 conferences, where NOTED was executing actions when it detected congestion on a given link and dynamically reconfigured the network topology by using an SDN (Software-defined Network) service. Recently, NOTED has been integrated with the CERN NMS (Network Monitoring System) to increase even more its capabilities and be driven by congestion. In this way, NOTED brings the capability to identify which WLCG sites are congesting the network, both in LHCOPN (Tier 0 to Tier 1's network) and LHCONE (Tier 1's to Tier 2's network) networks, and execute an action in the network to reconfigure it by adding capacity. This new version of NOTED has been tested during SC23 and will be used at scale during WLCG DC24 (Data Challenge) in which the NREN's and WLCG sites performs the testing at 25% of rate that will be used by HL-LHC (High Luminosity LHC) to accomplish the requirements by 2029.

International Symposium on Grids and Clouds (ISGC) 2024,
24th - 29th March 2024
Academia Sinica Taipei, Taiwan

[*] Speaker

PoS(ISGC2024)013

1. Introduction

The large scientific data transfers generated by the Large Hadron Collider (LHC) experiments can saturate network links, while alternative paths may remain underutilised. Owing to the agnostic characteristics of routing protocols towards network load, we often encounter scenarios where certain links experience congestion and other expensive links are left idle.

Network Optimised Transfer of Experimental Data (NOTED) aims to improve network utilisation and better exploit the available bandwidth. The project was introduced in 2020 and presented at several international conferences. First achievements and outcomes are outlined in articles [1, 2]. Furthermore, a comprehensive study on traffic forecasting has been conducted, using a machine learning approach with Long Short Term Memory (LSTM) neural networks, as detailed in [3].

The capabilities and functionalities of NOTED have undergone testing and demonstration at various international conferences. As outlined in [4], during SC22, independent instances of NOTED at CERN (Switzerland) and KIT (Germany) monitored large data transfers generated by the ATLAS experiment, between these sites and TRIUMF (Canada). The paper provides an overview of the NOTED architecture and highlights how it detects link congestion or a significant surge in network utilisation over an extended period of time. When deemed necessary, the system can automatically reconfigure the network topology by introducing an additional or alternative and better-performing path through dynamic circuit provisioning systems such as Software-Defined Networking (SDN) for End-to-End Networked Science at the Exascale (SENSE) [5–7].

Having set out an introduction, describe the motivation and objectives of the project as well as referring to previous work, we briefly describe NOTED characteristics, in section 2; the modes of operation and execution states of NOTED are detailed in sections 3 and 4 respectively; the identification process of WLCG federations is described in section 5; the results of the intensive tests during SC23 and the WLCG DC24 in section 6 and, finally, our conclusions and plans for future developments in section 7.

2. NOTED in a nutshell

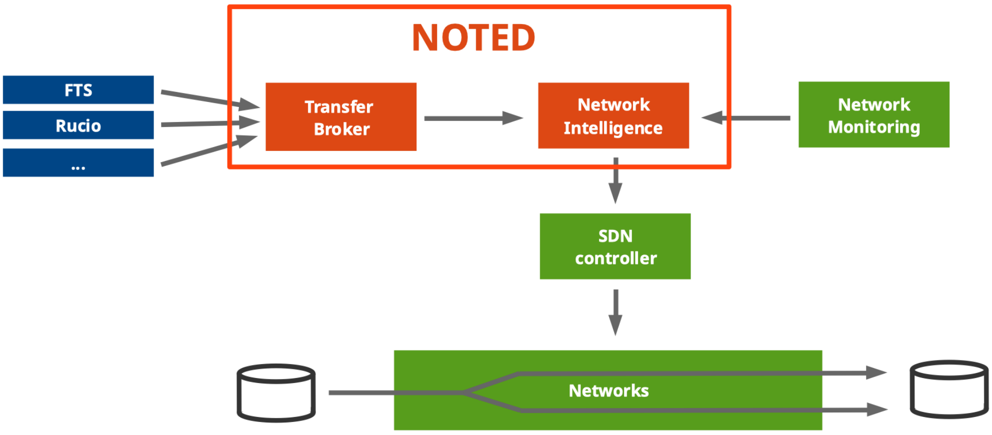


Figure 1: NOTED architecture and components.

The architecture of NOTED is illustrated in Figure 1, emphasising its principal components:

- Transfer Broker: Functioning as the interface interacting with data transfer applications to acquire data. The File Transfer Service (FTS) [8] is the service monitored to understand the network traffic. FTS is the data transfer service employed by LHC experiments for data distribution across sites within the Worldwide LHC Computing Grid (WLCG).

- Network Intelligence: Maintains a comprehensive understanding of network topology and executes network actions based on required bandwidth. The Computing Resource Information Catalogue (CRIC) [9–11] is the primary service used for topology information, forming the basis for network optimisation decisions. CRIC serves as a database employed by WLCG to declare and expose information about available computing resources at a given site.

NOTED queries FTS at one-minute intervals to gather information about on-going and queued transfers. This data is then scrutinised to estimate transfer duration and assess whether actions can be taken to optimise network utilisation, such as redirecting traffic to alternative links. As the network itself is agnostic to topology, NOTED relies on the CRIC database to obtain a comprehensive overview of network elements, encompassing endpoints, sites, and federation.

The interaction with FTS involves querying the optimiser parameters [12] to retrieve data and analyse on-going and queued data transfers. These parameters are crucial for the network decision-making process, which uses information about network utilisation, and source-destination pairs of on-going data transfers to calculate the aggregated transfer flow. Parameters such as throughput, file size, amount of data, number of parallel transfers and submitted transfers that are still waiting in the queue are used to compute network utilisation and estimate the duration of the transfers.

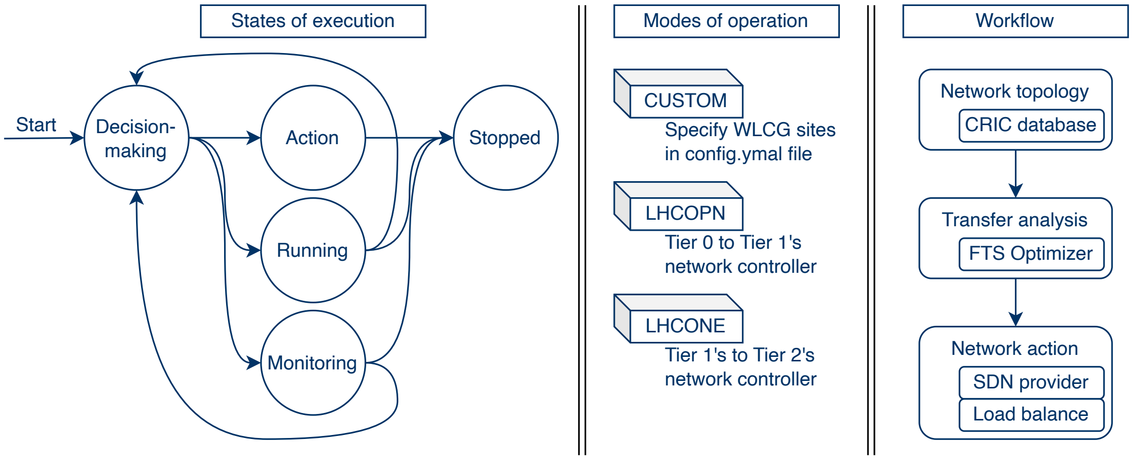


Figure 2: NOTED states of execution, modes of operation and workflow.

Figure 2 exposes a diagram that summarises the states of execution, modes of operation and the workflow of NOTED, which is divided into three stages:

- Network Topology: The network intelligence component queries the CRIC database to gain a comprehensive understanding of the network topology, identifying relevant endpoints associated with the given source-destination pairs for data transfers.

- Transfer Analysis: The transfer broker component analyses the on-going and queued data transfers in FTS every minute, computing the overall network utilisation.

- Network Action: NOTED makes network decisions when congestion is detected on a link. It may provide a dynamic circuit using an SDN provider like SENSE to dynamically allocate additional capacity or divert some load to existing idle circuits.

The NOTED project has effectively showcased its capability to dynamically allocate network links, thereby augmenting the effective bandwidth available for FTS-driven transfers among endpoints, such as WLCG sites, by inspecting on-going data transfers and so identifying those experiencing bandwidth-limited for a long period of time. Recently, the architecture of NOTED has undergone recent enhancements, and the software has been streamlined for easy distribution.

3. Modes of operation

NOTED was initially designed to inspect FTS data transfers between endpoints at WLCG sites, referred to here as CUSTOM mode of operation, and this has been demonstrated at SC22 and reported at various conferences. However, to enhance and abstract the network administrator from manual interventions and parameter definitions, NOTED has undergone integration into the CERN Network Monitoring System (NMS). This integration empowers NOTED to autonomously and automatically identify network congestion within the LHC Optical Private Network (LHCOPN) and LHC Open Network Environment (LHCONE) networks [13, 14], introducing two additional modes of operation:

- CUSTOM Mode: Operation in this mode relies on input parameters specified in a configuration YAML file provided by the network administrator. The parameters, detailed in [4], define the monitoring of FTS data transfers between predefined source and destination endpoints within WLCG sites. For instance, during SC22, two configuration files were utilised to monitor FTS data transfers between CH-CERN (Switzerland) and CA-TRIUMF (Canada), as well as DE-KIT (Germany) and CA-TRIUMF. Characteristics of each link, such as maximum and minimum threshold values and parameters related to the SENSE dynamic circuit provisioning system, were included.

- CERN NMS Mode: In this mode, when the CERN internal monitoring system generates an alarm on interfaces of the LHCOPN (Tier 0 to Tier 1's links) and LHCONE (Tier 1's to Tier 2's links) border routers. NOTED then autonomously initiates a trace and identification process to determine whether the congested interface belongs to a Tier 1 or a Tier 2 WLCG site and, based on this, NOTED enters either LHCOPN or LHCONE operation mode. At this point, it can be mentioned that the entire process is carried out automatically without the need for a YAML configuration file from the network administrator. This process of identifying peers susceptible to network congestion is detailed below in section 5.

    ⋄ LHCOPN Mode: NOTED identifies traffic causing congestion within the LHCOPN network, originating from data transfers between Tier 1 WLCG sites or between CERN (Tier 0) and one or more Tier 1's. Subsequently, it monitors all data transfers in FTS for Tier 1's source and destination endpoints, making network decisions to increase network capacity when deemed necessary. It is noteworthy that NOTED acquires and inspects FTS data transfers for all possible combinations among the 13 Tier 1's within the LHCOPN network.

- ⋄ LHCONE Mode: In this mode, NOTED responds to network congestion caused by data transfers between Tier 1's and/or Tier 2's WLCG sites. It monitors FTS data transfers for all conceivable source and destination endpoints, making decisions to increase network capacity when required. This process is more intricate and intense compared to LHCOPN, considering the extensive network coverage of LHCONE, involving over 100 institutions and hundreds of potential source-destination endpoints contributing to network congestion.

4. States of execution

NOTED manages several execution states that are carried out regardless of the operating mode. Transitions between states are governed by parameters derived from the FTS optimiser, encompassing throughput, transmitted data volume, and queued transfers.

- Running State: This is the initial state entered upon NOTED execution. This initiates an examination of FTS to identify transfers pertinent to the designated link, filtering out those unrelated to the source and destination endpoints, as well as idle transfers. In the absence of on-going data transfers, NOTED remains in this state, consistently monitoring FTS at 5-minute intervals until the link saturation alarm from CERN NMS is cleared. Conversely, upon detecting on-going data transfers, NOTED transitions to either the monitoring or decision-making state, contingent on the estimated network usage.

- Monitoring State: Within this execution state, NOTED is executing and overseeing FTS data transfers that are currently on-going. However, the network usage falls below the predetermined bandwidth threshold required to raise a NOTED alarm for the specified link. Under this circumstance, if the network usage persists below the established bandwidth threshold, NOTED will remain in this state, continuously operating and scrutinising FTS at 1-minute intervals. This will continue until the CERN NMS link saturation alarm is cleared, prompting a transition to the stopped state and subsequently concluding its execution. Conversely, if the network usage surpass the predefined bandwidth threshold, NOTED will switch to the action state, initiating the process of reconfiguring the network.

- Decision Making State: This state is triggered by congestion on a designated link, this state is entered to assess whether to execute an action on the network. This state involves monitoring the FTS queue for a defined number of events, ensuring that the on-going large data transfer lasts longer than 30 minutes, thereby justifying the execution of a network action. It is important to mention that short-duration on-going data transfers are not considered, as the objective is to ensure that network reconfigurations resulting from actions persist for a substantial period of time.

- Action State: Within this execution state, the detection of high network usage exceeding the established threshold prompts the initiation of a network action. This action may involve requesting the provision of a dynamic circuit through an SDN provider to enhance capacity or executing load balancing on existing links. During this phase, the FTS queue is monitored at 1-minute intervals until the on-going data transfer concludes and congestion ends. Subsequently, a transition to the stopped state occurs if the CERN NMS saturation alarm has been cleared. Otherwise, a return to the monitoring state takes place if the alarm persists, anticipating the potential new data transfers.

- Stopped State: Serving as the final phase before concluding NOTED execution, this state is accessed when there are no on-going data transfers in FTS or when the network usage descends below the predefined threshold value, coupled with the resolution of the link saturation alarm in CERN NMS.

5. Identification process of WLCG federations

The identification of WLCG federations triggered by congestion primarily relies on the routing forwarding tables of the LHCOPN and LHCONE border routers. This process encompasses three key stages:

1. Network Monitoring and Alarm Polling: CERN NMS generates diverse alarm types based on network events, encompassing categories such as BGP and DNS. However, for this particular application, the focus is on IN/OUT LOAD THRESHOLD EXCEEDED alarms. Consequently, alarms are polled, and the data is organised within a DataFrame structure. It includes details related to the device, alarm type, interface, severity, and the timestamp at which the event was generated.

```
   Start            Alarm             Device        Severity  Interface
 1710092333    IN LOAD EXCEEDED  BORDER-ROUTER-1     MINOR     irb.3530
 1710082439   OUT LOAD EXCEEDED  BORDER-ROUTER-2     MINOR     irb.3529
```

2. Border Router Forwarding Table:

   - Identifying the IP of the Next-Hop Router: The subsequent stage entails identifying the prefixes routed through the alarmed interface. Given the existing record about the alarmed interface and its associated device, the JUNOS API [15] is used in this process to establishes a connection to the border router, enabling access to the interface description and facilitating the retrieval of IP prefixes associated with that interface.

```
BORDER-ROUTER-1> 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
```

- Identifying the Network Prefixes: The subsequent step involves obtaining the network prefixes associated with the alarmed interface. This is accomplished by once again using the JUNOS API to connect to the border router and access the description of the next-hop route.

```
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
```

3. Retrieve the WLCG Federation: Despite having the IP of the next-hop router and the associated network prefixes, discerning the specific federation to which the alarmed interface belongs remains unknown, as the network is agnostic towards managing entities and thus we depend on the assurance offered by the CRIC database. This database holds extensive records of WLCG-affiliated sites, encompassing information such as network prefixes, federation, country, virtual organisation, AS number, and other network parameters. Consequently, a search within CRIC is conducted based on the network prefixes of the alarmed interface to ascertain the corresponding WLCG site.

6. Results

This section details the results of NOTED in two different perspectives and configurations, the first strategy providing dynamic circuits through an SDN provider and the second performing load balancing on existing links.

6.1 NOTED at SC23: dynamic circuits

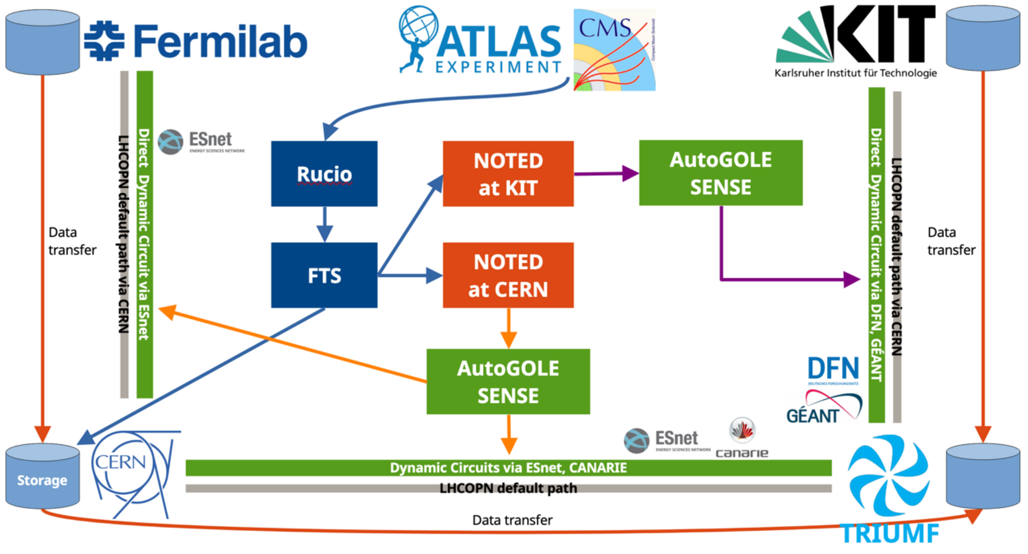
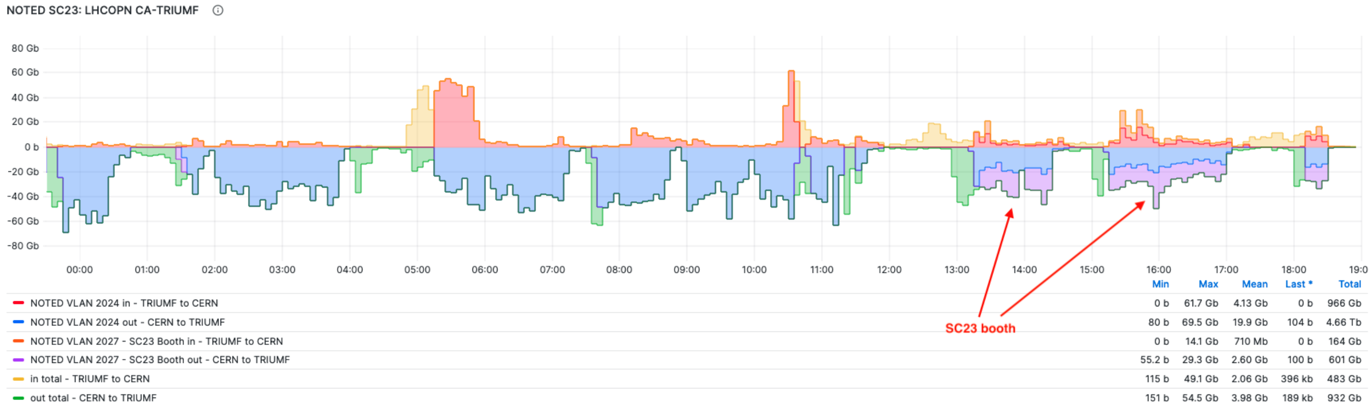

Fermilab
ATLAS EXPERIMENT
CMS
KIT
Karlsruher Institut für Technologie
ESnet
Direct Dynamic Circuit via ESnet
LHCOPN default path via CERN
Data transfer
Rucio
FTS
NOTED at KIT
NOTED at CERN
AutoGOLE SENSE
AutoGOLE SENSE
Direct Dynamic Circuit via DFN, GÉANT
LHCOPN default path via CERN
Data transfer
DFN
GÉANT
ESnet
canarie
Storage
CERN
Dynamic Circuits via ESnet, CANARIE
LHCOPN default path
Data transfer
TRIUMF

Figure 3: Diagram of the testbed for NOTED demonstration at SC23.

NOTED demonstrated a major success in tests conducted at the International Conference for High-Performance Computing, Networking, Storage, and Analysis, widely known as Super

Computing (SC) in 2023. This event took place during the week of 12th - 17th November 2023 in Denver (Colorado, United States of America).

Figure 3 illustrates the testbed configuration employed for the NOTED demonstration at SC23. During this demonstration, the ATLAS experiment orchestrated large data transfers between CH-CERN and CA-TRIUMF, as well as DE-KIT and CA-TRIUMF. Simultaneously, the CMS experiment managed large data transfers between FermiLab and CH-CERN. The diagram shows the parallel and independent execution of two instances of NOTED running at CH-CERN and DE-KIT. Furthermore, in response to network congestion detected by NOTED, three dynamic circuits were provided through the SENSE service, an SDN provider. These circuits allowed the rerouting of traffic from the default LHCOPN route to the new dynamic circuit, thereby dynamically reconfiguring the network architecture.



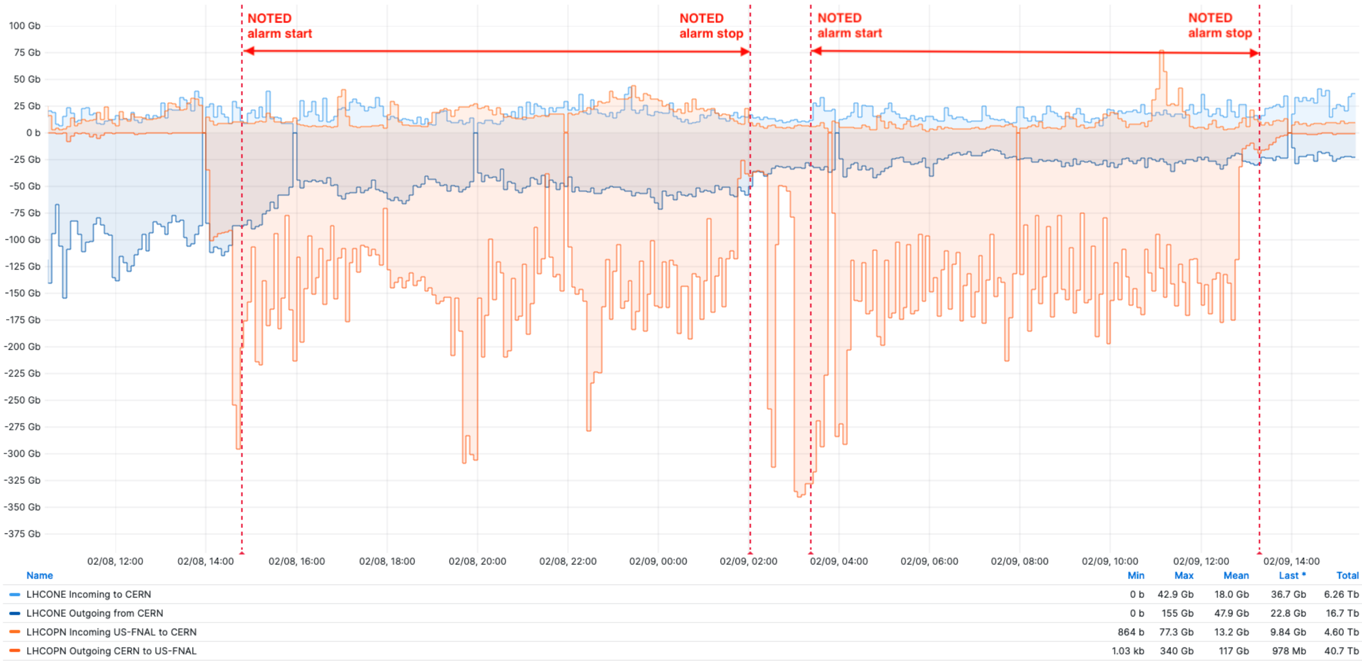
	Min	Max	Mean	Last *	Total
🟥 NOTED VLAN 2024 in - TRIUMF to CERN	0 b	61.7 Gb	4.13 Gb	0 b	966 Gb
🟦 NOTED VLAN 2024 out - CERN to TRIUMF	80 b	69.5 Gb	19.9 Gb	104 b	4.66 Tb
🟧 NOTED VLAN 2027 - SC23 Booth in - TRIUMF to CERN	0 b	14.1 Gb	710 Mb	0 b	164 Gb
🟪 NOTED VLAN 2027 - SC23 Booth out - CERN to TRIUMF	55.2 b	29.3 Gb	2.60 Gb	100 b	601 Gb
🟨 in total - TRIUMF to CERN	115 b	49.1 Gb	2.06 Gb	396 kb	483 Gb
🟩 out total - CERN to TRIUMF	151 b	54.5 Gb	3.98 Gb	189 kb	932 Gb

Figure 4: Network usage and data tranfers between CH-CERN and CA-TRIUMF at SC23.

Figure 4 shows the outcomes derived from the NOTED demonstration at SC23, presenting the network usage and data transfers instigated by the ATLAS experiment. Specifically, the graph showcases the activity on the CH-CERN link to CA-TRIUMF on the 14th November 2023. In the graph, the yellow and green lines represent the traffic associated with the default LHCOPN link, while the red and blue lines mean the traffic redirected through the dynamic circuit. NOTED triggered this redirection when it estimated network congestion on the given link. At this point, it is noteworthy to mention that two dynamic circuits were established for this particular test: the first constituting a direct circuit between CH-CERN and TRIUMF, tagged with VLAN 2024, and the second routed through the SC23 booth, tagged with VLAN 2027, intentionally configured to route the traffic through the SC23 booth at conference in Denver, serving as a demonstrative measure.

6.2 NOTED at WCLG DC24 pre-testing: load balancing

In the pre-tests of the WLCG Data Challenge 2024 (DC24), conducted during the week of 6th - 9th February 2024, NOTED operated in dry-run mode for the LHCOPN and LHCONE links. It generated alarms suggesting the implementation of load balancing for a specific link with other existing links.



Name	Min	Max	Mean	Last *	Total
LHCONE Incoming to CERN	0 b	42.9 Gb	18.0 Gb	36.7 Gb	6.26 Tb
LHCONE Outgoing from CERN	0 b	155 Gb	47.9 Gb	22.8 Gb	16.7 Tb
LHCOPN Incoming US-FNAL to CERN	864 b	77.3 Gb	13.2 Gb	9.84 Gb	4.60 Tb
LHCOPN Outgoing CERN to US-FNAL	1.03 kb	340 Gb	117 Gb	978 Mb	40.7 Tb

Figure 5: Network usage and data transfers for US-FNAL Tier 1 at WLCG DC24 pre-testing.

Figure 5 presents the outcomes of the NOTED demonstration at WLCG DC24, showcasing network usage and data transfers for the US-FNAL Tier 1 link on 8th February 2024. The graph illustrates the traffic corresponding to the LHCONE network in blue and the traffic of the LHCOPN network in orange. Observing the graph, a remarkable surge in network usage is evident at 14:00h on 8th February, coinciding with the initiation of a large data transfer. Shortly thereafter, NOTED raises an alarm, suggesting load balancing on this link since it estimated network congestion for a long period of time and thus suggesting the necessity to reconfigure the network and enhance the capacity of the concerned link. Subsequently, after several hours, around 2:00h on the 9th February, the large data transfer concludes, prompting NOTED to suggest discontinuing the load balancing as the congestion is not anticipated to persist.

7. Conclusions

This paper introduces the new implemented features in NOTED, an intelligent network controller driven by congestion to improve the throughput of large data transfers in FTS. The operational modes, execution states, the identification process of WLCG federation, and the results of the extensive testing derived from the SC23 conference and WLCG DC24 pre-tests.

In conclusion, NOTED has demonstrated its capability to detect congestion and dynamically reconfigure the network under various network configurations. This includes the provision of dynamic circuits facilitated by an SDN provider like SENSE and the execution of load balancing among existing links. As part of future work, our focus will be on implementing traffic forecasting by using machine learning to enhance decision-making processes when proposing network actions to optimise on-going large data transfers.

References

[1] 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*, 2020.

[2] 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*, 2021. 10.1051/epjconf/202125102049.

[3] J. Waczynska, E. Martelli, S. Vallecorsa, E. Karavakis and T. Cass, *Convolutional lstm models to estimate network traffic*, 08, 2021. 10.1051/epjconf/202125102050.

[4] 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*, 2023.

[5] I. Monga, C. Guok, J. MacAuley, A. Sim, H. Newman, J. Balcas et al., *Software-defined network for end-to-end networked science at the exascale*, 2020. https://doi.org/10.1016/j.future.2020.04.018.

[6] J. Guiang, A. Arora, D. Davila, J. Graham, D. Mishin, I. Sfiligoi et al., *Integrating end-to-end exascale sdn into the lhc data distribution cyberinfrastructure*, 2022. 10.1145/3491418.3535134.

[7] T. Lehman, X. Yang, C. Guok, F. Wuerthwein, I. Sfiligoi, J. Graham et al., *Data transfer and network services management for domain science workflows*, 2022.

[8] Karavakis, Edward, Manzi, Andrea, Arsuaga Rios, Maria, Keeble, Oliver, Garcia Cabot, Carles, Simon, Michal et al., *Fts improvements for lhc run-3 and beyond*, 2020. 10.1051/epjconf/202024504016.

[9] Anisenkov, Alexey, Andreeva, Julia, Di Girolamo, Alessandro, Paparrigopoulos, Panos and Vasilev, Boris, *Cric: Computing resource information catalogue as a unified topology system for a large scale, heterogeneous and dynamic computing infrastructure*, 2020. 10.1051/epjconf/202024503032.

[10] Anisenkov, Alexey, Andreeva, Julia, Di Girolamo, Alessandro, Paparrigopoulos, Panos and Vedaee, Aresh, *Cric: a unified information system for wlcg and beyond*, 2019. 10.1051/epjconf/201921403003.

[11] M. Alandes, J. Andreeva, A. Anisenkov, G. Bagliesi, S. Belforte, S. Campana et al., *Consolidating wlcg topology and configuration in the computing resource information catalogue*, oct, 2017. 10.1088/1742-6596/898/9/092042.

[12] CERN, "Fts optimiser documentation." https://fts3-docs.web.cern.ch/fts3-docs/docs/optimizer/optimizer.html.

[13] E. Martelli and S. Stancu, *Lhcopn and lhcone: Status and future evolution*, dec, 2015. 10.1088/1742-6596/664/5/052025.

[14] E. Martelli, *Evolving the lhcopn and lhcone networks to support hl-lhc computing requirements*, 2023.

[15] Juniper, "Junos api: Pyez." https://junos-pyez.readthedocs.io/en/2.6.4/.