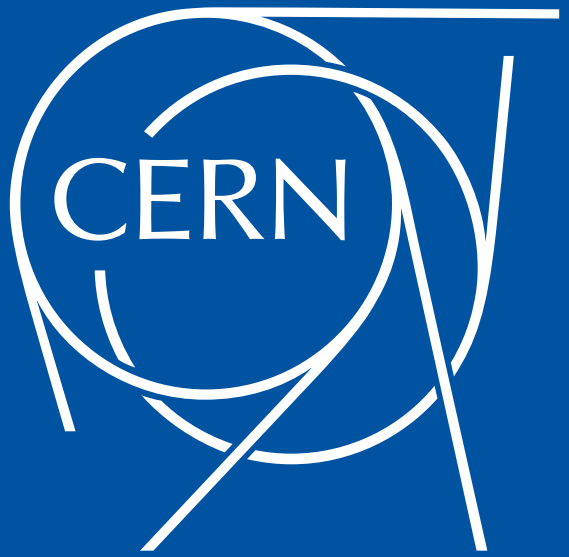

CERN

NOTED: Network Optimized Transfer of Experimental Data

CERN Data Center
IT-CS-NE Department

Carmen Misa Moreira
Edoardo Martelli

Outline

History - Motivation

From an idea discussed at the LHCONE meeting #37 at BNL in 2017.
Project started by CERN in 2018, funded by WLCG.
Problems:

- Large data transfers can saturate network links.
- Routing protocols don't take into account the link utilization.
- Alternative paths may be left idle.

Traffic engineering challenges:

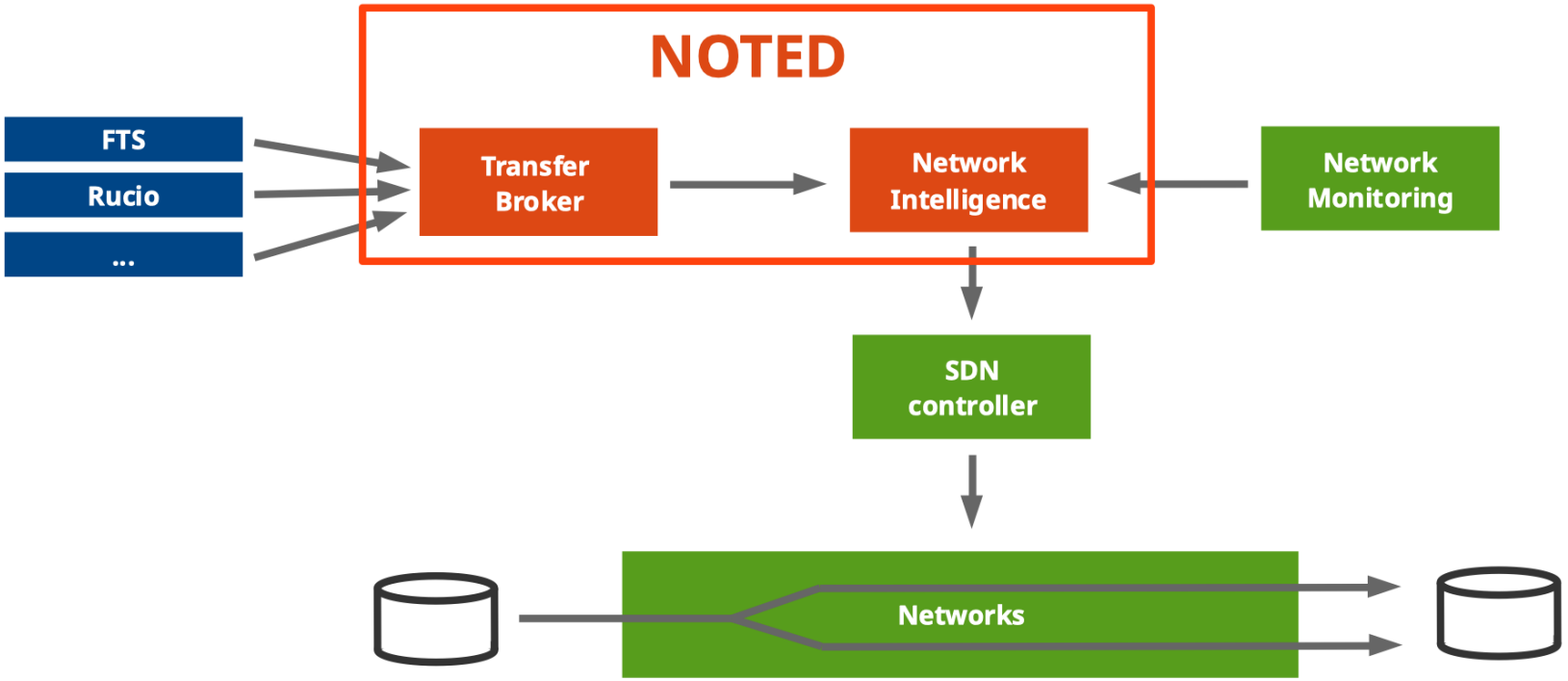
- How long a large transfer will last? Is it worth to change the routing?
- When a corrective action should be taken?

NOTED approach: predict the behaviour of the applications.

- Make effective use of bandwidth, especially on the expensive transoceanic links and reduce idle periods.
- Some links get congested, while parallel ones run low $\rightarrow$ optimized data transfers of LHC to detect link saturation and execute an action.

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.

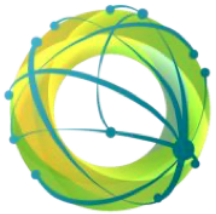
Elements

FTS (File Transfer Service):

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

CRIC (Computing Resource Information Catalog):

- Enriched from CRIC database to get an overview and knowledge of the network topology.
- Rcsite query → get IPv4/IPv6 addresses.
- Service query → get endpoints, rcsite and federation.



FTS
File Transfer Service

CRIC
Computing Resource Information Catalog

elasticsearch

Interaction with FTS

query monit_prod_fts_raw_queue* $\rightarrow \sim$ 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",
},
```

Interaction with CRIC

query service


```
"FZK-LCG2-CE-HTCONDOR-CE-htcondor-ce-1-kit.gridka.de": {
    "arch": "",
    "country": "Germany",
    "country_code": "DE",
    "endpoint": "htcondor-ce-1-kit.gridka.de:9619",
    "federation": "DE-KIT",
    "flavour": "HTCONDOR-CE",
    "is_ipv6": false,
    "is_monitored": true,
    "name": "FZK-LCG2-CE-HTCONDOR-CE-htcondor-ce-1-kit.gridka.de",
    "rcsite": "FZK-LCG2",
    "rcsite_state": "ACTIVE",
    "resources": {
      "condor": {
        "id": 1318,
        "name": "condor",
        "state": "ACTIVE",
        "status": "production",
        "usage": {
          "lhcb": [
            {
              "is_monitored": true,
              "site": "LCG.GRIDKA.de",
              "status": ""
            }
          ]
        }
      }
    },
```

```
"usage": {
  "alice": [
    {
      "is_monitored": true,
      "site": "FZK",
      "status": ""
    }
  ],
  "atlas": [
    {
      "is_monitored": true,
      "site": "FZK-LCG2",
      "status": "production"
    }
  ],
  "cms": [
    {
      "is_monitored": true,
      "site": "T1_DE_KIT",
      "status": ""
    }
  ],
  "lhcb": [
    {
      "is_monitored": true,
      "site": "LCG.GRIDKA.de",
      "status": "production"
    }
  ]
```

Dataset structure and workflow

Configuration given by the user $\rightarrow$ a list of {src_rcsite, dst_rcsite} pairs.

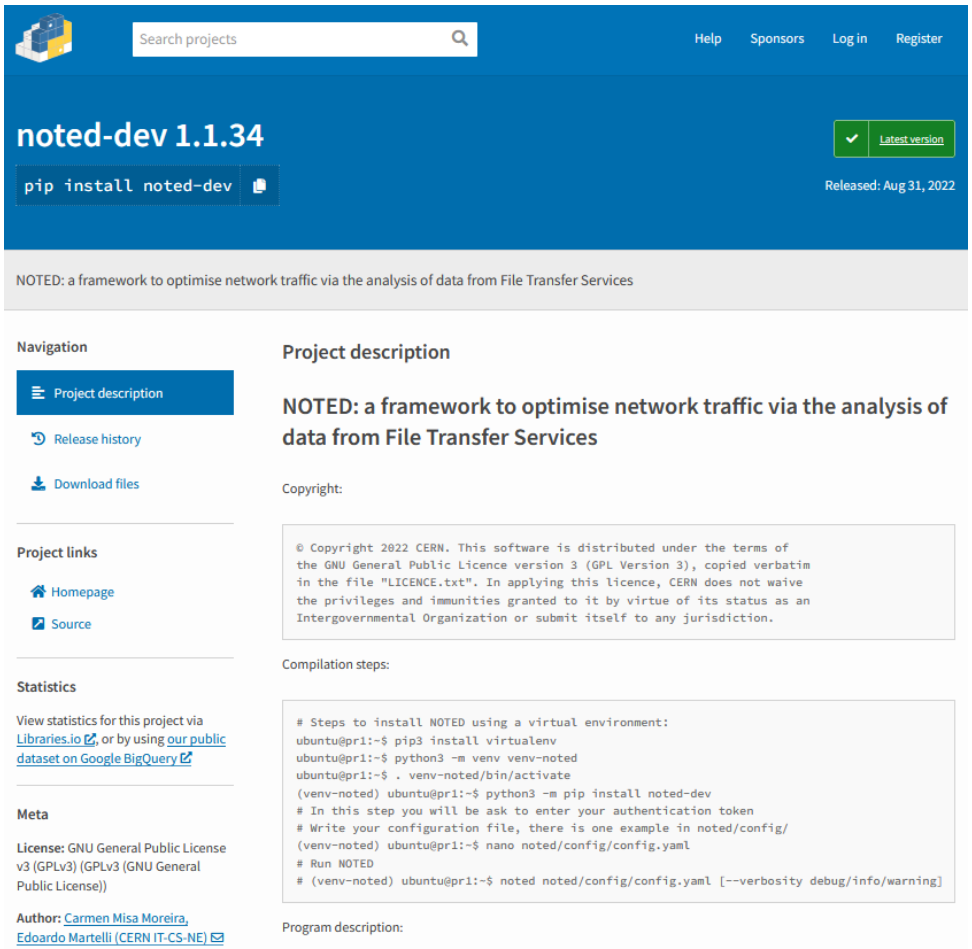
1. Enrich NOTED with the topology of the network:
   - Query CRIC database $\rightarrow$ get the endpoints (α_i, β_i) that could be involved in the transfers for the given {src_rcsite, dst_rcsite} pairs.
2. Analyse on-going and upcoming data transfers:
   - Query FTS recursively $\rightarrow$ get the 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 $\rightarrow$ provides a dynamic circuit via Sense/AutoGOLE.

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

Status of the project and tests

Package distribution and installation

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



Common steps:

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

Ubuntu installation:

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

CentOS installation:

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

Configuration file

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

```
noted -h

positional arguments:
  config_file the name of the configuration file [config-example.yaml]

optional arguments:
  -h, --help show this help message and exit
  -v VERBOSITY, --verbosity VERBOSITY defines logging level [debug, info, warning]
```

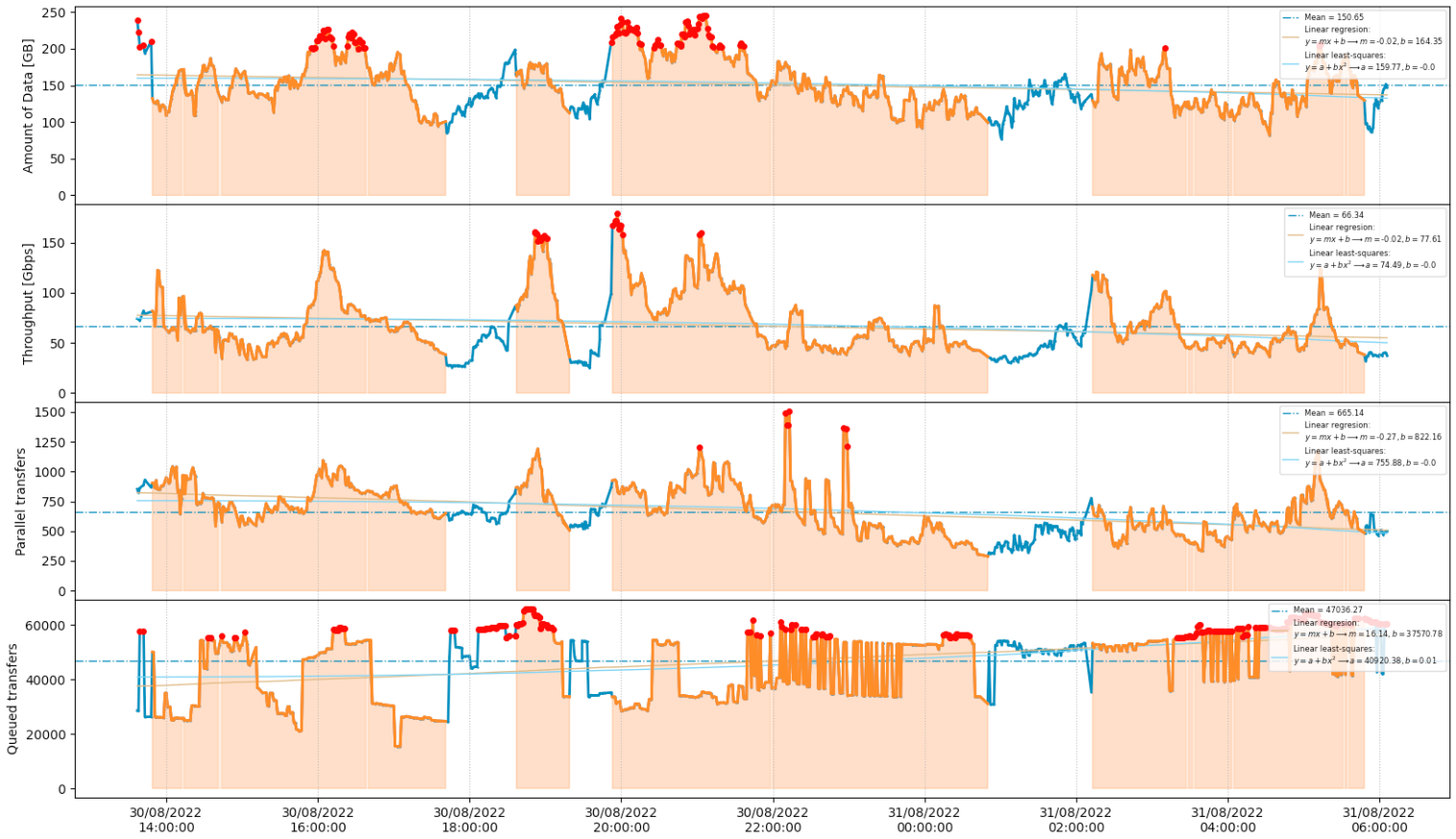
- Example of config.yaml:

```
src_rcsite: ['rc_site_1', 'rc_site_2', 'rc_site_3', 'rc_site_4'] # Source RC_Sites
dst_rcsite: ['rc_site_1', 'rc_site_2', 'rc_site_3', 'rc_site_4'] # Destination RC_Sites
events_to_wait_until_notification: 5 # Events to wait until email notification
max_throughput_threshold_link: 80 # If throughput > max_throughput -> START
min_throughput_threshold_link: 20 # If throughput < min_throughput -> STOP
unidirectional_link: False # If False both TX and RX paths will be monitoring
number_of_dynamic_circuits: 2 # Number of dynamic circuits
sense_uuid: 'sense_uuid_1' # Sense-o UUID dynamic circuit
sense_vlan: 'vlan_description_1' # VLAN description
sense_uuid_2: 'sense_uuid_2' # Sense-o UUID dynamic circuit
sense_vlan_2: 'vlan_description_2' # VLAN description
from_email_address: 'email_1' # From email address
to_email_address: 'email_1, email_2' # To email address
subject_email: 'subject' # Subject of the email
message_email: "message" # Custom message
auth_token: auth_token # Authenticantion token
```

Transfers of WLCG sites in LHCONE

Test carried out on the 31st of August 2022:

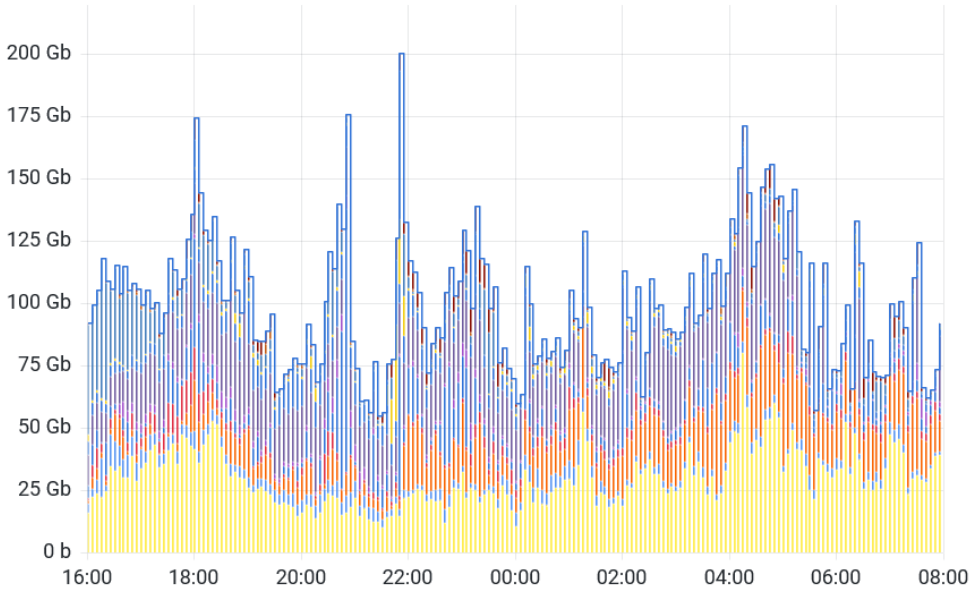
- Raise an alert if throughput $>$ 80 GB/s $\rightarrow$ dynamic circuit is provided.
- Raise an alert if throughput $<$ 40 GB/s $\rightarrow$ dynamic circuit is cancelled.



Transfers of WLCG sites in LHCONE

Test carried out on the 31st of August 2022.
Results:

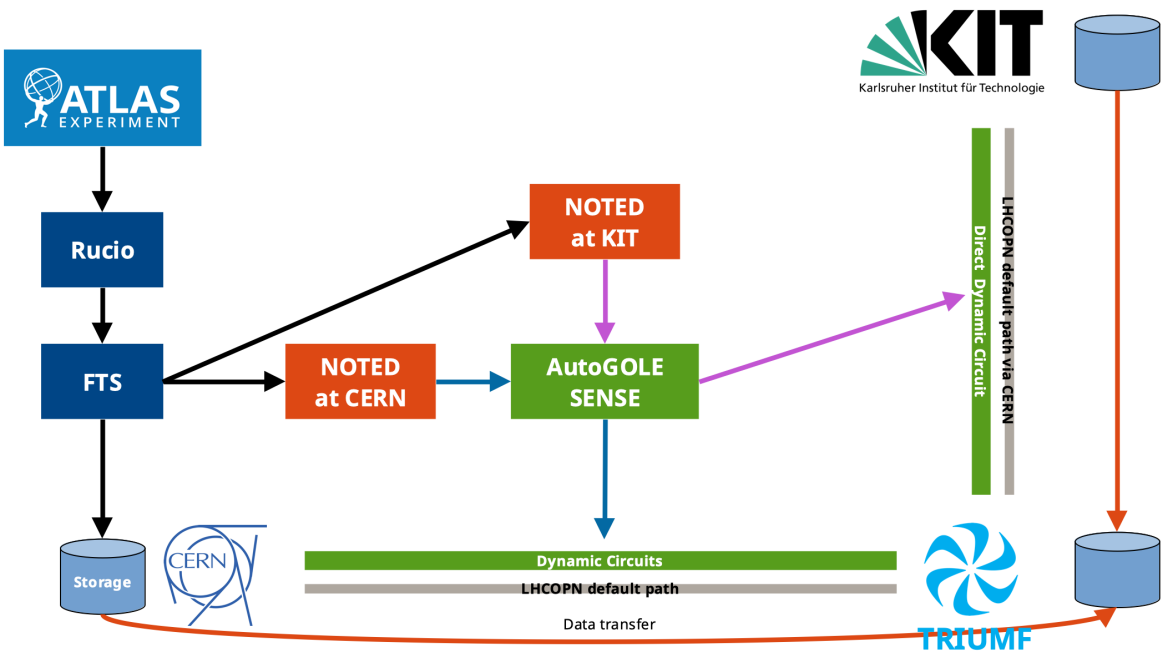
- Observations of NOTED about the network utilization correspond with the reported ones in Grafana by LHCONE/LHCOPN production routers.
- Therefore, by inspecting FTS data transfers is possible to get an understanding of the network usage and improve its performance by executing an action in the topology of the network.



	Mean	Last *	Max	Total
Outgoing DE-KIT	30.0 Gb	39.4 Gb	80.4 Gb	5.76 Tb
Outgoing KR-KISTI	42.6 Mb	10.3 Mb	328 Mb	8.18 Gb
Outgoing RU-T1	4.18 Gb	1.52 Gb	15.6 Gb	802 Gb
Outgoing FR-IN2P3	13.9 Gb	11.9 Gb	40.5 Gb	2.68 Tb
Outgoing NDGF	3.52 Gb	3.07 Gb	14.0 Gb	675 Gb
Outgoing NL-T1	8.34 Gb	2.80 Gb	106 Gb	1.60 Tb
Outgoing TW-ASGC	2.07 Gb	1.87 Gb	7.91 Gb	397 Gb
Outgoing IT-INFN-CNAF	20.2 Gb	26.6 Gb	45.3 Gb	3.88 Tb
Outgoing UK-RAL	5.13 Gb	2.80 Gb	17.5 Gb	985 Gb
Outgoing CA-TRIUMF	1.54 Gb	841 Mb	38.8 Gb	296 Gb
Outgoing US-BNL	7.43 Gb	729 Mb	59.5 Gb	1.43 Tb
Outgoing US-FNAL	503 Mb	445 Mb	5.25 Gb	96.6 Gb
Outgoing ES-PIC	2.82 Gb	347 Mb	9.81 Gb	542 Gb
Total	99.7 Gb	92.3 Gb	201 Gb	19.1 Tb

Supercomputing 2022 demo

Components and WLCG sites involved in SC22



1. NOTED looks in FTS for large data transfers.
2. When 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.

SENSE provisioning system

SENSE (SDN for E2E Networked Science at the Exascale): provision system that dynamically builds end-to-end virtual guaranteed networks across administrative domains without manual intervention.

- Provisioning automation: bring-up and management of services without human involvement.
- Multi-domain: multiple administrative domains, independent policies and AUP (Acceptable Use Policy).
- Resource orchestration: allocation and reservation of resources including compute, storage and network.
- End-to-end: DTN NIC to DTN NIC, across Science DMZ (Demilitarized zone), WANs, Open exchange points...

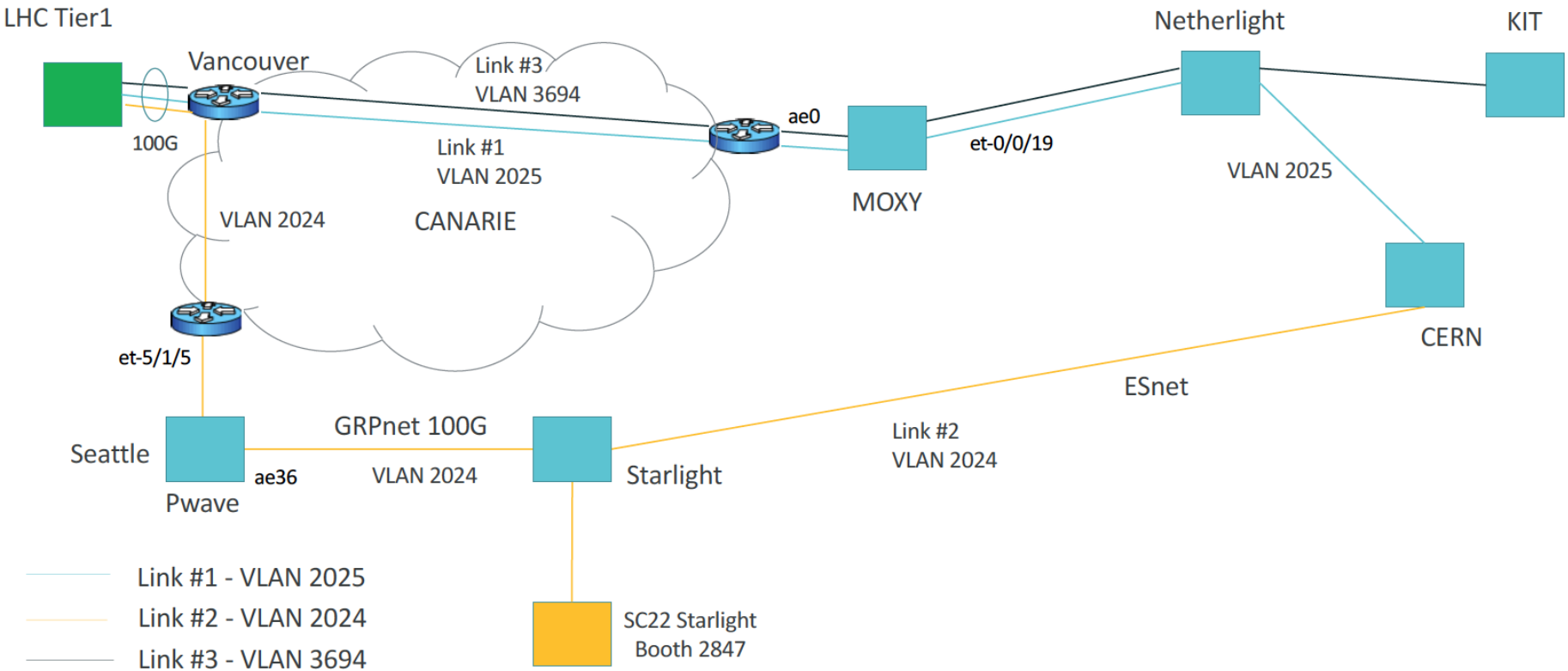
Components and participants

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:

SC22 demo logical connections

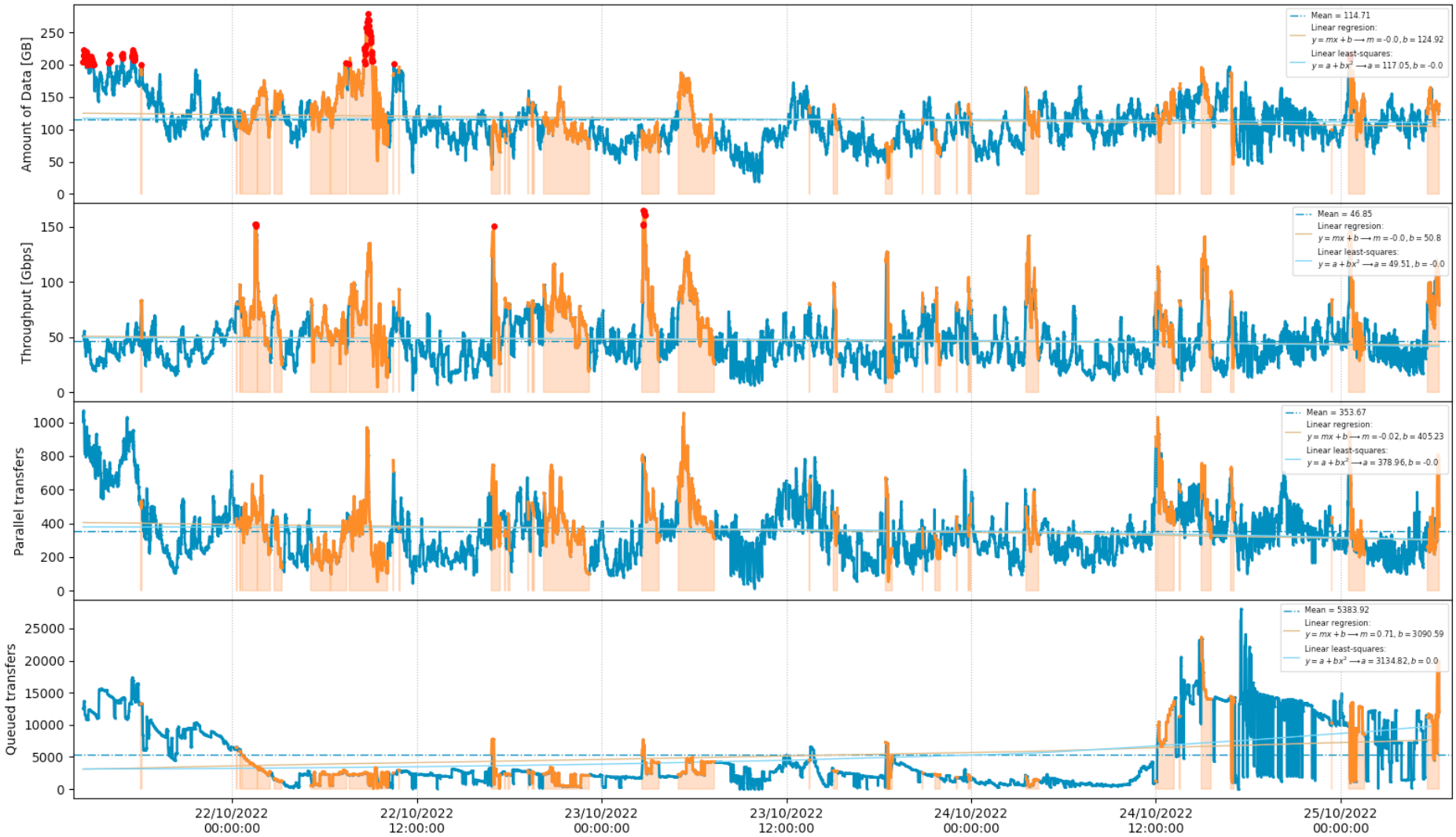


Future lines

Future lines

- Improve decision-making to raise an alert:
  - Improve provision and cancellation of dynamic circuits.
- Traffic forecasting:
  - LSTM (Long Short-Term Memory) recurrent neural network.
  - Predict the duration of large data transfers.
- Network monitoring integration.
- FTS integration.

Observations of NOTED during LHCONE Meeting #49



About yesterday night...

"I know a great IPv6 joke but I don't think you're ready for it".

Thanks for your attention!

NOTED: Network Optimized Transfer of Experimental Data

CERN Data Center
IT-CS-NE Department

Carmen Misa Moreira
Edoardo Martelli


CERN