Identifying and Understanding Scientific Network Flows

McKee, Shawn (Presenter, University of Michigan Physics), Babik, Marian (CERN), Chown, Tim (Jisc), Hanushevsky, Andrew (SLAC National Accelerator Laboratory), Sullivan, Tristan (University of Victoria), Hoeft, Bruno (Karlsruhe Institute of Technology (KIT)), Letts, James (University of California, San Diego), Carder, Dale (ESnet), Attebury, Garhan (University of Nebraska), Lambert, Michael (Pittsburgh Supercomputing Center), Newell, Karl (Internet2)

CHEP 2023, Norfolk, VA, USA

https://indico.jlab.org/event/459/

May 8, 2023

Presentation Overview

For High-Energy Physics (HEP), we have identified a need to better understand and optimize our network traffic to ensure we are using the network as effectively (for our science) as possible.

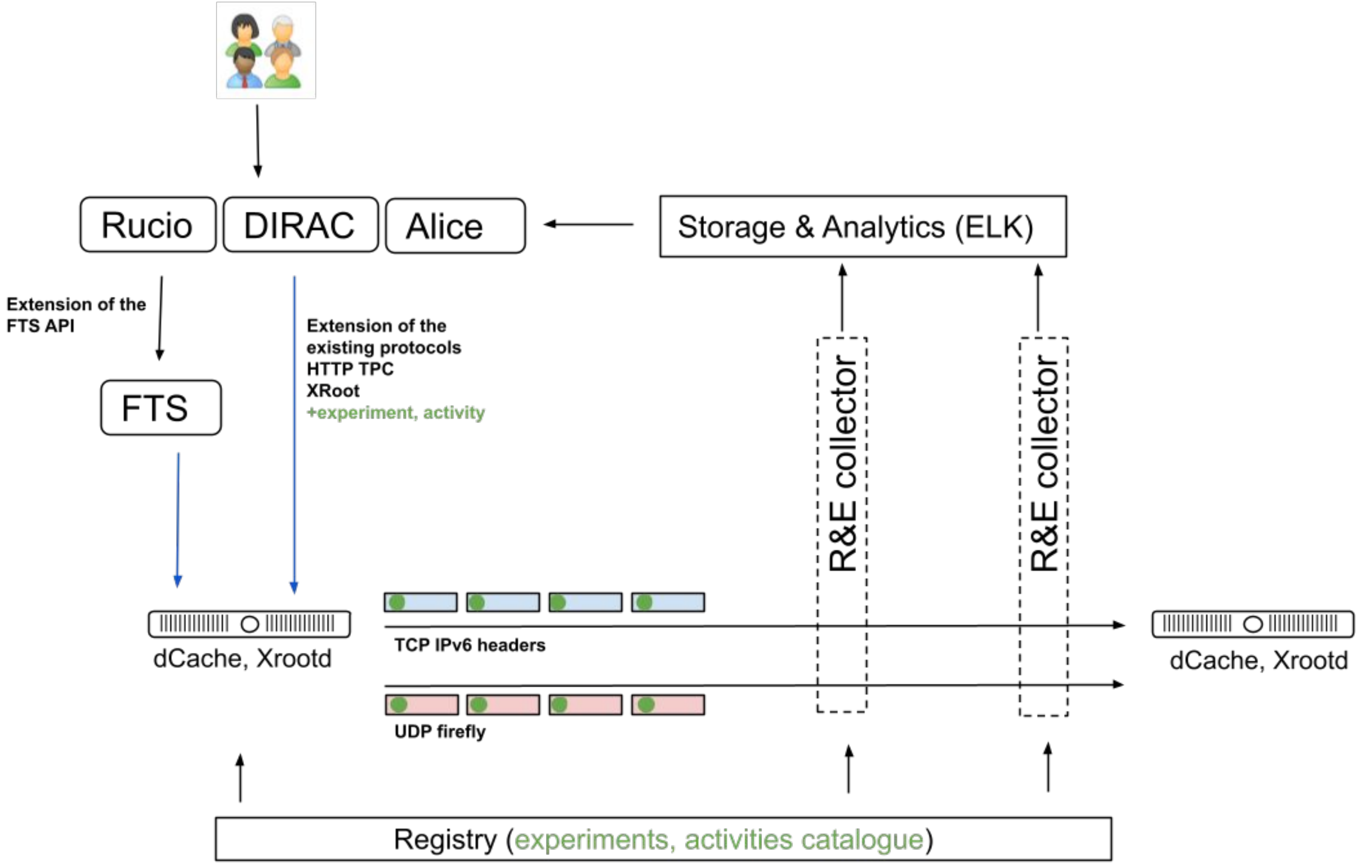
One of the challenges we have faced is being able to understand and identify the source of our traffic within the Research and Education (R&E) networks, especially when **critical links** are **overloaded**, impacting our workflows and data transfers.

This presentation will cover the ongoing work to understand and identify our scientific network flows.

Network Visibility and Scitags

- Scientific Network Tags (scitags) is an initiative promoting identification of the science domains and their high-level activities at the network level.
- Enable tracking and correlation of our transfers with Research and Education Network Providers (R&Es) network flow monitoring
  - Utilizing packet and flow marking to identify traffic owner/purpose.
- **Experiments** can better understand how their network flows perform along the path
  - Improve visibility into how network flows perform (per activity) within R&E segments
  - Get insights into how experiments are using the networks, get additional data from R&Es on behaviour of our transfers (traffic, paths, etc.)
- Sites could get visibility into how different network flows perform
  - We can add network flow stats to flow marking (augmenting experiment/activity info)

How scitags work



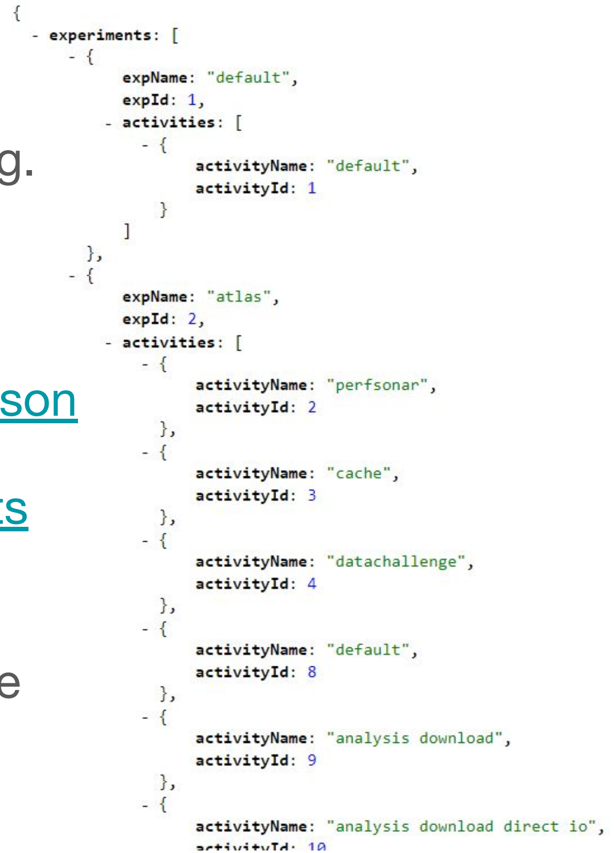
Registry

We have standardized the “experiment” and “activity” fields we use for both flow labeling and packet marking.

The scitags.org domain provides an API that can be consulted to get the standard values: https://api.scitags.org or https://www.scitags.org/api.json

The underlying source of truth is a set of Google sheets that are maintained and writeable by a few stewards.

Note: the API provides the defined values **but** how the values are used in packet marking are specified in our Google sheets (bit location in IPv6 flow label)



```
{
  - experiments: [
      - {
            expName: "default",
            expId: 1,
          - activities: [
              - {
                    activityName: "default",
                    activityId: 1
                }
            ]
        },
      - {
            expName: "atlas",
            expId: 2,
          - activities: [
              - {
                    activityName: "perfsonar",
                    activityId: 2
                },
              - {
                    activityName: "cache",
                    activityId: 3
                },
              - {
                    activityName: "datachallenge",
                    activityId: 4
                },
              - {
                    activityName: "default",
                    activityId: 8
                },
              - {
                    activityName: "analysis download",
                    activityId: 9
                },
              - {
                    activityName: "analysis download direct io",
                    activityId: 10
```

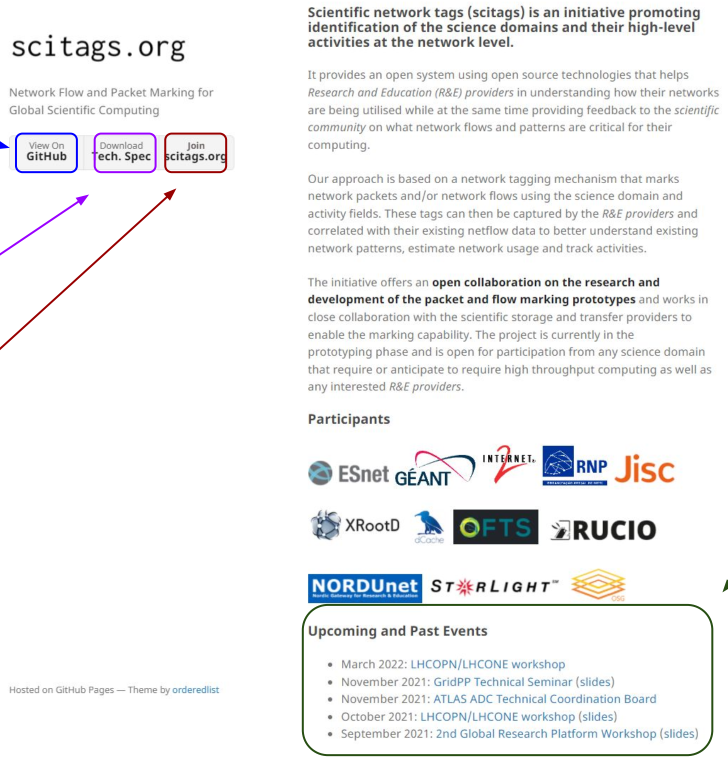
Finding More Information: https://scitags.org

Code

Technical Spec

Mailing List

Presentations



scitags.org

Network Flow and Packet Marking for Global Scientific Computing

Hosted on GitHub Pages — Theme by orderedlist

Scientific network tags (scitags) is an initiative promoting identification of the science domains and their high-level activities at the network level.

It provides an open system using open source technologies that helps *Research and Education (R&E) providers* in understanding how their networks are being utilised while at the same time providing feedback to the *scientific community* on what network flows and patterns are critical for their computing.

Our approach is based on a network tagging mechanism that marks network packets and/or network flows using the science domain and activity fields. These tags can then be captured by the *R&E providers* and correlated with their existing netflow data to better understand existing network patterns, estimate network usage and track activities.

The initiative offers an **open collaboration on the research and development of the packet and flow marking prototypes** and works in close collaboration with the scientific storage and transfer providers to enable the marking capability. The project is currently in the prototyping phase and is open for participation from any science domain that require or anticipate to require high throughput computing as well as any interested *R&E providers*.

Participants

ESnet GÉANT INTERNET2 RNP Jisc

XRootD dCache FTS RUCIO

NORDUnet STARLIGHT OSG

Upcoming and Past Events

- March 2022: LHCOPN/LHCONE workshop
- November 2021: GridPP Technical Seminar (slides)
- November 2021: ATLAS ADC Technical Coordination Board
- October 2021: LHCOPN/LHCONE workshop (slides)
- September 2021: 2nd Global Research Platform Workshop (slides)

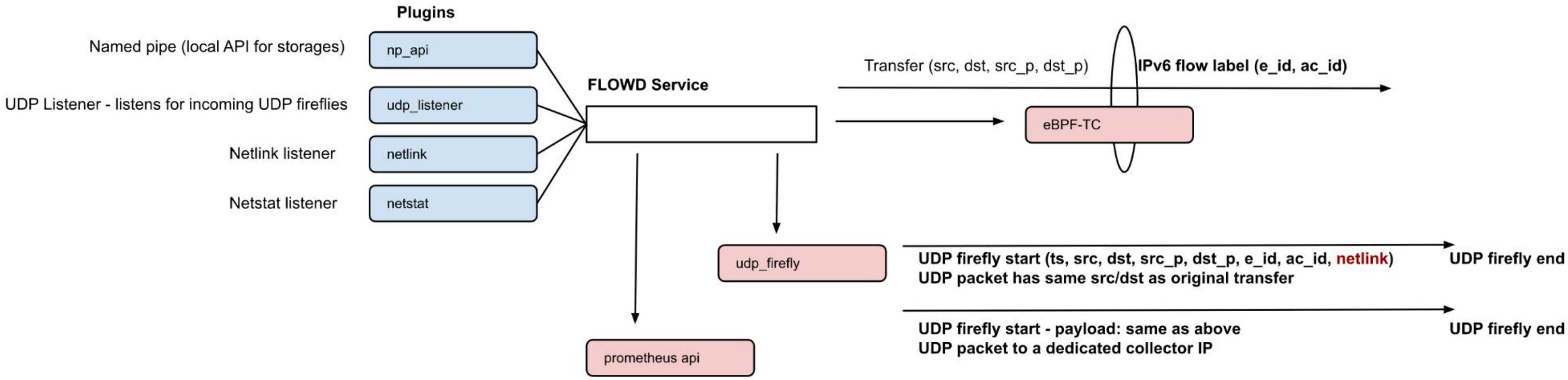
Technical Spec for Packet Marking/Flow Labeling

The detailed technical specifications are maintained on a Google doc

- The spec covers both **Flow Labeling** via UDP Fireflies and **Packet Marking** via the use of the IPv6 Flow Label.
    - **Fireflies** are UDP packets in Syslog format with a defined, versioned JSON schema.
        - Packets are intended to be sent to the same destination (port 10514) as the flow they are labeling and these packets are intended to be world readable.
        - Packets can also be sent to specific regional or global collectors.
        - Use of syslog format makes it easy to send to Logstash or similar receivers.
    - **Packet marking** is intended to use the 20 bit flow label field in IPv6 packets.
        - To meet the spirit of RFC6437, we use 5 of the bits for entropy, 6 for activity and 9 for owner/experiment.
- The document also covers methods for communicating owner/activity and other services and frameworks that may be needed for implementation.

End-system Utility: Flowd Service

- Flow and Packet Marking service developed in Python

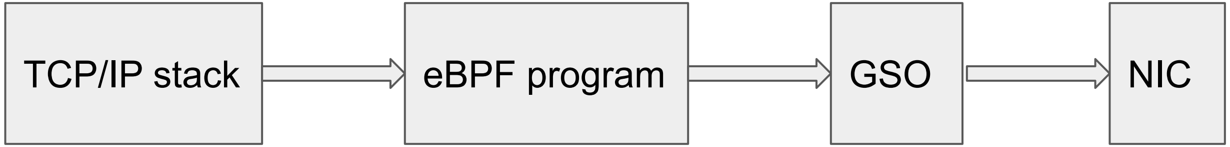

Plugins
Named pipe (local API for storages)
np_api
UDP Listener - listens for incoming UDP fireflies
udp_listener
Netlink listener
netlink
Netstat listener
netstat
FLOWD Service
Transfer (src, dst, src_p, dst_p)
IPv6 flow label (e_id, ac_id)
eBPF-TC
udp_firefly
UDP firefly start (ts, src, dst, src_p, dst_p, e_id, ac_id, netlink)
UDP packet has same src/dst as original transfer
UDP firefly end
UDP firefly start - payload: same as above
UDP packet to a dedicated collector IP
UDP firefly end
prometheus api

- Plugins provide different ways get connections to mark (or interact with storage)
  - New plugins were added to support netlink readout and UDP firefly consumer
- Backends are used to implement flow and/or packet marking
  - New backends were added to mark packets (via eBPF-TC) and expose monitored connection to Prometheus

Flowd: Packet Marking via eBPF-TC Backend

- eBPF is a general-purpose RISC instruction set that runs on an in-kernel VM; programs can be written in restricted C and compiled into bytecode that is injected into the kernel (after verification)
- Can sometimes replace kernel modules
- eBPF-TC programs run whenever the kernel receives (ingress) or sends (egress) a packet

Egress path:

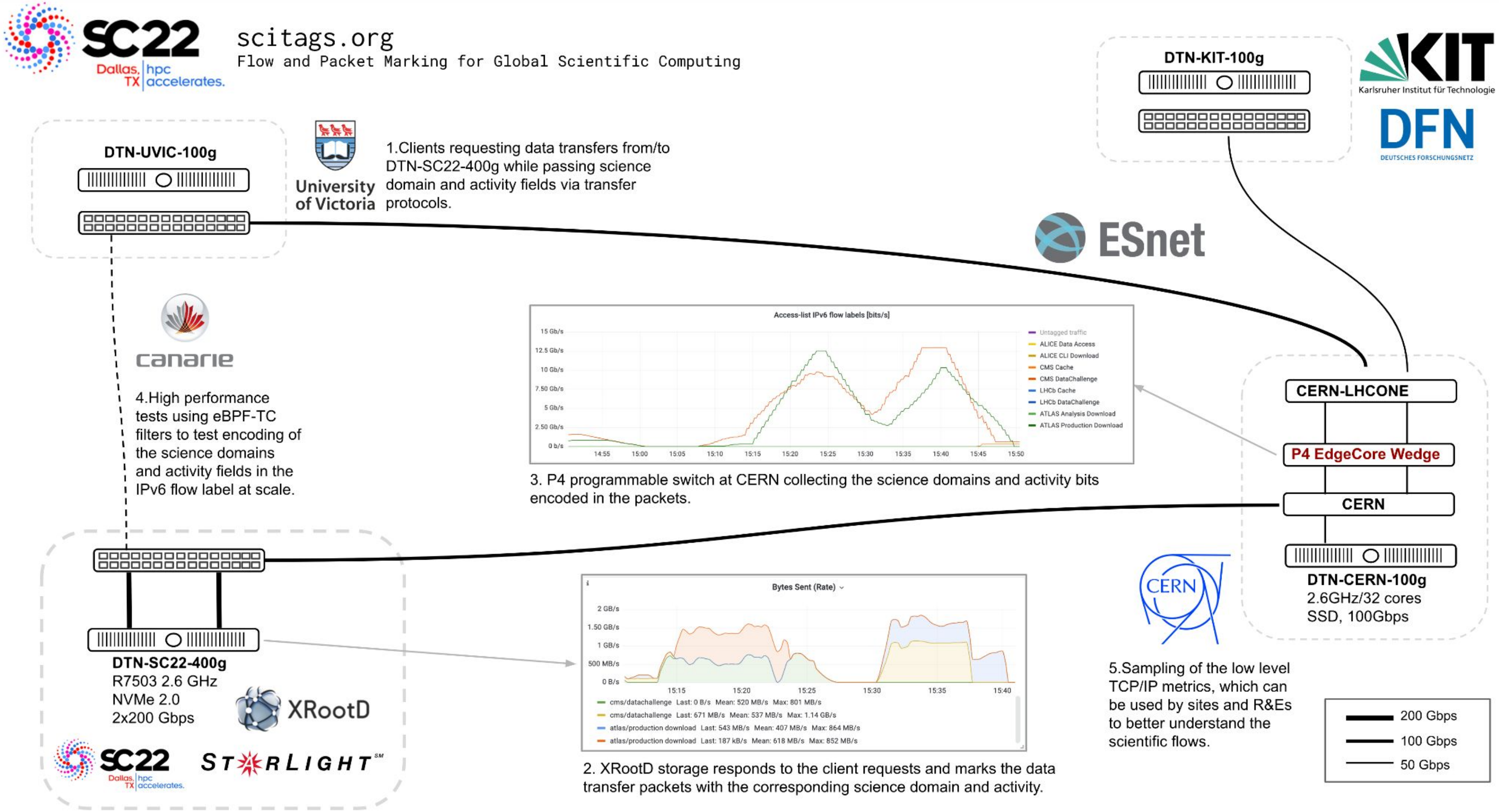


- The flowd backend maintains a hash table of flows to mark. The plugin sends the backend (src address, dst address, src port, dst port); this is used as the key in the hash, and the flow label to put on the packets is the value
- Each packet is inspected, and if the attributes match an entry in the hash, the corresponding flow label is put on the packet

Status

- **Flow Marking** (UDP firefly) implementations
  - Xrootd 5.4+ supports UDP fireflies
    - https://xrootd.slac.stanford.edu/doc/dev54/xrd_config.htm#_pmark
  - Flowd - This helper service to be more broadly deployed on storage endpoints
    - Handles both packet marking and flow marking (fireflies) for storage software
- **Collectors**
  - Initial prototype was developed by ESnet (available on scitags github)
  - ESnet and Jisc/Janet*
- **Registry**
  - Provides list of experiments and activities supported
  - Exposed via JSON at api.scitags.org
- Simplified deployment was tested during DC21, new version for DC24
  - Flowd + ESnet collector + Registry
  - **AGLT2, BNL, KIT, UNL and Caltech** participated
  - Brunel, Glasgow and QMUL interested to help with further testing
- New flowd version will be ready to be deployed shortly (building packages)

Packet and Flow Marking Demos



scitags.org
Flow and Packet Marking for Global Scientific Computing

1.Clients requesting data transfers from/to DTN-SC22-400g while passing science domain and activity fields via transfer protocols.

2. XRootD storage responds to the client requests and marks the data transfer packets with the corresponding science domain and activity.

3. P4 programmable switch at CERN collecting the science domains and activity bits encoded in the packets.

4.High performance tests using eBPF-TC filters to test encoding of the science domains and activity fields in the IPv6 flow label at scale.

5.Sampling of the low level TCP/IP metrics, which can be used by sites and R&Es to better understand the scientific flows.

DTN-SC22-400g
R7503 2.6 GHz
NVMe 2.0
2x200 Gbps

DTN-CERN-100g
2.6GHz/32 cores
SSD, 100Gbps

Scitag (Packet/Flow) Plans

We have a number of activities planned to get us from where we are to where we want to be for the Second WLCG Network Data Challenge (Feb/Mar 2024?):

- Plans (https://indico.cern.ch/event/1244448/)
  - Storages - engage more storage technologies to adopt Scitags
    - dCache implementation - target SC for production demo; AGLT2 has version to test now
    - Engage with EOS, Echo, StoRM to understand their plans and challenges
    - Flowd in production on multiple XRootd, dCache systems
  - Propagation of the flow identifier in WLCG DDM
    - FTS and Rucio implementations
    - Engage with DIRAC and Alice O2
  - Collectors/Receivers
    - Establish production level network of receivers (ESnet, Jisc, GEANT ?)
  - R&D
    - Routing and forwarding using flow label in P4 testbed (MultiONE)

Summary

The community's current focus is on the network traffic visibility through the work on flow labeling and packet marking for DC24.

- The **WLCG DOMA**, **HEPiX** and the **LHCOPN/LHCONE** communities are all engaged in better identifying and understanding our traffic flows.
- There remains a significant amount of work to do, especially regarding enabling packet marking on our storage infrastructure and in the area of collecting, aggregating and making visible the marked traffic.
- The **Scitags Initiative** is providing the common working umbrella to make R&E network traffic visible anywhere in our global networks.

Acknowledgements

We would like to thank the **WLCG**, **HEPiX**, **perfSONAR** and **OSG** organizations for their work on the topics presented.

In addition we want to explicitly acknowledge the support of the **National Science Foundation** which supported this work via:

- OSG: NSF MPS-1148698
- IRIS-HEP: NSF OAC-1836650

Questions?

Questions, Comments, Suggestions?

Useful URLs

RNTWG Google Folder

RNTWG Wiki

RNTWG mailing list signup

HEPiX NFV Final Report WG Report

RNTWG Meetings and Notes: https://indico.cern.ch/category/10031/

The scitags web page: https://scitags.github.io

Code at https://github.com/scitags/scitags.github.io

Backup slides

NOTE: SciTag Firefly Implications

One quick heads-up for sites and network providers: we are beginning to send **UDP fireflies** from some of our sites.

UDP fireflies (by default) are sent to the same destination as the data transfer flow. This means UDP packets arriving at storage servers on port 10514.

A site can choose to ignore, block or capture these packets

We are working on an informational RFC (target to publish Fall 2023)

One implication: if packets hit iptables, it may generate noise in the logging that may be a concern (fill /var/log?)

Recommendation is to open port 10514 for incoming UDP packets or explicitly ‘drop’ them.

FTS & XRootD

FTS and XRootD are key to reaching full potential in programmable networks

XRootD already provides SciTags implementation (from 5.0+)

- Enables using SciTags by R&E networks analytics (ESnet6 High-Touch)
- Currently looking for sites that would configure/test this in production

FTS/gfal2 needed to propagate SciTags to storages

- Extensions proposed for XRoot and HTTP-TPC

FTS as a transfer broker is key component for NOTED

- Understanding where/when on-demand network provisioning is needed
- Combined with analytics to determine duration, capacity, etc.

Programmable networks can be beneficial for FTS and XRootD to get better network performance, flexibility and monitoring