Packet and Flow Marking for Global Science Domains

Marian Babik / CERN

4th Global Research Platform, Cyprus, Oct 2023

https://indico.cern.ch/event/1234127/

October 10, 2023

Introduction

LHCOPN/LHCONE traffic accounts for about 40% of R&E network traffic.

Tracking and correlating data transfers with research and education (R&E) network flows is a significant challenge for WLCG. With increasing number of scientific communities using R&E networks this likely to become a more common problem.

Research Networking Technical Working Group (RNTWG) has researched, designed, and developed a comprehensive framework and explored technologies to achieve this goal

Research Networking Technical WG

This working group is focused on some specific, practical network efforts:

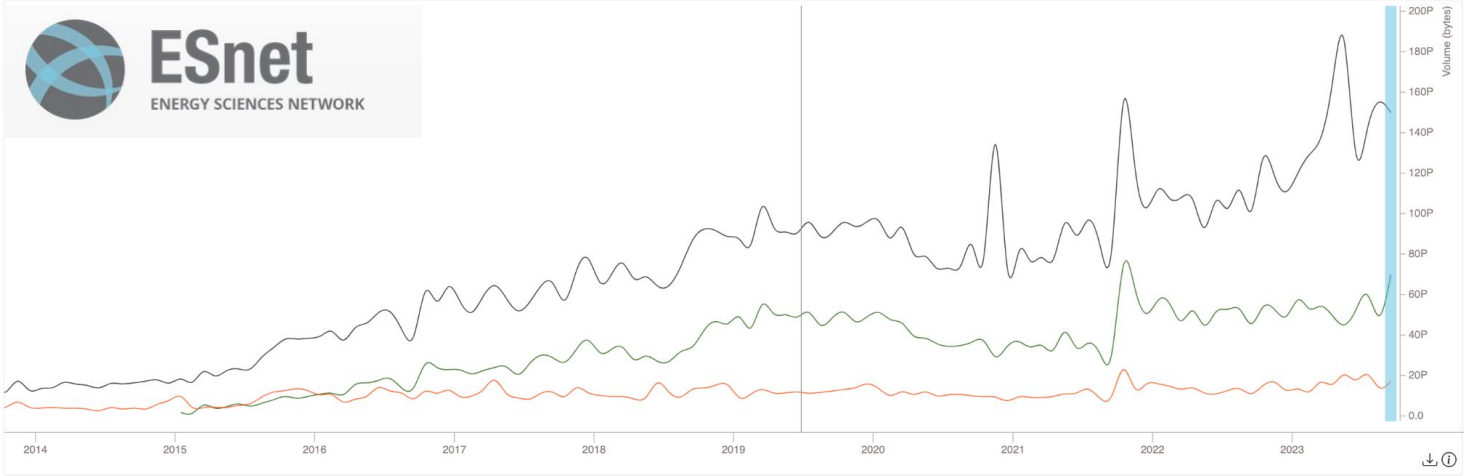
1. **Network visibility** via Packet Marking / Flow Labeling
2. **Network usage optimization** via Packet Pacing / Traffic Shaping
3. **Network orchestration** via Network Orchestration / GNA-G DIS / SENSE / NOTED

Charter for the main group is at https://zenodo.org/record/6470973#.YmamPNrMJD8

Are meetings are available in Indico: https://indico.cern.ch/category/10031/

To undertake the above efforts we have created three subgroups looking into each of the areas above.

Network Visibility



	Bytes	Percent of Total	One Month Change	One Year Change
OSCARS	17.16PB	11.5%	+22.2%	+55.7%
LHCONE	69.68PB	46.5%	+40.7%	+52.9%
Normal traffic	63.02PB	42.1%	-31.1%	+40.6%
Total	149.86PB		-3.35%	+47.8%

One of the challenges we're facing is being able to understand and identify the source of our traffic within the Research and Education (R&E) networks

Scitags Initiative

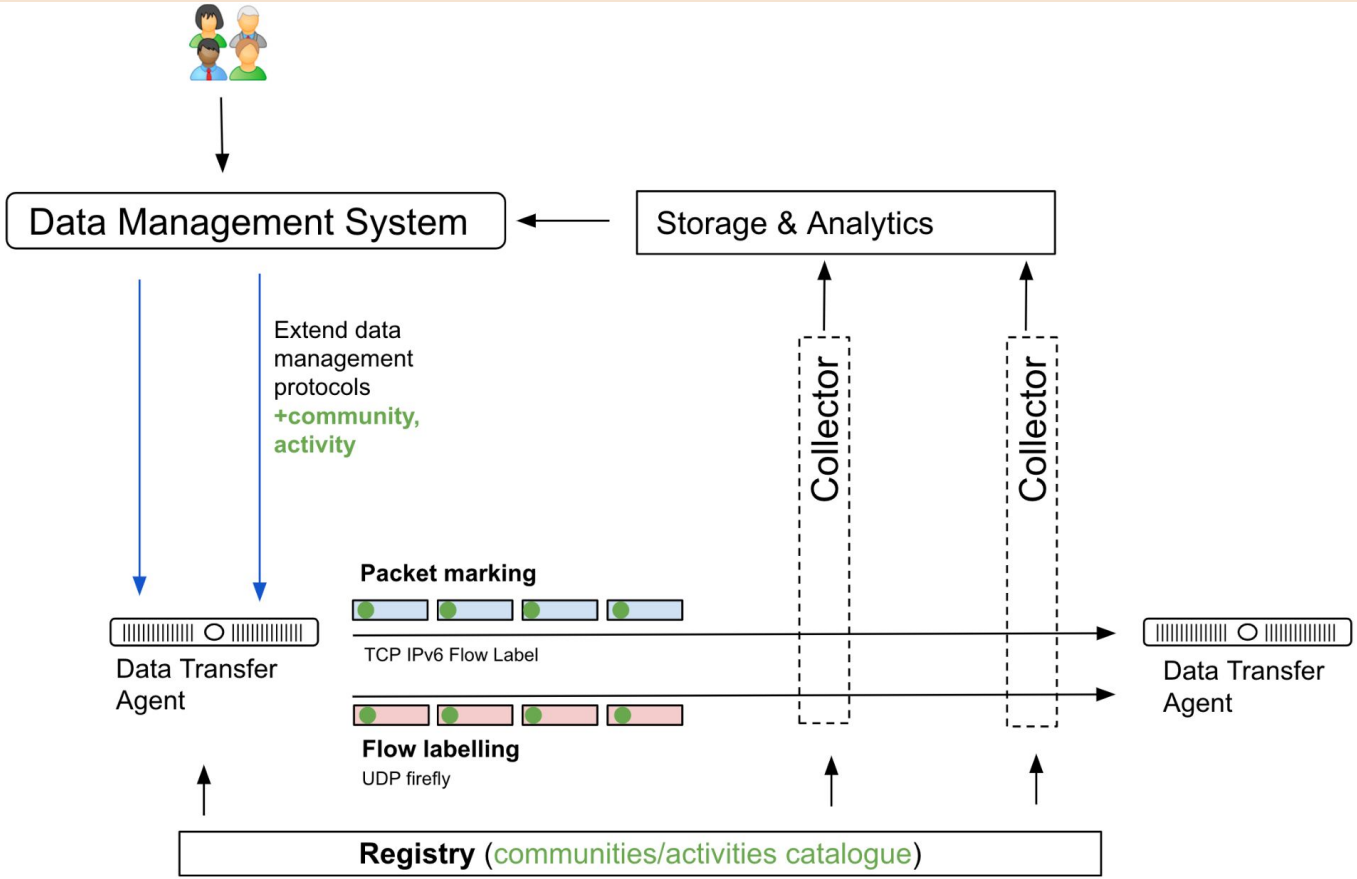
- Scientific Network Tags (Scitags) is an initiative promoting identification of the science domains and their high-level activities at the network level.

- Enable tracking and correlation of our transfers with Research and Education Network Providers (R&Es) network flow monitoring
- **Experiments** can better understand how their network flows perform along the path
  - Get insights into how experiment is using the networks, get additional data from R&Es on behaviour of our transfers (traffic, paths, etc.)
- Sites can get visibility into how different network flows perform
  - Network monitoring per flow (with experiment/activity information)

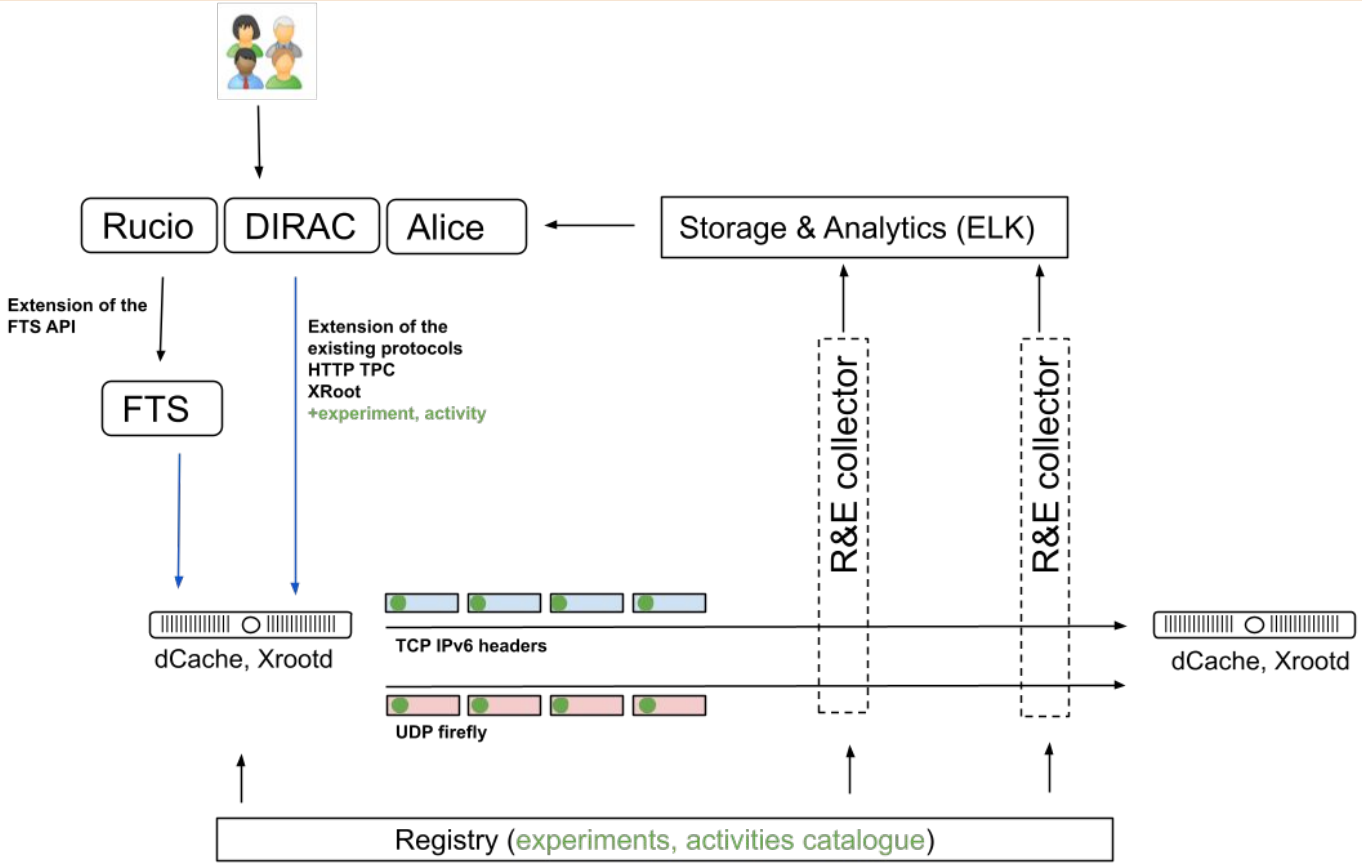
Scitags Framework Rationale

- **Open platform** that can be used by any data-intensive science community
- **Identify the owner and purpose of the traffic**
- Define a **standard** for exchange of information between scientific communities, sites and network operators
- Use coarse definitions of community/activity to provide insight into the aggregate
- **Enable tracking and correlation with existing network flow monitoring**
- Quantify global behaviour and analyse trade-offs at scale

Scitags Framework

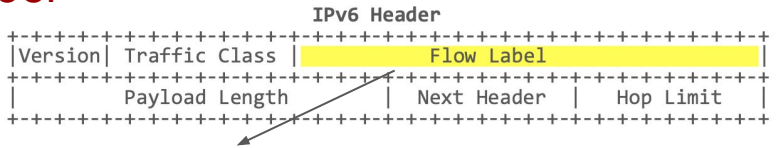


Scitags Framework - WLCG use case



Technical Spec for Packet Marking/Flow Labeling

Packet Marking via the use of the IPv6 Flow Label



Flow Label

01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
E	E	C	C	C	C	C	C	C	C	C	E	A	A	A	A	A	A	E	E

- (C) Community identifier: "Who are you affiliated with?"
- (A) Activity identifier: "What are you doing within your community?"
- (E) Entropy bits sprinkled throughout

IETF RFC-Informational Draft is available with more details

Technical Spec for Packet Marking/Flow Labeling

The detailed technical specifications are maintained on a Google doc

- **Flow Labeling** via UDP Fireflies:
  - **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.

- The document also covers methods for communicating owner/activity and other services and frameworks that may be needed for implementation.

Registry

We have standardized the “community” 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.

API is open to all R&Es and can be used by any data-intensive science community.


```
{
  - 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

Rucio & FTS & XRootD

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

Rucio already support Scitags from 32.4.0

XRootD provides SciTags implementation (from 5.0+)

- Enables using SciTags by R&E networks analytics (ESnet6 High-Touch)
- Already configured on a few production sites in UK and US, looking for volunteers

FTS/gfal2 propagates SciTags to storages

- FTS support from 3.2.10 and GFAL2 support from 2.21.0

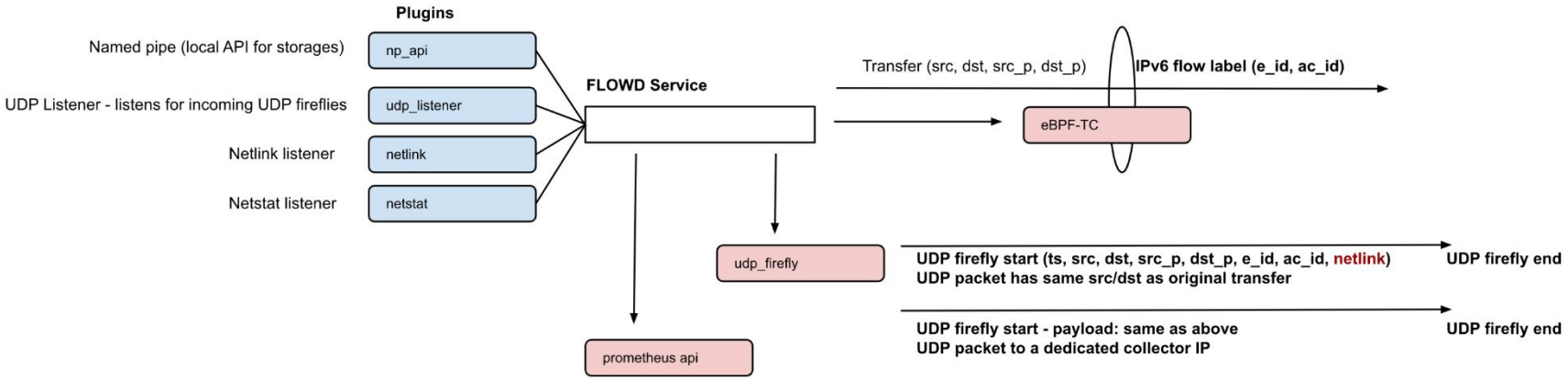
FTS as a transfer broker is key component for NOTED

- 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

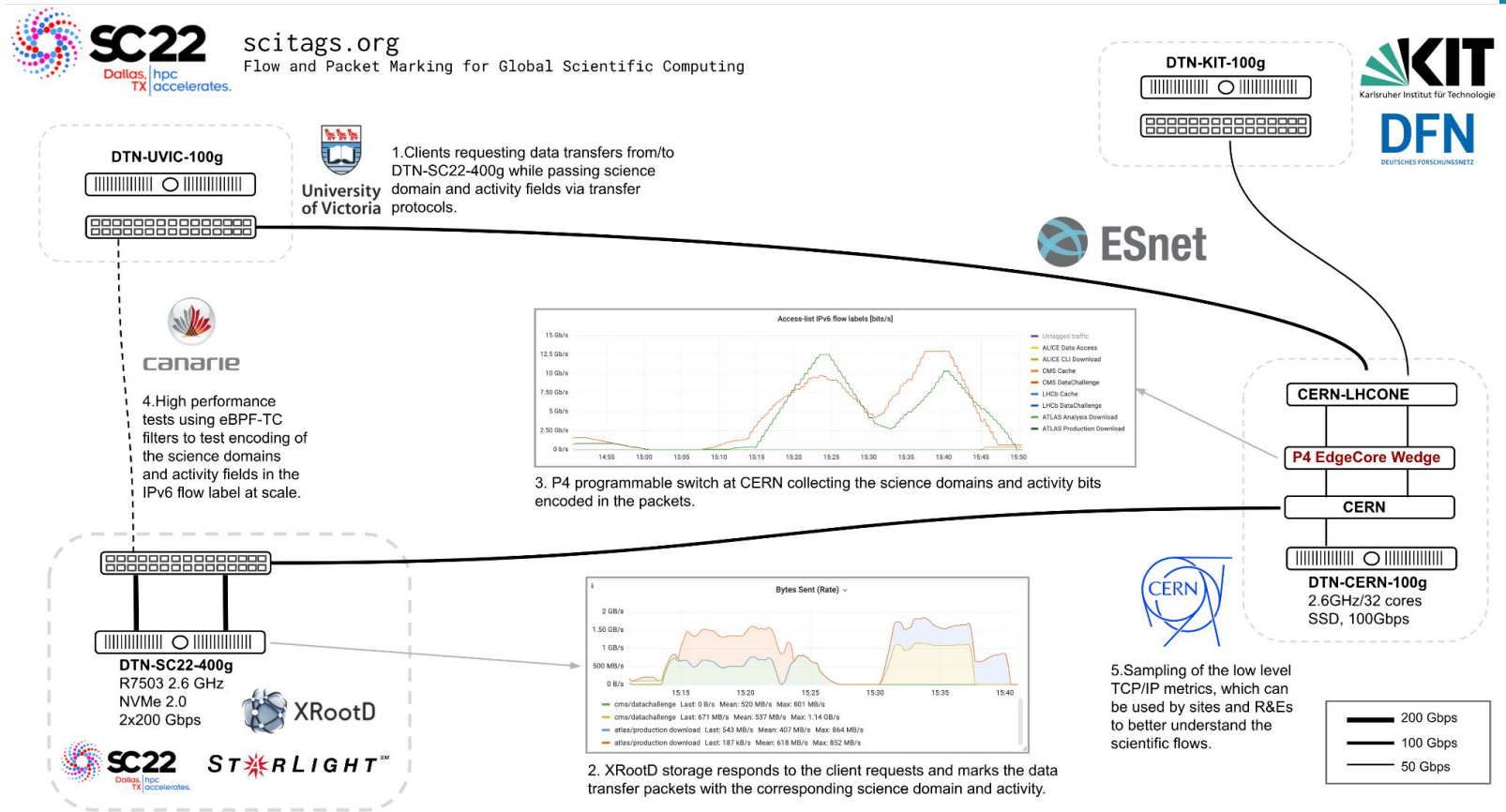
Flowd Service

- Flow and Packet Marking service developed in Python



- 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

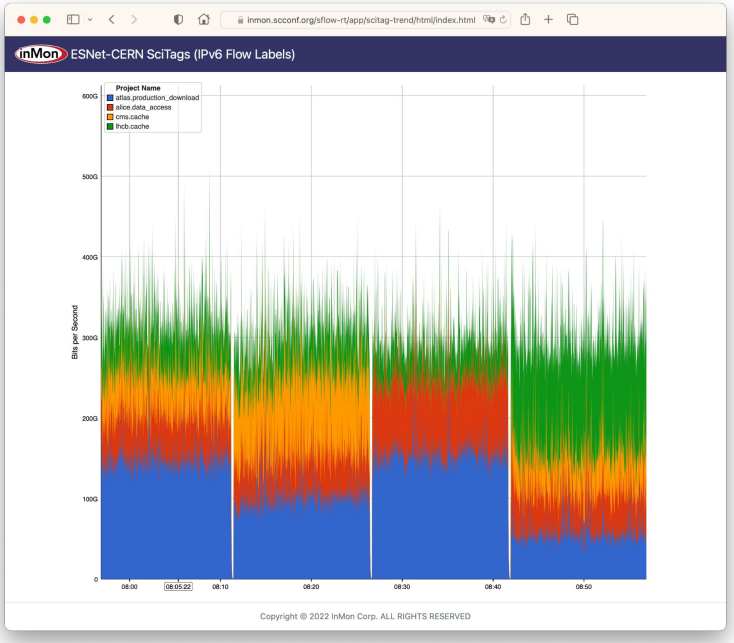
Packet and Flow Marking Demos



Packet and Flow Marking Demos

During Supercomputing 23 in Denver, we plan to demonstrate a number of aspects of our packet and flow marking work.

- Show **packet marking** at **400 Gbps** rates using **xrootd** and iperf3.
- Integration with **ESnet's High-Touch Service**
  - Analytics at the packet-level
- In collaboration with InMon, set up packet collectors via sflow and demonstrate real-time monitoring of flows by experiment and activity.
- Demos will also run on LHCONE using equipment in the SC23 booth, KIT, University of Victoria and Nebraska and CERN



Scitag (Packet/Flow) Plans

We have a number of activities planned:

- RNTWG plans
  - Storages - engage more storage technologies to adopt Scitags
    - dCache implementation - target SC for production demo
    - Engage with EOS, Echo, StoRM to understand their plans and challenges
  - Propagation of the flow identifier in WLCG DDM
    - Engage with DIRAC and Alice O2
  - Collectors/Receivers
    - Establish production level network of receivers (ESnet, Jisc, GEANT ?)
  - R&D
    - Investigate H2H as an alternative to flow label
    - Routing and forwarding using flow label in P4 testbed (MultiONE)
- Packet marking is part of the DC24 R&D projects

Summary

The RNTWG, driven by the needs and interests of the LHC, HEP and R&E networking communities, is implementing packet marking and flow labeling of network flows for all R&E network users

- We have a well defined program of work and strong collaboration with storage and transfer application providers, WLCG experiments and sites.

Our goal is to showcase packet marking and flow labeling in production during the upcoming WLCG Data Challenge

- Please consider joining our initiative !

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?

Pacing/Shaping WAN data flows

A challenge for HEP storage endpoints is to utilize the network efficiently and fully.

- An area of interest for the experiments is **traffic pacing**.
  - Without traffic pacing, network packets are emitted by the network interface in bursts, corresponding to the wire speed of the interface.
    - **Problem**: microbursts of packets can cause buffer overflows
    - The impact on TCP throughput, especially for high-bandwidth transfers on long network paths can be **significant**.
- Instead, pacing flows to match expectations [min(SRC,DEST,NET)] smooths flows and significantly reduces the microburst problem.
  - An important extra benefit is that these smooth flows are much friendlier to other users of the network by not bursting and causing buffer overflows.
  - Broad implementation of pacing could make it feasible to run networks at much higher occupancy before requiring additional bandwidth

This work has yet to have much effort; we plan to begin work during this summer!

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

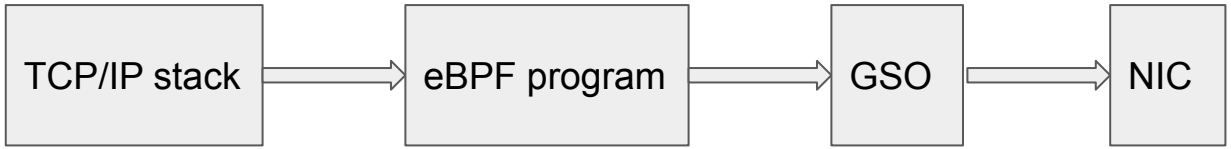
Reminder: WLCG Network Requirements

- Many WLCG facilities need **network** equipment refresh
  - Routers in many sites are End-Of-Life and moving out of warranty
  - Local area networking often has 10+ year old switches which are no longer suitable
- WLCG planning is including networking to a much greater degree than before
  - HL-LHC computing review: DOMA, dedicated networking section
  - HL-LHC Computing Conceptual Design Reports, highlight needs
  - Snowmass CompF4 has dedicated networking section
  - All include input from HEPiX, LHCONE/LHCOPN and WLCG working groups
- **Requirements Summary**
  - **Capacity**: Run-3 moving to multiple 100G links for big sites, Run-4 targeting Tbps links
  - **Capability**: WLCG needs to understand the impact of new features in networking (SDN/NFV) by testing, prototyping and evaluating impact. They will need to evolve their applications, facilities and computing models to meet the HL-LHC challenges; *it will take time*.
  - **Visibility**: As the ESnet Blueprinting meetings have shown, our ability to understand our WAN network flows is too limited. We need new methods to mark and monitor our network use
  - **Testing**: We need to be able to develop, prototype and test network features at suitable scale

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

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.

History

- HEPiX Network Functions Virtualisation Working Group
  - Working Group Report was published at the end of 2019 with three chapters
    - Cloud Native DC Networking
    - Programmable Wide Area Networks
    - Proposed Areas of Future Work
- LHCOPN/LHCONE workshop (January 2020)
  - Requirements on networks from the WLCG experiments
- Research Networking Technical Working Group
  - Formed after the workshop in response to the requirements discussion
  - 98 members from ~ 50 organisations have joined
  - Three main areas of work:
    - **Network traffic visibility**
    - **Network traffic pacing**
    - **Network traffic orchestration**

Network orchestration

- As we have seen this week, OpenStack and Kubernetes are being leveraged to create very dynamic infrastructures to meet a range of needs.
  - Critical for these technologies is a level of automation for the required networking using both software defined networking and network function virtualization.
  - For HL-LHC, important to find tools, technologies and improved workflows that may help bridge the anticipated gap between the resources we can afford and what will actually be required
- The ways we organize our computing / storage resources will need to evolve.
- This area is being led by the **GNA-G** (Global Network Advancement Group; https://www.gna-g.net/) and is exploring many options for traffic engineering, resource management and network-application interfaces.
  - The **SENSE** project is serving as a reference implementation
- The NOTED project is also an example of a practical way to effectively utilize available paths to better distribute network load.