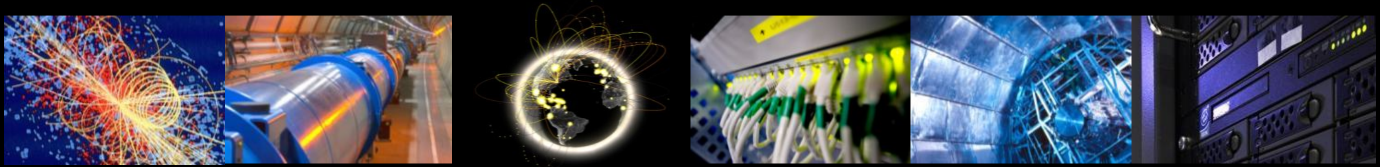
DC24: Packet Marking

Marian Babik / CERN

on behalf of the RNT Packet Marking working group

Data Challenge 2024 Workshop



WLCG
Worldwide LHC Computing Grid

Open Science Grid

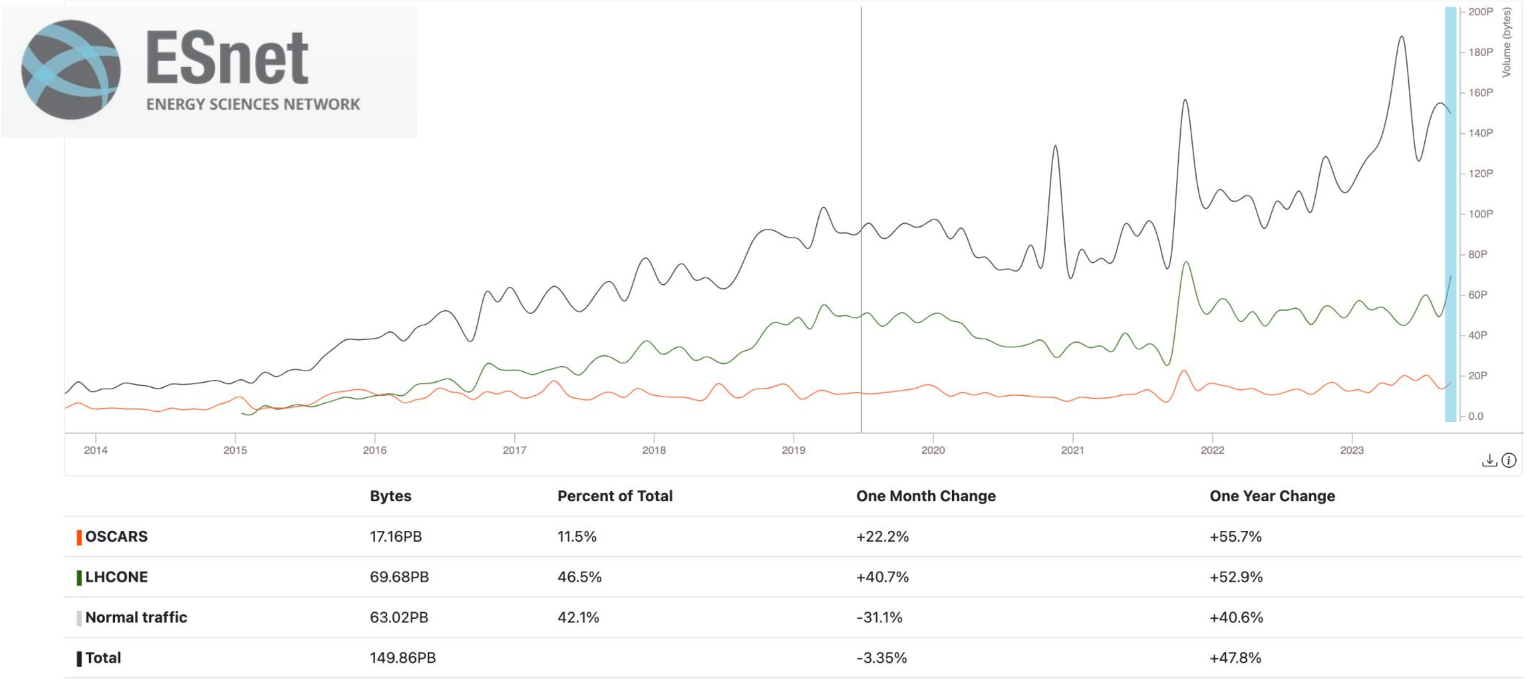
Introduction

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 on R&E networks this is 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.

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

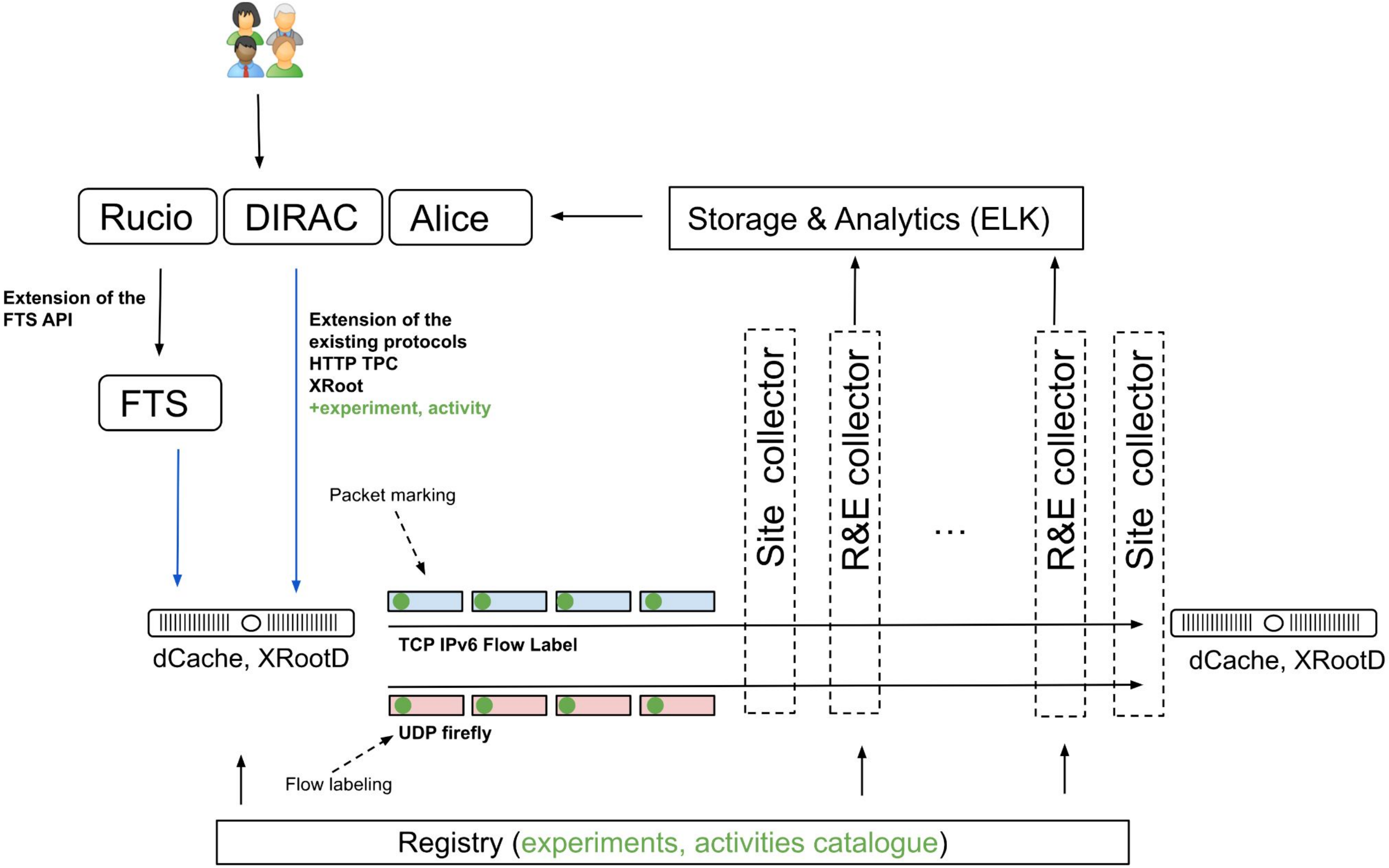
- Scientific Network Tags (Scitags) is an initiative promoting identification of the science domains and their high-level activities at the network level.

ESnet GÉANT INTERNET2 RNP ORGANIZAÇÃO SOCIAL DO MCTI Jisc NORDUnet Nordic Gateway for Research & Education STARLIGHT℠ OSG

XRootD dCache FTS RUCIO

- 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 - WLCG use case



Technical Spec for Packet Marking/Flow Labeling

Packet Marking via the use of the IPv6 Flow Label

```
                          IPv6 Header
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version| Traffic Class |           Flow Label                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Payload Length        |  Next Header  |   Hop Limit   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

```
                          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 are sent to the destination of the transfer, but can also be sent to specific regional or global collectors.
    - Use of syslog format makes it easy to process by Logstash or similar receivers.

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

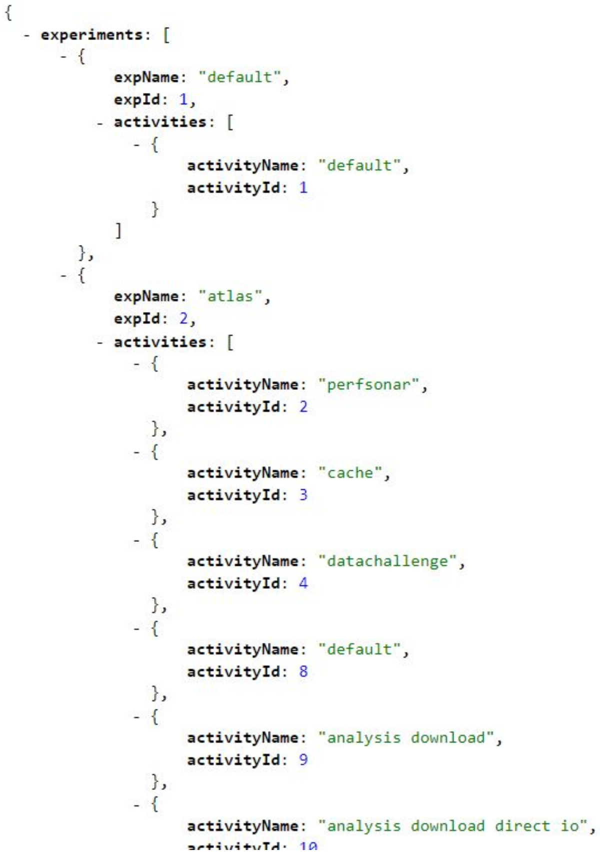
Registry

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

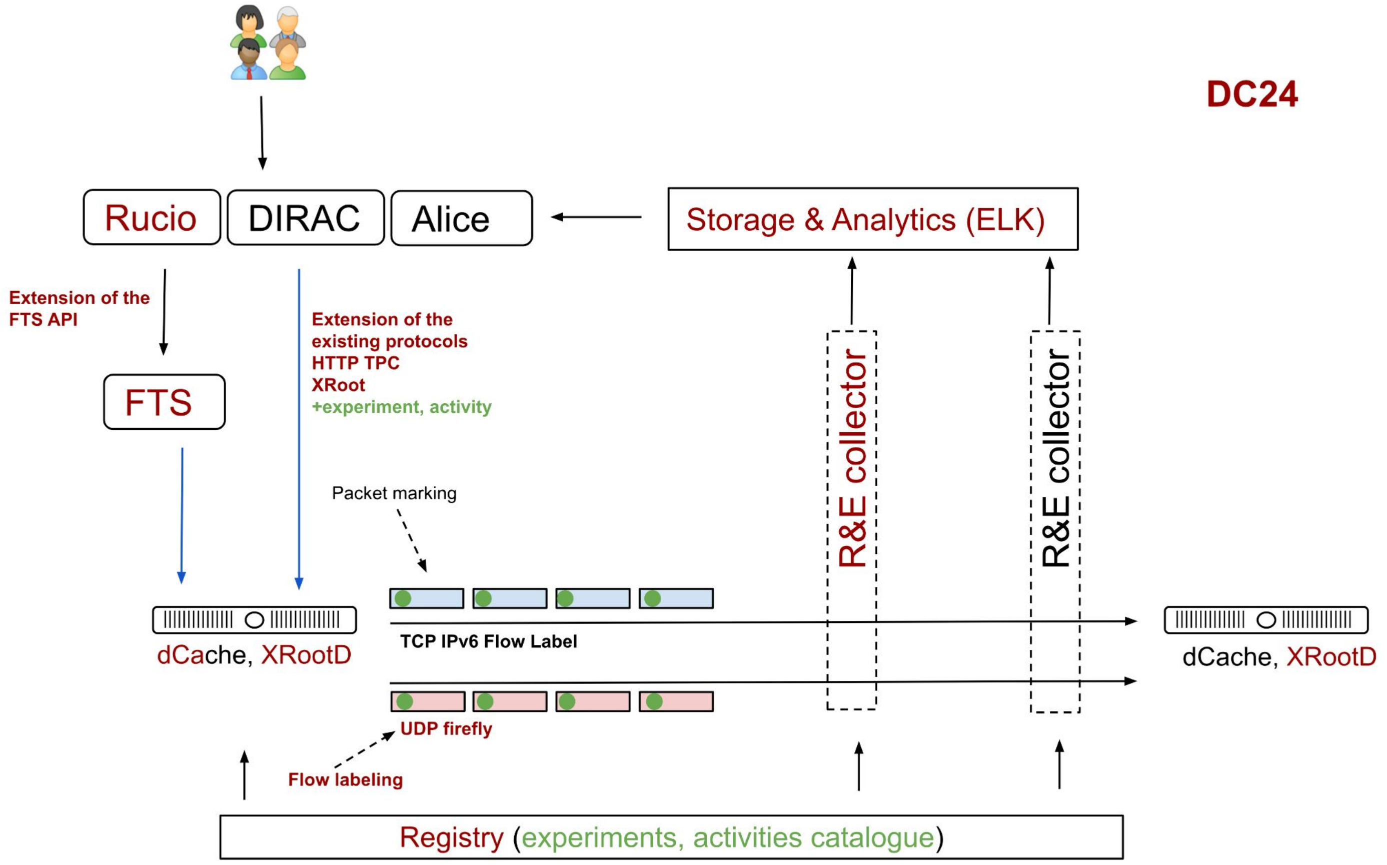
The underlying source of truth is a set of Google sheets that are maintained and writeable by a few stewards.

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
```

DC24 Scitags Overview



DC24: Current status

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

Rucio added support for Scitags in 32.4.0

FTS/gfal2 now also supports Scitags

- Support from FTS 3.12.11 and GFAL2 2.22.0 (HTTP-TPC headers)

XRootD provides SciTags implementation (from 5.0+)

- Support for flow labeling (UDP fireflies) - configurable for a single collector
- Already configured on a few production sites in UK and US, looking for volunteers
- Support for HTTP-TPC headers coming soon

dCache prototype available (in testing at AGLT2)

Flowd service - 1.0.2

- Supports integration with storage (via local UDP fireflies listener)
- Adds packet marking (for IPv6) and can enrich UDP fireflies with TCP metrics

Collector(s) - 1.0.0

- Collects UDP fireflies and integrates with ELK stack

DC24 Plans (updated)

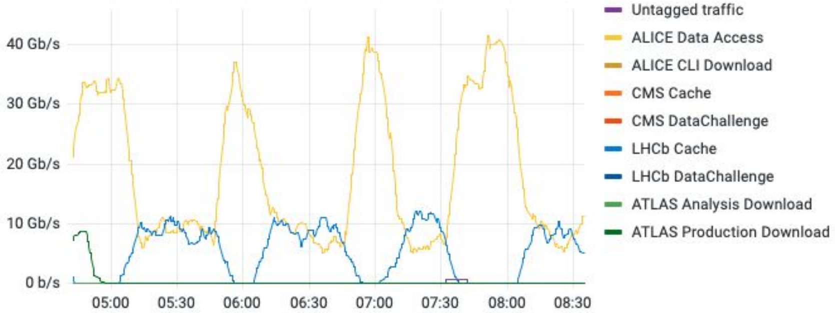
- Aug 2023: Demonstrate **flow labeling** from two or more production sites
- Sep 2023: Show initial accounting of flows via the ESnet Stardust (**collector**)
- Sep 2023: Flowd service available for EL8/EL9 (packet marking service)
- Nov 2023: Demonstrate **packet marking** in IPv6 packets from two or more production sites
- Nov 2023: Demonstrate accounting for packet marking and flow labelling (site and collectors)
- Jan 2024: Identify and demonstrate traffic volume accounting by VO and activity globally and at one or more net locations.

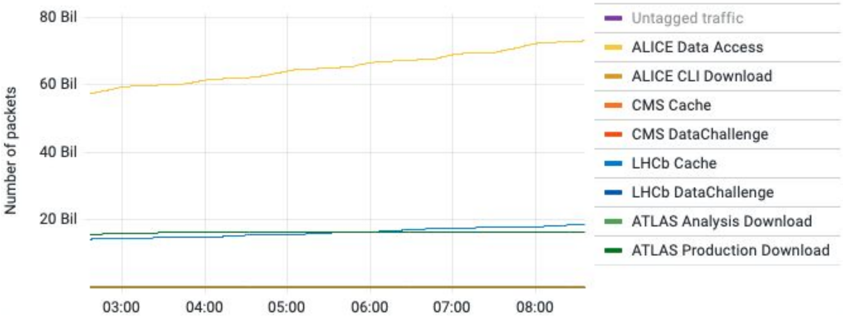
DC24 Status

- Aug 2023: Demonstrate *flow labeling* from two or more production sites ✅
  - Flow labeling deployed at UNL, BNL, AGLT2, UK sites (TBD)
- Sep 2023: Show initial accounting of flows via the ESnet Stardust ✅
  - Deploy and test collector; at ESnet and at Jisc deployed and receiving UDP fireflies
- Sep 2023: Flowd service available for EL8/EL9 ✖
  - Delayed due to missing sponsor for EPEL, will work on introducing it in another repo
  - Service is needed to enable packet marking (for any storage) and to encode additional monitoring data in UDP fireflies (e.g. MTU/MSS, congestion algorithm used, buffers, etc.)
- Nov 2023: Demonstrate *packet marking* in IPv6 packets from two or more production sites
  - Needs previous step, but UNL already has the setup (both xrootd and flowd)
- Nov 2023: Demonstrate accounting for packet marking and flow labelling (site and collectors)
  - Accounting for flow labeling (UDP fireflies) that carries additional information
  - Demonstrate integration with advanced R&E service, e.g. ESnet High-touch service
  - Accounting for packet marking using network equipment (inMon SFlow-RT, CERN P4)

Packet marking accounting for DC24

- During SC22 we demonstrated the accounting of packet marked ✅
  - Generated dashboards in MONIT Grafana (see figure below)
- Accounting on P4 Tofino switch of traffic mirrored from LHCONE/LHCOPN
  - Demonstrated with QFX Juniper platform ✅
  - Current status: the LHCONE/LHCOPN border router was migrated to PTX Junos platform. There is a bug and port mirroring is not working so we cannot perform the accountability of packet marked for SC23 ❌
- Accounting on Juniper switches, if supported
  - Current status: the flow label matching is in the road map of Junos (potentially by February)
  - Contact with Broadcom for accounting in Trident 4 switches

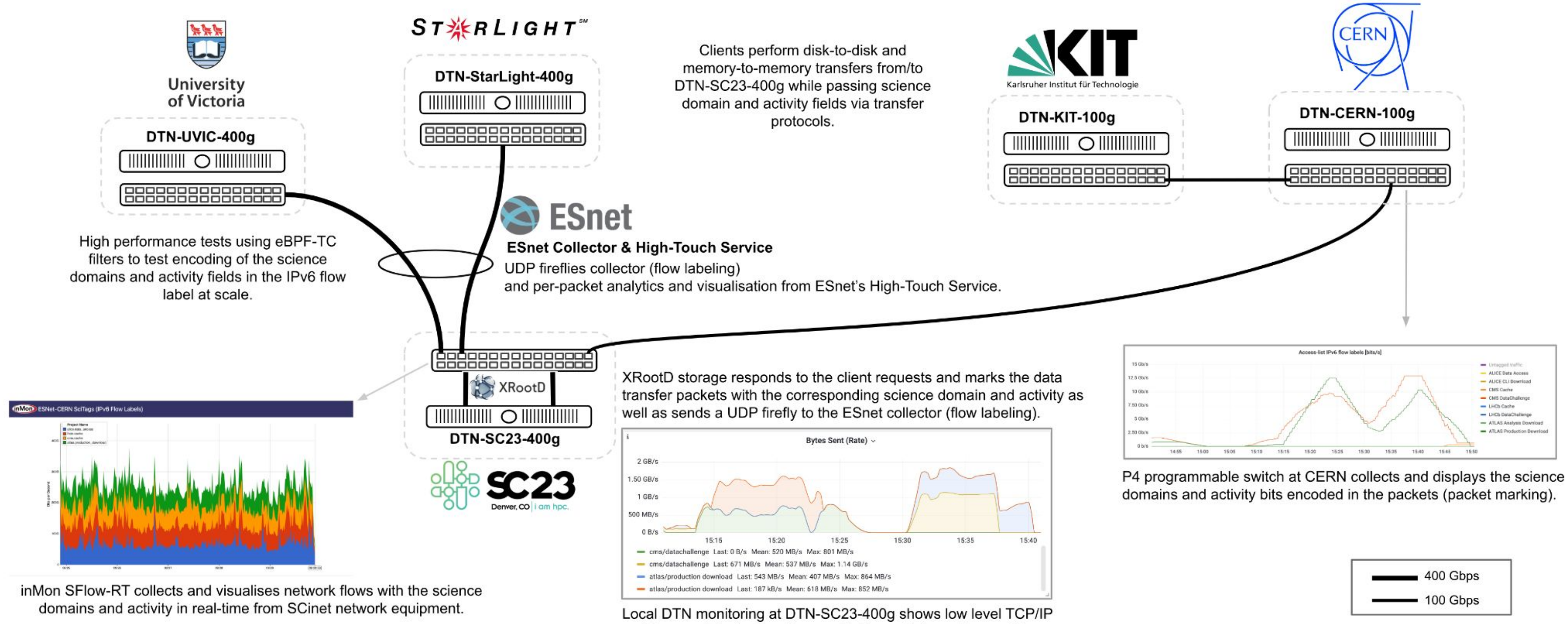




SC23: Packet and Flow Marking Demos

SC23 Denver, CO | i am hpc.

scitags.org
Packet Marking for Data Intensive Scientific Workflows



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

Clients perform disk-to-disk and memory-to-memory transfers from/to DTN-SC23-400g while passing science domain and activity fields via transfer protocols.

ESnet Collector & High-Touch Service
UDP fireflies collector (flow labeling)
and per-packet analytics and visualisation from ESnet's High-Touch Service.

XRootD storage responds to the client requests and marks the data transfer packets with the corresponding science domain and activity as well as sends a UDP firefly to the ESnet collector (flow labeling).

- cms/datachallenge Last: 0 B/s Mean: 520 MB/s Max: 801 MB/s
- cms/datachallenge Last: 671 MB/s Mean: 537 MB/s Max: 1.14 GB/s
- atlas/production download Last: 543 MB/s Mean: 407 MB/s Max: 864 MB/s
- atlas/production download Last: 187 kB/s Mean: 618 MB/s Max: 852 MB/s

inMon SFlow-RT collects and visualises network flows with the science domains and activity in real-time from SCinet network equipment.

Local DTN monitoring at DTN-SC23-400g shows low level TCP/IP metrics correlated with science activities, which can be used by sites and R&Es to better understand the scientific flows.

P4 programmable switch at CERN collects and displays the science domains and activity bits encoded in the packets (packet marking).

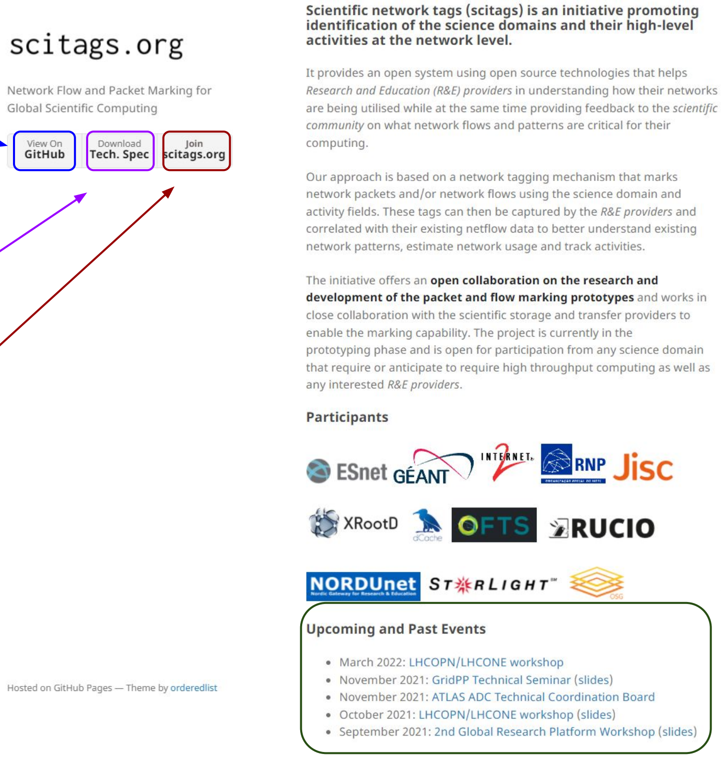
Finding More Information: https://scitags.org

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*.

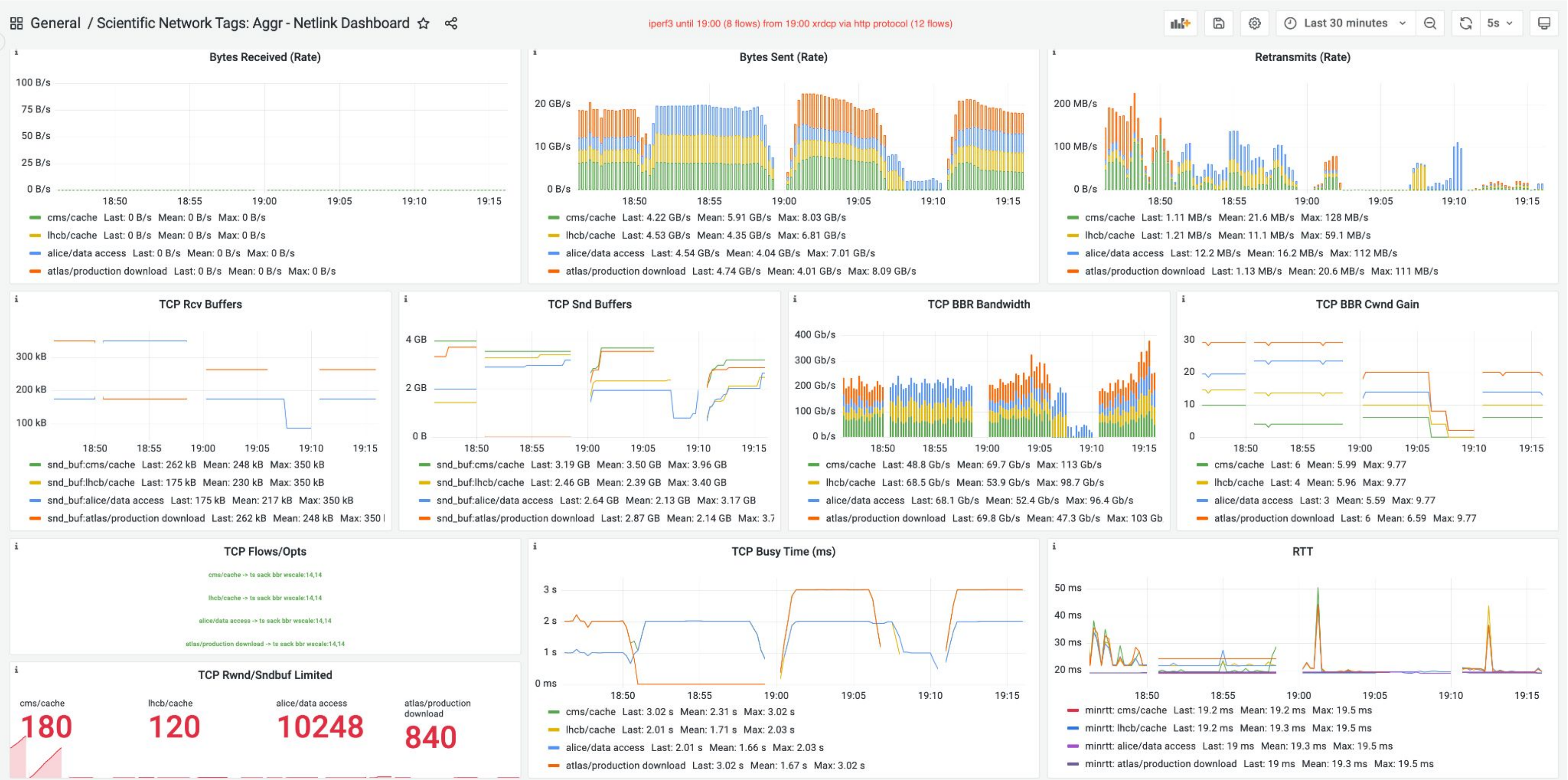
Participants

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)

Backup slides

Scitags combined with netlink



General / Scientific Network Tags: Aggr - Netlink Dashboard

iperf3 until 19:00 (8 flows) from 19:00 xrdcp via http protocol (12 flows)

Last 30 minutes · 5s

Bytes Received (Rate)

- cms/cache Last: 0 B/s Mean: 0 B/s Max: 0 B/s
- lhcb/cache Last: 0 B/s Mean: 0 B/s Max: 0 B/s
- alice/data access Last: 0 B/s Mean: 0 B/s Max: 0 B/s
- atlas/production download Last: 0 B/s Mean: 0 B/s Max: 0 B/s

Bytes Sent (Rate)

- cms/cache Last: 4.22 GB/s Mean: 5.91 GB/s Max: 8.03 GB/s
- lhcb/cache Last: 4.53 GB/s Mean: 4.35 GB/s Max: 6.81 GB/s
- alice/data access Last: 4.54 GB/s Mean: 4.04 GB/s Max: 7.01 GB/s
- atlas/production download Last: 4.74 GB/s Mean: 4.01 GB/s Max: 8.09 GB/s

Retransmits (Rate)

- cms/cache Last: 1.11 MB/s Mean: 21.6 MB/s Max: 128 MB/s
- lhcb/cache Last: 1.21 MB/s Mean: 11.1 MB/s Max: 59.1 MB/s
- alice/data access Last: 12.2 MB/s Mean: 16.2 MB/s Max: 112 MB/s
- atlas/production download Last: 1.13 MB/s Mean: 20.6 MB/s Max: 111 MB/s

TCP Rcv Buffers

- snd_buf:cms/cache Last: 262 kB Mean: 248 kB Max: 350 kB
- snd_buf:lhcb/cache Last: 175 kB Mean: 230 kB Max: 350 kB
- snd_buf:alice/data access Last: 175 kB Mean: 217 kB Max: 350 kB
- snd_buf:atlas/production download Last: 262 kB Mean: 248 kB Max: 350 I

TCP Snd Buffers

- snd_buf:cms/cache Last: 3.19 GB Mean: 3.50 GB Max: 3.96 GB
- snd_buf:lhcb/cache Last: 2.46 GB Mean: 2.39 GB Max: 3.40 GB
- snd_buf:alice/data access Last: 2.64 GB Mean: 2.13 GB Max: 3.17 GB
- snd_buf:atlas/production download Last: 2.87 GB Mean: 2.14 GB Max: 3.7

TCP BBR Bandwidth

- cms/cache Last: 48.8 Gb/s Mean: 69.7 Gb/s Max: 113 Gb/s
- lhcb/cache Last: 68.5 Gb/s Mean: 53.9 Gb/s Max: 98.7 Gb/s
- alice/data access Last: 68.1 Gb/s Mean: 52.4 Gb/s Max: 96.4 Gb/s
- atlas/production download Last: 69.8 Gb/s Mean: 47.3 Gb/s Max: 103 Gb

TCP BBR Cwnd Gain

- cms/cache Last: 6 Mean: 5.99 Max: 9.77
- lhcb/cache Last: 4 Mean: 5.96 Max: 9.77
- alice/data access Last: 3 Mean: 5.59 Max: 9.77
- atlas/production download Last: 6 Mean: 6.59 Max: 9.77

TCP Flows/Opts

cms/cache -> ts sack bbr wscale:14,14

lhcb/cache -> ts sack bbr wscale:14,14

alice/data access -> ts sack bbr wscale:14,14

atlas/production download -> ts sack bbr wscale:14,14

TCP Rwnd/Sndbuf Limited

cms/cache	lhcb/cache	alice/data access	atlas/production download
180	120	10248	840

TCP Busy Time (ms)

- cms/cache Last: 3.02 s Mean: 2.31 s Max: 3.02 s
- lhcb/cache Last: 2.01 s Mean: 1.71 s Max: 2.03 s
- alice/data access Last: 2.01 s Mean: 1.66 s Max: 2.03 s
- atlas/production download Last: 3.02 s Mean: 1.67 s Max: 3.02 s

RTT

- minrtt: cms/cache Last: 19.2 ms Mean: 19.2 ms Max: 19.5 ms
- minrtt: lhcb/cache Last: 19.2 ms Mean: 19.3 ms Max: 19.5 ms
- minrtt: alice/data access Last: 19 ms Mean: 19.3 ms Max: 19.5 ms
- minrtt: atlas/production download Last: 19 ms Mean: 19.3 ms Max: 19.5 ms

WLCG DOMA Project Details

Description of objective for DC24: We intend to mark a significant amount of WLCG network traffic via flow marking (UDP Fireflies) and/or packet marking (for IPv6 traffic). Significant means at least 5% of traffic during DC24 for two or more participating VOs. The goal is providing visibility of all R&E traffic anywhere in the network, giving experiments feedback to optimize net-use and networks info/control

Timeline: We will need to have mini-challenges to develop the capabilities and metrics required

- August 2023: We will demonstrate marked flows from two or more production sites using dCache and Xrootd
- Who: Shawn McKee, Marian Babik

- September 2023: We will show initial accounting of flows via the ESnet Stardust system.
- Who: Andy Lake, Shawn McKee

- September 2023: Flowd available in EPEL for EL8/EL9
- Who: Steve Traylen, Marian Babik

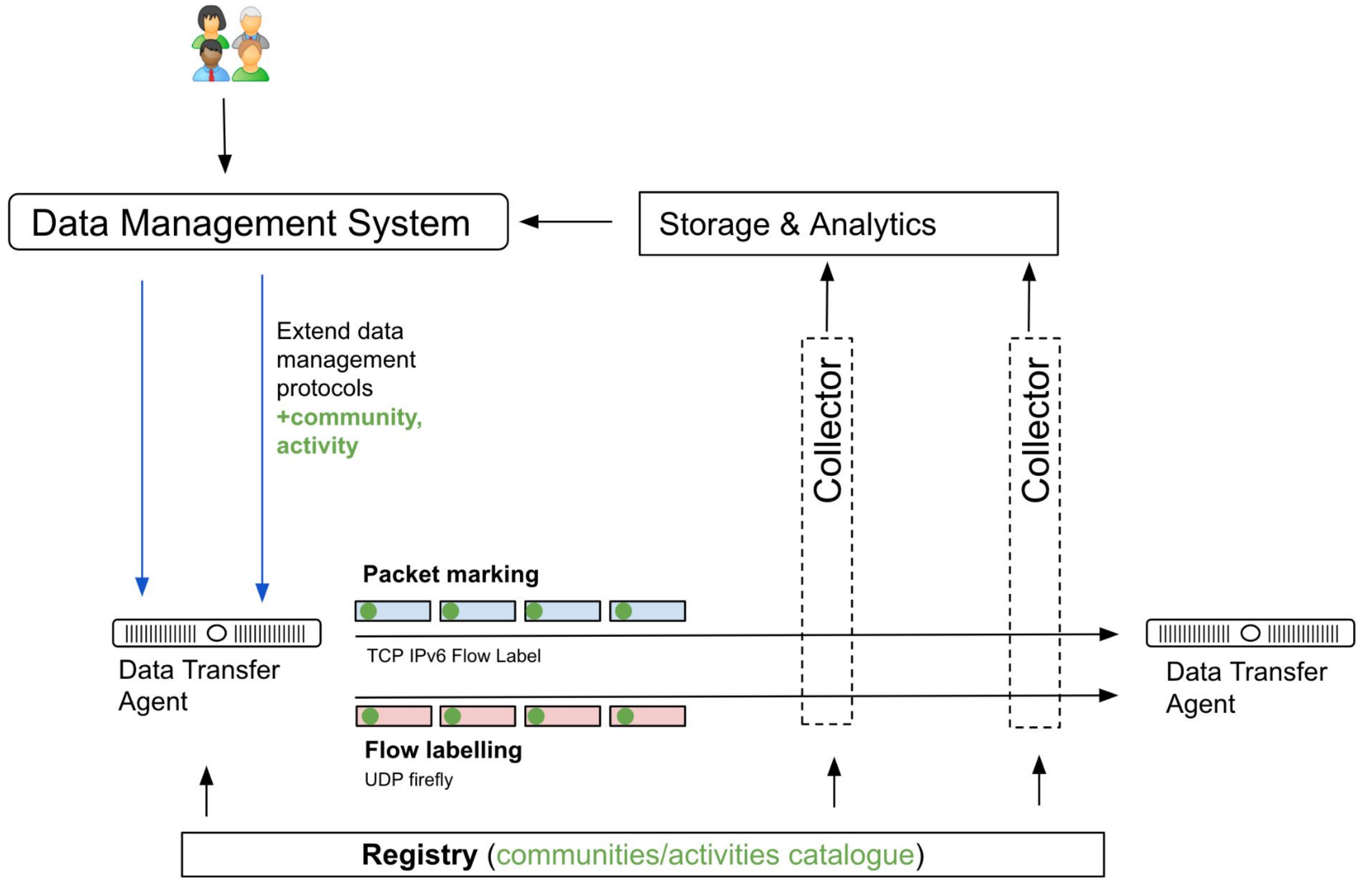
- October 2023: We will demonstrate marked IPv6 packets from two or more production sites using dCache and Xrootd.
- Who: Garhan Attebury, Marian Babik

- October 2023: We will demonstrate the use of packet accounting of marked IPv6 packets.
- Who: Carmen Misa, Edoardo Martelli, Yatish Kumar(? Or other ESnet), Jeronimo Bezerra(?)

- December 2023: Identify and demonstrate traffic volume accounting by VO and activity globally and at one or more net locations.
- Who: Marian Babik, Andy Lake, GEANT

Metrics: During DC24 we intend to track
The amount of traffic that is marked by packets (MarkedPacketTotal [TB]) and by flow (MarkedFlowTotal[TB])
The number of UDP fireflies sent (estimated) and received (#Sent, #Recieved)
The total size of traffic, broken out by owner and activity (by flow or by packet) at one or more network locations (ESnet, GEANT?, Internet?) and globally via the ESnet Stardust receiver. Estimate December 2023

Scitags Framework

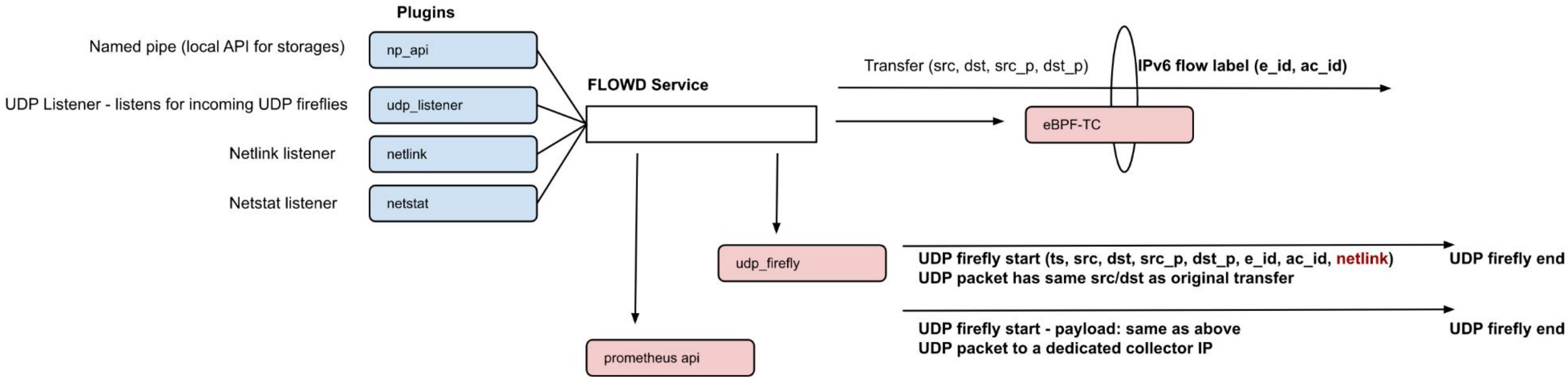


Scitags Platform Rationale

- **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

Flowd Service

- Flow and Packet Marking service developed in Python
  - Can be used to support/extended functionality provided by dCache


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

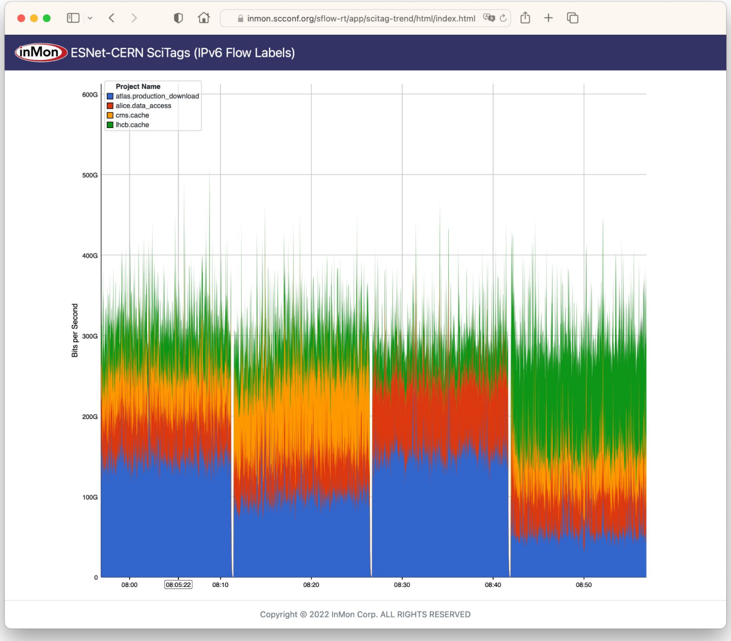
- Plugins provide different ways get connections to mark (or interact with storage)
  - New plugins were added to support netlink readout and UDP firefly consumer
- Backends are used to implement flow and/or packet marking
  - New backends were added to mark packets (via eBPF-TC) and expose monitored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

Packet and Flow Marking Demos

SC23 DENVER NOV 12-17

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 community/activity**.
- Demos will also run on LHCONE using equipment in the SC23 booth, KIT, University of Victoria and Nebraska and CERN



inmon.scconf.org/sflow-rt/app/scitag-trend/html/index.html

inMon ESNet-CERN SciTags (IPv6 Flow Labels)

Copyright © 2022 InMon Corp. ALL RIGHTS RESERVED

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

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/

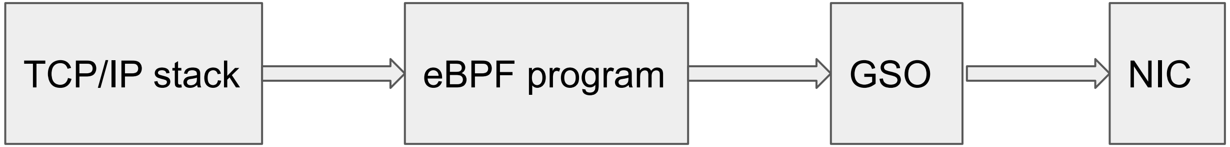
The scitags web page: https://scitags.github.io

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

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.