RNTWG: Packet Marking and Pacing Update

Shawn McKee / University of Michigan, Marian Babik / CERN

#52 LHCOPN/LHCONE Meeting, Catania, Italy

https://indico.cern.ch/event/1349135/

Apr 11, 2024

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 management** 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/

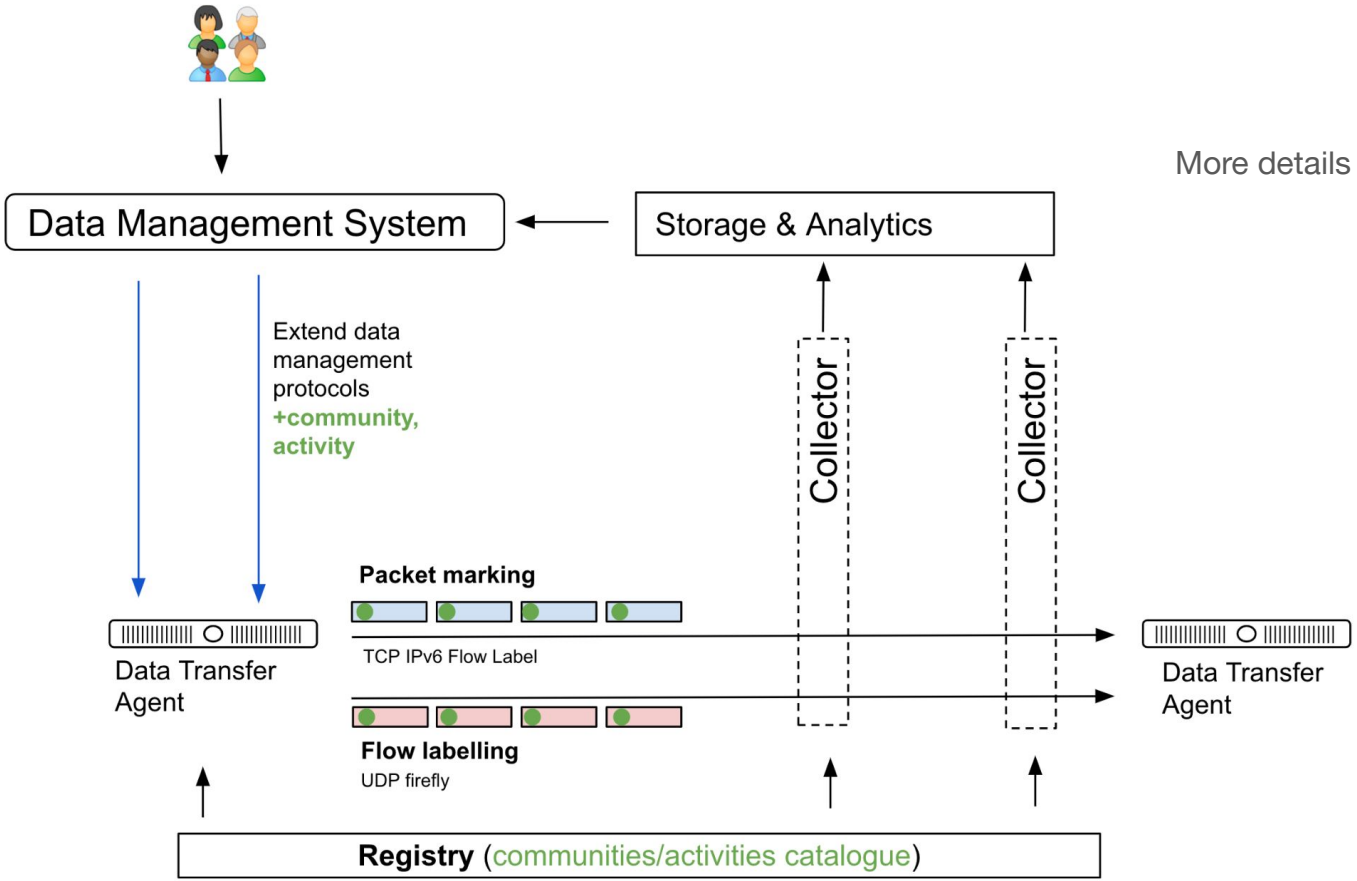
To undertake the above efforts we have created three subgroups looking into each of the areas above.

Scitags

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

- **Goals**
  - Provide **standardised means of information exchange** on network flows between experiments, sites and network providers.
  - **Improve** experiments' and sites' **visibility** into how network flows perform within network segments.
  - Get insights into how experiments are using the networks and **benefit from additional data from the network providers**.
  - Make **network performance tuning and troubleshooting** easier and more effective by gaining insights into how different network configurations impact performance

How Scitags Work



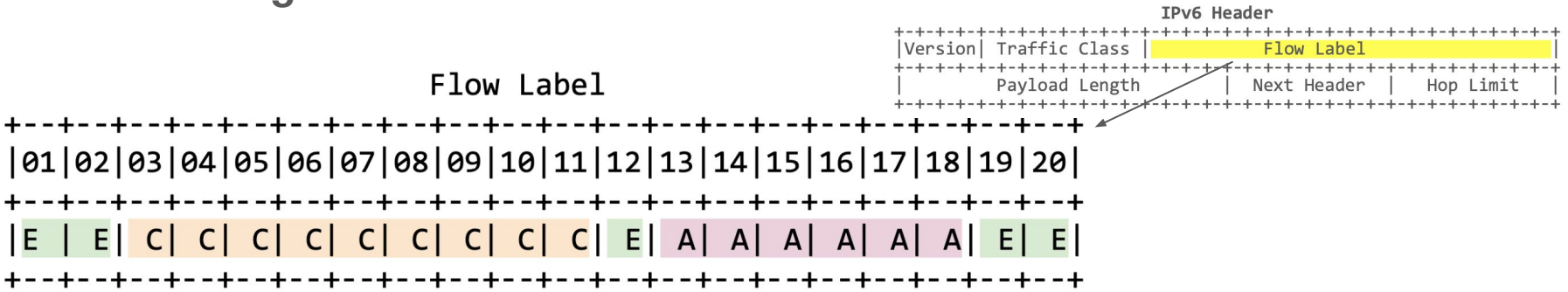
More details in CHEP paper

Scitags Framework Rationale

- **Open platform** that can be used by any data-intensive science community
- **Identify the owner (experiment) and purpose (activity) of the traffic**
- Define a **standard(s)** for exchange of information between scientific communities, sites and network operators
  - Packet marking - encoding exp/activity directly in packets
  - Flow labeling - sending a separate UDP packet (firefly) with metadata
- **Enable tracking and correlation with existing network flow monitoring and existing monitoring systems deployed by R&E networks**
- Quantify global behaviour and analyse trade-offs at scale

Technical Spec for Packet Marking

Packet Marking via the use of the IPv6 Flow Label



- (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

Started exploring HbH option as an alternative (eBPF-PDM, eBPF-extHeaders)

Technical Spec for Flow Labeling

- **Flow Labeling** via UDP packets (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.
    - Works for IPv4 and IPv6; content is not limited (as long as it fits in a single frame)
      - Apart from exp/act we now have also usage (bytes sent/rcv) and RTT in fireflies

- The detailed technical specifications are maintained on a Google doc
- The document also covers methods for communicating owner/activity and other services and frameworks that may be needed for implementation.

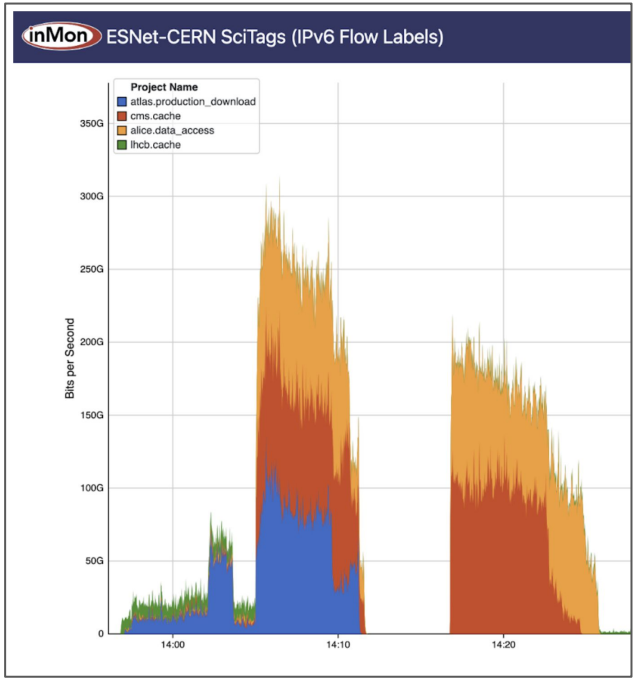
Collectors

- Different types depending on what is being collected
  - **HW/on-the-wire** to collect UDP fireflies and/or IPv6 flow label
  - **SW/network of receivers** - collecting UDP fireflies sent to them
    - Working on update to the current architecture to introduce a message bus - to interconnect different (N)REN collectors and also allow to subscribe
  - **SW/Collectors**
    - Site-collector - forwards fireflies via UDP, optional local storage
    - Regional collector - receives fireflies from sites, stores locally and publishes to message bus
    - Global collector - receives all fireflies (directly or via bus), global store
    - Experiments collector - subscribes to the bus for specific fireflies

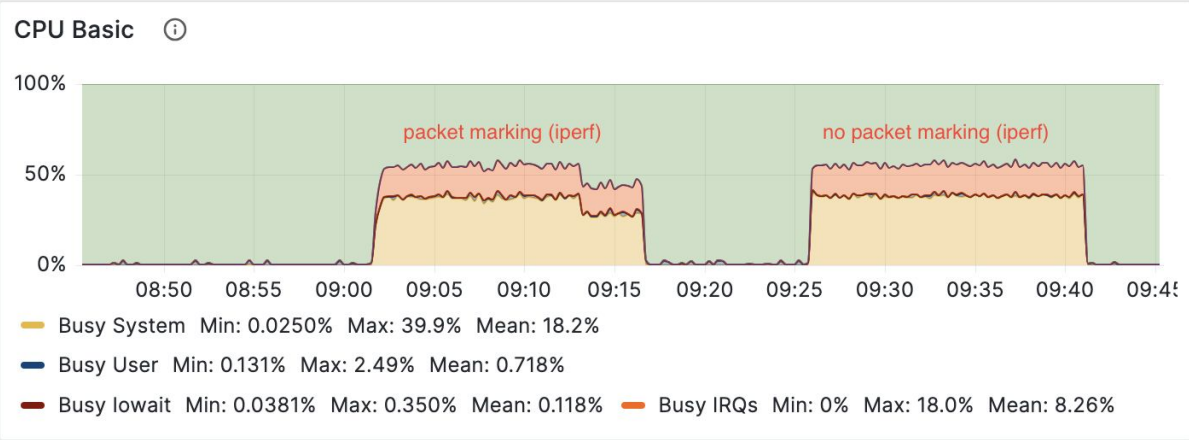
Packet Marking Demo

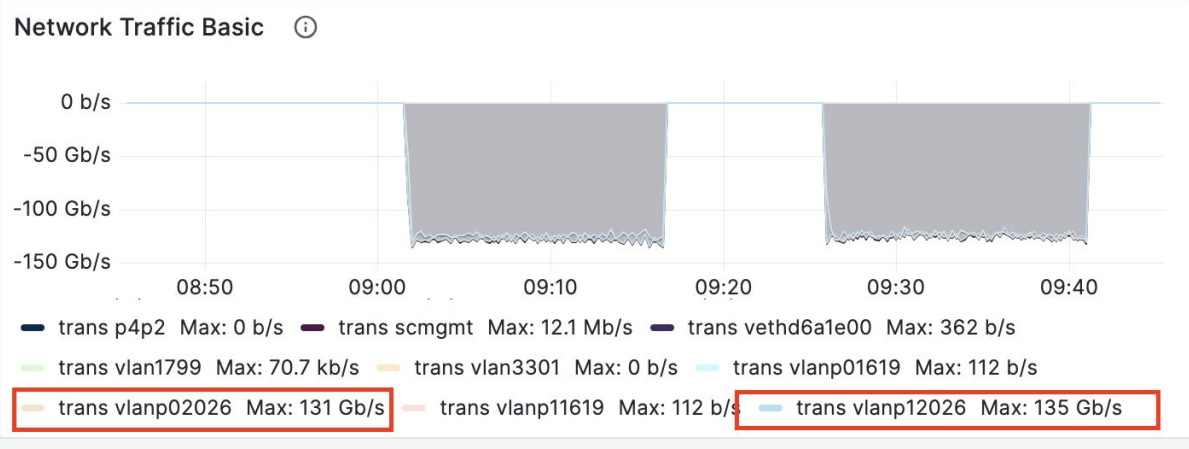
During Supercomputing 23 in Denver, we demonstrated a number of aspects of our packet and flow marking work.

- Showed **packet marking** at **300 Gbps** rates using **xrootd**/iperf3 (with just two nodes; using eBPF).
- 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**.
- Demo was run in collaboration with Starlight, ESnet, KIT, University of Victoria, University of Nebraska and CERN



Packet and Flow Marking Demos

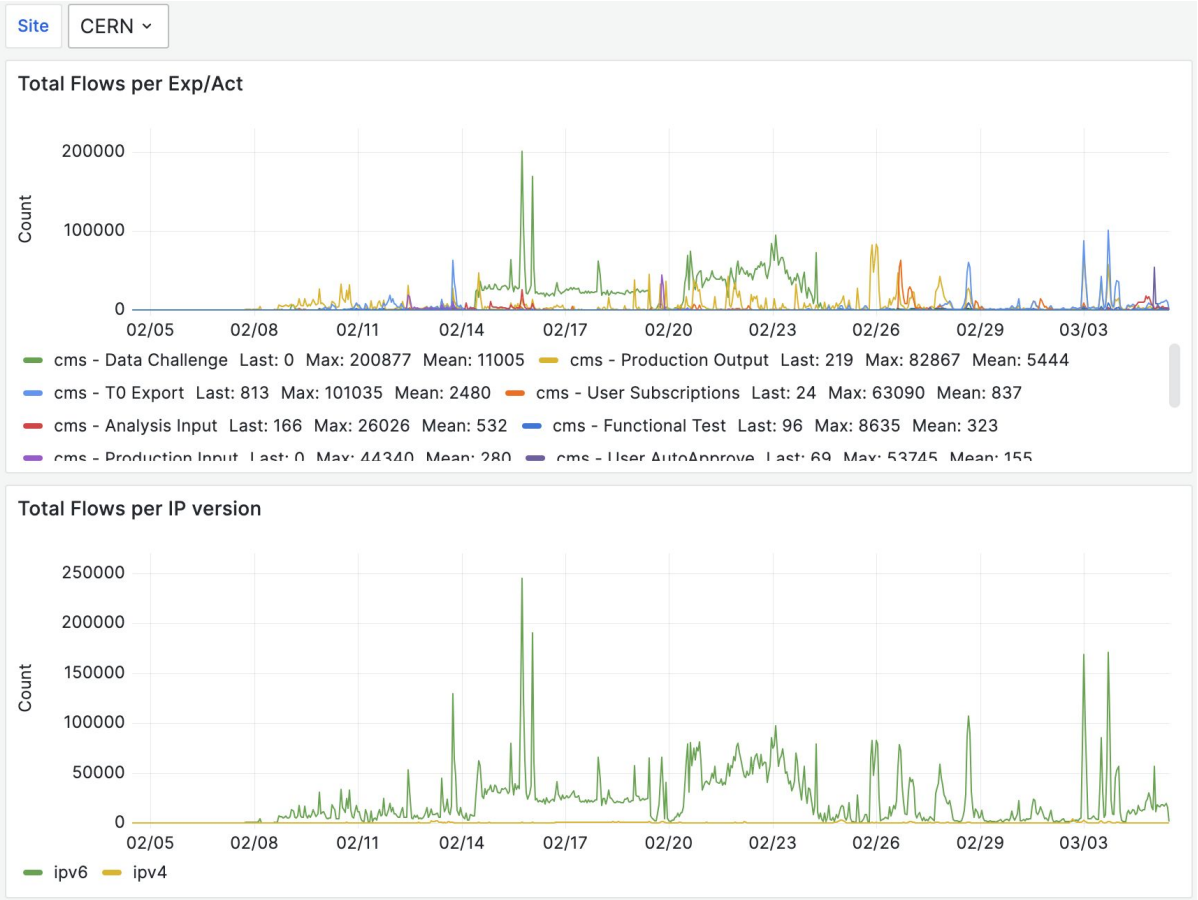




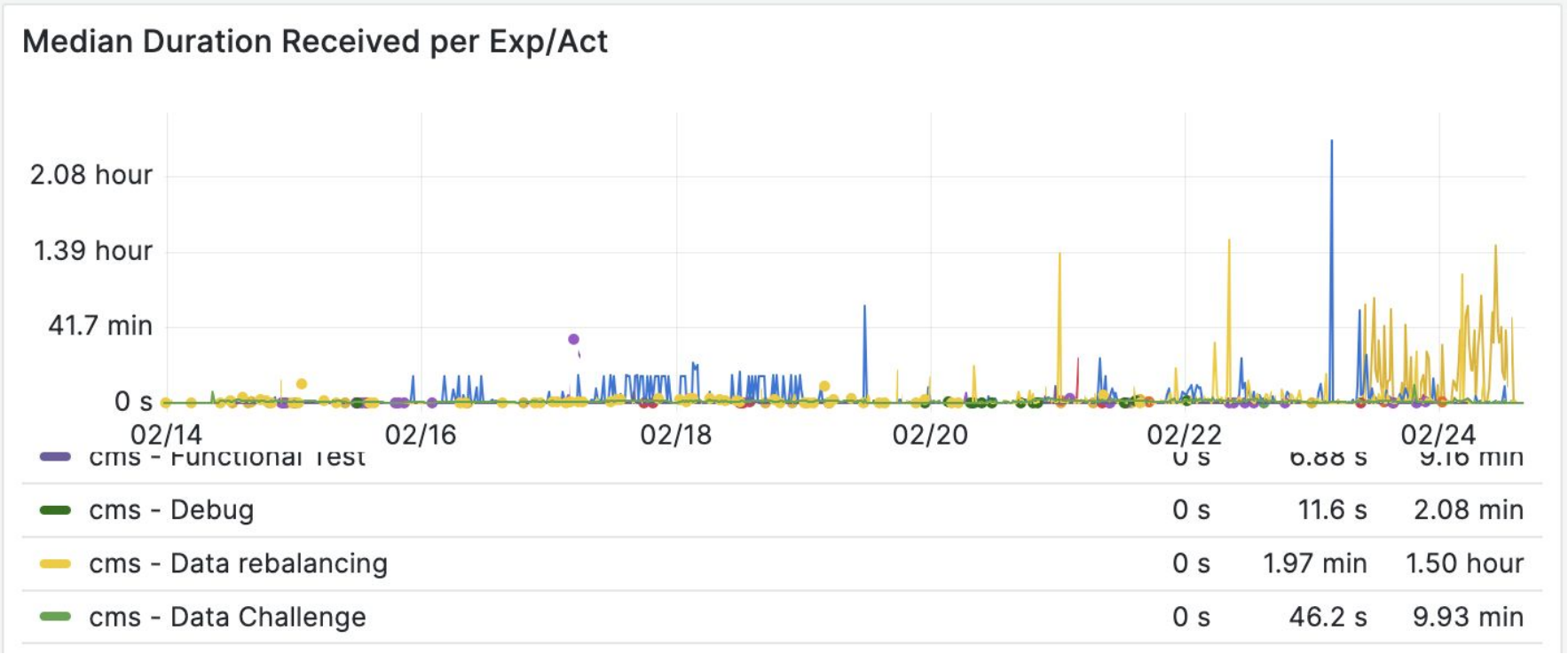
Data Challenge 24

- **Scitags Deployment**
  - 80% of EOS CMS (production), UNL production storage
  - Flow labeling functionality (fireflies)
- **Results**:
  - **Confirmed the capability to propagate Scitags all the way to the storages (for both ATLAS and CMS)**
  - Sending fireflies (from XRootd, EOS storages)
  - **Collection and visualisation at ESnet collector**
    - Results shown in live dashboard
- Issues:
  - We hit an issue with xrootd crashing when receiving scitags http headers
    - This had impact on ATLAS testing and availability of the ATLAS Xrootd storages
  - The issue was fixed quickly but we were unable to rollout (as DC was already running)

CERN EOS CMS plot showing split by experiment/activity and IP versions



CERN EOS CMS
Median duration of flows split by Exp/Activity
Shows duration of DC flows was quite short wrt. production/rebalancing



Current Status

Implementation status:

Propagation:

- **Rucio** supports Scitags from 32.4.0
- **FTS/gfal2** support Scitags from 3.2.10/2.21.0

Storages:

- **XRootD** provides Scitags implementation (from 5.0+)
- **EOS** provides Scitags support from 5.2.19+
  - Working on a project for production rollout at CERN (for WLCG)
- **dCache** prototype exists, roadmap for release pending
- Also working with StoRM and Pelican

Collectors:

- Production deployments at ESnet and Jisc

Summary

- Scitags (flow labelling) ready for production
  - Expecting sites and experiments will gradually enable it during this year
    - **ATLAS, CMS and ALICE ready to enable in production**
    - Sites ramp-up can be quick once it starts
    - Plan to enable fireflies at CERN T0
  - SW/Collector network will need to be ready and scale
    - Network providers are encouraged to deploy a collector to benefit from the initiative
  - Scitags facilitate collaboration with experiments
    - Reporting of issues and follow up becomes easier
- Significant progress in packet marking R&D
    - Will benefit from flow labelling deployment and production

Finding More Information: https://scitags.org

Code

Technical Spec

Mailing List

Presentations

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.

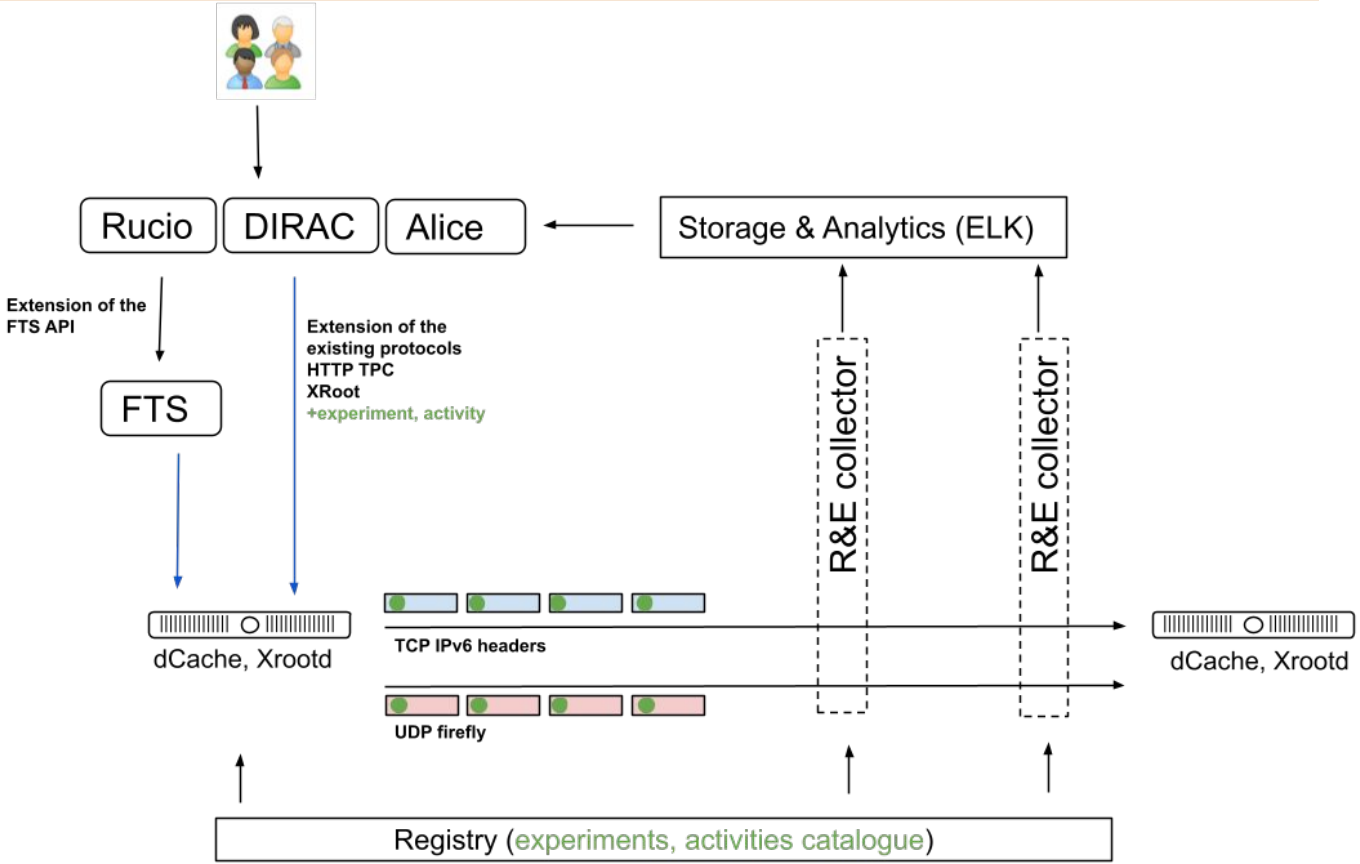
We want to update you on the technical working group, which is focused on addressing some specific areas of interest to HEP that are relevant for the broader R&E community globally.

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
- **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 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)
    - E.g. RTT, retransmits, segment size, congestion window, etc. all per flow

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
```

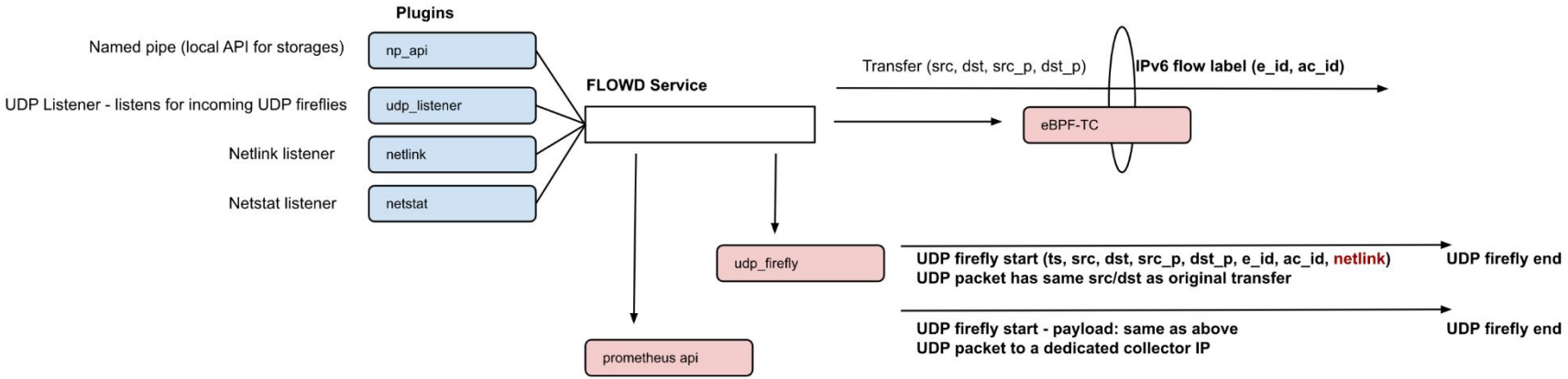
Technical Spec for Packet Marking/Flow Labeling

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.

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

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

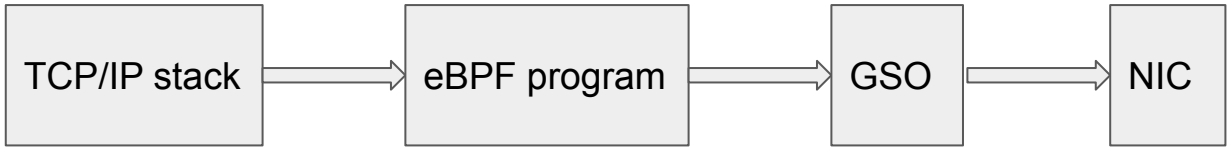
- Understanding where/when on-demand network provisioning is needed
- Combined with analytics to determine duration, capacity, etc.

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

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:

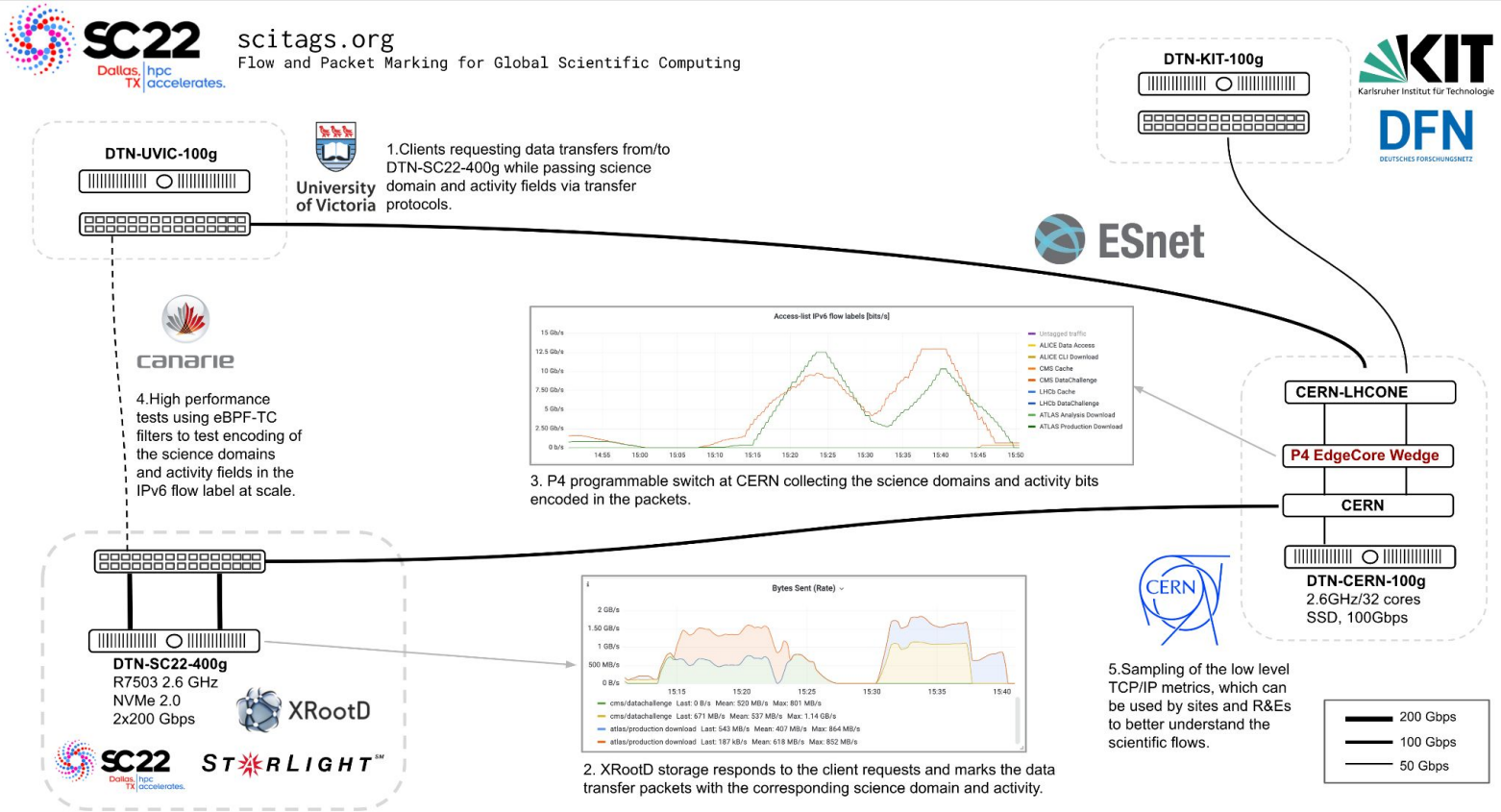


- The flowd backend maintains a hash table of flows to mark. The plugin sends the backend (src address, dst address, src port, dst port); this is used as the key in the hash, and the flow label to put on the packets is the value
- Each packet is inspected, and if the attributes match an entry in the 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
    - **map2exp** - can be used to map particular path to an experiment
    - **map2act** - can be used to map particular user/role to an activity
  - Flowd - prototype service
    - Issue fireflies from netstat for a given experiment (only for dedicated storages)
- **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
  - 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)

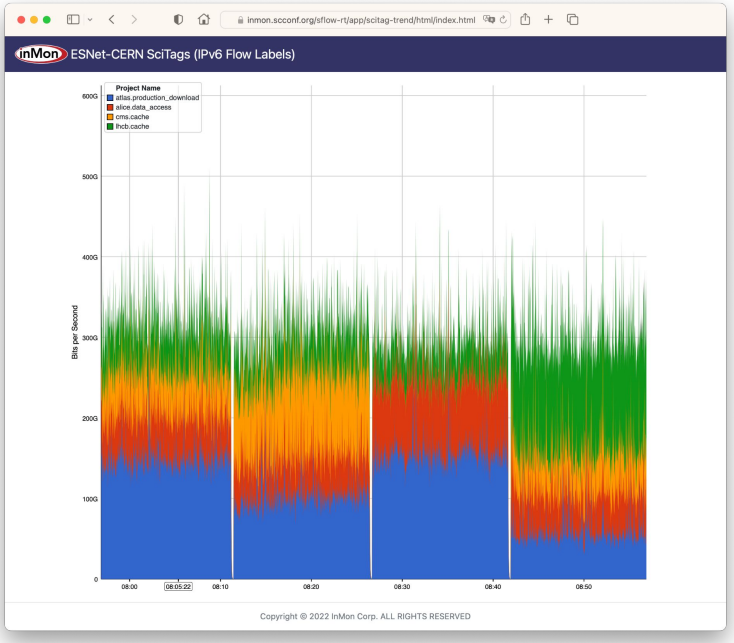
Packet and Flow Marking Demos



Packet and Flow Marking Demos

During Supercomputing 22 in Dallas, we demonstrated a number of aspects of our packet and flow marking work.

- We showed **packet marking** at **200 Gbps** rates using flowd with both **xrootd** and iperf3.
- Scinet and ESnet set up packet collectors via sflow and demonstrated real-time monitoring of packets by experiment and activity.
- Demos were also run on LHCONE using equipment in the SC22 booth, KIT, UVic and CERN where packet marking for all transfers was monitored using a P4 programmable switch.



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!

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?):

- RNTWG plans (https://indico.cern.ch/event/1244448/)
  - 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
    - 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)
- USATLAS DC24 Draft plan for networking objectives and milestones

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.

Summary

The RNTWG has made significant progress in the identified network priority focus areas for the WLCG community. The current focus is on the network traffic visibility through the work on flow labeling and packet marking for DC24.

- 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.

We have additional near-term work to pursue in traffic pacing:

- While network orchestration has significant activity underway, we need to find new effort interested in developing, prototyping and evaluating traffic pacing for science data flows.
- See Eli’s upcoming talk later in this session…

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

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

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.