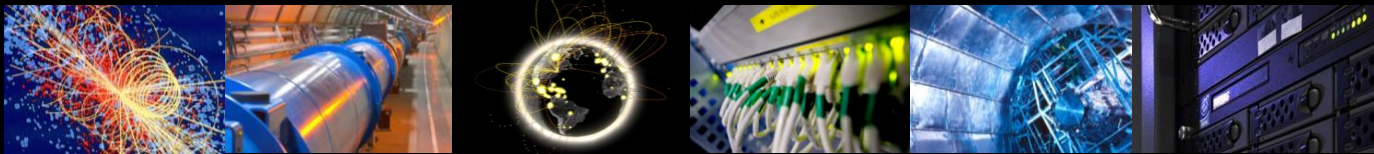


Packet and Flow Marking for Global Science Domains

Shawn McKee / University of Michigan

3rd GRP, October 10, 2022

on behalf of the Research Networking Technical Working Group



Introduction to Packet Marking and Flow Labeling

To start the presentation I would like to motivate and define the goal the **Research Networking Technical Working Group (RNTWG)** is trying to achieve.

Motivation: The poor experience for WLCG trying to understand network flows, especially across the Atlantic, using just end site transfer and ESnet stats.

GOAL: To be able to identify the owner and purpose of any research and education network flow anywhere in the network.

WHY??: Many reasons:

- Network links can become congested and it is vital to understand the sources of the traffic involved and work with users to better orchestrate.
- R&E networks want to understand their users and associated flows and optimize how they are served.
- Science collaborations are often unaware of the impact tuning or changes to their workflows have on the wide area network and the possible detrimental effects they are causing.



Review of Existing Network Monitoring and Management

The work to identify R&E network traffic is just part of a broader set of efforts in R&E networking monitoring and management for science. Here are a few to note:

- **WLCG data challenge monitoring:**
<https://monit-grafana.cern.ch/d/W2Uj1gDnz/wlcg-transfers-playground?orgId=20>
- The **Global Network Advancement Group** (GNA-g) working on traffic orchestration and computing model integration <https://www.gna-g.net/join-working-group/data-intensive-science/>
- **ESnet network monitoring** - ESnet created a monitoring page specifically for our WLCG Network Data Challenge:
<https://public.stardust.es.net/d/lkFCB5Hnk/lhc-data-challenge-overview?orgId=1>
- **The NetSage project** has LHC specific information
<https://lhc.netsage.global/grafana/d/xk26IFhmk/flow-data-for-circuits?orgId=2>
- **perfSONAR global deployment** and associated analytics
 - Details documented in many presentations; see last [LHCONE meeting](#)
- **Research Networking Technical Working Group** ([RNTWG](#))
 - Has working areas in packet marking / flow labeling, traffic shaping / packet pacing and network orchestration.



The SciTags Initiative

To manage our packet marking and flow labeling efforts, we started the **Scientific Network Tags** (scitags): an initiative promoting identification of the science domains and their high-level activities at the network level.

The initiative is managed by the RNTWG and is working to:

- Enable tracking and correlation of network transfers with Research and Education Network Providers (R&Es) network flow monitoring.
- Supporting collaborations to better understand network use and impact
 - 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.)
- Allow sites and end users to get detailed 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



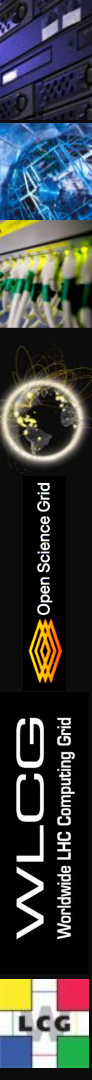
Scitags Overview

Our goal is to “instrument” any research network flows such that they are identifiable at any location along the path.

Challenges:

- **IPv4** has no good location in the packet header for marking
- **IPv6** has the “Flow Label” but our use is somewhat in tension with the RFC
- Linux kernels prior to 5.x lack good support to modify the Flow Label
- The node transferring the data needs to mark or label but this may require modifying all transfer software
- How does information about owner and purpose get to the transfer node?

To address these issues we have implemented a program of work which includes both “packet marking” (for IPv6) and “flow labeling” via “**Fireflies**” for IPv4 as well as a “*flowd*” service to encapsulate the needed capabilities via plugins

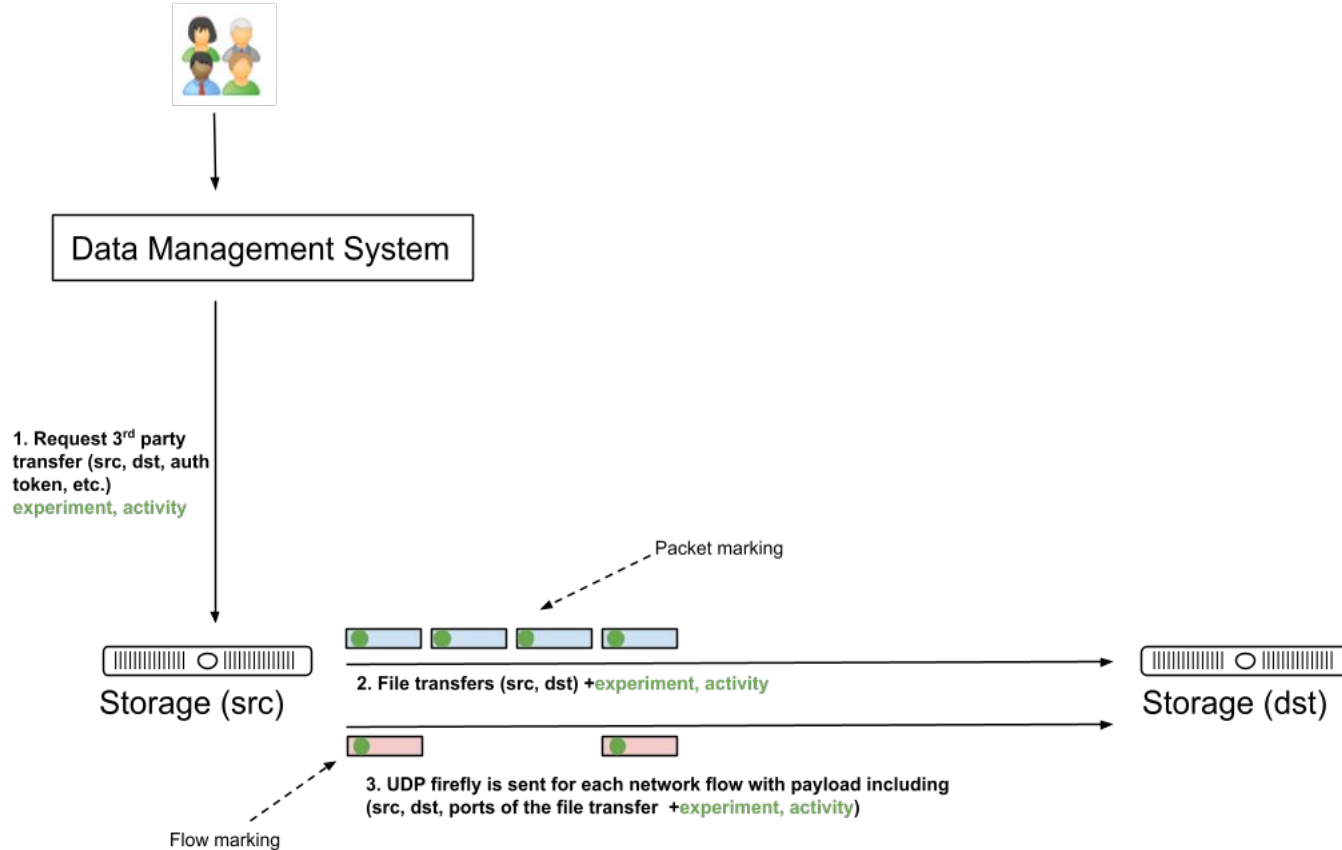


Technical Spec for Packet Marking/Flow Labeling

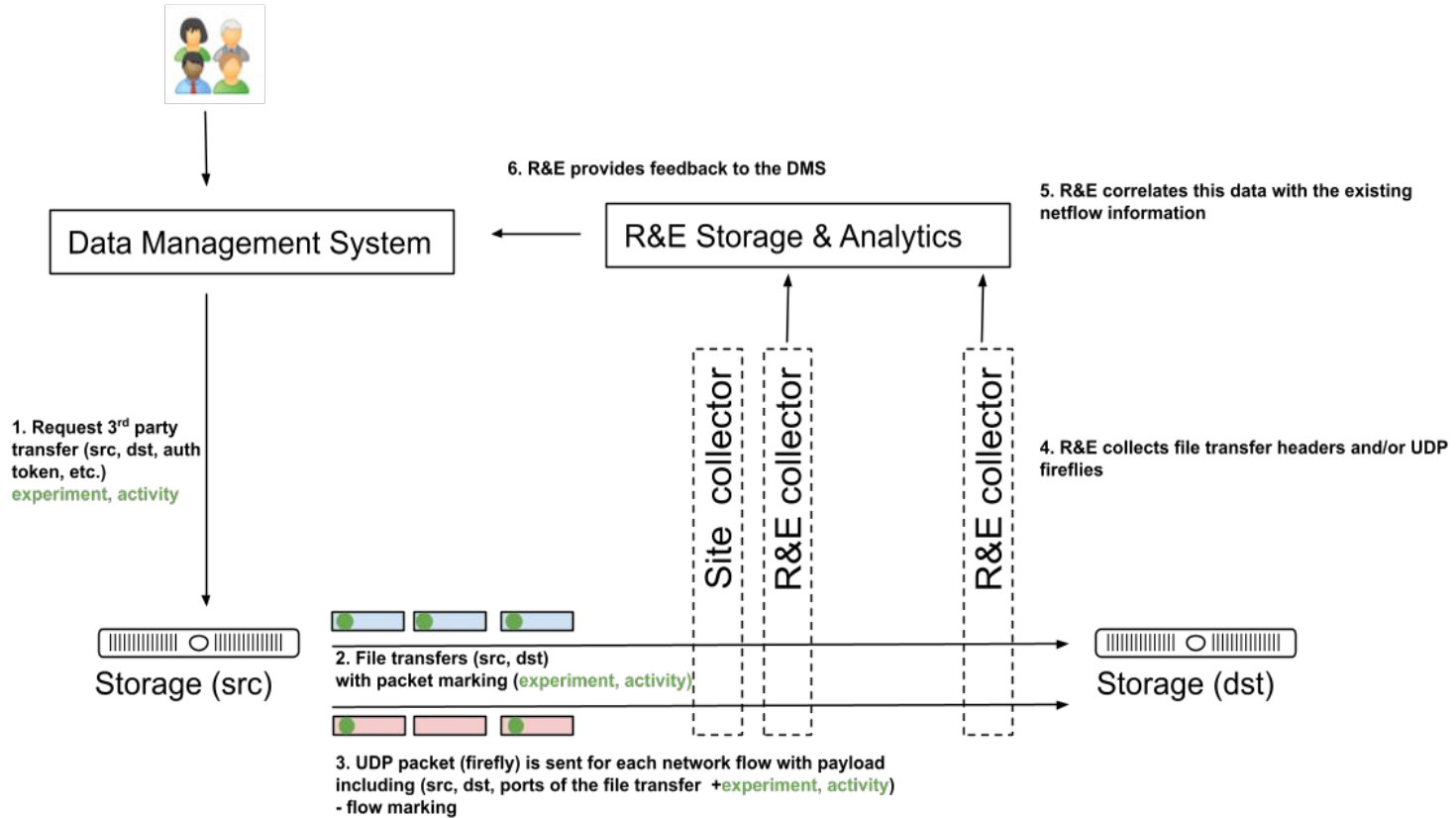
The detailed technical specifications are maintained on a [Google doc](#)

- 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.

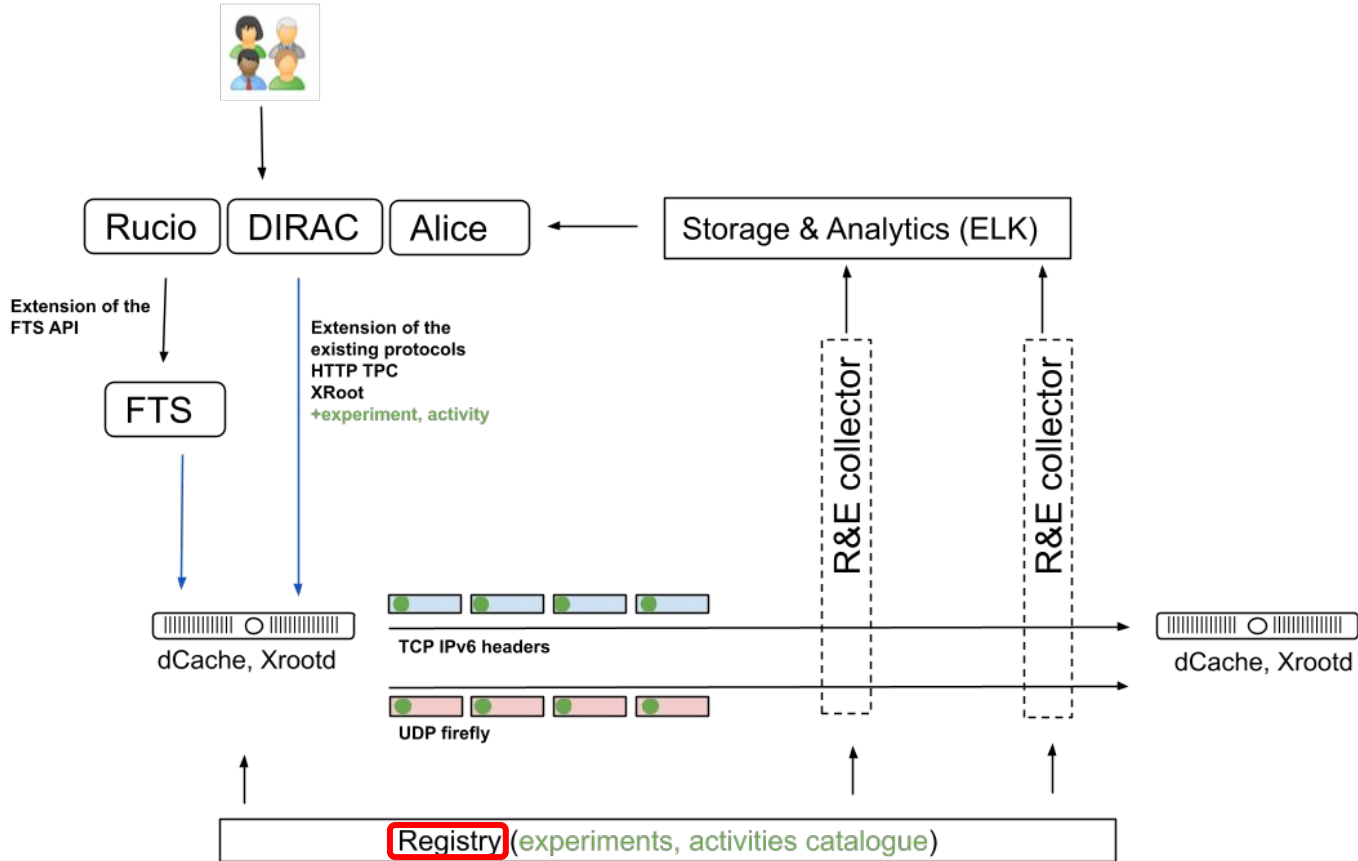
How scitags work



How scitags work



How scitags work



Registry Details

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

scitags.org

Network Flow and Packet Marking for
Global Scientific Computing



Technical Spec

Mailing List

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)

Presentations

Hosted on GitHub Pages — Theme by [orderedlist](#)

Status Summary

- Scitags initiative piloting two approaches
 - **Flow marking** - UDP firefly - marking flows using a covert channel
 - **Packet marking** - marking flows using IPv6 header flow label
- Technical specification is now available
 - Defines flow identifier as a tuple (experiment, activity)
 - Describes the entire lifecycle of the flow identifier
- Approach was validated to work during the last DC, i.e.
 - We were able to issue UDP fireflies, ESNNet run a collector, captured the fireflies and was able to correlate with their netflow data
- WLCG Experiments (+Belle II, LST, SKA and Dune) were contacted to provide list of activities that would be interested to track
 - In order to continue we need to suggest a technical approach for them to distribute the flow identifiers into the system (so Rucio -> FTS -> Storage) - draft in the technical spec.
- XRootd implementation (5.4+)
 - Support for UDP fireflies
 - Support for IPv6 header flow label (implemented but not released yet)



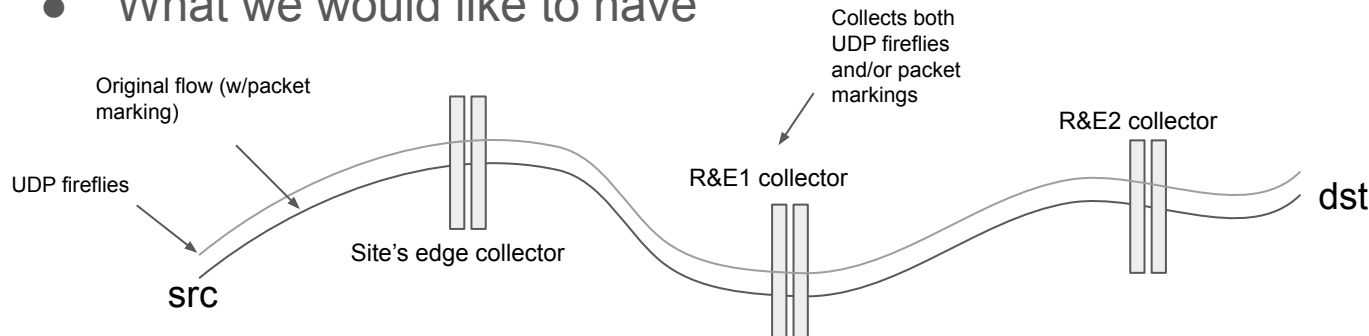
The Other Side of the Coin

To be useful, it is not enough to just put packet marking and flow labeling in place; we need to be able to **collect**, **analyze** and **display** the data.

We have discussed this in a few of our working group meetings and the following few slides will show our thinking and recommendations in this area.

Collector Requirements

- What we would like to have



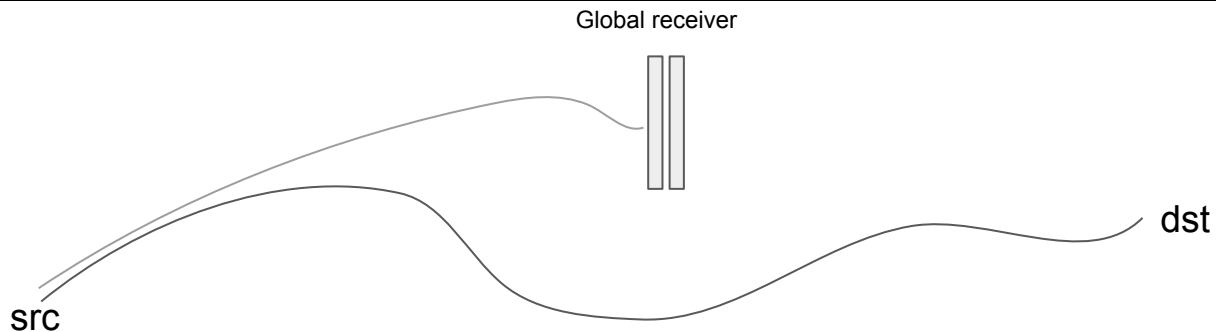
- Enable collection of packet and flow markings along the path
 - In order to extend R&E netflow information with flow identifier (experiment + activity)
 - UDP firefly packets needs to be collected and relayed to ensure they reach all collectors
- Each R&E can setup and operate one or more collectors
- Sites have an option to set up their own collector at the edge

Collector Design

Based on the requirements we can think of two types of collectors:

- **Hardware/inline collectors** which can collect **UDP fireflies and packet markings**
 - Can be two separate systems, but they're capable to collect/reflect packets on the wire
 - Collectors can work independently of each other (no functional dependencies btw collectors)
- **Software collectors (receivers)** can receive UDP fireflies and either store them or relay/propagate them further
 - Support only subset of activities (UDP fireflies) and can only try to approximate "collection along the path"
 - Receivers are endpoints which listen to UDP fireflies
 - They're missing the principal capability to "collect" UDP packets on the wire
 - Dependencies will likely exist between different receivers (wrt. availability, reachability, etc.)
- We will need to have a way to support both approaches, i.e. have a way how to provide gradual transition from one system to the other

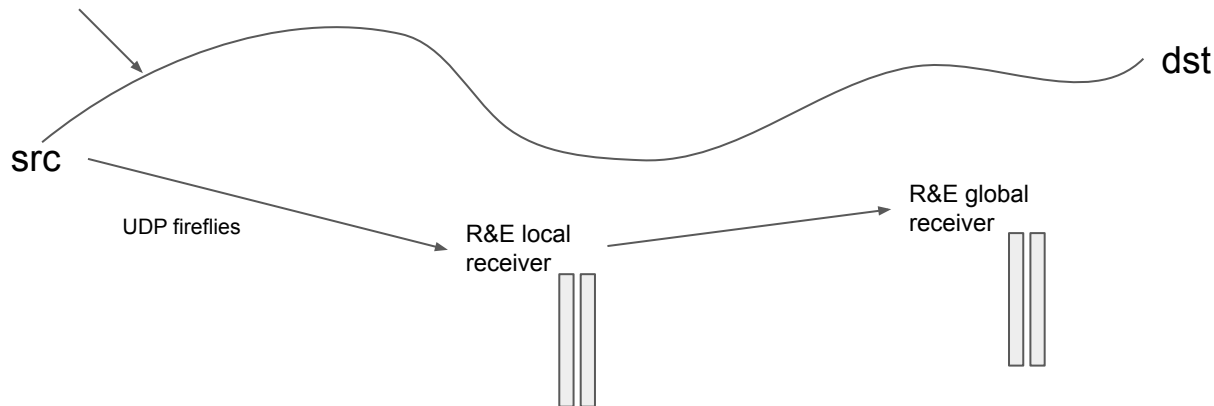
Topology: Global Receiver



- Keep just one global receiver
 - Assumes performance would be sufficient for the entire R&E traffic
 - Also assumes that some R&E would volunteer to host it and all the others would agree to use it under specific conditions (as defined by the host)
- Challenges
 - Reachability might be an issue (TCP might be a better option)
 - Unclear how to bring in R&E netflow data (privacy/security concerns)
 - Would require an interface to stream/download R&E traffic (how do we do this ? transit ?) in order for R&Es to perform correlation locally
 - Resulting data would need to be uploaded/streamed back
 - High availability/failover would be needed

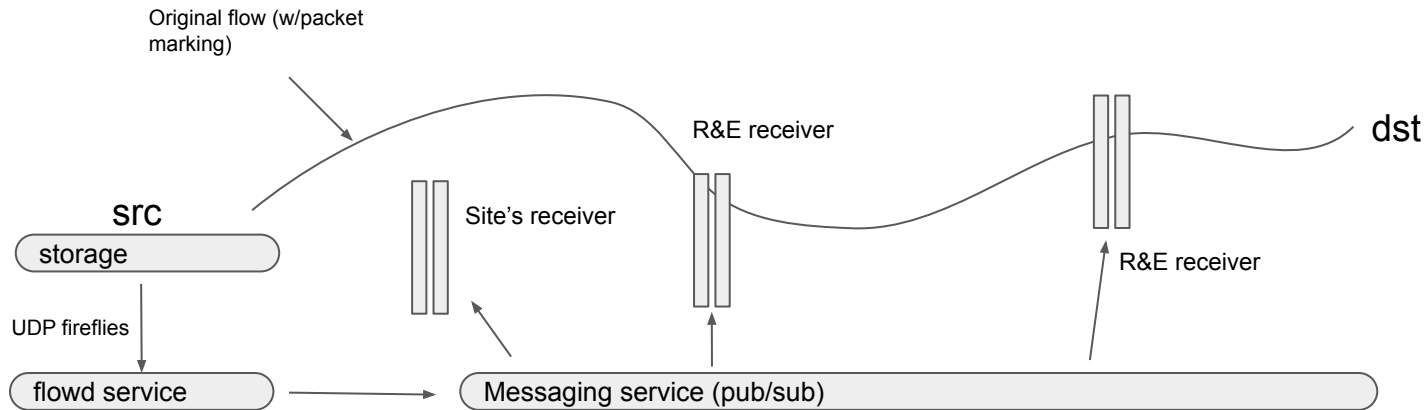
Network of Receivers - Anycast Design Draft

Original flow (w/packet marking)



- Src sends fireflies to an anycast address
- R&E can configure (via BGP) anycast address pointing to a local receiver (by giving it shorter distance)
- Local receiver will relay to a global one (while adding netflow + originating R&E id)
- Challenges:
 - Requires allocation of anycast address across all R&Es
 - Coordinated BGP configuration as well as hosting global receivers
 - Doesn't guarantee that local receiver will be used (can be complex to debug)

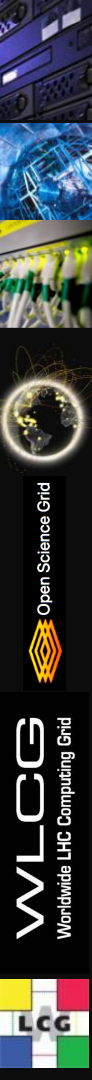
Network of Receivers - Messaging



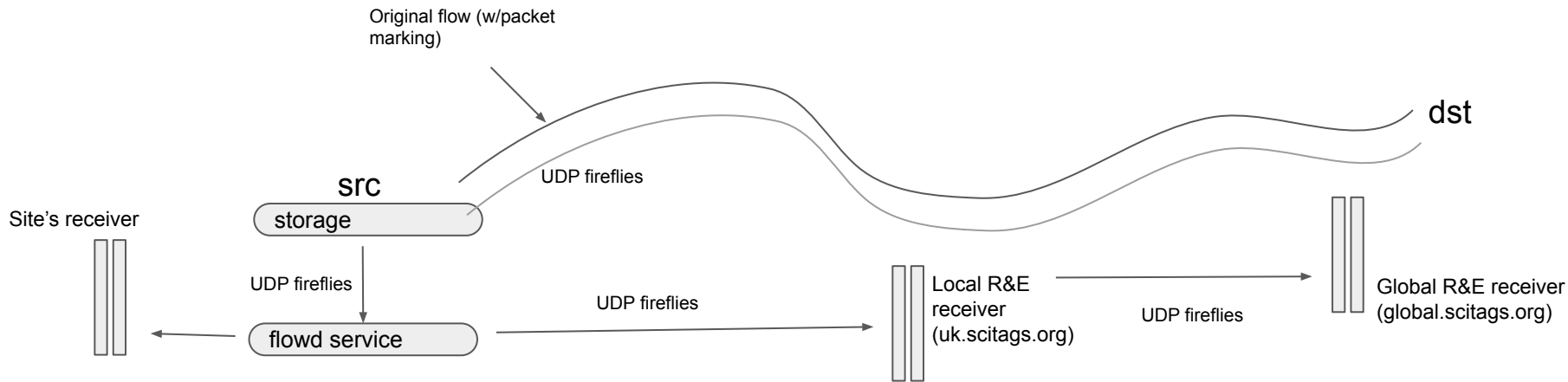
- Storage sends firefly to flowd, which then uses messaging service to distribute the fireflies
 - Messaging service would need to be hosted by R&Es
 - Can also support experiment-specific receivers
- Challenges
 - Cross domain messaging service is a **significant** challenge (multiple R&Es would need to host and coordinate)
 - High operational complexity

Recommendation

- Our recommendation is to use hardware/in-line collectors where possible
 - Requires port mirroring or other means to capture the fireflies
 - Easiest to organise and operate as there is no need for a separate collector network
 - Only way to capture flow markings along the path



Recommendation II - Network of Receivers



- Storages are configured with predefined DNS aliases (based on region; hosted by scitags.org)
 - Flowd service will expose API for site's local receivers and will also forward UDP fireflies to R&E collector (storage will send fireflies along the path)
 - Local R&E collector can be established (optional) and will need to pass all received fireflies to the global one (can switch to TCP)
- Works with inline/hardware collectors (which can be setup in parallel)
- Easy way to setup local R&E receiver (and correlate with local netflow)
- Lightweight - should be easy to operate, but requires some development in flowd and in the R&E collector
- DNS aliases will give us flexibility to make changes in the future (e.g. move to anycast)

Recent Work

As shown on previous slides, we have created a **flowd** service (see <https://github.com/scitags/flowd>) to more easily support packet marking and flow labeling, especially for applications that don't have direct socket access

- Firefly packets have recently be extended to include TCP linux stack statistics for the associated flow.
- A new eBPF (extended Berkeley Packet Filter) plugin was created by Tristan Sullivan to mark IPv6 packets.

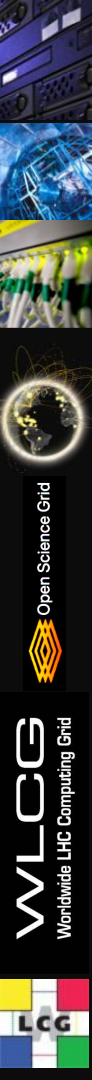
The working group has opened tickets and had focused meetings with our transfer tool and storage application providers to enable the whole software chain WCLG uses to support marking and labeling.

There has also been testing by University of Michigan undergraduate Austin Kim on the SLATE Kubernetes platform to determine what best practices should be for using these tools with containers.

Activities Till the Next Data Challenge

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 2024?):

- **Supercomputing 2022**: We have two demos planned showing packet marking and its accounting via P4 switches and flow labeling capture all using 100 Gbps interfaces.
- **Mini-Data Challenges**: We are participating with **WLCG DOMA**, **LHCONE/LHCOPN** and others in planning ~quarterly mini-data challenges starting in spring 2023 to ensure our deployment, tools, methods and monitoring are ready.
- **Application Integration**: We are working with the **Rucio** (distributed data management), **FTS** (data transfer) and storage systems (**Xrootd**, **dCache**) to enable the needed changes to support packet marking and flow labeling for WLCG experiments and other collaborations.



Conclusion

- 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 have large scale packet marking and flow labeling in place by the time of the next WLCG Data Challenge
- Please consider participating in the work!

Acknowledgements

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

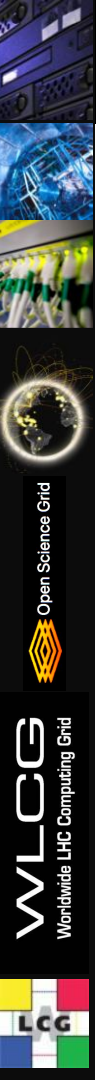
In addition we want to explicitly acknowledge the support of the **National Science Foundation** which supported this work via:

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



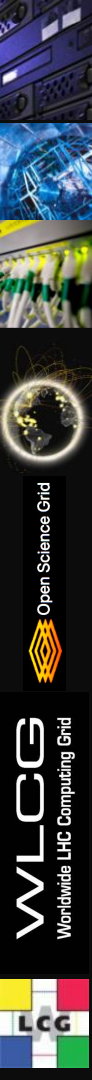
Questions / Discussion

Questions, Comments, Suggestions?



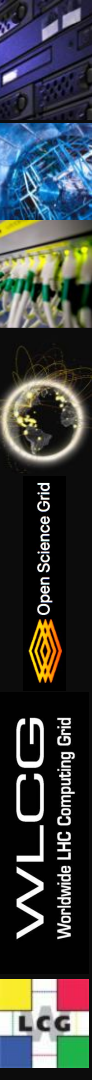
Useful Networking URLs

- OSG/WLCG Networking Documentation
 - <https://opensciencegrid.github.io/networking/>
- SciTags
 - <https://scitags.org>
- The RNTWG
 - <https://docs.google.com/document/d/1aAnsuipZnxn3oIU9JZxcw0ZpoJNVXkHp-Yo5oj-B8U/edit?usp=sharing>
- perfSONAR Central Configuration
 - <https://psconfig.opensciencegrid.org/>
- Toolkit information page
 - <https://toolkitinfo.opensciencegrid.org/>
- Grafana dashboards
 - <http://monit-grafana-open.cern.ch/>
- ATLAS Alerting and Alarming Service: <https://aaas.atlas-ml.org/>
- The pS Dash application: <https://ps-dash.uc.ssl-hep.org/>
- ESnet WLCG DC Dashboard:
<https://public.stardust.es.net/d/lkFCB5Hnk/lhc-data-challenge-overview?orgId=1>



Backup Slides Follow

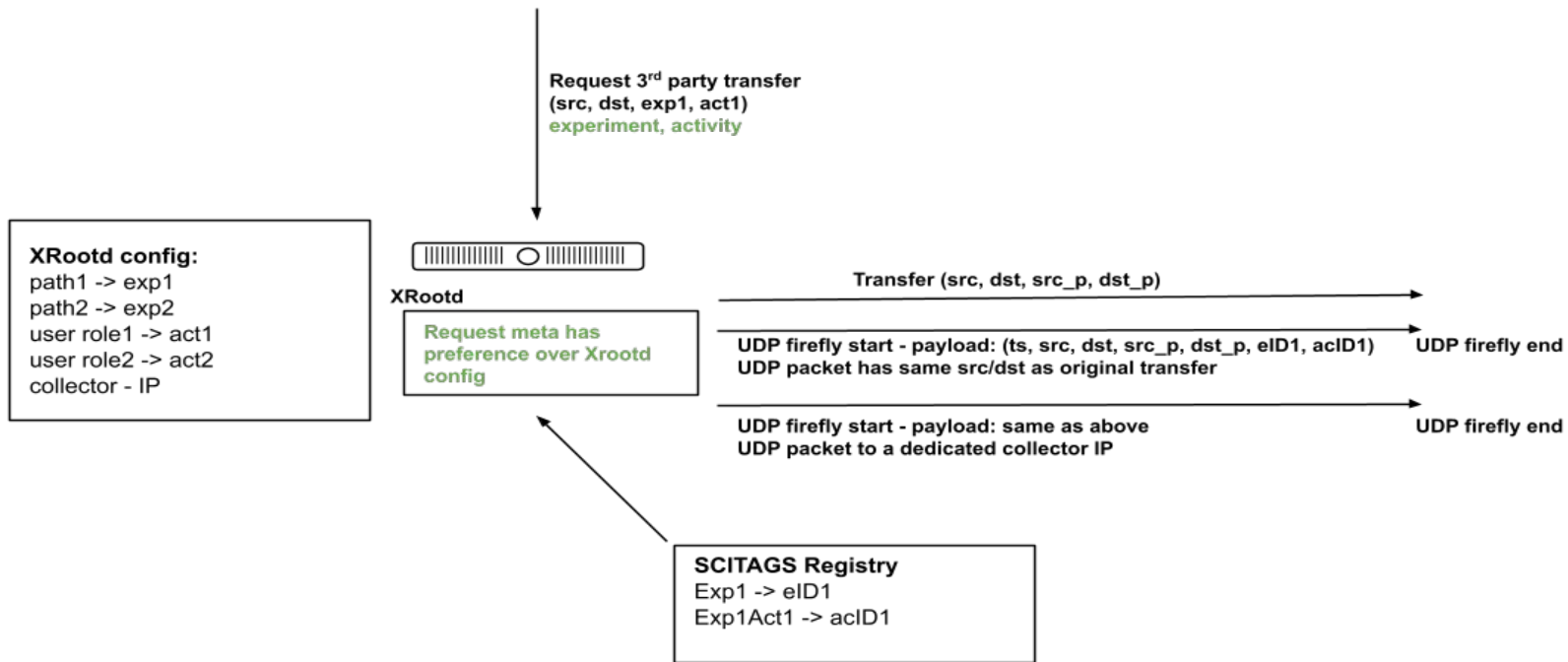
- **Flow Marking** (UDP firefly) implementations
 - Xrootd 5.4+ supports UDP fireflies
 - https://xrootd.slac.stanford.edu/doc/dev54/xrd_config.htm#_pmark
 - **map2exp** - can be used to map particular path to an experiment
 - **map2act** - can be used to map particular user/role to an activity
 - Flowd - 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 the last Data Challenge (& still operating)
 - Flowd + ESnet collector + Registry
 - **AGLT2, BNL, KIT, UNL and Caltech** participated
 - Brunel, Glasgow and QMUL interested to help with further testing



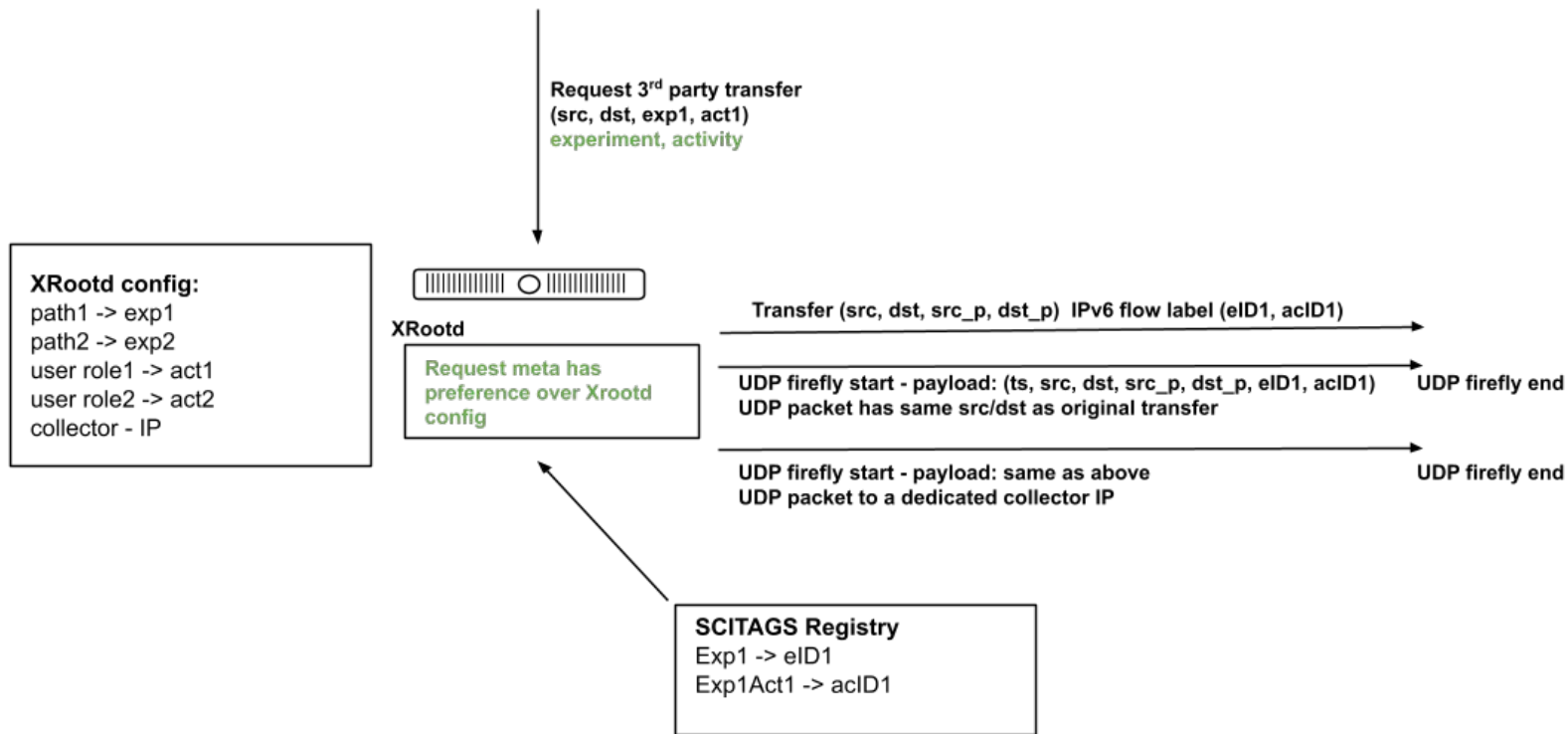
Flow identifier lifecycle

- Flow identifiers stored in registry
 - Stored and managed in a [Google Sheet](#)
 - api.scitags.org - JSON encoded list of experiments/activities
- Rucio
 - Already has both experiment and activity and is already passing this to the storage(s) for certain applications (ATLAS Data Carousel)
- FTS
 - Proposal is to add this as part of the file metadata (which is accessible via FTS REST API) or via protocols
- Propagation via protocols
 - HTTP TPC proposal
 - XRoot proposal

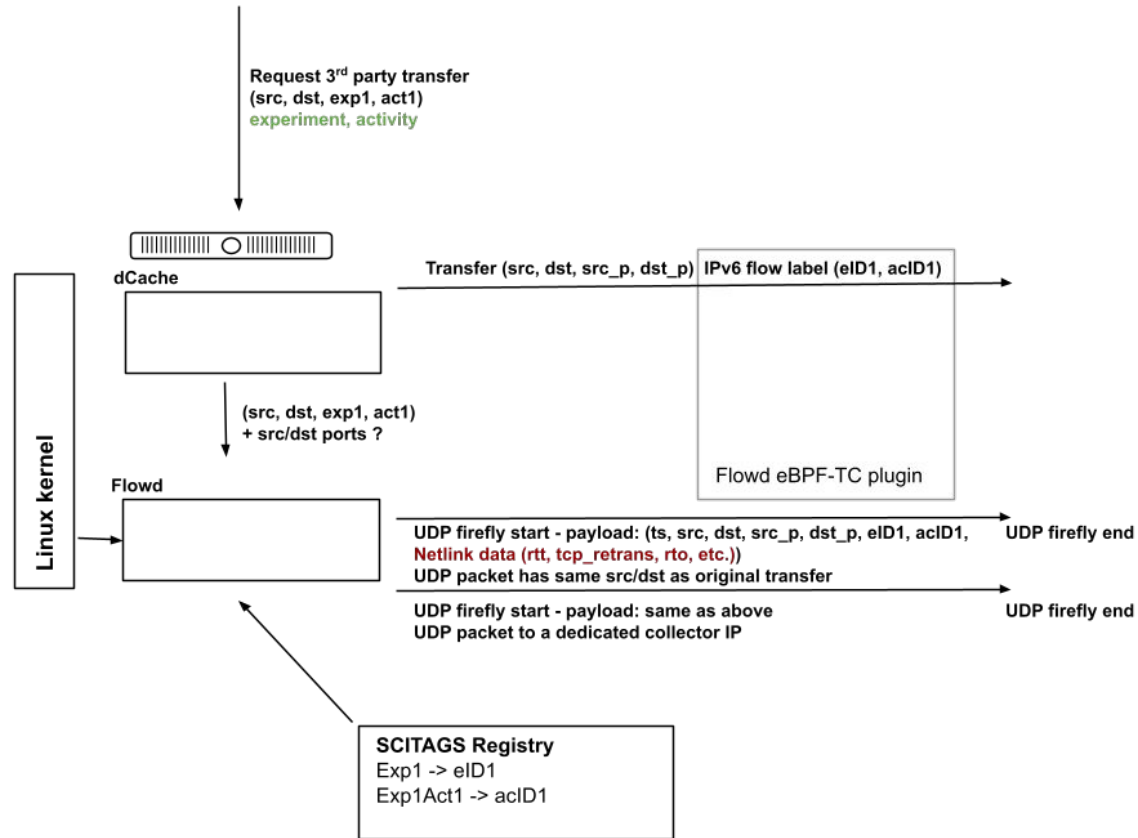
XRootd Implementation



XRootd Implementation (flow label)



dCache Implementation



ESnet Monitoring for WLCG Data Challenge

ESnet created a very nice [monitoring dashboard](#)

LHC Data Challenge / LHC Data Challenge Overview

Last 6 hours

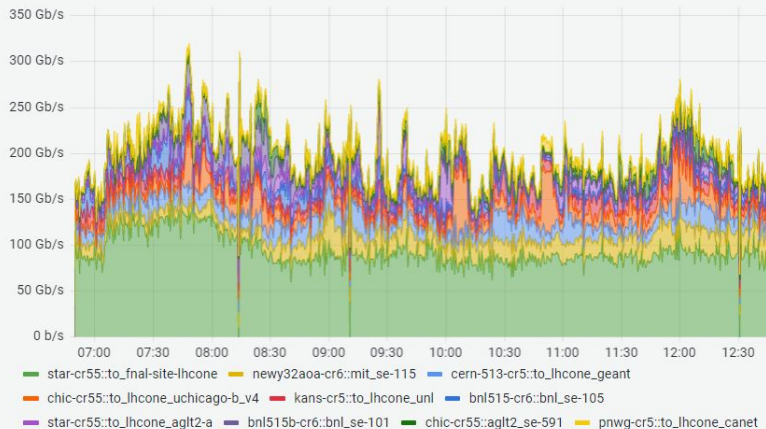
LHC Data Challenge Overview

Menu: Overview | Interfaces | Sites | Transatlantic | LHCOPN

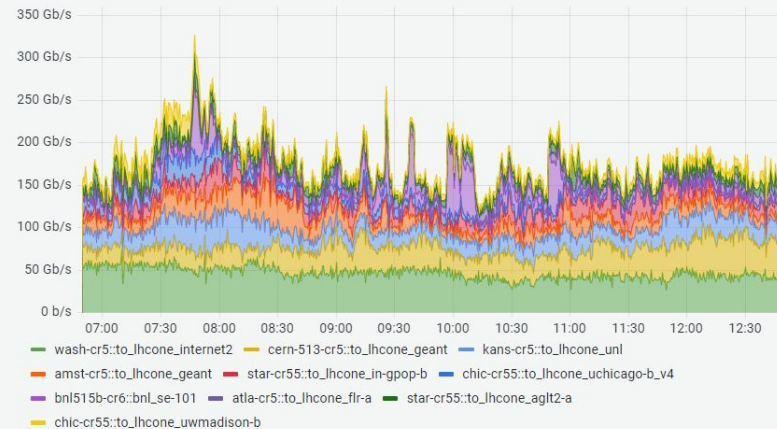
This dashboard shows an overview of statistics relevant to the LHC data challenge. It contains a combination of SNMP and flow statistics from ESnet's Stardust measurement system. Use the navigation menu above this text or links in the data below to move to other dashboards that provide different views of the data.

SNMP Statistics

Top 10 Interfaces by Incoming Rate (SNMP)



Top 10 Interfaces by Outgoing Rate (SNMP)



Top Interfaces by Incoming Volume (SNMP)

Top Interfaces by Outgoing Volume (SNMP)

WLCG Network Throughput Support Unit

Support channel where sites and experiments can report potential network performance incidents:

- Relevant sites, (N)RENs are notified and perfSONAR infrastructure is used to narrow down the problem to particular link(s) and segment. Also [tracking past incidents](#).
- Feedback to WLCG operations and LHCOPN/LHCONE community

Most common issues: MTU, MTU+Load Balancing, routing (mainly remote sites), site equipment/design, firewall, workloads causing high network usage

As there is no consensus on the MTU to be recommended on the segments connecting servers and clients, LHCOPN/LHCONE working group was established to investigate and produce a recommendation. (See coming [talk](#) :))

