



High-speed IPv6-only Scitag marking of CERN experiment traffic

Tim Chown, Jisc tim.chown@jisc.ac.uk



Acknowledgements

Worldwide Large Hadron Collider Computing Grid (WLCG) Research Networking Team (RNT WG) Scitag project members

- Includes but not limited to:
- Shawn McKee (University of Michigan Physics)
- Marian Babik (CERN)
- Tim Chown (Jisc)
- Andrew Hanushevsky (SLAC National Accelerator Laboratory)
- Andy Lake (ESnet)
- Tristan Sullivan (University of Victoria)
- Bruno Hoefft (Karlsruhe Institute of Technology (KIT))
- Dale Carder (ESnet)
- Garhan Attebury (UNL)
- Michael Lambert (Pittsburgh Supercomputing Center)
- Joe Mambretti (Northwestern University)
- Karl Newell (Internet2)
- Pablo Soto (UAM)
- et al.

What's the problem being addressed?

Identifying CERN Experiment and other science traffic

- (Inter)National R&E networks (NRENs) carry a large volume of science traffic
- It is useful to understand the nature of the traffic for many reasons:
 - Knowing the science **experiment** and network **activity** involved
 - Efficient network use, traffic steering, future provisioning and capacity planning
 - Performance measurement - flows, data rates, ...
 - Site traffic profiling, or traffic accounting for shared inter-continental links
- Current tooling to achieve this?
 - Network flows - but flow data **the IPs / ports in themselves don't reveal much**
 - Overlay networks - but a single L3VPN may have broad granularity, e.g., LHCONE

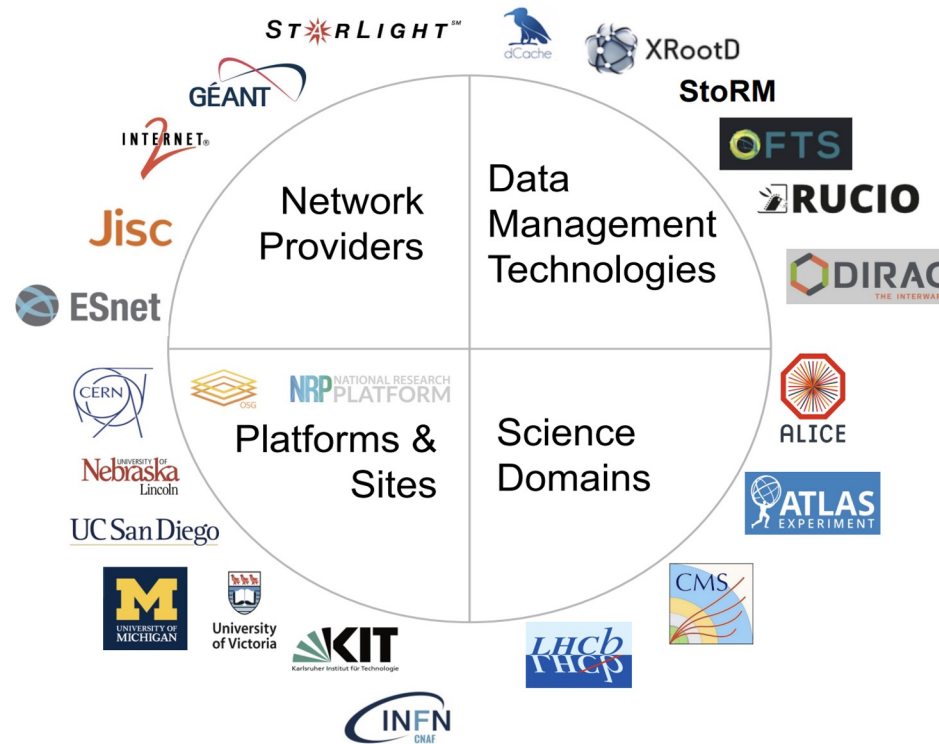
A solution - Scitags

Per-flow and per-packet marking - <https://scitags.org/>

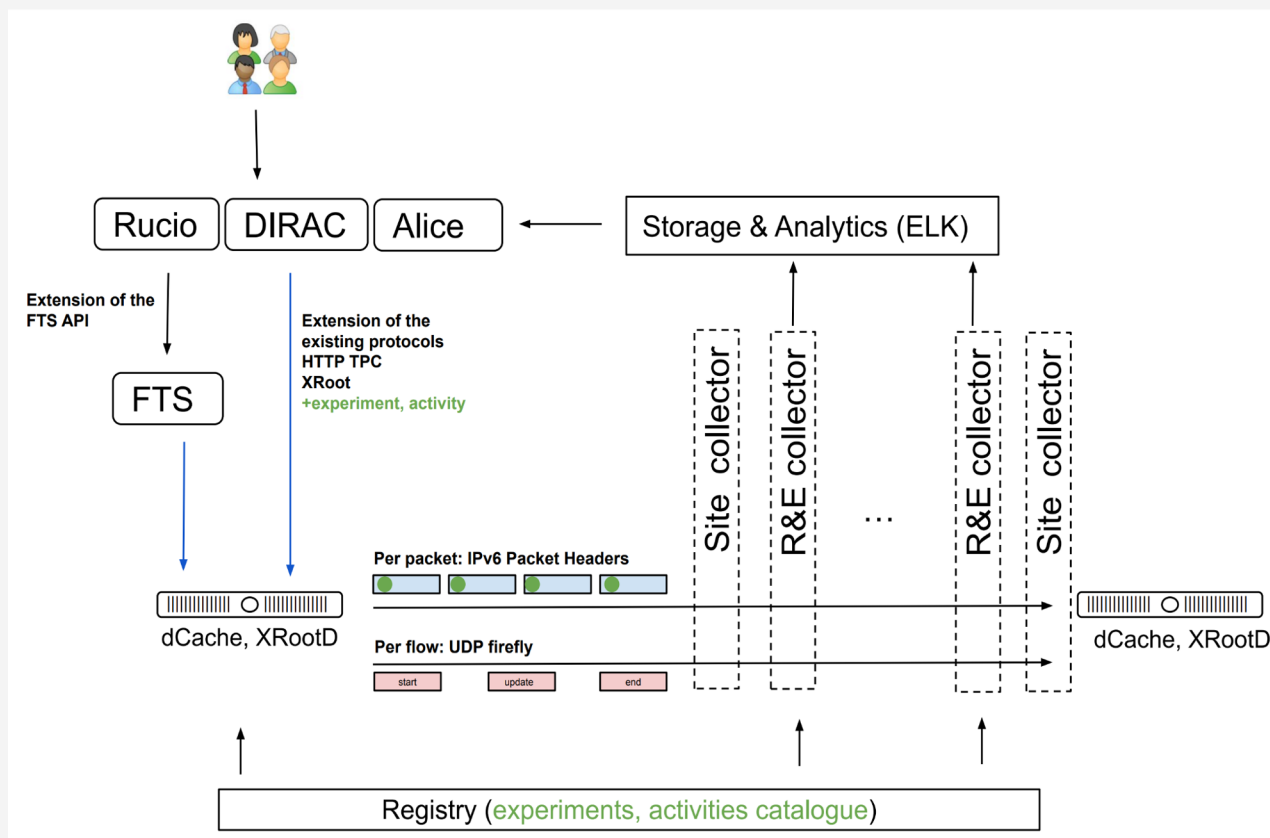
- **Open platform** available to any (data-intensive) science community
- **Identifies the owner (experiment) and purpose (activity) of traffic**
 - Uses a new, public registry of identifiers for experiments and activities
- Defines **standard(s)** for the exchange of such information between scientific communities, sites and network operators
 - By packet marking - encoding experiment/activity directly in packets
 - Uses IPv6 Flow Label header or an IPv6 Destination Option header (**so IPv6-only**)
 - By flow marking - sending a separate IPv4/IPv6 UDP packet (firefly) with metadata
 - Typically sent at the start and end of a flow to configured collector(s)
- Enables tracking/correlation with existing network flow monitoring tools
- Provides useful insights for the NRENs and projects / VOs

scitags.org

Network Flow and Packet Marking for
Global Scientific Computing



The architecture...



Per-flow marking

Fireflies

- The initial focus has been on per-flow fireflies
- Works for IPv4 and IPv6, just a UDP packet in syslog format
- Participating sites configure their software to send fireflies to a specific collector
- Collectors can share data via a Kafka message bus – the main global collector is at ESnet
 - A dashboard allows viewing and analysis of the firefly data, e.g., for throughput performance
- Support in the CERN experiment tooling is good
 - Rucio, DIRAC, ALICE, XRootD, EOS, StORM, dCache, ...
- Adoption by CERN experiments is wide
 - CMS, ATLAS, ALICE, LHCb
- And potential for adoption by other science projects and communities
 - SKA, Belle II, Vera Rubin, etc., especially when they adopt the WLCG tooling

Firefly coverage



Currently processing around 2M fireflies/day vs. 6-8M transfers/day seen in WLCG FTS

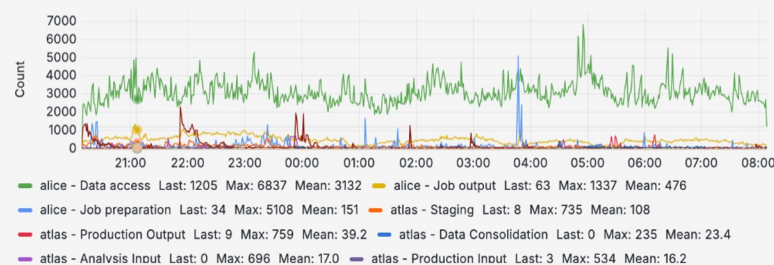
ESnet host the production [Scitag firefly dashboard](#)

Scientific Network Tags

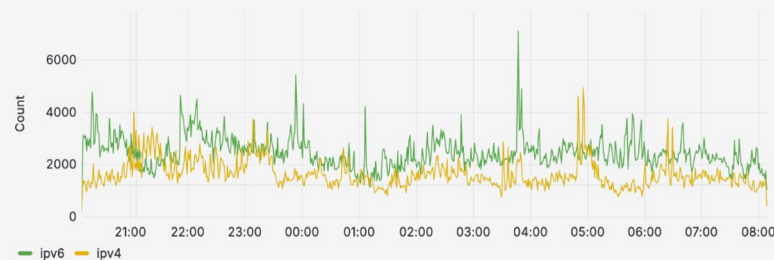
Menu: [Overview](#) | [Interfaces](#) | [Sites](#) | [Regionals](#) | [Transatlantic](#) | [LHCOPN](#) | Scientific Network Tags

This dashboard shows statistics related to flows marked with [Scientific Network Tag \(scitags\)](#) and sent to the ESnet firefly collector.

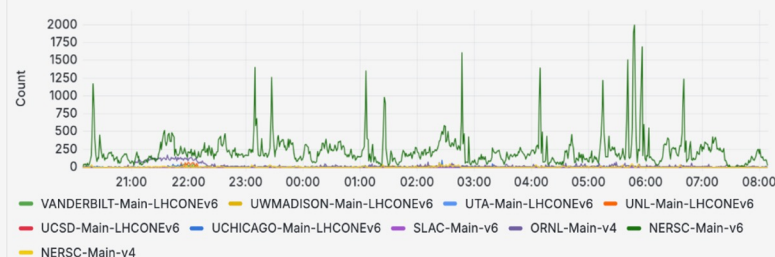
Total Flows per Exp/Act



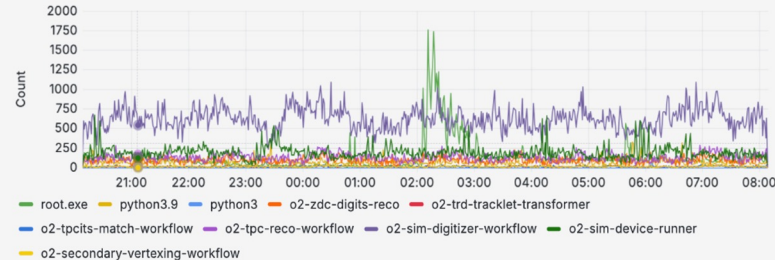
Total Flows per IP version



Total Flows per Prefix



Total Flows per Application (top10)



But what about per-packet marking?

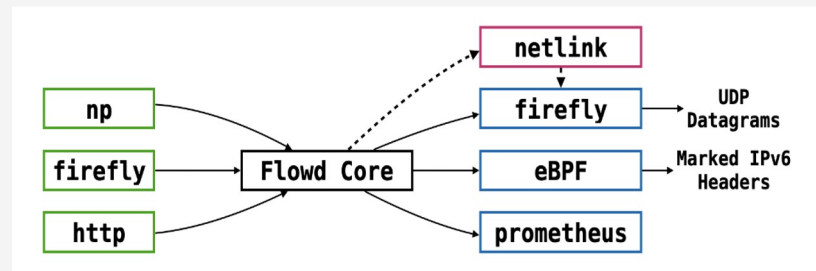
An IPv6-only approach...

- How can we leverage IPv6 for per-packet marking?
- Noting the WLCG traffic is already >90% IPv6 (see previous talks here by Dave Kelsey)
- Option 1 – the 20-bit Flow Label
 - Include 9 bits for Experiment, 6 bits for Activity, and 5 bits of Entropy
 - See <https://datatracker.ietf.org/doc/draft-cc-v6ops-wlwg-flow-label-marking/>
 - BUT the IPv6 spec requires the Flow Label doesn't use semantics, so...
- Option 2 – an IPv6 Hop-by-Hop or Destination Option extension header
 - Insert the HbH or DO in the packet directly, taking care with MTU
 - This implies a small performance penalty – overhead of the marking bits in the packet

Enter flowd

Packet Marking and Flow Labelling service

- Service and library to prototype and test various approaches to per-packet marking and firefly flow marking
- flowd-go – by Pablo Soto (UAM)
 - Pluggable architecture
 - Can work alongside storages
 - Portable eBPF deployment
 - Container and package distribution
- Can enrich fireflies with TCP/IP stack information
 - Supports export via both eBPF netlink APIs
- Supports packet marking with either Flow Label or HbH/DO EHs



Flowd implementation

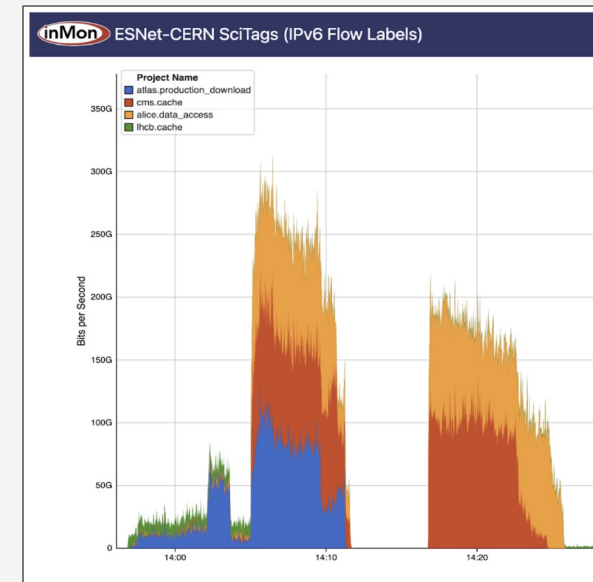
Specifics...

- The software is using 8-byte EHs for both HbH and DOs, using the type 0x1E
- Can combine both EHs in one datagram if desired, with HbH appearing before the DO
- The eBPF program is protected against MTU overflows – don't add the EH if would make the layer 2 frame larger than the output interface's MTU
 - But, e.g., you can configure MSS as MTU -20 -8 or MTU -20 -16
- The encoded data is 3 bytes long (experiment and activity – no need for entropy) + a single Pad1
- The eBPF avoids tc() overhead by using an appropriate eBPF program type
- The eBPF code is optimised at the assembly level to reduce memory accesses when populating the EHs
- The eBPF program communicates with user space through a hash where the key is a three-tuple of destination IPv6 address and source and destination port

Ongoing Scitag work

What's happening at the moment?

- Demo at SuperComputing - [SC25](#)
 - At up to 1.2Tbps (3x400Gbps data transfer nodes)
 - flowd-go, via Flow Label and EHs
- Enabling more sites, talking to other projects, deploying more collectors
- Exploring on-path capture/export of EHs
- Looking at testing EH transparency
 - Using the [perfSONAR toolkit](#) to mark iperf or latency/loss test traffic
- IETF 6man WG discussion on EH processing – **are they usable?**
 - The specs for processing EHs are being debated (again)
 - We have RFC 8200 (IPv6), RFC 8504 (node requirements), RFC 9673 (HbH processing)
 - A more controversial draft failed at IESG review - draft-ietf-6man-eh-limits-18



Finding more...

Code

Technical Spec

Mailing List

scitags.org

Network Flow and Packet Marking for Global Scientific Computing

[View On GitHub](#) [Download Tech. Spec](#) [Join scitags.org](#)

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 INTERNET.2 RNP Jisc XRootD dCache OITS RUCIO NORDUnet STARLIGHT CDS

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)

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

Presentations

Useful URLs

See...

[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>


Thank you

Tim Chown

Network Development Manager, Jisc

tim.chown@jisc.ac.uk

 help@jisc.ac.uk

 0300 300 2212

 jisc.ac.uk



Except where otherwise noted, this work is licensed under CC-BY-NC-ND.



 [@jisc.bsky.social](https://bsky.app/profile/@jisc.bsky.social)

 [@jiscsocial](https://www.instagram.com/jiscsocial)

 [linkedin.com/company/jisc](https://www.linkedin.com/company/jisc)

Jisc