BCNET CONNECT
HIGHER ED & RESEARCH TECH SUMMIT

Scitags: An Initiative to Improve R&E Networking Visibility

Tristan Sullivan

Randall Sobie
HEPNET, University of Victoria

Outline

- Background
- Packet Marking
- Flow Marking and flowd
- SC23 Demo
- DC24
- Future plans

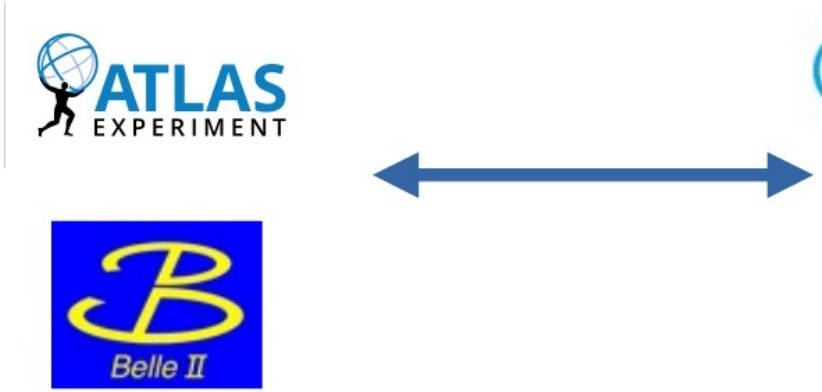
BCNET CONNECT

Background

Scitags: An Initiative to Improve R&E Networking Visibility

HEPNet

Acts as liaison between physics experiments and network providers (NRENs)



Director:
Randy Sobie
rsobie@uvic.ca

Technical contact:
Tristan Sullivan (me)
tssulliv@uvic.ca

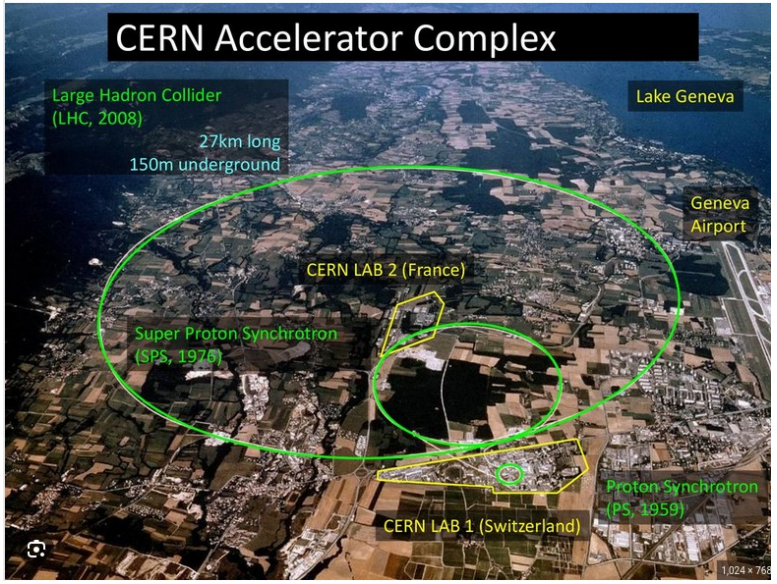
Contributes to network R&D projects

scitags.org

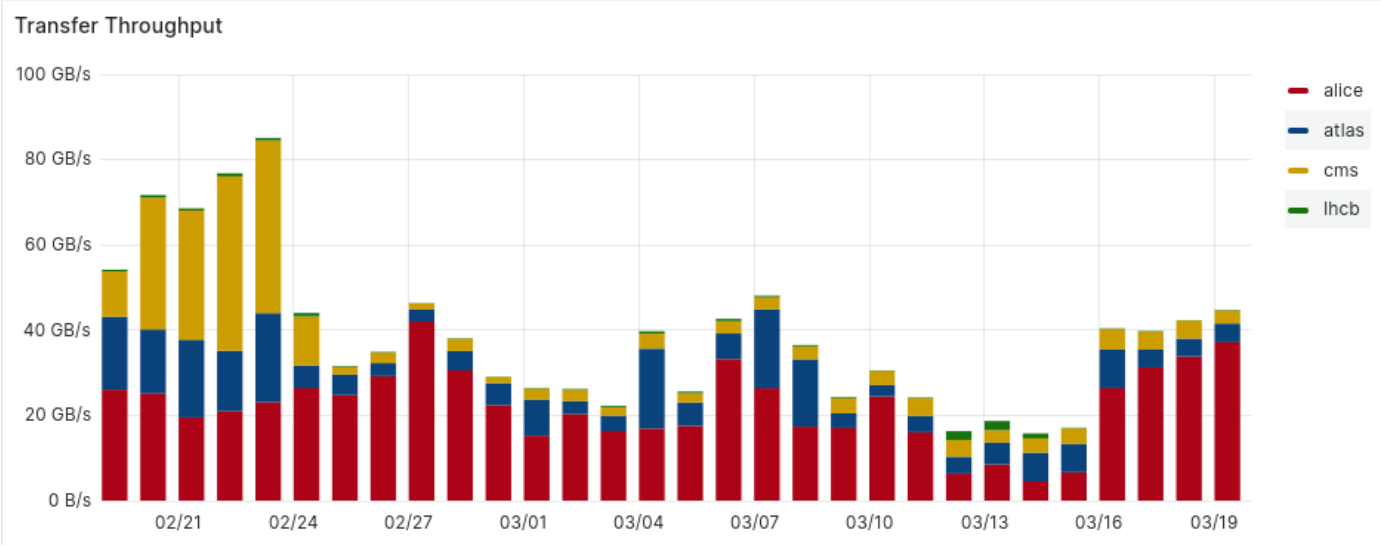
BCNET CONNECT

Background

- CERN: World's largest particle physics laboratory, located in Geneva, Switzerland
- Site of Large Hadron Collider (LHC)
- Hosts four major experiments (and many smaller ones): ATLAS, CMS, LHCb, ALICE
- WLCG: Worldwide LHC Computing Grid. Worldwide grid of about 200 computing sites in service of LHC experiments

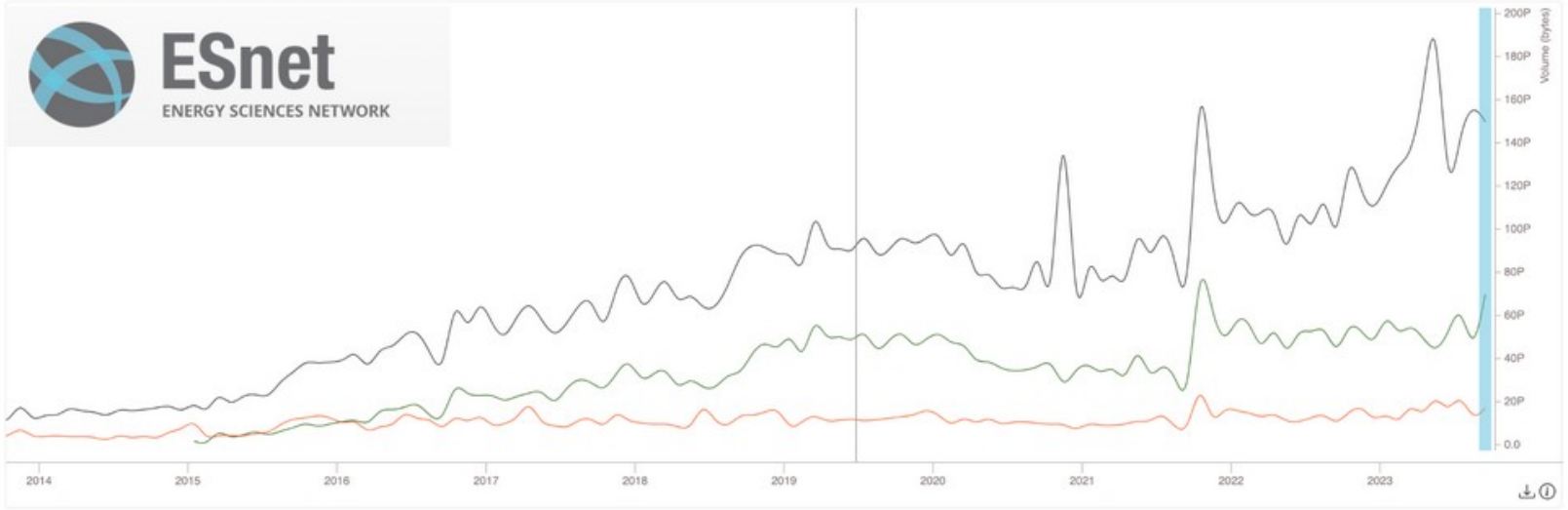


Transfers with CERN as source



Background

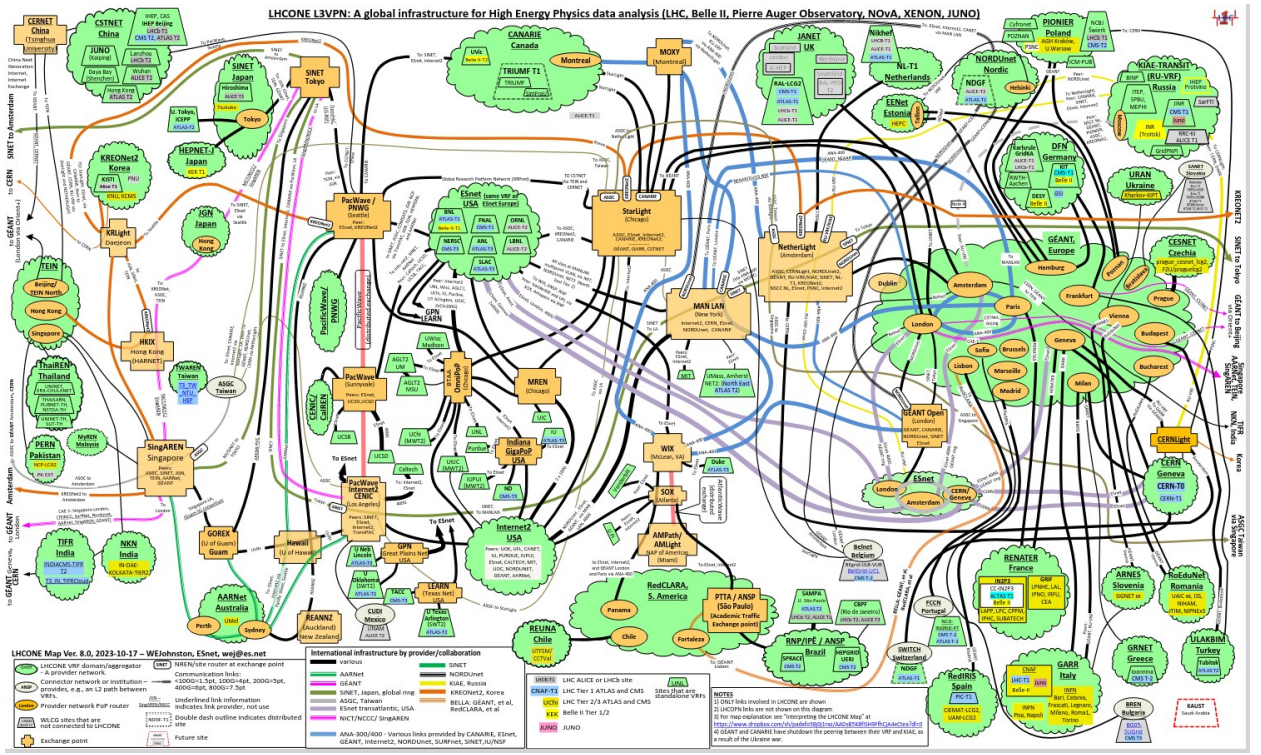
- Tier-1: Site that hosts raw data for one of the four main experiments. In Canada, only TRIUMF
- Tier-2: Site that provides computing and storage for processed data: UVic, SFU, Waterloo
- LHCOPN: International layer 2 network connecting CERN to “Tier-1” sites
- LHCONE: Layer 3 VRF overlay on R&E network, connecting CERN and Tier-1’s to “Tier-2” sites



	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%

Background

- LHCONE, in particular, is a large and complicated network crossing many international boundaries
- Difficult to correlate flows seen by NRENs with experiment data transfers
- Research Network Technical Working Group (RNTWG) started by CERN to address this challenge
- Concept of Scitags is to improve monitoring capability by marking traffic at the packet level with owner, activity



scitags.org

Packet Marking

Scitags: An Initiative to Improve R&E Networking Visibility

Packet Marking

- When the RNTWG was formed, multiple approaches were considered: using IPv6 address, hop-by-hop headers, flow label
- It was decided to focus on the flow label, for simplicity
- 20 bit field in the IPv6 header
- Several uses have been proposed, none universally adopted

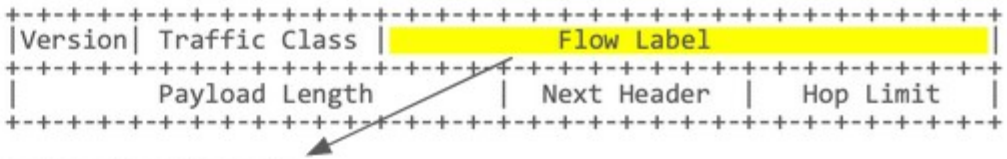
RFCs (10 hits)		
RFC 1809 **Using the Flow Label Field in IPv6**	1995-06 6 pages	Informational RFC
RFC 3595 *(was draft-ietf-ops-ipv6-flowlabel)* **Textual Conventions for IPv6 Flow Label**	2003-09 6 pages	Proposed Standard RFC
RFC 3697 *(was draft-ietf-ipv6-flow-label)* **IPv6 Flow Label Specification**	2004-03 9 pages	Proposed Standard RFC Obsoleted by RFC6437
RFC 6294 *(was draft-hu-flow-label-cases)* **Survey of Proposed Use Cases for the IPv6 Flow Label**	2011-06 18 pages	Informational RFC
RFC 6436 *(was draft-ietf-6man-flow-update)* **Rationale for Update to the IPv6 Flow Label Specification**	2011-11 13 pages	Informational RFC
RFC 6437 *(was draft-ietf-6man-flow-3697bis)* **IPv6 Flow Label Specification**	2011-11 15 pages	Proposed Standard RFC
RFC 6438 *(was draft-ietf-6man-flow-ecmp)* **Using the IPv6 Flow Label for Equal Cost Multipath Routing and Link Aggregation in Tunnels**	2011-11 9 pages	Proposed Standard RFC
RFC 7098 *(was draft-ietf-intarea-flow-label-balancing)* **Using the IPv6 Flow Label for Load Balancing in Server Farms**	2014-01 13 pages	Informational RFC

Document	Date	Status
Active Internet-Drafts (2 hits)		
draft-filsfils-6man-structured-flow-label-00 **Structured Flow Label**	2021-03-16 12 pages	I-D Exists New

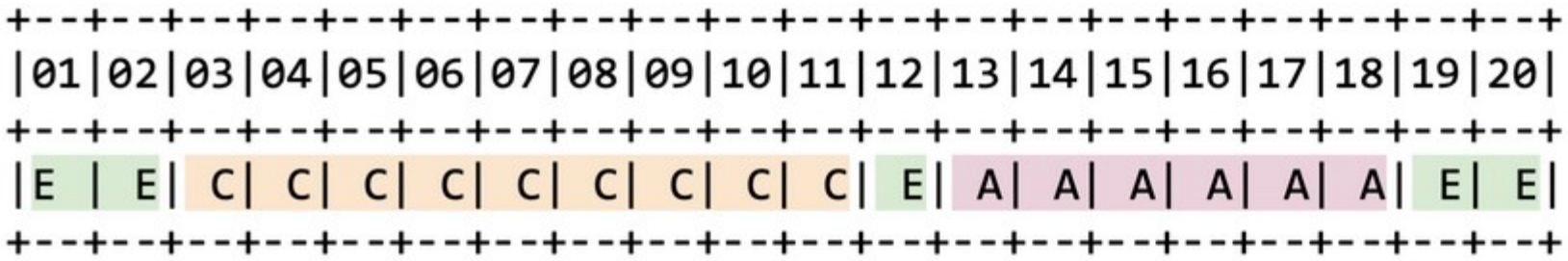
Flow label RFCs as of 2021

What To Mark Packets With?

IPv6 Header



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

What To Mark Packets With?

	LSB								MSB
ScienceDomain	Hdr Bit 14	Hdr Bit 15	Hdr Bit 16	Hdr Bit 17	Hdr Bit 18	Hdr Bit 19	Hdr Bit 20	Hdr Bit 21	Hdr Bit 22
	Bit 17	Bit 16	Bit 15	Bit 14	Bit 13	Bit 12	Bit 11	Bit 10	Bit 9
Reserved	0	0	0	0	0	0	0	0	0
DEFAULT	1	0	0	0	0	0	0	0	0
ATLAS	0	1	0	0	0	0	0	0	0
CMS	1	1	0	0	0	0	0	0	0
LHCb	0	0	1	0	0	0	0	0	0
ALICE	1	0	1	0	0	0	0	0	0
BelleII	0	1	1	0	0	0	0	0	0
SKA	1	1	1	0	0	0	0	0	0
DUNE	0	0	0	1	0	0	0	0	0
LSST	1	0	0	1	0	0	0	0	0
ILC	1	0	1	0	0	0	0	0	0
AUGER	0	1	0	1	0	0	0	0	0

Plenty of room for more projects

Mainly particle physics so far, also a couple of astronomy experiments

	MSB					LSB
Application	Hdr Bit 24	Hdr Bit 25	Hdr Bit 26	Hdr Bit 27	Hdr Bit 28	Hdr Bit 29
	Bit 7	Bit 6	Bit 5	Bit 4	Bit 3	Bit 2
Reserved	0	0	0	0	0	0
DEFAULT	0	0	0	0	0	1
perfSONAR	0	0	0	0	1	0
Cache	0	0	0	0	1	1
DataChallenge	0	0	0	1	0	0

Only activities common to all projects are listed in this table

Each experiment is also free to come up with their own list of up to 64 activities

How To Mark Packets?

What is eBPF?

Feature in the Linux kernel that allows user code to be injected into the kernel at runtime

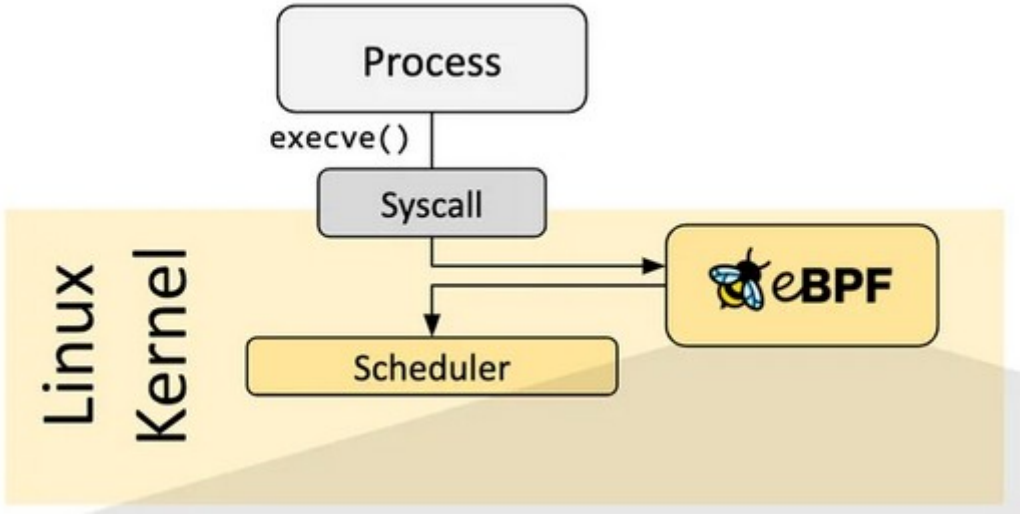
Can be a simpler replacement for kernel modules

How does it work?

First, the code is run through a verifier to ensure it is safe to execute

Then it is attached to some kernel hook

In our case, the code is run whenever a packet is sent



What does the code do?

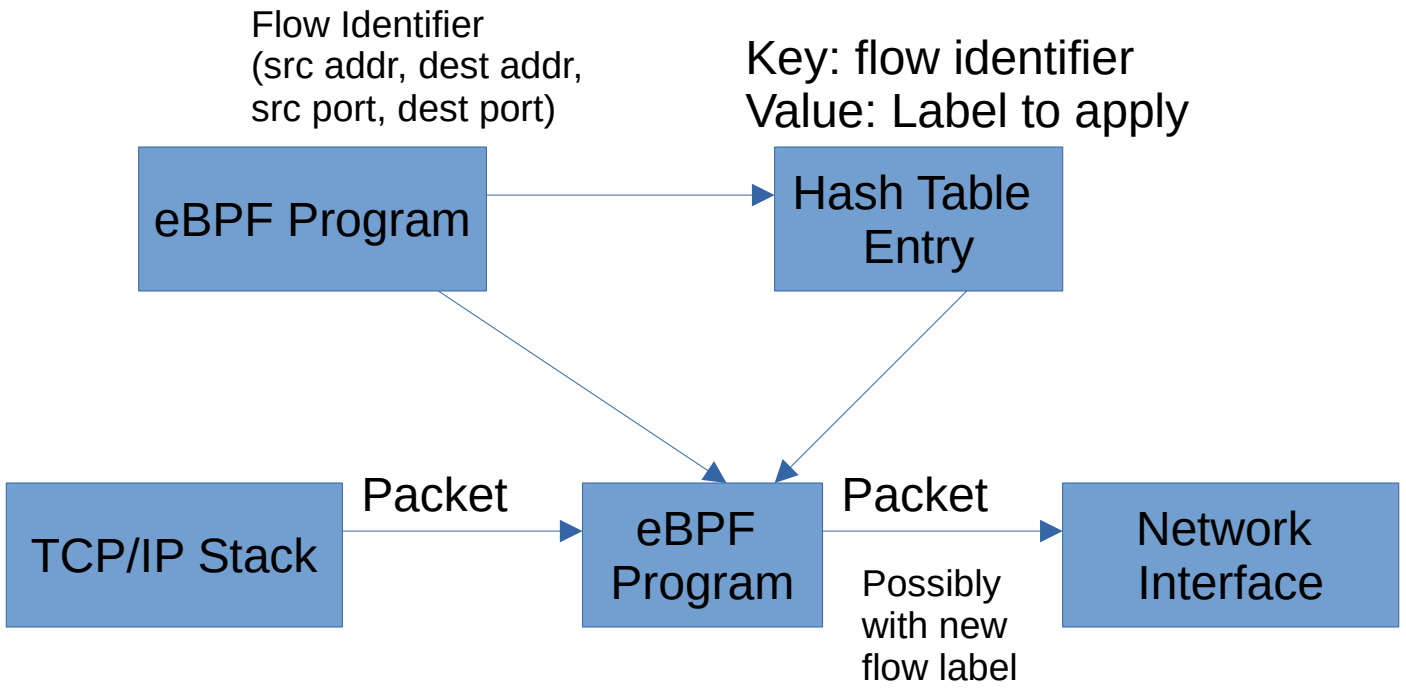
Flows to be marked identified by flow identifier: (src addr, dest addr, src port, dest port)

eBPF map: can pass information between userspace code and eBPF code

Our eBPF map is key/value pair

Flow identifier is key, label to be applied is value

Each packet is inspected; if flow identifier matches, eBPF program overwrites flow label with desired value

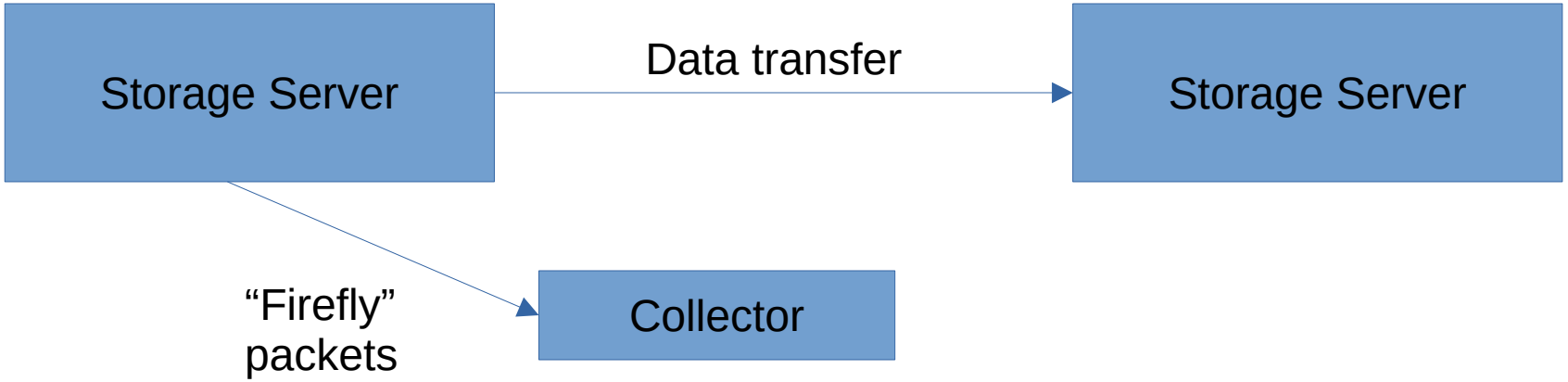


Flow Marking and Flowd

Scitags: An Initiative to Improve R&E Networking Visibility

Flow Marking

- Alternative to direct packet marking
- Sends information about flows to dedicated collector machine
- No requirement for kernel version or IPv6



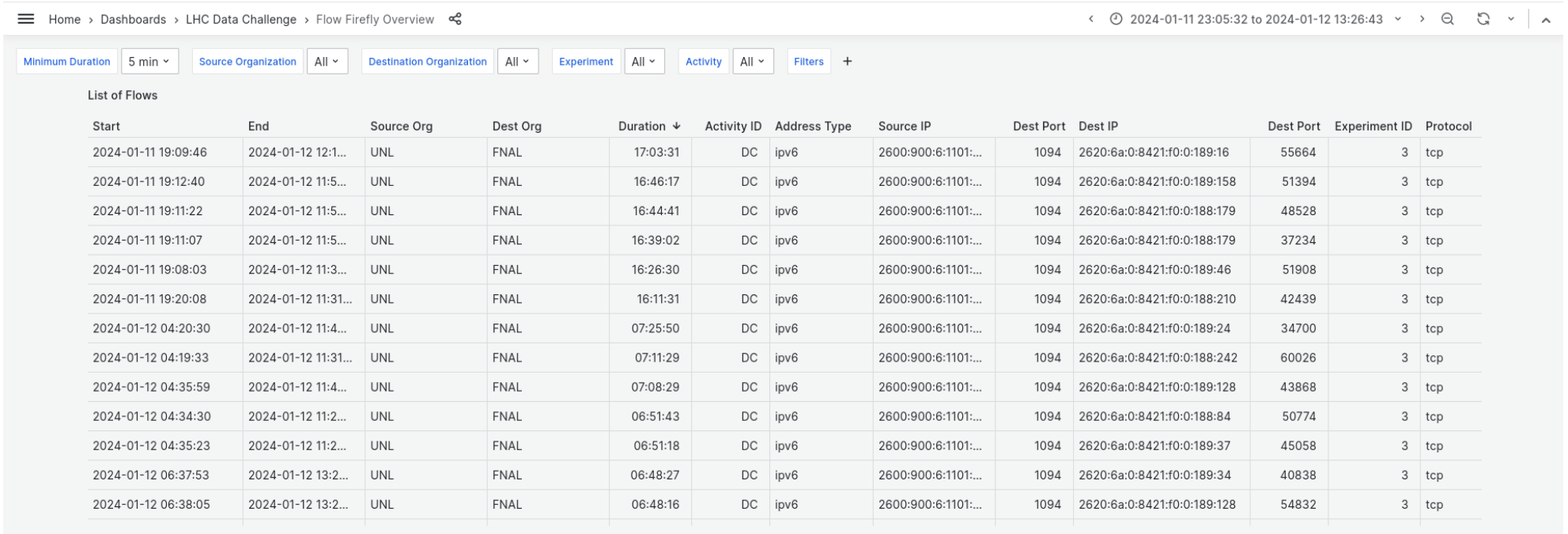
Fireflies are UDP packets in syslog format; one is sent at the beginning of the transfer, one at the end

Initial packet contains experiment and activity; other information can be added as well, no byte limit

Final packet also contains bytes transferred

Flow Marking Implementation

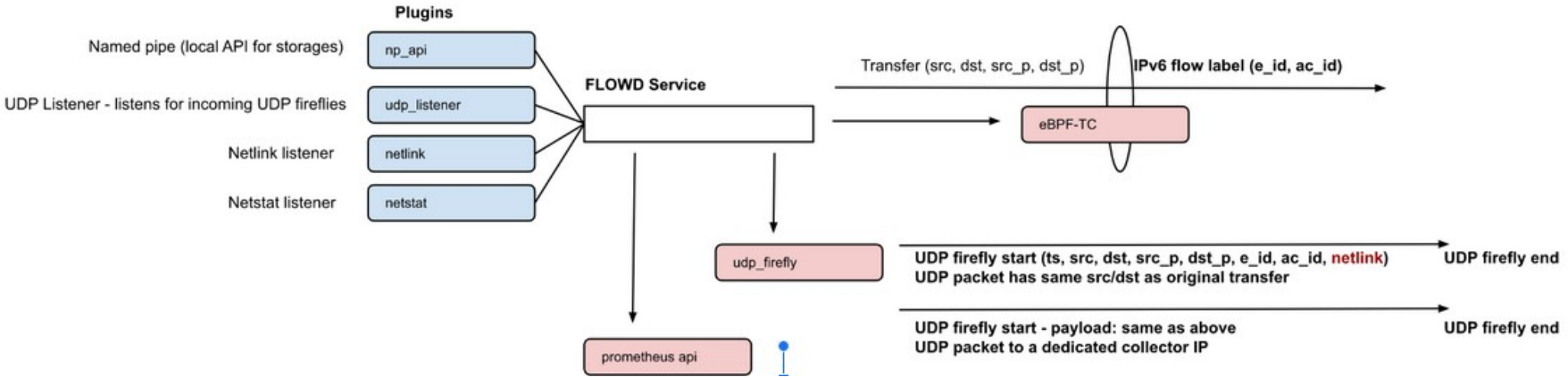
- Xrootd: one of the storage software packages used for particle physics
- UDP firefly supported added to a recent xrootd release
- ESNet set up dedicated collector machine
- In production at University of Nebraska WLCG computing site



Start	End	Source Org	Dest Org	Duration ↓	Activity ID	Address Type	Source IP	Dest Port	Dest IP	Dest Port	Experiment ID	Protocol
2024-01-11 19:09:46	2024-01-12 12:1...	UNL	FNAL	17:03:31	DC	ipv6	2600:900:6:1101:...	1094	2620:6a:0:8421:f0:0:189:16	55664	3	tcp
2024-01-11 19:12:40	2024-01-12 11:5...	UNL	FNAL	16:46:17	DC	ipv6	2600:900:6:1101:...	1094	2620:6a:0:8421:f0:0:189:158	51394	3	tcp
2024-01-11 19:11:22	2024-01-12 11:5...	UNL	FNAL	16:44:41	DC	ipv6	2600:900:6:1101:...	1094	2620:6a:0:8421:f0:0:188:179	48528	3	tcp
2024-01-11 19:11:07	2024-01-12 11:5...	UNL	FNAL	16:39:02	DC	ipv6	2600:900:6:1101:...	1094	2620:6a:0:8421:f0:0:188:179	37234	3	tcp
2024-01-11 19:08:03	2024-01-12 11:3...	UNL	FNAL	16:26:30	DC	ipv6	2600:900:6:1101:...	1094	2620:6a:0:8421:f0:0:189:46	51908	3	tcp
2024-01-11 19:20:08	2024-01-12 11:31...	UNL	FNAL	16:11:31	DC	ipv6	2600:900:6:1101:...	1094	2620:6a:0:8421:f0:0:188:210	42439	3	tcp
2024-01-12 04:20:30	2024-01-12 11:4...	UNL	FNAL	07:25:50	DC	ipv6	2600:900:6:1101:...	1094	2620:6a:0:8421:f0:0:189:24	34700	3	tcp
2024-01-12 04:19:33	2024-01-12 11:31...	UNL	FNAL	07:11:29	DC	ipv6	2600:900:6:1101:...	1094	2620:6a:0:8421:f0:0:188:242	60026	3	tcp
2024-01-12 04:35:59	2024-01-12 11:4...	UNL	FNAL	07:08:29	DC	ipv6	2600:900:6:1101:...	1094	2620:6a:0:8421:f0:0:189:128	43868	3	tcp
2024-01-12 04:34:30	2024-01-12 11:2...	UNL	FNAL	06:51:43	DC	ipv6	2600:900:6:1101:...	1094	2620:6a:0:8421:f0:0:188:84	50774	3	tcp
2024-01-12 04:35:23	2024-01-12 11:2...	UNL	FNAL	06:51:18	DC	ipv6	2600:900:6:1101:...	1094	2620:6a:0:8421:f0:0:189:37	45058	3	tcp
2024-01-12 06:37:53	2024-01-12 13:2...	UNL	FNAL	06:48:27	DC	ipv6	2600:900:6:1101:...	1094	2620:6a:0:8421:f0:0:189:34	40838	3	tcp
2024-01-12 06:38:05	2024-01-12 13:2...	UNL	FNAL	06:48:16	DC	ipv6	2600:900:6:1101:...	1094	2620:6a:0:8421:f0:0:189:128	54832	3	tcp

Flowd

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

Supercomputing 23 Demo

Scitags: An Initiative to Improve R&E Networking Visibility

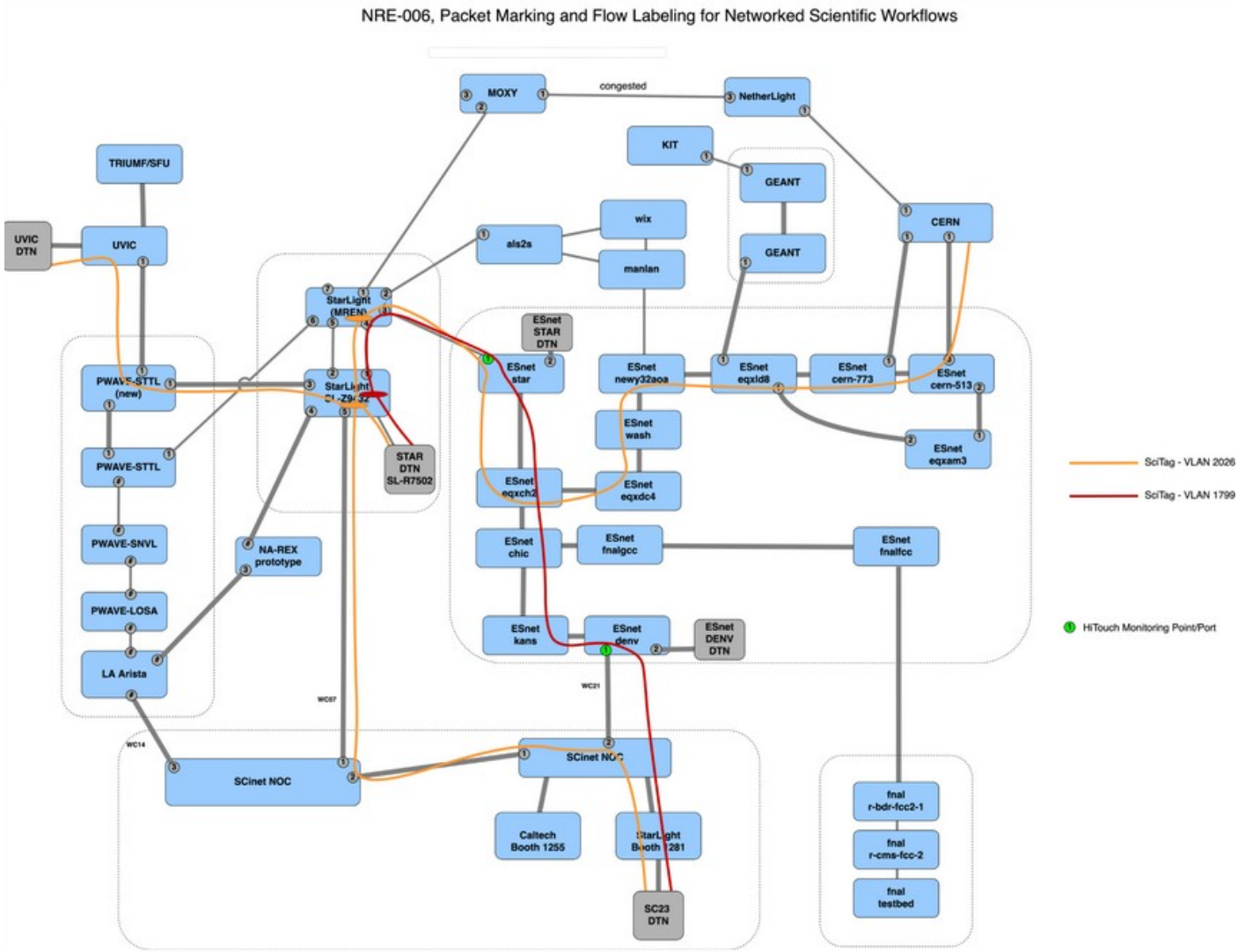
Supercomputing 23

International Conference for High Performance Computing, Networking, Storage and Analysis, is the annual conference established in 1988 by the Association for Computing Machinery and the IEEE Computer Society

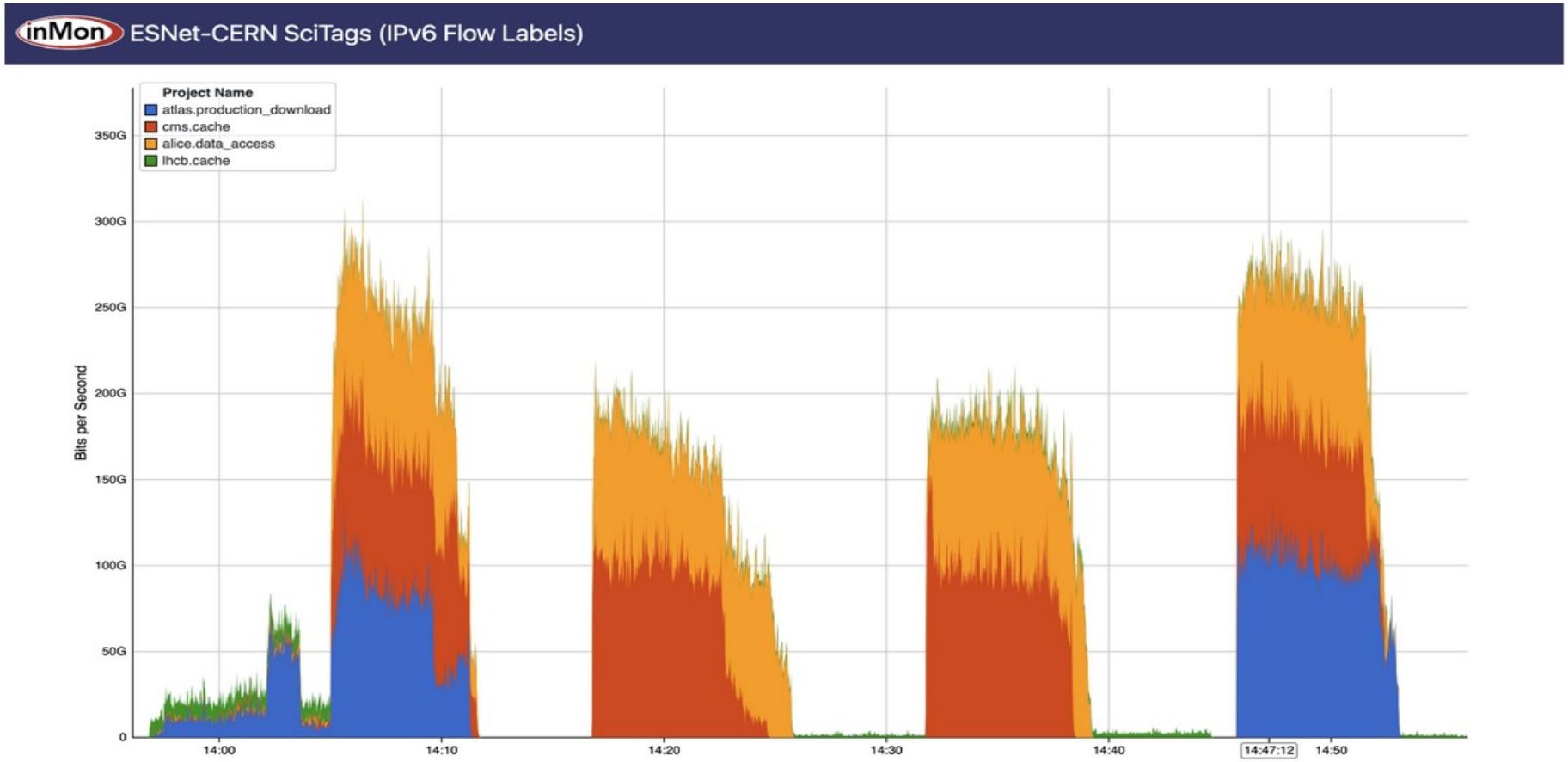
Participating sites in Scitags demo: CERN, UVic, SFU, Starlight, SC23 show floor (Denver)

UVic and SFU Servers supplied by Dell and Lenovo

400G NICs supplied by Nvidia

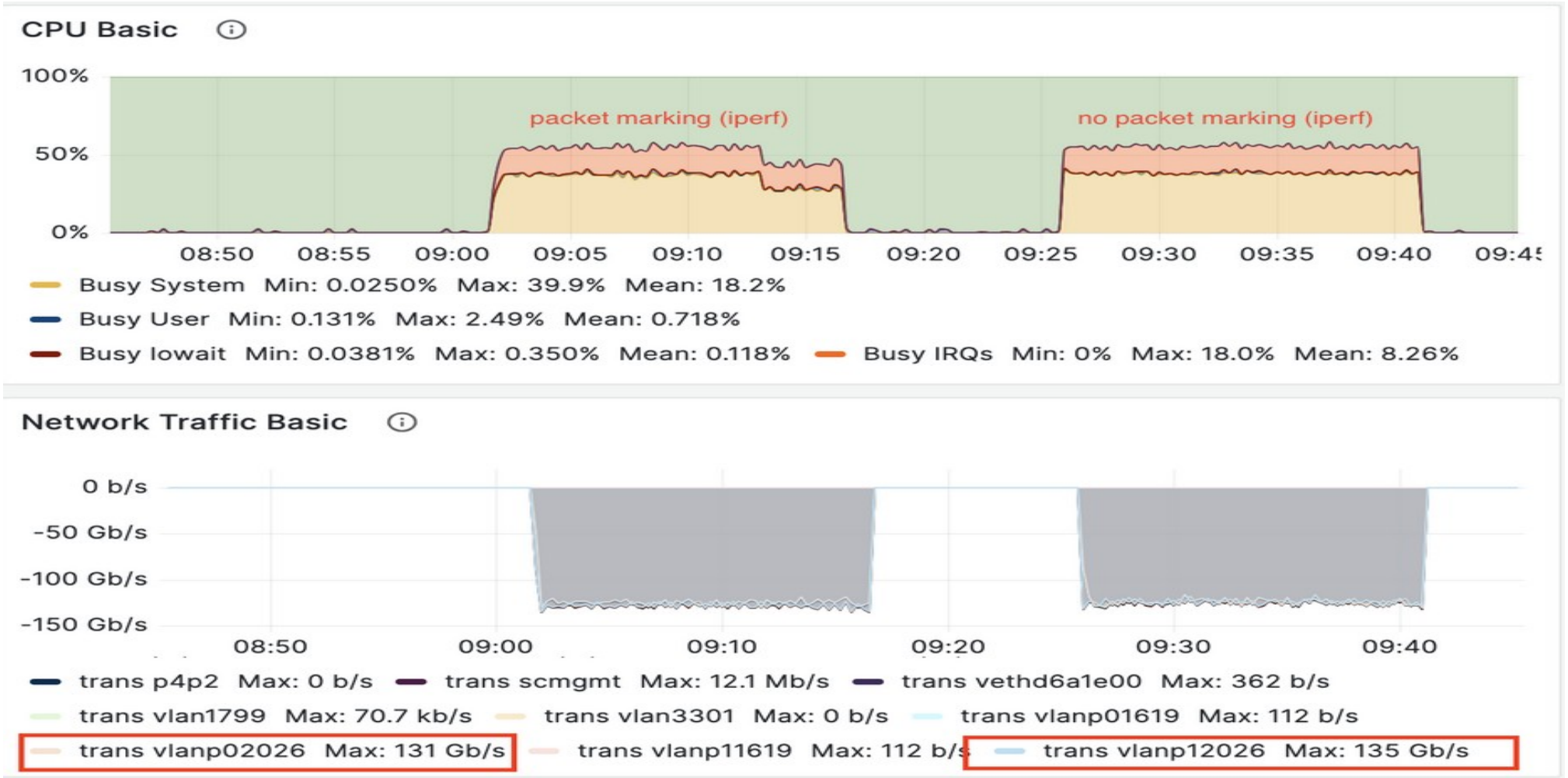


Supercomputing 23



Data collected from switch on show floor using sflow

Supercomputing 23



BCNET CONNECT

DC24

Scitags: An Initiative to Improve R&E Networking Visibility

Data Challenge 24

- Data rates from CERN experiments expected to increase by factor of ~10 in 2029
- CERN has organized a series of “data challenges” to ensure the infrastructure can handle the increased rate
- DC21: 10%
- DC24: 25%
- DC26: 50%
- DC28: 100%

Data Challenge 24

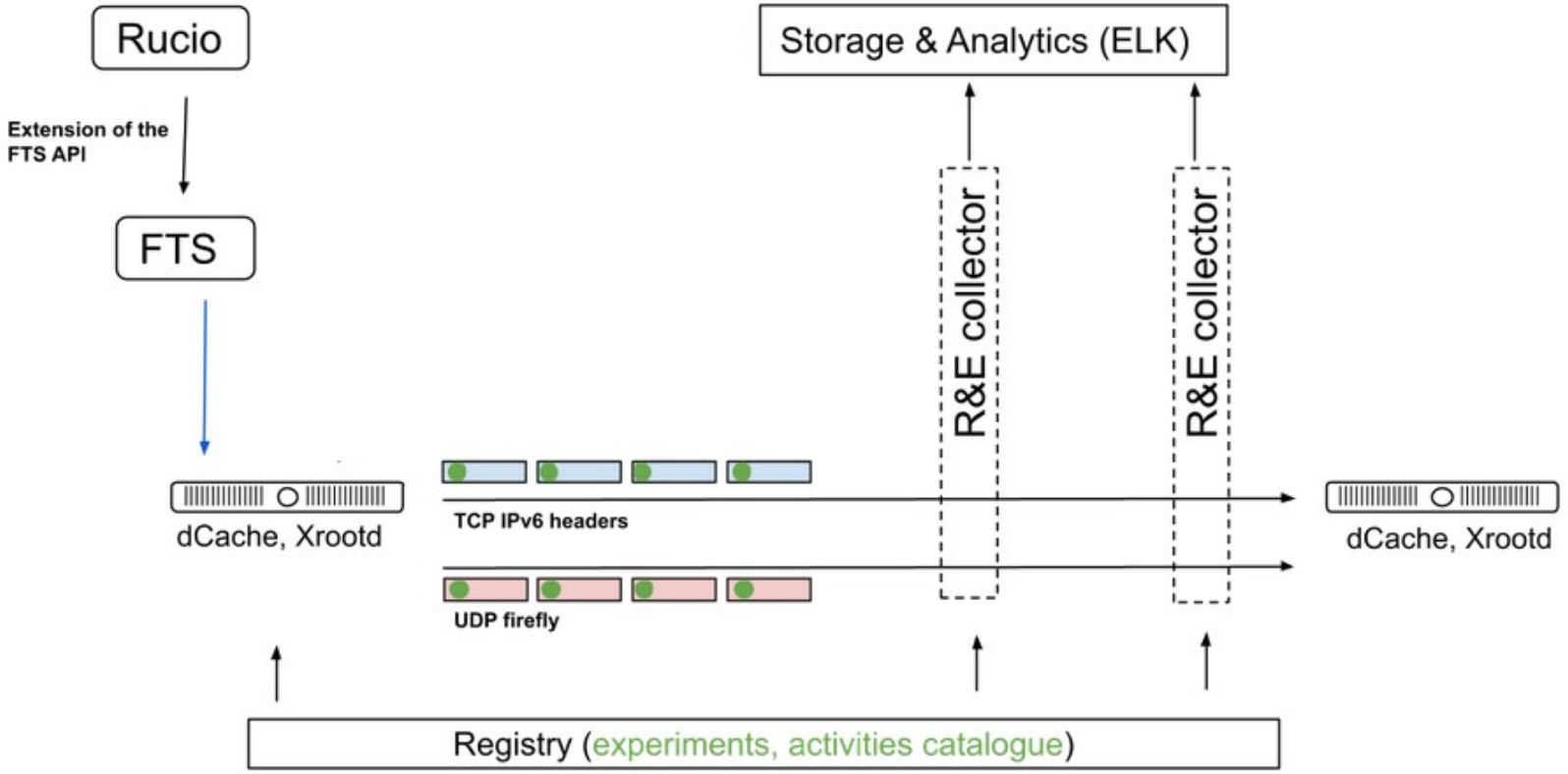
Rucio: data management software

FTS: File Transfer Service

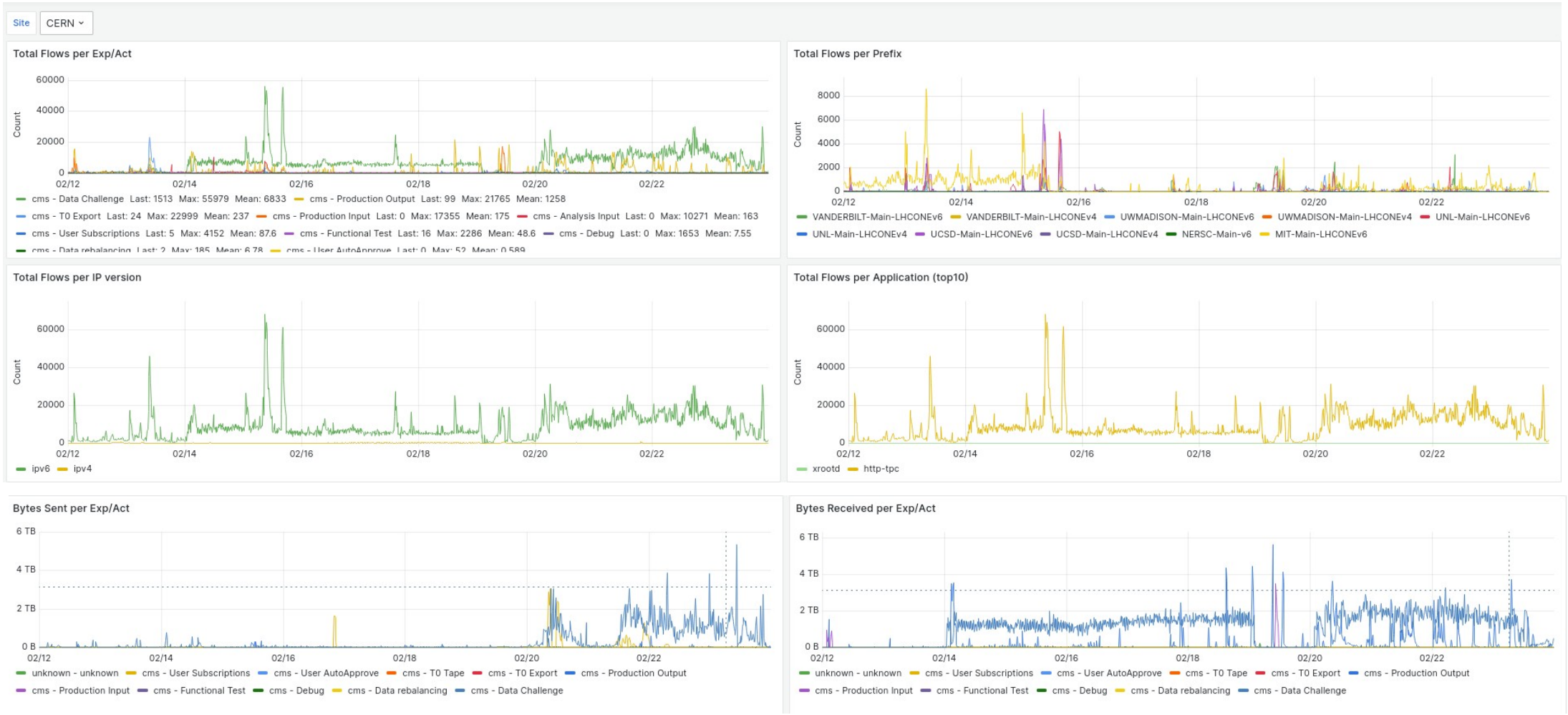
dCache, Xrootd: storage software

Rucio is the piece that knows the correct experiment and activity labels

Full chain (Rucio → FTS → Storage) enabled for the first time at DC24



Data Challenge 24



Future Plans

Scitags: An Initiative to Improve R&E Networking Visibility

Future Plans

- Work with dCache developers to add support for Scitags
- Explore possibility of using hop-by-hop headers rather than flow label
- Work on deploying flow marking more widely
- Encourage sites to install flowd and enable packet marking
- Encourage NRENs to develop the capability to make use of the marking

Thank you!