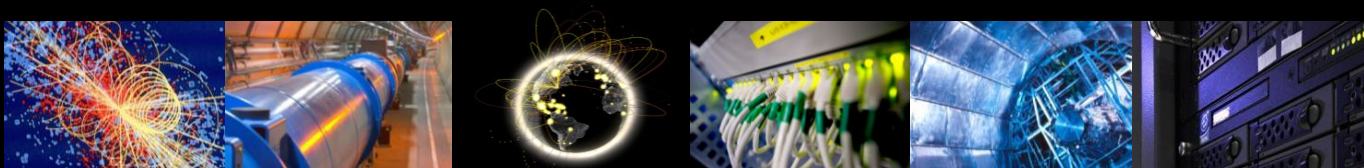


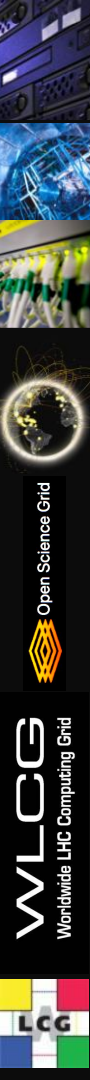
LHCOPN/LHCONE Update

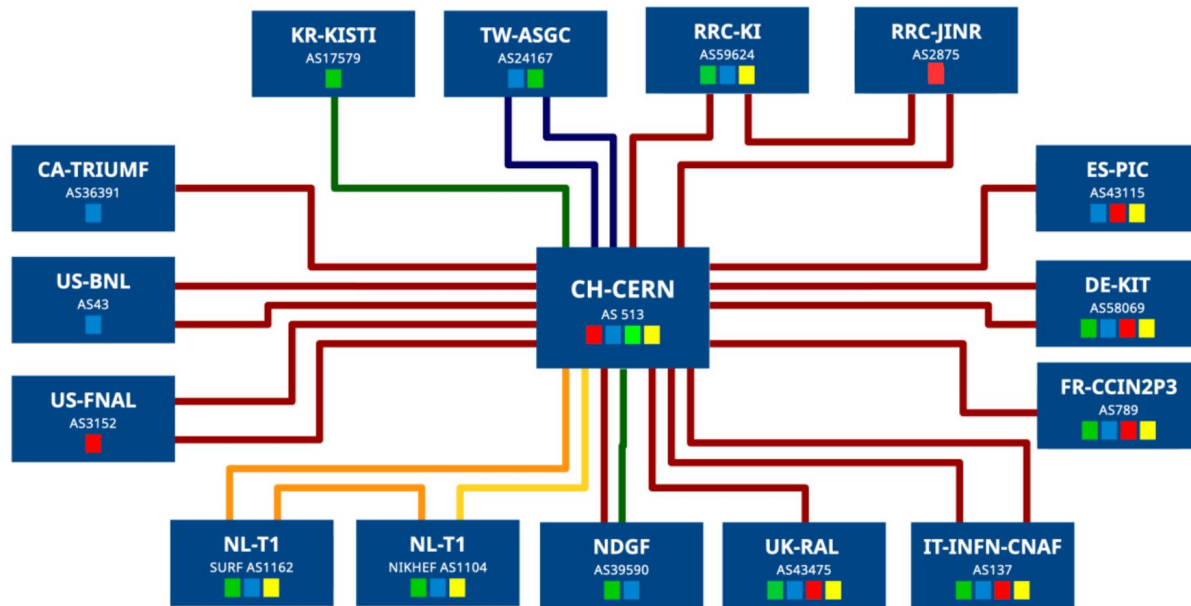
Marian Babik / CERN, Edoardo Martelli / CERN
XRootD & FTS Workshop @JSI



Outline

- LHCOPN Updates
- LHCONE Updates
- LHCONE R&D
 - Network Orchestration
 - NOTED, AutoGOLE/SENSE, Transport Control Networks
 - Network visibility and pacing
 - RNTWG, Scitags
 - Network routing and forwarding beyond LHCONE
 - MultiONE



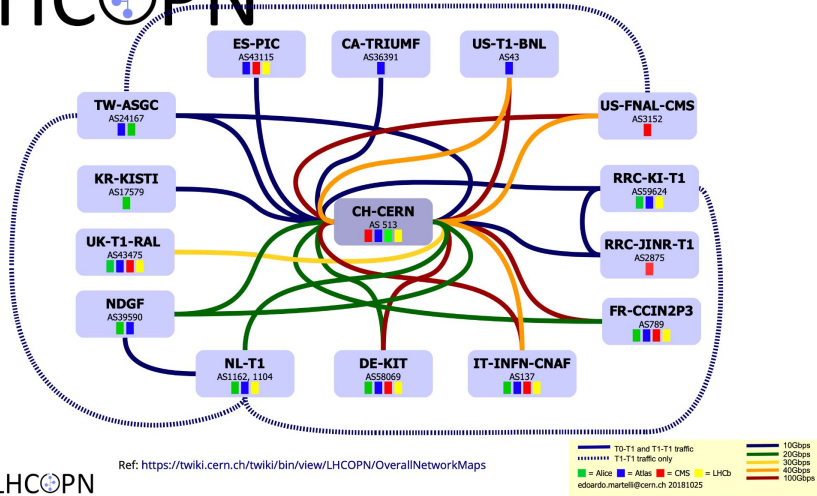


Numbers

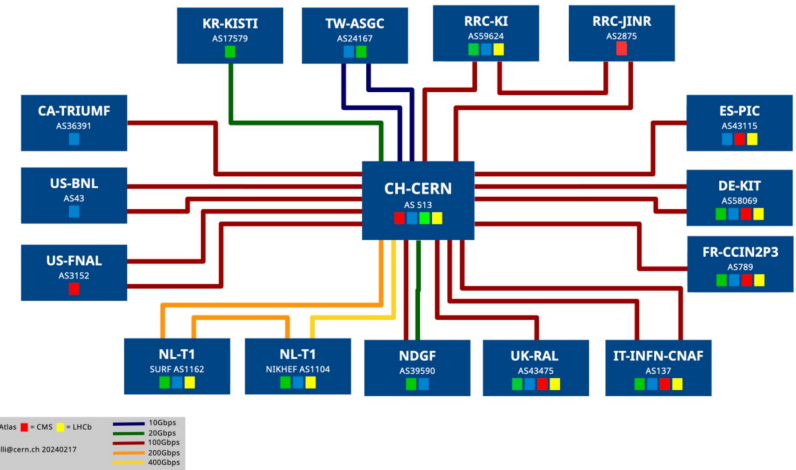
- 14 Tier1s + 1 Tier0
- 12 countries in 3 continents
- Dual stack IPv4-IPv6
- 1.9 Tbps to the Tier0

LHCOPN Evolution

LHCOPN 2018

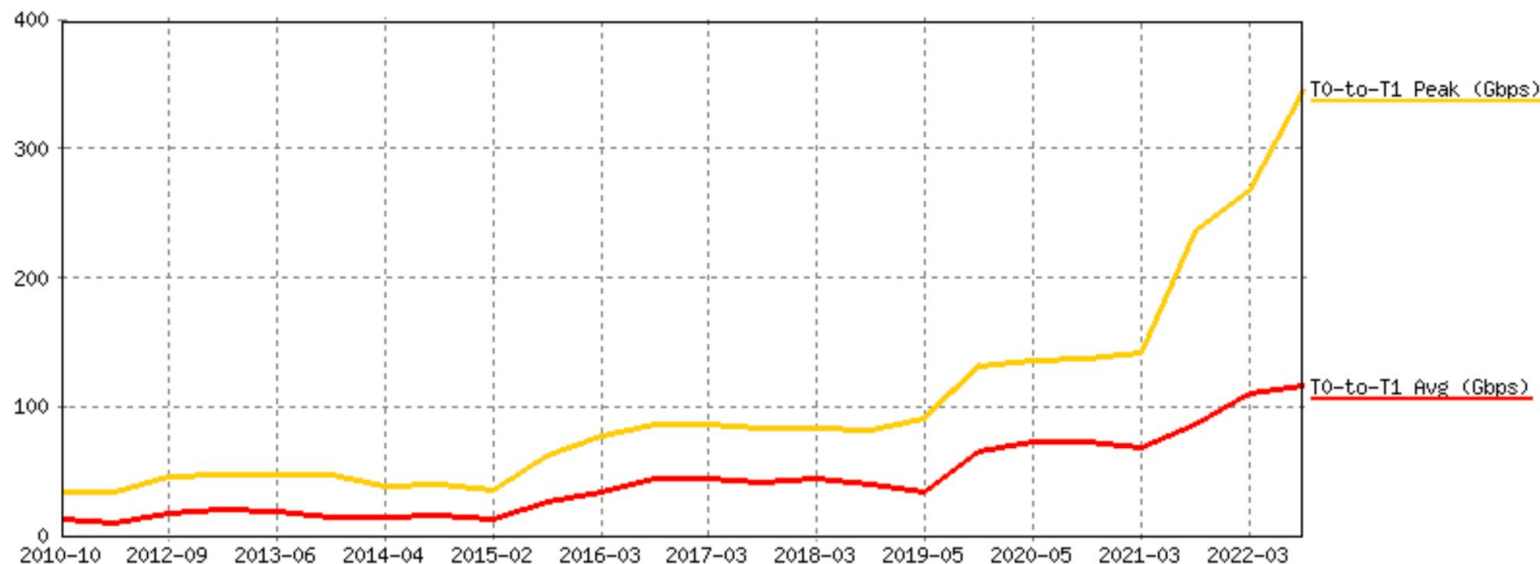


LHCOPN 2023



LHCOPN Long-term Growth

LHCOPN network traffic from CERN Tier0 to all the aggregated Tier1s



Run1: 2010-12 LS1:2013-14 Run2: 2014-2018 LS2: 2019-21 Run3: 2022

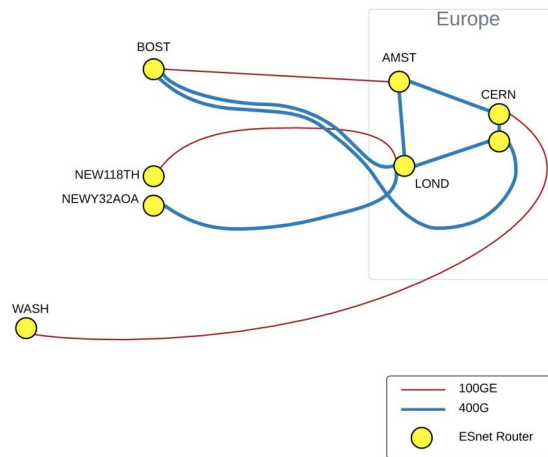
Y-Axis: Gbps (Giga bit per second)

Out: direction Tier0 to all Tier1s

Avg: average network bandwidth on the previous 12 months

Peak maximum peak network bandwidth on the previous 12 months

- Planned/recent capacity upgrades
 - NL-T1: 400Gbps
 - ES-PIC: 100Gbps
 - UK-RAL: deploying second 100Gbps
- New T1s status
 - New Tier1s: IHEP (CN) and NCBJ (PL)
 - TW-ASGC becoming Tier2
- Trans-Atlantic (ESNet) - currently underway:
 - 400G Boston - London
 - 400G Boston - CERN
 - 400G New York - London
 - 400G Europe Ring
- Trans-Atlantic capacity targets*
 - 500G now – 1.5T in Q3 2023 – ...
 - 3.2T in 2027, well in advance of Run 4



* targets and plans are indicative only, assume continuous funding

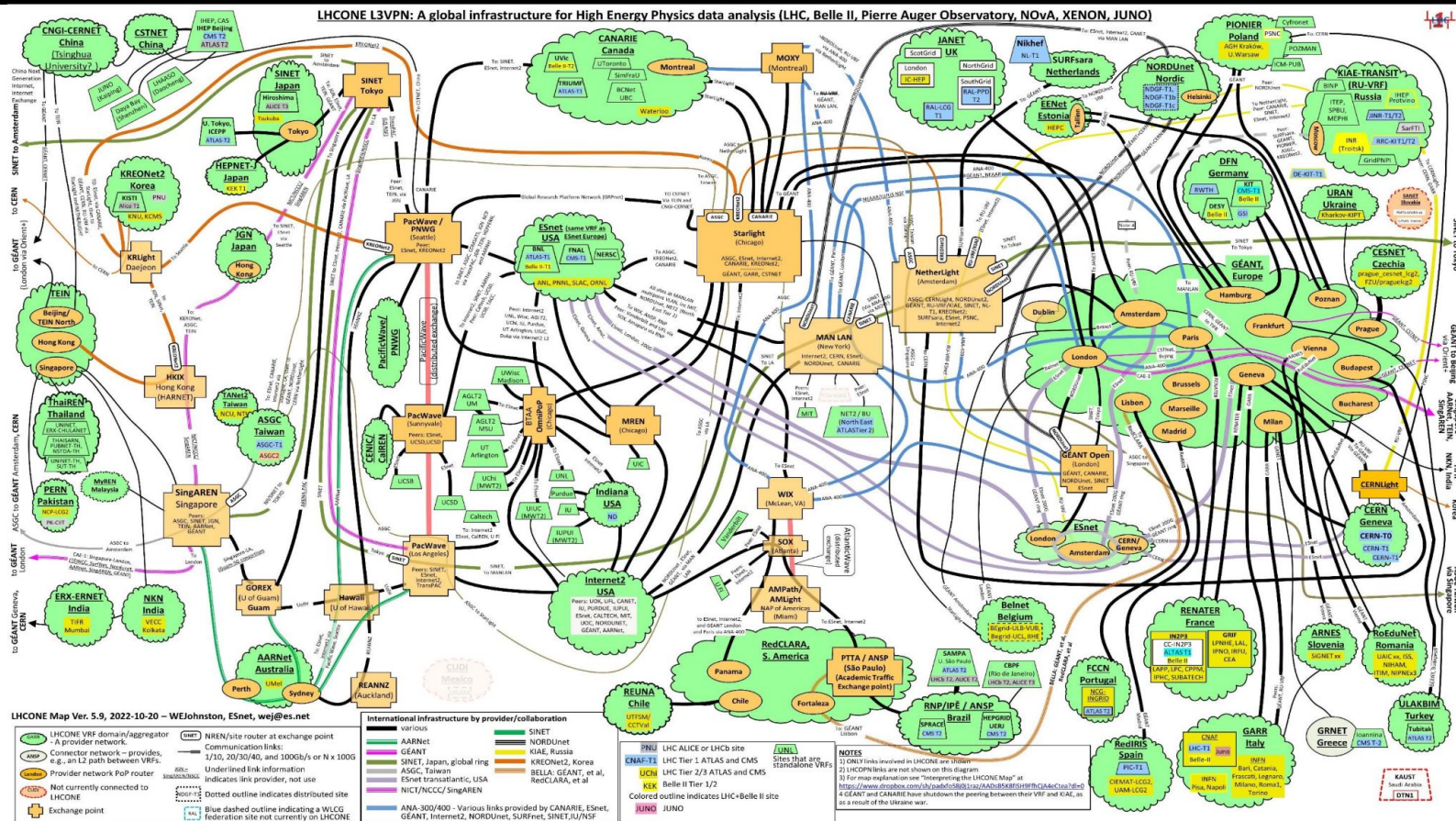
LHCONE



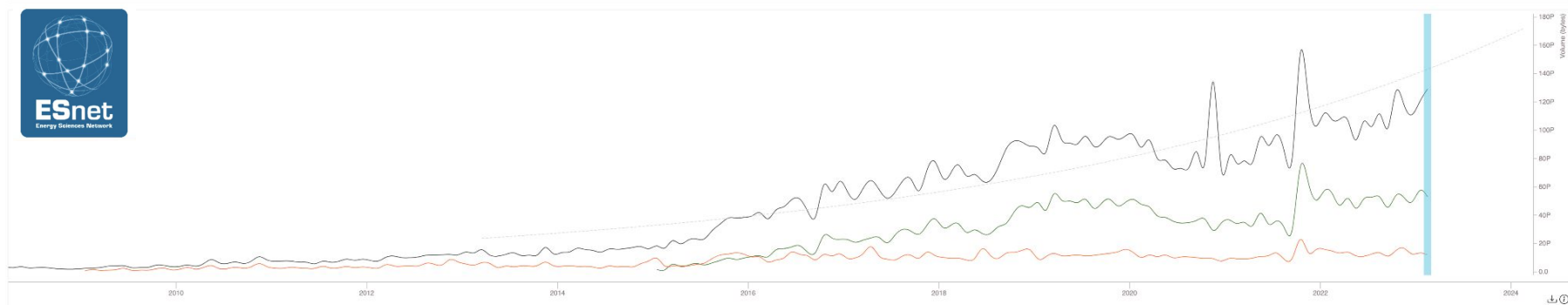


WLCG
Worldwide LHC Computing Grid

LCG



LHCONE Growth and Updates



	Bytes	Percent of Total	One Month Change	One Year Change
OSCARS	11.98PB	9.28%	-9.79%	-18.1%
LHCONE	52.97PB	41.0%	-7.61%	-4.16%
Normal traffic	64.18PB	49.7%	+29.4%	+71.3%
Total	129.14PB		+7.43%	+20.3%

Updates:

- KREOnet (KR) will give LHCONE transit to other VRFs
- CERN LHCONE access to grow to 1.6Tbps (2x400G with ESnet and 2x 400G with GEANT)
- DUNE will use LHCONE

ESnet6 High-touch Services

Precision Network Telemetry Services

Flow (Feature) Distillation

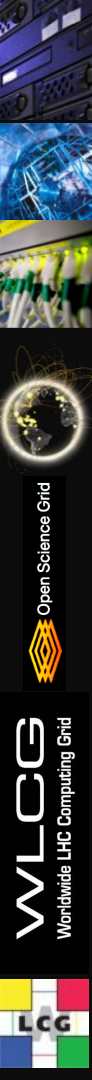
- Provides flow summaries of packets entering the network, thereby enabling full visibility of network traffic without large storage requirements.

Network Microscope

- Provides the capability to dynamically select network flow(s) for replication (of only the packet header and not user data contents), augmented with timestamps, to be redirected to compute resources for further processing, e.g., security analysis, feature extraction, etc.

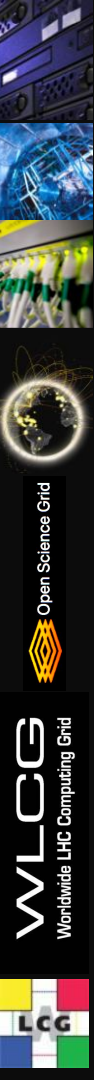
Benefits

- We can get detailed insights into how the network is behaving.
- We can profile how flows are performing in our network and take proactive action.
- We can use the detailed flow information for traffic engineering, capacity planning, or anomaly detection (e.g., AI/ML applications)



LHCONE R&D

- Network Orchestration
 - NOTED, AutoGOLE/SENSE
- Network visibility and pacing
 - RNTWG, Scitags
- Network routing and forwarding beyond LHCONE
 - MultiONE

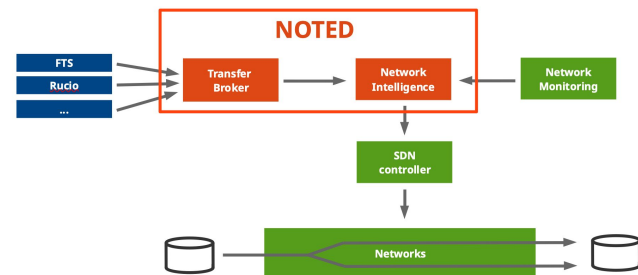


Network Orchestration

NOTED is a framework that can detect large FTS data transfers and trigger network optimization actions to speed up the execution of the transfers

Already tested with production transfers:

- CERN-PIC with LHCOPN-LHCONE load balancing
- CERN-TRIUMF and KIT-TRIUMF with the activation of dynamic circuits



AutoGOLE: Infrastructure which provides “end-to-end” network services in a fully automated manner

- Open-source software framework based on:
 - **Network Service Interface (NSI):** multi-domain network provisioning
 - **SENSE:** end-system provisioning and real-time integration with network services
- AutoGOLE, NSI and SENSE work together to provide the mechanisms for complete end-to-end services that include network and attached End Systems DTNs
- Circuit provisioning functionality used by NOTED during SC22

Network Visibility and Pacing

Research Networking Technical Working Group

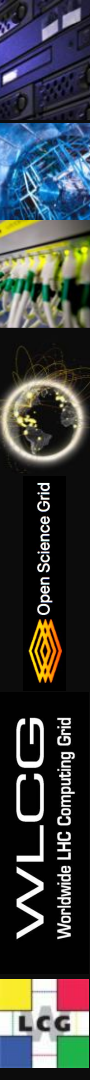
- Formed after the workshop in response to the requirements discussion
- 98 members from ~ 50 organisations have joined
- Three main areas of work:
 - **Network Visibility: Packet and Flow Marking** - viewed as the appropriate first step; regular meetings every ~2 months since summer 2020
 - [Packet Marking Document](#)
 - Outlines available technologies, standards and stakeholders perspectives
 - This has led to Scientific Network Tags (scitags) initiative, which is presented today
 - [Traffic Shaping](#) - Using techniques like packet pacing to achieve consistent throughput.
 - [Network Orchestration](#) - followed up by NOTED and AutoGOLE/GNA-G/SENSE

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:

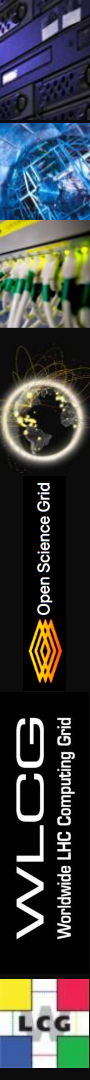
- Enable tracking and correlation of network transfers with Research and Education Network Providers (R&Es) network flow monitoring.
- Supporting collaborations to better understand network use and impact
 - Improve visibility into how network flows perform (per activity) within R&E segments
 - Get insights 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



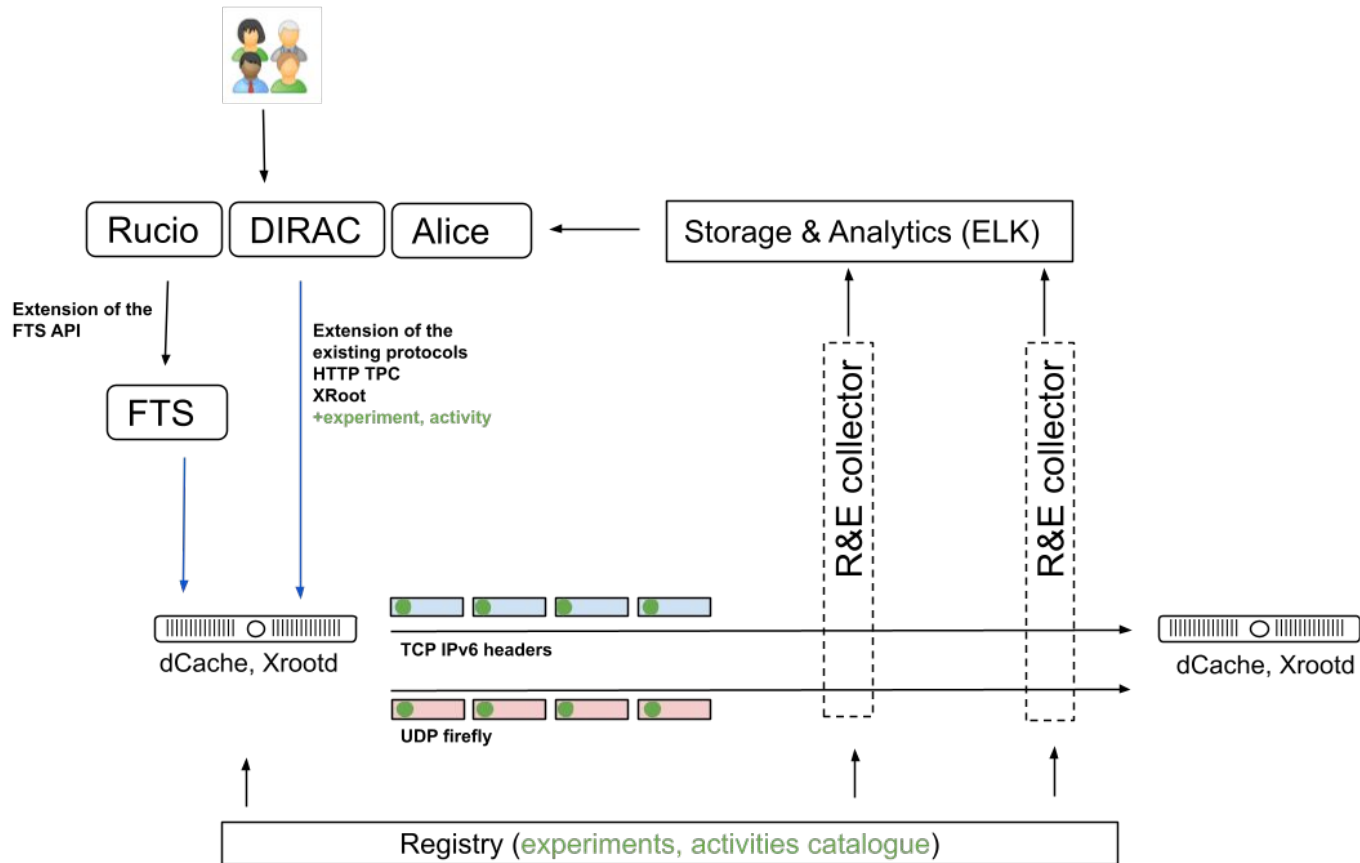
Technical Spec for Packet Marking/Flow Labeling

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



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](#)

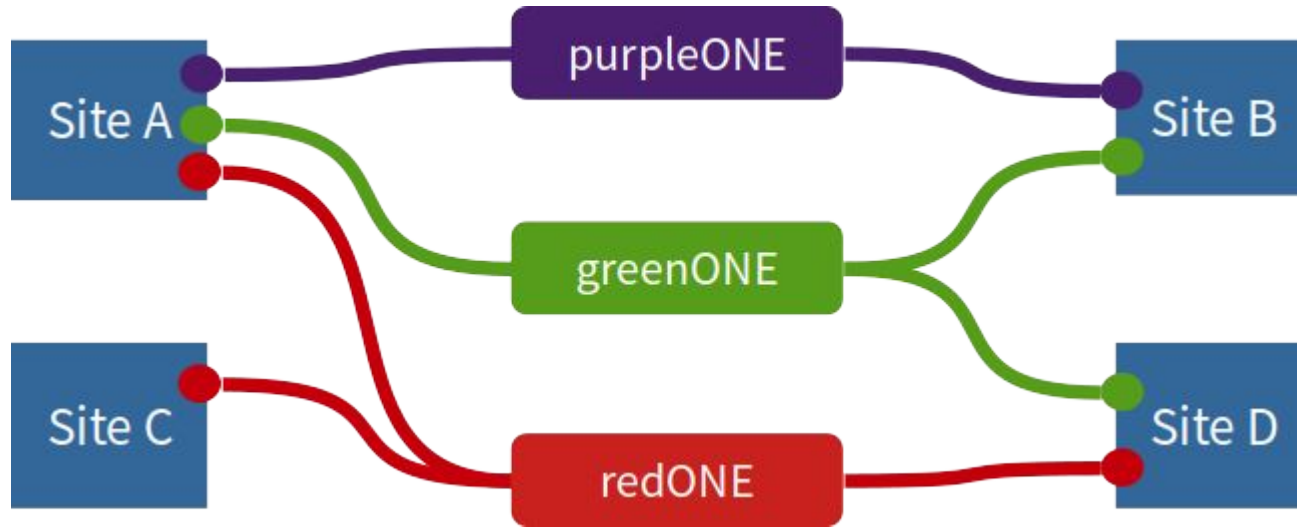
Network Pacing

A challenge for HEP storage endpoints is to utilize the network efficiently and fully.

- An area of interest for the experiments is **traffic shaping/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 $[\min(\text{SRC}, \text{DEST}, \text{NET})]$ 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!

Network Routing and Forwarding (multiONE)

- Each site joins only the VPNs it is collaborating with
- Benefit: reduce exposure of data-centres/Science-DMZs
- Current challenge: how to correctly route data traffic at sites that joins multiple VPNs?
 - Exploring ways how to use SciTags/MPLS for this



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

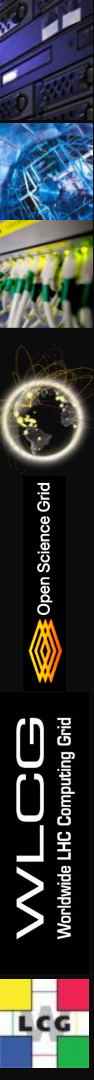
- Understanding where/when on-demand network provisioning is needed
- Combined with analytics to determine duration, capacity, etc.

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

Summary

- Significant capacity increases on LHCOPN/LHCONE
 - Important to ensure this capacity becomes available end-to-end (WAN/LAN)
- Number of ongoing R&D projects
 - Network Orchestration
 - NOTED and SENSE are leading projects in programmable networks
 - Network Visibility and Pacing
 - Technology available to make all our flows visible to R&E networks
 - Plan to start deployment and increase adoption as we approach DC24
 - Network Pacing activity will start, aiming to improve performance of the end-to-end transfers
 - Network Routing and Forwarding
 - Investigating technologies that can provide networking beyond LHCONE
- Still a lot of work ahead in many areas, important to continue the efforts
- We welcome additional contributions/project ideas/experiments; contact us if you are interested!
 - [LHCOPN/LHCONE Workshop 18-19 April in Prague](#)

Questions ?

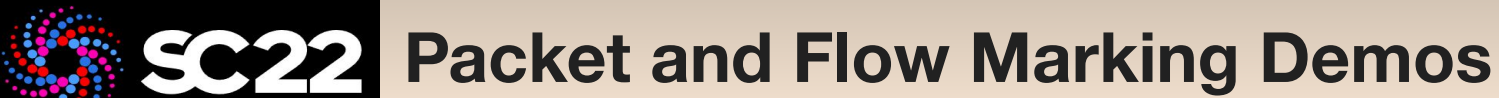


Backup Slides Follow

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.





scitags.org

Flow and Packet Marking for Global Scientific Computing

University
of Victoria

1. Clients requesting data transfers from/to DTN-SC22-400g while passing science domain and activity fields via transfer protocols.



canarie

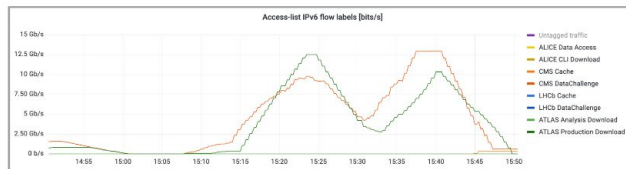
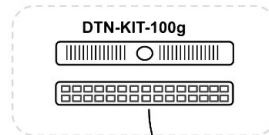
4.High performance tests using eBPF-TC filters to test encoding of the science domains and activity fields in the IPv6 flow label at scale.



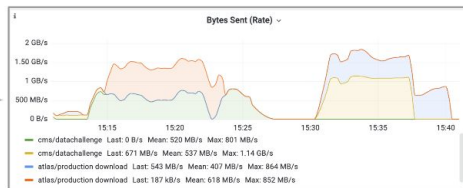
DTN-SC22-400g
R7503 2.6 GHz
NVMe 2.0
2x200 Gbps



STARLIGHTSM



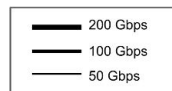
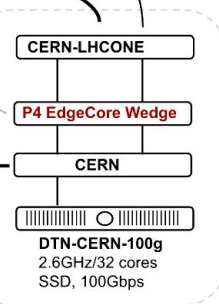
3. P4 programmable switch at CERN collecting the science domains and activity bits encoded in the packets.



2. XRootD storage responds to the client requests and marks the data transfer packets with the corresponding science domain and activity.



5. Sampling of the low level TCP/IP metrics, which can be used by sites and R&Es to better understand the scientific flows.





SC22

Packet and Flow Marking Demonstrations

General / Scientific Network Tags: Aggr - Netlink Dashboard ☆ ↻

2022-11-15 16:53:37 to 2022-11-15 17:57:38



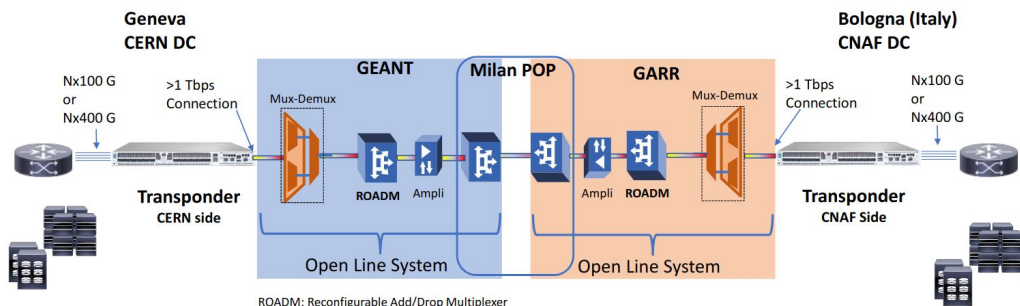
Network Visibility Plans

- Near-term objectives
 - Continue rollout and testing of Xrootd implementation
 - Detect flow identifiers from storage path/url, activities from user role mapping
 - Finalise development and deploy [network of receivers](#)
 - Instrument Rucio/FTS to pass flow identifiers to the storages
 - Good discussions with dCache last week covering SciTags design/implementation
 - Work with the [WLCG Monitoring TF](#) to improve site network monitoring
- Engage other R&Es and explore available technologies for collectors
 - Deploy additional collectors and perform R&D in the packet collectors
 - Improve existing data collection and analytics
- Test and validate ways to propagate flow identifiers
 - Engage experiments and data management systems
 - Validate, test protocol extensions and FTS integration
 - Explore other possibilities for flow identifier propagation, e.g. tokens
- Plan to submit IETF Informational RFC

CNAF-CERN DCI

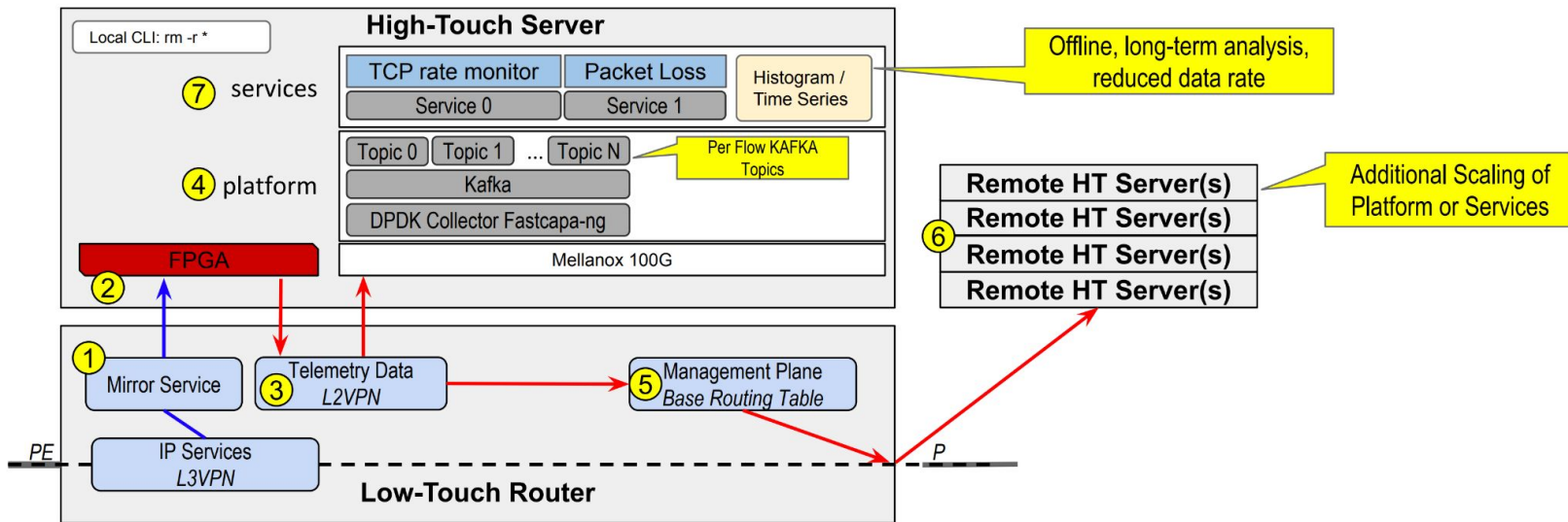
Pilot for a Tbps Data-Centre interconnections between INFN-CNAF (Bologna, IT) and CERN (Geneva CH)

- to meet HL-LHC requirements at an affordable cost
- made using transmission devices at CNAF and CERN
 - no need to use expensive routers on the providers' networks
- using optical channels on GARR and GEANT dark fibres
 - pilot also for new GEANT Spectrum Sharing service
- Equipment installed and connected
- Reached 400Gbps on single wavelength Geneva-Bologna (~1000km)



ESNet High-Touch Service

ESnet6 High-Touch Architecture Overview



1. Mirror Service - Allows selective flows in the dataplane to be duplicated and sent to the FPGA for processing.
2. Programmable Dataplane (DP) - Appends meta-data, timestamps and repackages packet for transmission to Platform code.
3. Telemetry Data L2VPN - Connect Dataplane and Platform, possibly on different High-Touch Servers.
4. Platform - Reads telemetry packets from the network and distributes information to High Touch Services.
5. Management Plane Base Routing Table - Provides connectivity to Remote Servers.
6. Remote Server - Hosts Platform components or Services (but not a Dataplane). Telemetry data can be directed to Remote Servers.
7. Service - Reads data from the Platform and performs real-time analysis as well as inserts selected telemetry data into database.



Protocol Extensions

- XRoot protocol extension uses flow.scitag cgi
 - `//path/?scitag.flow=flow_id`
 - `flow_id = (exp_id << 6) | act_id`
 - `exp_id` - experiment id; `act_id` - activity id (both as seen in registry)
- HTTP-TPC extension
 - Adds additional HTTP request headers in the COPY request
 - `COPY /path/to/destination HTTP/1.1`
Host: `destination.example.org`
Source: <https://source.example.org/path/to/source>
TransferHeaderAuthorization: Bearer ABCD...
TransferFlowExperiment: `exp_id`
TransferHeaderFlowExperiment: `exp_id`
TransferFlowActivity: `act_id`
TransferHeaderFlowActivity: `act_id`
- Both extensions to be implemented in **GFAL2 libs**
 - Xroot extension already implemented in xrootd clients v5.0+

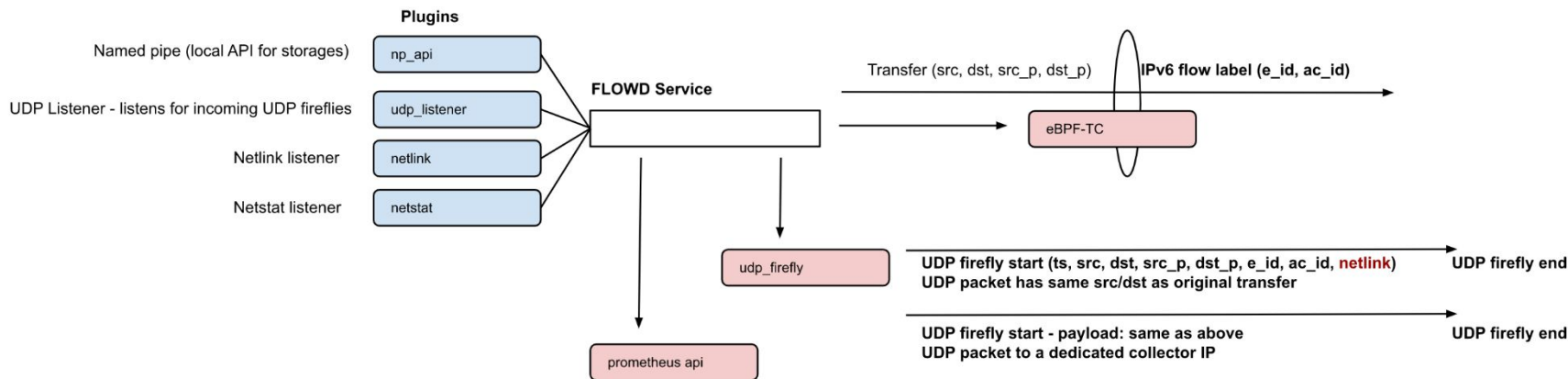
AutoGOLE and SENSE

AutoGOLE: Infrastructure which provides “end-to-end” network services in a fully automated manner

- Open-source software framework based on:
 - **Network Service Interface (NSI)**: multidomain network provisioning
 - **SENSE**: end-system provisioning and realtime integration with network services
- Persistent Infrastructure, somewhere in between production and a testbed
- AutoGOLE, NSI and SENSE work together to provide the mechanisms for complete end-to-end services that include network and attached End Systems DTNs
- Circuit provisioning functionality used by NOTED and Scitags demo for SC22

Flowd Service

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



- 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

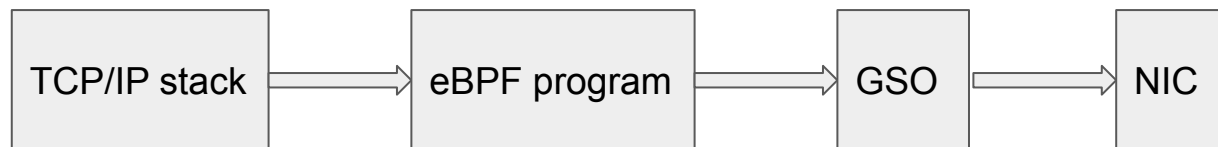
Status

- **Flow Marking** (UDP firefly) implementations
 - **Xrootd** 5.0+ supports UDP fireflies
 - Site admins can configure mapping of paths to experiments and user/roles to activities
 - Xroot protocol extension via **scitag** URL flag
 - **dCache** PoC now ready, supports UDP fireflies
 - Testing deployment at AGLT2 (backported to 7.2) with issues reported back to Tigran
- **Flow and Packet Marking**
 - **Flowd** - packet and flow marking service
 - Independent service that can mark flows and packets for 3rd party services
- **Collectors/Receivers**
 - Initial receiver prototype developed by ESnet (available on [scitags github](#))
- **Registry**
 - Provides list of experiments and activities supported
 - Exposed via JSON at [api.scitags.org](#)
- **Flow id propagation**
 - ~~Work needed has been agreed with Rucio and FTS (tickets were submitted to follow~~

Flowd: 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

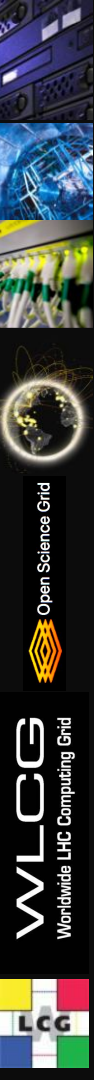
Egress path:



- 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

Questions / Discussion

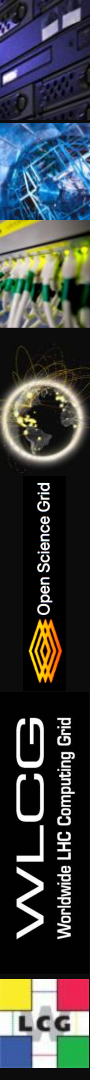
Questions, Comments, Suggestions?



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/Mar 2024?):

- **Supercomputing 2022**: Next week 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

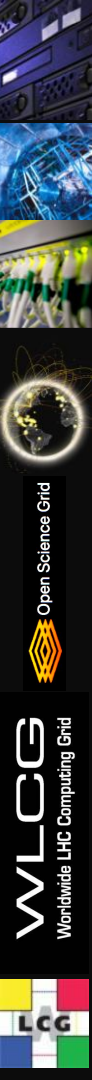
- The **RNTWG**, driven by the needs and interests of the **LHC**, **HEP** and **R&E networking** communities, is implementing **packet marking** and **flow labeling** of network flows for **all** R&E network users
 - We have a well defined program of work and strong collaboration with storage and transfer application providers, WLCG experiments and sites.
 - Rucio will play a central role in the process.
- Our goal is to have large scale packet marking and flow labeling in place by the time of the next WLCG Data Challenge

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:

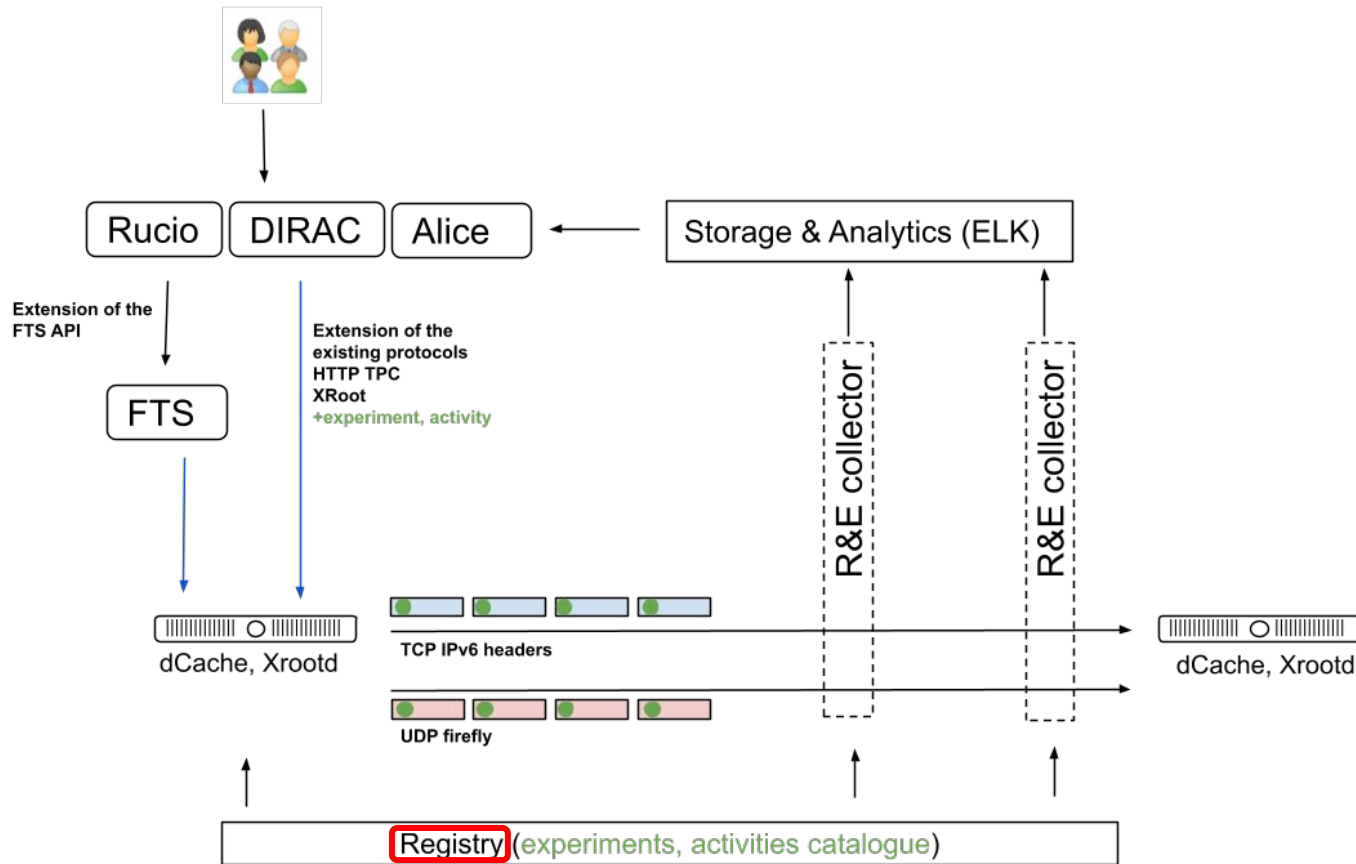
- OSG: NSF MPS-1148698
- IRIS-HEP: NSF OAC-1836650



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/1kFCB5Hnk/lhc-data-challenge-overview?orgId=1>

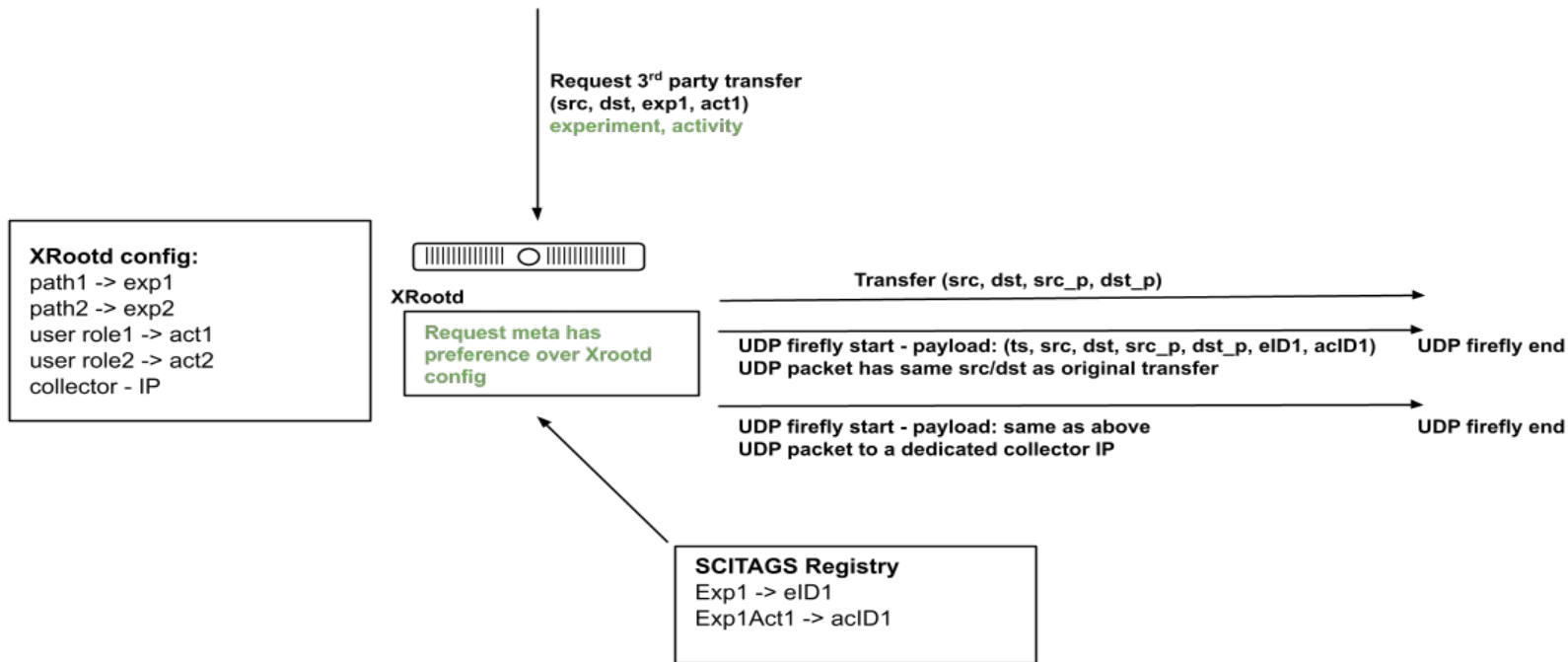
How scitags work



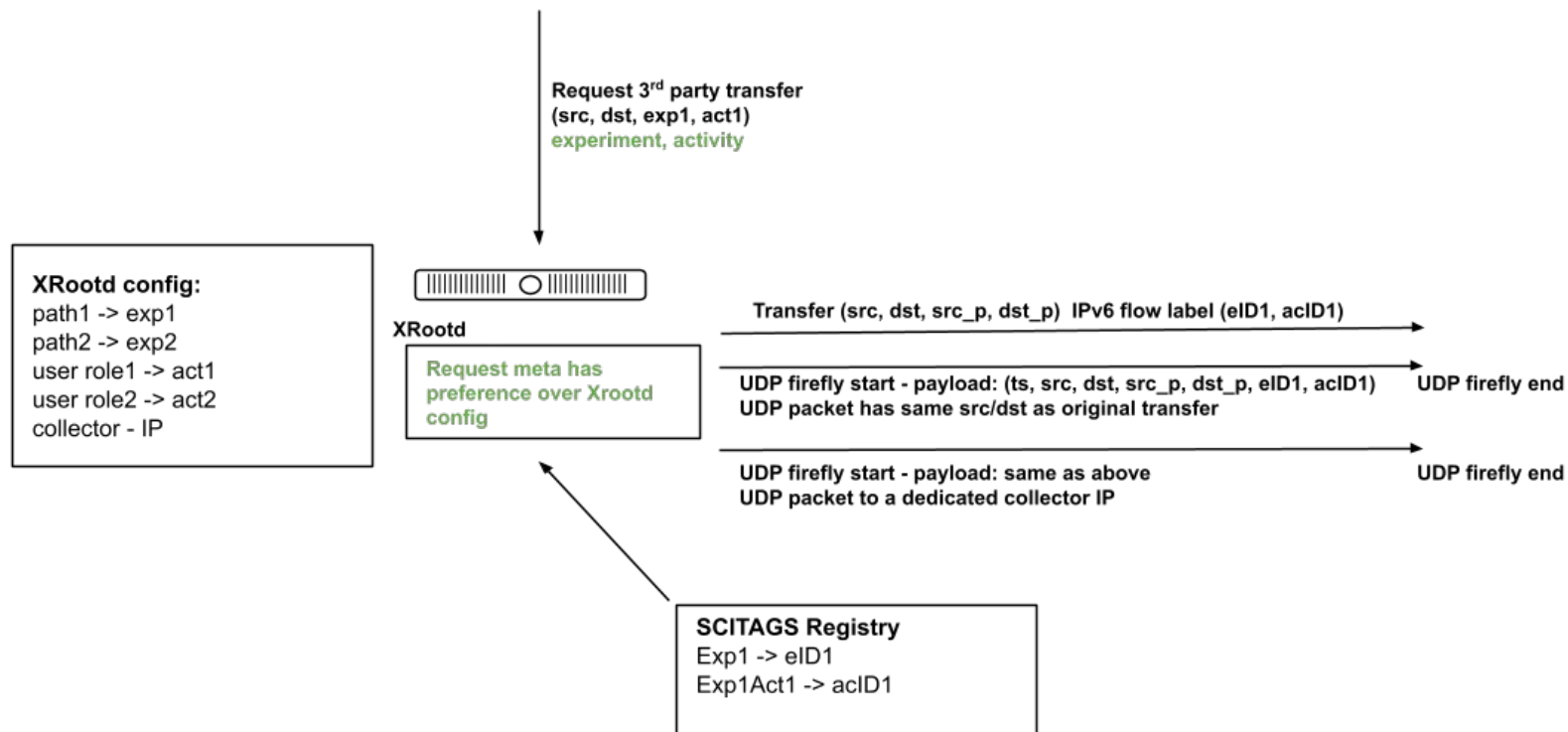
Flow identifier lifecycle

- 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

