Research Networking Technical WG Status & Plans

Shawn McKee / University of Michigan, Marian Babik / CERN

Spring 2022 HEPiX Meeting https://indico.cern.ch/event/1123214/

April 26, 2022

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.

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 (spring 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**

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

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

Network Visibility Motivation

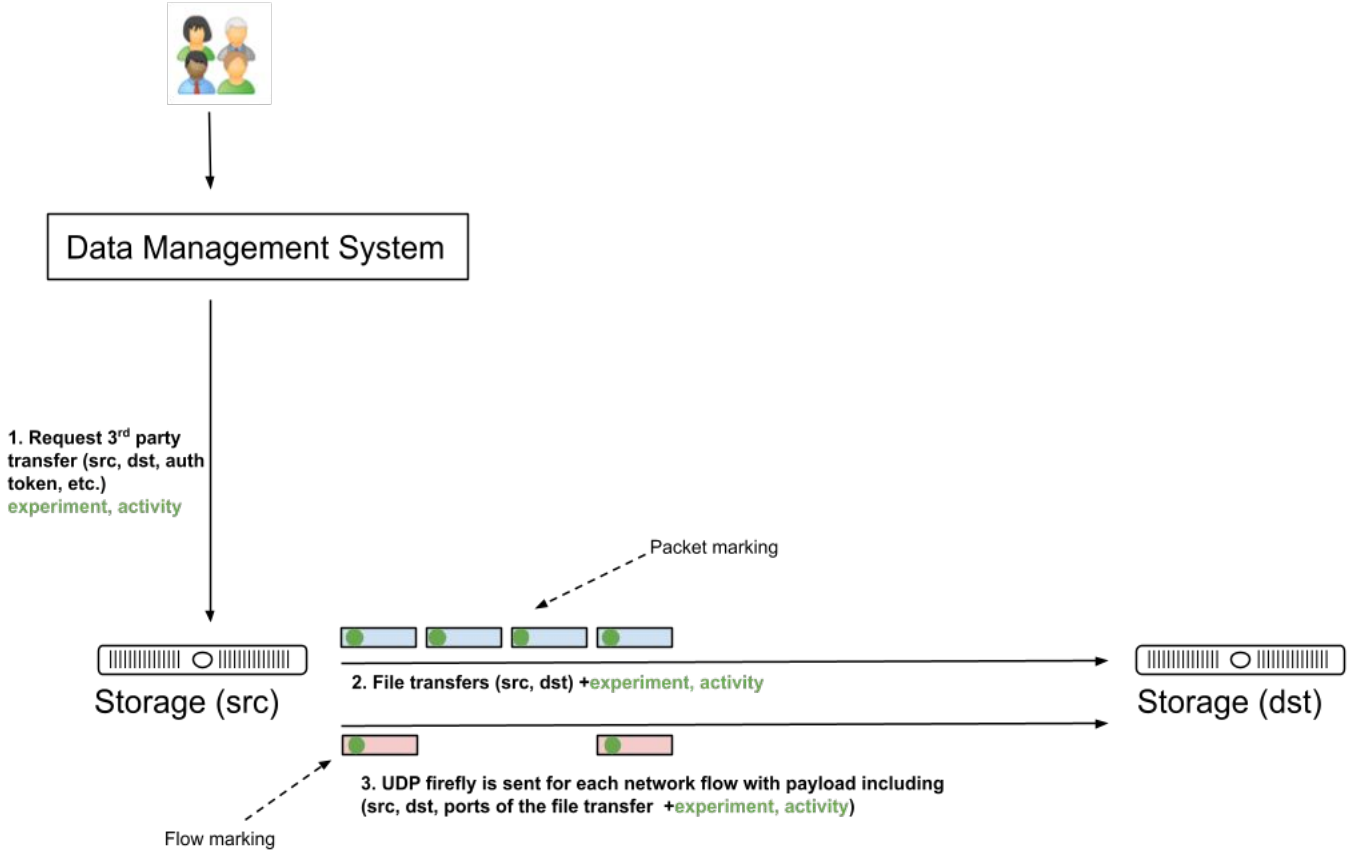
- Networks are becoming more programmable and capable with technologies such as P4, SDN, virtualisation, eBPF, etc.
- But with less and less context about the traffic they carry.
  - Cloud deployments, Kubernetes, encryption, tunneling, privacy, etc.
- Understanding scientific traffic flows in detail is critical for understanding how our complex systems are actually using the network.
  - Current monitoring/logging tell us where data flows start and end, but is unable to understand the data in flight.
  - Dedicated L3VPNs can be created to track high throughput science domains, but with more domains requiring high throughput this will become expensive, it won't scale, won't work at big sites having to support multiple domains at the same time.
- In general the monitoring we have is experiment specific and very difficult to correlate with what is happening in the network. We suggest this is a general problem for users of the Research and Education Networks (RENs).

Network Visibility and Scitags

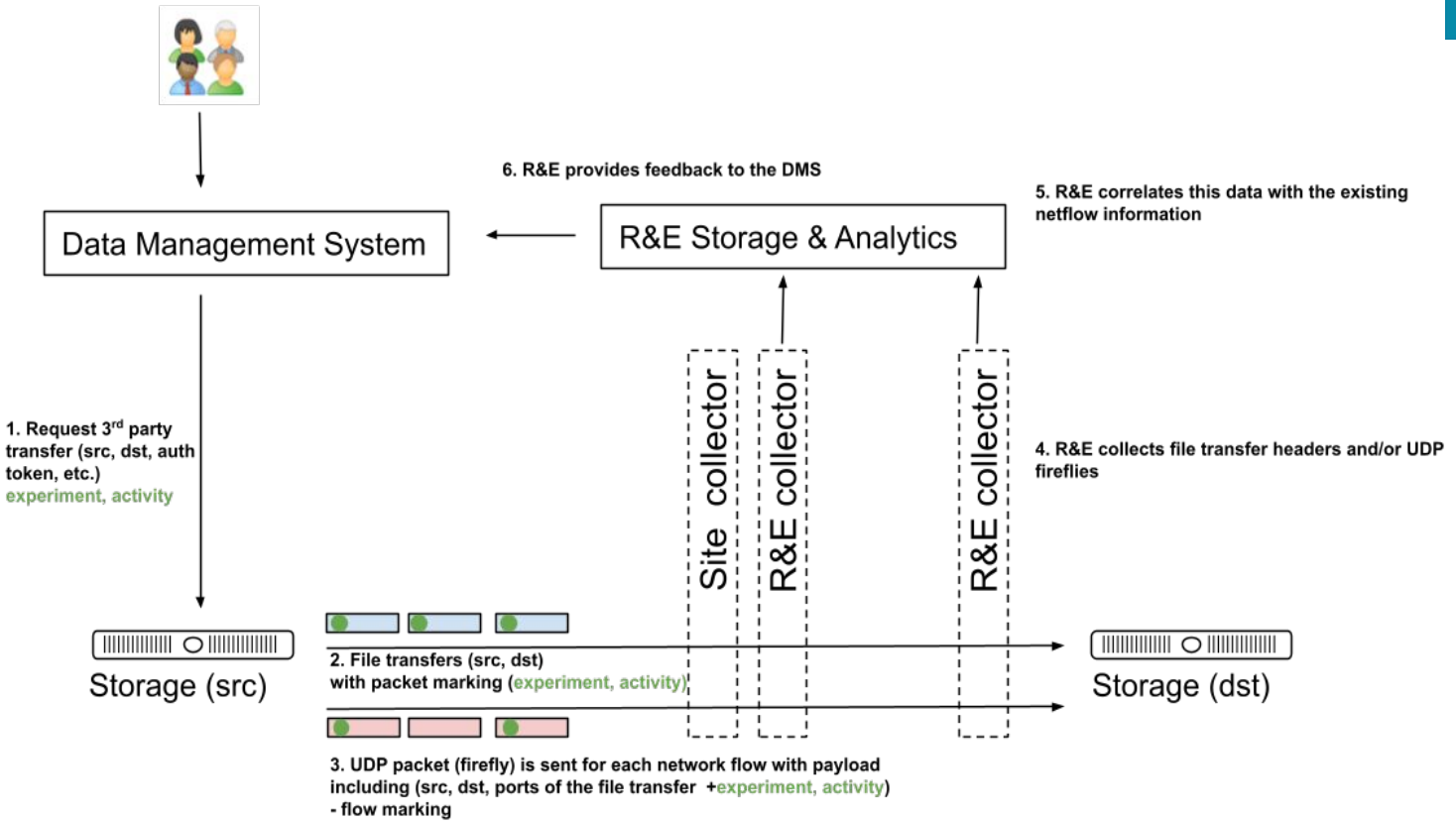
- Scientific Network Tags (scitags) is an initiative promoting identification of the science domains and their high-level activities at the network level.

- 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

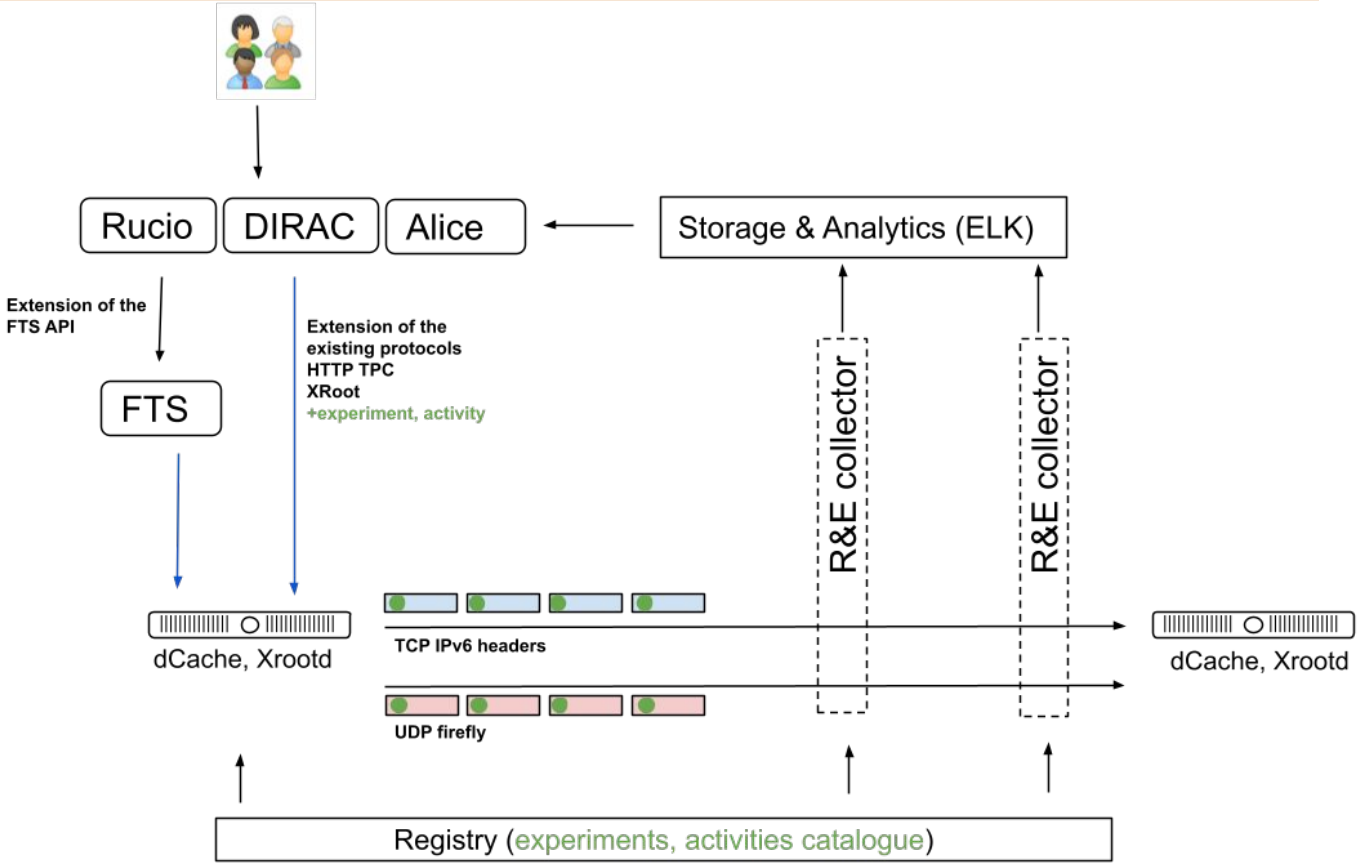
How scitags work



How scitags work



How scitags work



Finding More Information: https://scitags.org

Code

Technical Spec

Mailing List

Presentations

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.

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 (will be available on scitags github soon)
  - 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 DC (& still operating)
  - Flowd + ESnet collector + Registry
  - **AGLT2, BNL, KIT, UNL and Caltech** participated
  - Brunel, Glasgow and QMUL interested to help with further testing

Registry

We need to standardize 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
```

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 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(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!

Network orchestration

- 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.

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 (backup slides)
  - Instrument Rucio/FTS to pass flow identifiers to the storages
  - Involve other storage systems (dCache, etc.); discuss possible 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
- R&D activities
  - **Packet marking** - further testing and validation is required for IPv6 flow label implementation (next meeting or two)

Plans: Network Pacing & Orchestration

The RNTWG has primarily focused on the network visibility area due to limited manpower, but we have two additional areas the are part of the overall group goal: traffic shaping and network orchestration

Traffic Shaping:

- The WLCG experiments would like to explore traffic shaping/packet pacing.
  - Without packet 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

Network Orchestration:

- This effort is being led by the GNA-g and includes work from the SENSE and FABRIC projects. Current focus is on integration with the experiments via **RUCIO to demonstrate value and capability.**

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.

- 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 shaping:

- While network orchestration has significant activity underway, we need to find new effort interested in developing, prototyping and evaluating traffic shaping

Interested? We are always looking for additional members to join the effort!

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/

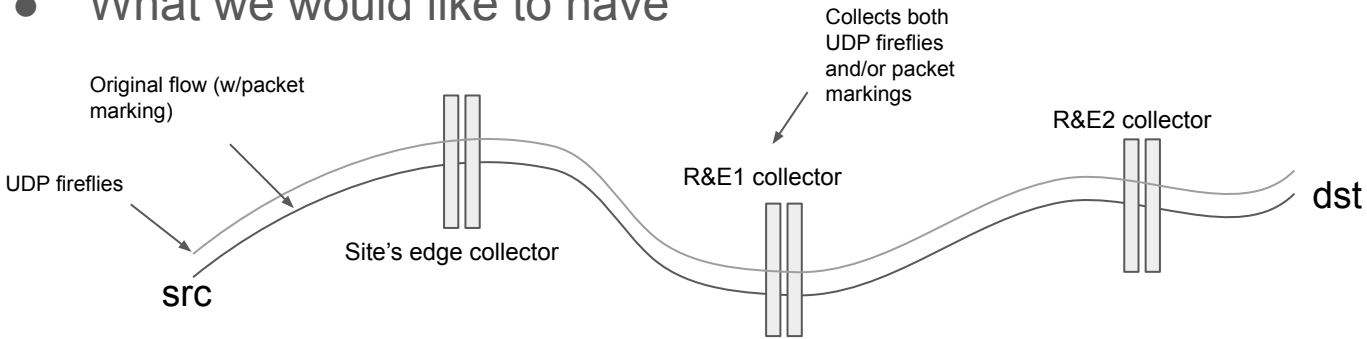
The scitags web page: https://scitags.github.io

Code at https://github.com/scitags/scitags.github.io

Backup slides

Collectors

- What we would like to have

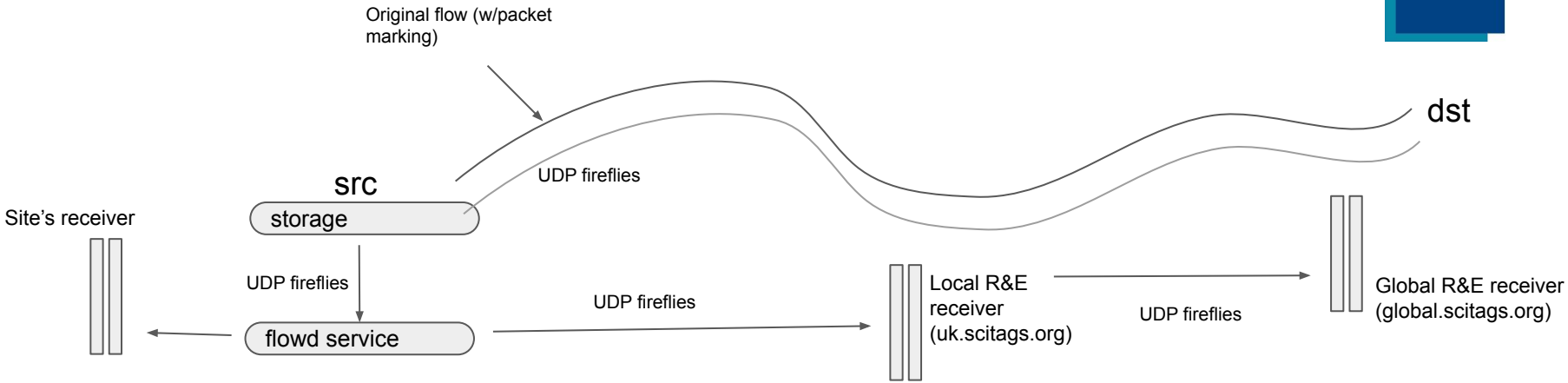


- Enable collection of packet and flow markings along the path
  - In order to extend R&E netflow information with flow identifier (experiment + activity)
  - UDP firefly packets needs to be collected and relayed to ensure they reach all collectors
- Each R&E **can** setup and operate one or more collectors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

Collectors

- 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.
- However, in-line collectors require the ability to either selectively identify and capture fireflies or the ability to capture IPv6 flow labels from packets
  - Many possible ways to implement.
  - Strategy and technology to implement will depend on the R&E, their topology and hardware.
  - Would be great to get example implementations that can be shared between R&E network operators.

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)

Packet Marking - Storage Elements

The primary challenge here is in two areas:

1. Augmenting the existing storage system to be able to set the appropriate bits in the network packets
2. Communicating the appropriate bits as part of a transfer request
    a. Likely need some protocol extension to support this
    b. Other ideas?

Packet Marking - Jobs

As jobs source data onto the network OR pull data into the job, we should try to ensure the corresponding packets are marked appropriately

- Containers and VMs may allow this to be easily put in place
- Still need configuration options that specify the right bits
- Signalling to the “source” about what those bits are also needs to be in place

Packet Marking - IPv6

IPv6 incorporates a “Flow Label” in the header (20 bits)

Fixed header format

Offsets	Octet	0								1								2								3							
Octet	Bit	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
0	0	Version				Traffic Class								Flow Label																			
4	32	Payload Length																Next Header								Hop Limit							
8	64	Source Address																															
12	96																																
16	128																																
20	160																																
24	192	Destination Address																															
28	224																																
32	256																																
36	288																																

Packet Marking - IPv4

IPv4 incorporates a “Options” in the header (allowing to add more 32 bit words)

IPv4 Header Format

Offsets	Octet	0								1								2								3							
Octet	Bit	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
0	0	Version				IHL				DSCP						ECN		Total Length															
4	32	Identification																Flags				Fragment Offset											
8	64	Time To Live								Protocol								Header Checksum															
12	96	Source IP Address																															
16	128	Destination IP Address																															
20	160	Options (if IHL > 5)																															
24	192																																
28	224																																
32	256																																

Network Functions Virtualisation WG

Mandate: Identify use cases, survey existing approaches and evaluate whether and how Software Defined Networking (SDN) and Network Functions Virtualisation (NFV) should be deployed in HEP.

Team: 60 members including R&Es (GEANT, ESNet, Internet2, AARNet, Canarie, SURFNet, GARR, JISC, RENATER, NORDUnet) and sites (ASGC, PIC, BNL, CNAF, CERN, KIAE, FIU, AGLT2, Caltech, DESY, IHEP, Nikhef)

Monthly **meetings** started in Jan 2018 (https://indico.cern.ch/category/10031/)

NFV WG Report

NFV WG produced an interim-report that could serve as one of the inputs for the LHCOPN/LHCONE feedback

Executive summary for NFV Phase 1 report is at

https://docs.google.com/document/d/1w7XUPxE23DJXn--j-M3KvXlfXHUnYgsVUhBpKFjyjUQ/edit#heading=h.flthknqgm3ub

Report has **3 chapters**:

- Cloud Native DC Networking
- Programmable WAN
- Proposed Areas of Future Work

Future (phase 2) is partially the work of this RNT WG, but we may end up separating out a more focused NFV/SDN group.

NFV Report Conclusions

The primary challenge we face is ensuring that WLCG and its constituent collaborations will have the networking capabilities required to most effectively exploit LHC data for the lifetime of the LHC. To deliver on this challenge, automation is a must. The dynamism and agility of our evolving applications, tools, middleware and infrastructure require automation of at least part of our networks, which is a significant challenge in itself. While there are many technology choices that need discussion and exploration, **the most important thing is ensuring the experiments and sites collaborate with the RENs, network engineers and researchers to develop, prototype and implement a useful, agile network infrastructure that is well integrated with the computing and storage frameworks being evolved by the experiments as well as the technology choices being implemented at the sites and RENs.**

Research Networking Technical WG

Charter:
https://docs.google.com/document/d/1I4U5dpH556kCnoIHzyRpBI74IPc0gpgAG3VPUp98Io0/edit#

Mailing list:
http://cern.ch/simba3/SelfSubscription.aspx?groupName=net-wg

Members (79 as of today, in no particular order):

Christian Todorov (Internet2) Frank Burstein (BNL) Richard Carlson (DOE) Marcos Schwarz (RNP) Susanne Naegele Jackson (FAU) Alexander Germain (OHSU) Casey Russell (CANREN) Chris Robb (GlobalNOC/IU) Dale Carder (ESnet) Doug Southworth (IU) Eli Dart (ESNet) Eric Brown (VT) Evgeniy Kuznetsov (JINR) Ezra Kissel (ESnet) Fatema Bannat Wala (LBL) Joseph Breen (UTAH) James Blessing (Jisc) James Deaton (Great Plains Network) Jason Lomonaco (Internet2) Jerome Bernier (IN2P3) Jerry Sobieski Ji Li (BNL) Joel Mambretti (Northwestern) Karl Newell (Internet2) Li Wang (IHEP) Mariam Kiran (ESnet) Mark Lukasczyk (BNL) Matt Zekauskas (Internet2) Michal Hazlinsky (Cesnet) Mingshan Xia (IHEP) Paul Acosta (MIT) Paul Howell (Internet2) Paul Ruth (RENCI) Pieter de Boer (SURFnet) Roman Lapacz (PSNC) Sri N () Stefano Zani (CNAF) Tamer Nadeem (VCU) Tim Chown (Jisc) Tom Lehman (ESnet) Vincenzo Capone (GEANT) Wenji Wu (FNAL) Xi Yang (ESnet) Chin Guok (ESnet) Tony Cass (CERN) Eric Lancon (BNL) James Letts (UCSD) Harvey Newman (Caltech) Duncan Rand (Jisc) Edoardo Martelli (CERN) Shawn McKee (Univ. of Michigan) Simone Campana (CERN) Andrew Hanushevsky (SLAC) Marian Babik (CERN) James William Walder () Petr Vokac () Alexandr Zaytsev (BNL) Raul Cardoso Lopes () Mario Lassnig (CERN) Han-Wei Yen () Wei Yang (Stanford) Edward Karavakis (CERN) Tristan Suerink (Nikhef) Garhan Attebury (UNL) Pavlo Svirin () Shan Zeng (IHEP) Jin Kim (KISTI) Richard Cziva (ESnet) Phil Demar (FNAL) Justas Balcas (Caltech) Bruno Hoeft (FZK)