Networking Areas of Interest for HEP

Shawn McKee / University of Michigan

Large Scale Networking (LSN) Workshop on Huge Data: A Computing, Networking and Distributed Systems Perspective

April 14, 2020

Overview

There are two goals with this presentation:

1) **Inform** everyone on the recent work to document and plan network efforts - NFV WG report https://doi.org/10.5281/zenodo.3565562

2) **Describe** a new networking R&D effort focused on needs identified by the High-Energy Physics (HEP) community and the collaborating National Research & Education Networks

(HEPiX is a forum which meets twice per year to discuss practical experiences with Cyberinfrastructure for HEP and beyond)

Motivation: Why Worry about Networks?

- High Energy Physics (HEP) has significantly benefited from strong relationship with Research and Education (R&E) network providers
  - Thanks to LHCOPN/LHCONE community and NREN contributions, experiments enjoy almost “infinite” capacity at relatively low (or no-direct) cost
  - NRENs have been able to continually expand their capacities to overprovision the networks relative to the experiments needs and use
- Other data intensive sciences are coming online soon (SKA, LSST, etc.)
- Network provisioning will need to evolve
  - Focusing not only on network capacity, but also on other network capabilities
- DC networking is evolving in reaction to containers/virtual/cloud resources
- It’s important that we explore new technologies and evaluate how they could be useful to our future computing models
  - While it’s still unclear which technologies will become mainstream, it’s already clear that software (software-defined) will play major role in networks in the mid-term

NFV WG Report

NFV WG produced an interim-report describing the current practice, challenges and needed future work.

The report for NFV Phase 1 report is at
https://doi.org/10.5281/zenodo.3565562

Three main topics are covered

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

Future Work for Experiments/NRENs

The NFV report proposed areas of future work, primarily motivated by HEP and NREN needs, but targeting the broad R&E users of our global networks

The three areas proposed for work are:

1. Making our network use visible (**marking**)
2. Shaping WAN data flows (**pacing**)
3. Orchestrating the network to enable multi-site infrastructures (**orchestrating**)

This was presented to the WLCG experiments and NRENs during the **January 2020 LHCONE/LHCOPN** meeting and discussed in detail. We achieved a ***strong consensus*** that this work needed to move forward ASAP!

New Research Networking Technical WG

We are now organizing a new **Research Networking Technical Working Group**, focused on addressing the identified needs of HEP and the NRENs (and others!)

Charter for the group is at https://docs.google.com/document/d/1l4U5dpH556kCnoIHzyRpBl74IPc0gpgAG3VPUp98lo0/edit?usp=sharing

Kickoff meeting planned for week of April 20-24th. If you are interested, please:

- **Join** our group http://cern.ch/simba3/SelfSubscription.aspx?groupName=net-wg
- **Respond** to our Doodle poll https://doodle.com/poll/xmiqntndu6td8xiw

References

Research Networking Technical Working Group charter
https://docs.google.com/document/d/1l4U5dpH556kCnoIHzyRpBl74IPc0gpgAG3VPUp98Io0/edit?usp=sharing

HEPiX Network Function Virtualization WG Report: https://doi.org/10.5281/zenodo.3565562

HEPiX: http://www.hepix.org

WG Meetings and Notes: https://indico.cern.ch/category/10031/

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

Backup slides

Making our network use visible

Understanding HEP 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 we are unable to easily understand the data in flight.

- The proposed work here is to identify how we might label our traffic at the packet level to indicate which **experiment** and **activity** it is a part of.
  - Important for sites which support many experiments
  - With a standardized way of marking traffic, any NREN or end-site could quickly provide detailed visibility into HEP traffic to and from their site.
- The technical work would encompass how to mark traffic at the network level, defining a standard set of markings and providing the tools to the experiments to make it easy for them to participate.
  - VMs/containers will make marking traffic easier where they are in use.

Pacing/Shaping WAN data flows

It remains a challenge for HEP storage endpoints to utilize the network efficiently and fully.

- An area of potential interest to 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

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 in which we may organize our computing and storage resources will need to evolve.
- Data Lakes, federated or distributed Kubernetes and multi-site resource orchestration will certainly benefit (or require) some level of WAN network orchestration to be effective.
  - We would suggest a sequence of limited scope proof-of-principle activities in this area would be beneficial for all our stakeholders.

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																																

Packet Marking Overview (Feasibility)

The proposal is to provide a mechanism to mark our network packets with the **experiment** and **activity**

- Both **IPv4** and **IPv6** support optional headers, IPv6 has 20 bits for “flow labeling”. We should be able to get 20 bits in either version (via options or flow labeling)
- The target is the “source” emitting the packets: job, application, storage element.
- Goal is that at any point in the R&E network, we can identify/account/monitor traffic details and this helps both networks and experiments:
  - NRENs can easily quantify what science they supported
  - Experiments can quickly understand how changes get expressed in the use of the network
- Use libnet: https://github.com/libnet/libnet