NFW WG Network Discussion

Shawn McKee / University of Michigan
Marian Babik / CERN

LHCOPN/LHCONE workshop

CERN, January 13-14, 2020

Overview

There are two goals with this session:

1) Inform everyone on the recent work to document and plan network efforts - NFV WG report
2) Discuss how we want to continue with the networking R&D efforts in general as well as specifically how we organise the work/effort

HEPiX NFW WG report was primarily focused on sites, today we'd like to discuss potential areas of future work focused on the experiments and their use cases

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

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 last year (https://indico.cern.ch/category/10031/)

Mailing list: https://listserv.in2p3.fr/cgi-bin/wa?SUBED1=hepix-nfv-wg

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-M3KvXlfXHUnYgsVUhBpKFjvjUQ/edit#heading=h.flthknqgm3ub (**Feel free to comment on the whole report!**)

Report has **3 chapters**:

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

Future (phase 2) will be discussed at the upcoming HEPiX in Taiwan!

Future Work for Experiments/NRENs

The report proposes areas of future work with the experiments

- Open for discussion and **more importantly your feedback**

Yesterday we heard consistent interest in making network use more visible (all VOs), more effective (CMS pacing, others) and orchestrated (managed, controlled). This matches what we identified:

Areas discussed in the document (pages 53-56):

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

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.

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.**

Discussion

We'd like to understand which areas of future work our community should focus on that would have clear benefits for the experiments.

How do we organise the work and effort?

Our goal today is to **agree on how best to proceed**, not to go into technical details (which can be solved by dedicated TF) and also understand how best to follow up on a collaboration between sites, experiments and R&Es.

Let's Discuss

What things can we agree are worthwhile to pursue?

- Agreement to pursue implies we will have effort from the experiments, sites and NRENs!!

For activities that are pursued we should setup technical working groups and timelines…

Questions, Comments, Suggestions?

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

References

WG Report: https://docs.google.com/document/d/1w7XUPxE23DJXn--j-M3KvXlfXHUnYgsVUhBpKFjyjUQ/edit#
WG Meetings and Notes: https://indico.cern.ch/category/10031/
SDN/NFV Tutorial: https://indico.cern.ch/event/715631/
2018 IEEE/ACM Innovating the Network for Data-Intensive Science (INDIS) – http://conferences.computer.org/scw/2018/#!/toc/3
OVN/OVS overview: https://www.openvswitch.org/
GEANT Automation, Orchestration and Virtualisation (link)
Cloud Native Data Centre Networking (book)
MPLS in the SDN Era (book)

Backup slides

Organisations (Who, What)

HSF (Experiments, software)

HEPiX (Sites, best practices, technologies)

WLCG (Sites, Experiments, operations)

DOMA (Experiments, prototyping, HL-LHC)

LHCOPN/LHCONE (NRENs, networking)

We need aspects of all of these for possible work. Where should the “home” of our network efforts be?

Proposing Activities of Interest

What is useful? Feasible? Possible?

The idea of marking, shaping and orchestration are potential steps in order of assumed difficulty and time-to-implement

Marking and shaping/pacing must happen on the source

Orchestration is much more feasible once marking is 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																																

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

Cloud Native DC Networking

Cloud Native DC Networking

- Discusses current trends in industry and their potential impact on HEP
  - Cloud native networking - as a response to the paradigm shift in compute (VMs, containers) - rethinking network design of the DCs
  - **Network disaggregation** - cloud providers push for open network environments (ONIE, **OCP**, etc., related Linux kernel developments)
  - **Network virtualisation** - surveys a number of approaches that offer ways to build scalable, robust and cost-effective DC networks
  - **DC edge services**, Data Center Interconnect (DCI) tech, etc.
- Aimed at site managers, network engineers and experiments
  - Introduce new networking tech and trends to wider audiences
  - Comprehensive overview of the cloud native networking landscape*

Programmable Wide-Area Networks

Programmable WAN

Focuses on the future of network provisioning and operations

Programmable Networks for Data-intensive Sciences

- Surveys existing R&D projects in the area of SD/WAN (incl. SDX), Network Orchestrators and Network-aware Transfer Systems
- Projects such as SENSE, multi-ONE, NOTED, BigDataExpress, SDX projects, etc.

R&E Networks Programmable Services

- Surveys R&E plans on higher-level network services
- ESnet6, FABRIC and GEANT's OAV plans

Outlines core challenges and outlook

Outlook

What's next ?

The report presents number of approaches and technologies that could help us establish global scientific networks, improve network efficiency and monitoring as well as provide scalable models for Cloud Native DCs

We will need to invest in networking R&D to understand how we can design, deploy and test these technologies to match our use cases so that they become solutions (and at some point also converge on a set approaches that will be interoperable). Such activities would very likely require a cross-collaboration btw sites, experiments and R&Es.

How can we best organise our efforts (wrt WGs/TFs and more importantly wrt R&D projects/funding) ? The work required will likely need to be done across the entire infrastructure landscape (i.e. R&Es, sites and experiments). We currently have only few examples where such cross-collaborations were successful.

and also ...

Where do we see LHCONE in 5 years ? Currently, we seems to be pursuing 3 different directions:

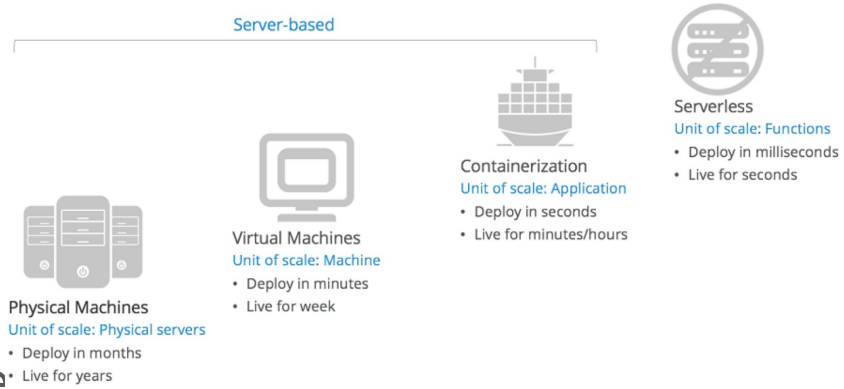
1. Fully automated networking with no/minimal user interaction (SDX, etc.)
2. Network orchestrators (NOTED, SENSE, etc) with different levels of decentralisation (and some level of interaction with end-user systems)
3. WAN traffic separation

We have a number of existing activities (mostly with the exact same people participating in different WGs/TFs, etc.), some already have networking (DOMA), only two have active participation of R&Es (this WG and LHCOPN/LHCONE), how can be better involve/collaborate - do we continue with a focused group ?

Paradigm Shift in Computing

Moving from static physical machines to very dynamic models with VMs, containers, clusters of containers, etc.
This has major impact on networking requirements in DC

- **Node lifecycle in msecs**
- East-west traffic increases
- **Nodes can migrate**
- Multiple orchestration methods
- **Networking across stacks and within needs to perform**



This transition has already started, we already have experiments running payloads in containers, services bundled in K8s pods, physics analysis in K8s has been demoed recently

Programmable WAN Use Cases

Network capabilities in WAN have a number of use cases:

- Traffic engineering
  - Additional capacity exists and can be provisioned by steering traffic via alternate paths
- Network provisioning
  - With DC networking moving towards WAN protocols, there is an opportunity to leverage this to find alternative ways how to organise/manage current L3VPNs/LHCONE (multi-ONE)
- Provide QoS transfers
  - We have been running two dedicated networks (LHCOPN and LHCONE) which mainly differ in QoS provided. Other experiments will likely come up with similar requirements.
- Improve network to storage performance
  - Currently there is often a mismatch between target storage and network performance
- Capacity sharing
  - Network as a resource is becoming likely in the future (like compute/storage today)
- Effective use of HPCs and Clouds

Networking Challenges

- Capacity/share for data intensive sciences
  - No issues wrt available technology, however
  - What if N more HEP-scale science domains start competing for the same resources ?
- Remote data access proliferating in the current DDM design
  - Promoted as a way to solve challenges within experiment's DDM
  - Different patterns of network usage emerging
    - Moving from large streams to a mix of large and small frequent event streams
- Integration of Commercial Clouds
  - Impact on funding, usage policies, security, etc.
- Technology evolution
  - Software Defined Networking (SDN)/Network Functions Virtualisation (NFV)

Technology Impact

- Increased importance to oversee network capacities
  - Past and anticipated network usage by the experiments, including details on future workflows
- New technologies will make it easier to transfer vast amounts of data
  - HEP quite likely no longer the only domain that will need high throughput
- Sharing the future capacity will require greater interaction with networks
  - While unclear on what technologies will become mainstream (see later), we know that software will play a major role in the networks of the future
  - We have an opportunity here
- It's already clear that software will play major role in networks in the mid-term
- Important to understand how we can design, test and develop systems that could enter existing production workflows
  - **While at the same time changing something as fundamental as the network that all sites and experiments rely upon**
  - We need to engage sites, experiments and (N)REN(s) in this effort

Software Defined Networks (SDN)

- Software Defined Networking (SDN) are a set of new technologies enabling the following use cases:
  - **Automated service delivery** - providing on-demand network services (bandwidth scheduling, dynamic VPN)
  - **Clouds/NFV** - agile service delivery on cloud infrastructures usually delivered via Network Functions Virtualisation (NFV) - underlays are usually Cloud Compute Technologies, i.e. OpenStack/Kubernetes/Docker
  - **Network Resource Optimisation (NRO)** - dynamically optimising the network based on its load and state. Optimising the network using near real-time traffic, topology and equipment. This is the core area for improving end-to-end transfers and provide potential backend technology for DataLakes
  - **Visibility and Control** - improve our insights into existing network and provide ways for smarter monitoring and control
- Many different point-to-point efforts and successes reported within LHCOPN/LHCONE
  - **Primary challenge is getting end-to-end!**
- While it's still unclear which technologies will become mainstream, it's already clear that software will play major role in networks in the mid-term
  - Massive network automation is possible - in production and at large-scale
- HEPiX SDN/NFV Working Group was formed to bring together sites, experiments, (N)RENs and engage them in testing, deploying and evaluating network virtualization technologies

Network Operations

- Deployment of perfSONARs at all WLCG sites made it possible for us to see and debug end-to-end network problems
  - OSG is gathering global perfSONAR data and making it available to WLCG and others
- A group focusing on helping sites and experiments with network issues using perfSONAR was formed - WLCG Network Throughput
  - Reports of non-performing links are actually quite common (almost on a weekly basis)
  - Most of the end-to-end issues are due to faulty switches or mis-configurations at sites
  - Some cases also due to link saturation (recently in LHCOPN) or issues at NRENs
- Recent network analytics of LHCOPN/LHCONE perfSONAR data also point out some very interesting facts:
  - Packet loss greater than 2% for a period of 3 hours on almost 5% of all LHCONE links
- Network telemetry (real-time network link usage) likely to become available in the mid-term (but likely not from all NRENs at the same time)
- It is increasingly important to focus on site-based network operations