Research Networking Technical Working Group

Shawn McKee, Marian Babik

on behalf of the RNTWG

LHCONE/LHCOPN Meeting (virtual)

May 13, 2020

Presentation Overview

From our last meeting in January, the various LHC/HEP experiments described their networking needs, interests and use-cases

The experiments reinforced what the HEPiX NFV phase I report suggested were useful areas to focus effort upon:

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

In this presentation I want to cover the recent work to create a working group to push the outlined work forward

WLCG Network Requirements

- Many WLCG facilities need network equipment refresh
  - Current routers in some sites are End-Of-Life and moving out of warranty
  - Local area networking often has 10+ year old switches which are no longer suitable for new nodes or operating at our current or planned scale.
- WLCG experiment's planning is including networking to a much greater degree than before
  - HL-LHC computing review: DOMA, dedicated networking section
  - ATLAS HL-LHC Computing Conceptual Design Report, highlights needs
  - Both 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**: ATLAS 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

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 (80 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)

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 is unable to understand the data in flight. **In general the monitoring we have is experiment specific and very difficult to correlate with what is happening in the network.**

- 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.

Packet Marking	Science Domain												Traffic Type							
Bits (Assume 20)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

(See next slide example)

- The technical work would encompass how to **mark traffic** at the network level, defining a standard set of markings, **provide the tools** to the experiments to make it easy for them to participate and define how the NRENs can **monitor/account** for such data.

Packet Marking Overview (Example)

The proposal is to provide a mechanism to mark our network packets with the **experiment/owner** 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

Packet Marking	Science Domain												Traffic Type							
Bits (Assume 20)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
ATLAS-any	0	0	0	0	0	0	0	0	0	0	0	1	x	x	x	x	x	x	x	x
perfSONAR	x	x	x	x	x	x	x	x	x	x	x	x	0	0	0	0	0	0	0	1
CMS-remote-xroot	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0

- **The target**: any “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

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.

Straw man proposal for work plan

We already identified areas of work, so the proposed work plan would be (per area):

- Identify who is interested in participating
- Identify concrete technologies we'd like to look at
- Perform feasibility study (for each technology)
  - Evaluate tasks/work necessary for adoption across stack
    - Experiments applications, Network equipment support, Application support (Linux kernel support, libraries), Deployment aspects, etc.
- Implement prototype, perform initial tests
- Identify tasks/work needed for broader adoption and seek approval/effort/funding for this

Goal: finish prototype marking stage by EoY (or Q1 2021)*

Packet Marking Sub Group

Since Packet Marking was first on the list, we have a soon-to-be-announced document focused on organizing this work

See draft here

Join the mailing list to participate

My goal would be to have some amount of WLCG traffic being labeled by the end of this calendar year and we should discuss this.

Packet Marking Challenges

We would like this to be applicable for ALL significant R&E network users/science domains, not just HEP

- Requires us to think broadly during design

How best to use the number of bits we can get?

- Need to standardize bits and publish and **maintain**!!
- Can we agree on some standard “type” bits?

What can we rely on from the Linux network stack and what do we need to provide?

What can the network operators provide for accounting?

Let's Discuss!

We have identified packet marking as important for WLCG
How do we enable it for all (most) of our data sources?

- Storage elements, jobs, applications

We really need a broad range of expertise involved: network programming, standardization experience, experiment software expertise, storage software expertise, NRENs, documentation experience, monitoring, accounting, etc.

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
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)
RNTWG Google Folder
RNTWG Wiki
RNTWG mailing list signup

Backup slides

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 - 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?

Some Important Notes

Network monitoring needs to continue and evolve

- IRIS-HEP OSG-LHC maintains and develops the perfSONAR infrastructure for our sites and networks
- SAND is focused on exploitation of the collected metrics but ends this year

We have a good collaboration with ESnet, who provides our primary connectivity for WLCG traffic between North America and Europe. We have Monthly meetings to analyze our use and help ESnet plan how best to support our future needs.

The new IRNC testbed option will be important for our prototyping

High Level Notes

What is useful? Feasible? Possible?

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

Marking and shaping/pacing **must happen on the source**

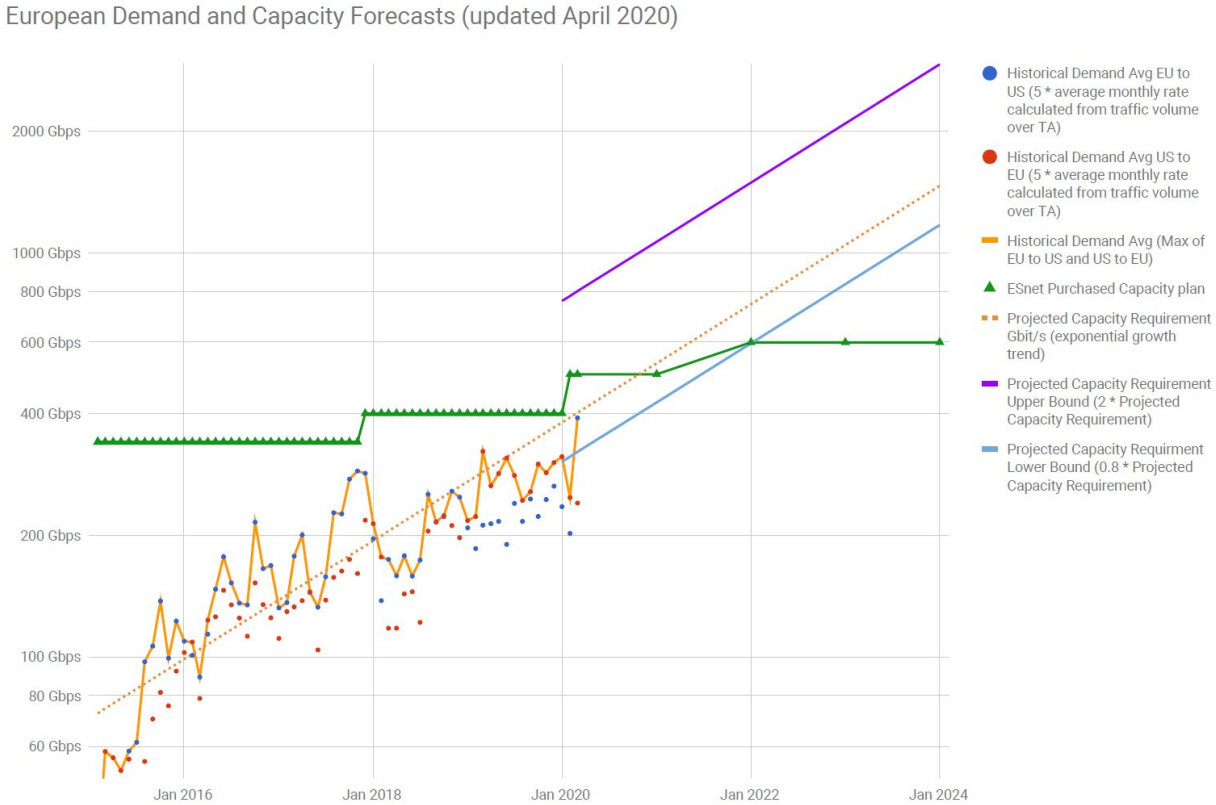
Orchestration is much more feasible once marking is in place

ESnet TransAtlantic Capacity Forecast

What ESnet can afford for $2M/year is the green line.

Capacity evolution for our **terrestrial** networks looks reasonable in terms of technology up to the HL-LHC

- Uncertainties arise from other users
- Undersea capacity will be a **big** challenge due to optics range



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/)

Future Work for Experiments/NRENs

The report proposes areas of future work with the experiments

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

During the LHCONE/LHCOPN meeting 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 proposed for this WG (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)

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.**