IPv6 and RNTWG Sub-groups

Shawn McKee, Marian Babik *on behalf of the RNTWG*

HEPiX IPv6 Working Group

June 3, 2020

Presentation Overview

I want to discuss the Research Networking Technical Working Group activities in relation to IPv6.

Summary: It looks like IPv6 will be a necessary component of our plans to mark packets.

First some context and background...

Research Networking Technical WG

Charter:
https://docs.google.com/document/d/1l4U5dpH556kCnoIHzyRpBl74IPc0gpgAG3VPUp98lo0/edit#

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

Members (88 as of June 2, 2020, in no particular order):

Tony Cass, Eric Christian Lancon, James Letts, Harvey Newman, Duncan Rand, Rolf Seuster, Edoardo Martelli, Shawn McKee, Simone Campana, Andrew Hanushevsky, Marian Babik, James Walder, Petr Vokac, Alexandr Zaytsev, Bruno Hoeft, Raul Lopes, Mario Lassnig, Han-Wei Yen, Wei Yang, Edward Karavakis, Tristan Suerink, Ian Collier, Garhan Attebury, Vitaliy Kondratenko, Justas Balcas, Pavlo Svirin, Shan Zeng, Mihai Patrascoiu, Jin Kim, Richard Cziva, Ajinkya Fotedar, Alexander Germain, Brian Carpenter, Buseung Cho, Casey Russell, Chris Rapier, Chris Robb, Chris Tengi, Dale Carder, Doug Southworth, Eli Dart, Eric Brown, Evgeniy Kuznetsov, Ezra Kissel, Fatema Bannat Wala, Hans Yodzis, Joe Breen, James Blessing, James Deaton, Jan Erik Sundermann, Jason Lomonaco, Jerome Bernier, Jerry Sobieski, Ji Li, Joe Mambretti, Julio Ibarra, Karl Newell, Li Wang, Marco Marletta, Mariam Kiran, Mark Lukasczyk, Matt Zekauskas, Mian Usman, Michael Lambert, Michal Hazlinsky, Mingshan Xia, Paul Acosta, Paul Howell, Paul Ruth, Phil Demar, Pieter de Boer, Roman Lapacz, Sri N, Stefano Zani, Tamer Nadeem, Tim Chown, Tom Lehman, Vincenzo Capone, Wenji Wu, Xi Yang, Chin Guok, Lars Fischer, Christian Todorov, Dale Finkelson, Frank Burstein, Rich Carlson, Marcos Schwarz, Susanne Naegele-Jackson

Work Areas for the RNTWG

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 in the HEPiX NFV working group report, so we proposed three areas of work:

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

We are organizing these three areas into sub-groups focused on defining, prototyping and testing what would work for the R&E community.

Packet Marking Overview (Initial Thoughts)

The proposal is to provide a mechanism to mark our network packets with the **experiment/owner** and **activity.** 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 flows start and end, but is unable to understand the data in flight.

- 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.

Packet Marking Sub Group

Since Packet Marking was first on the list, we have a document focused on organizing this work

See evolving draft here

Join the RNTWG mailing list to participate

Our first Packet Marking sub-group meeting will be this Thursday, June 4, 11-Noon Eastern time (5-6 PM CEST).

https://indico.cern.ch/event/925729/

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 Challenges (Where?)

- IPv4: It seems ‘option headers’ is the only viable location. **Problems**:
  - No support in stock kernels for this
  - Some network device don’t handle option headers and drop traffic
  - Adding 20 bits (anything >16) would require adding 2 32-bit header words
- IPv6: New option because of ‘flow label’. Good news:
  - Flow labels are supported in kernels (varies).
    - RHEL 8 will have /sys interface.
    - RHEL 7 systems can already set bits
  - **perfSONAR** testing was successful in getting bits end-to-end
  - **Problem**:
    - Flow label acceptable use was changed from original RFC. New use is **only** for adding entropy for use in path selection. Bits should be pseudo-random :(

Using Alternative IPv6 Options

With IPv6, we have other options besides the flow-label

The Destination option header could be used for marking

- Pro: less conflict with pre-existing intentions from standards bodies
- Con: Linux implementation in particular requires root or CAP_NET_RAW
- Con: Extracting bits via routers and exporting it via IPFIX may not be practical.

Use Hop-by-Hop option, discussed in draft-krishnan-ipv6-hopbyhop-05.

- Pro: Draft that proposes its use - draft-ietf-6man-mtu-option-02.
- Con: Used to signal routes should process, if not configured may DROP
- Con: Can devices access/process these bits?

Use IPv6 addressing to indicate owner and purpose

- If we knew every IPv6 SRC had a range of 20 zeros somewhere…

In general things are very complicated (see the document)

Let's Discuss!

For **packet marking**, I believe we will end up requiring IPv6

- Still lots of questions about where the bits go, how to get them there and how to utilize them

For **traffic shaping** and **network orchestration**, it is less clear about requiring IPv6.

- Does this working group have suggestions or considerations regarding needing/using IPv6 for this effort?

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 packet marking document
RNTWG Google Folder
RNTWG Wiki
RNTWG mailing list signup

Backup slides

News

- 88 members in the mailing list
- Update on the WG was presented at various meetings:
  - LHCOPN/LHCONE workshop
  - Internet2 Community Measurement, Metrics, and Telemetry
  - ATLAS DDM and joint USCMS/USATLAS meetings
- Drive area created and draft docs on all three areas:
  - Thanks for all your contributions (mainly to the packet marking)
  - We're still looking for volunteers to lead the proposed areas
- Tim Chown has invited Brian Carpenter to the WG
- Tim contacted authors of the other three relevant RFCs
  - 8250, 7837 and draft-fz-6man-ipv6-alt-mark-09

Work plan review from last meeting

Notes from last meeting

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
- Implement prototype, perform initial tests
- Identify tasks/work needed for broader adoption and seek approval/effort/funding for this

Organisation of work: Right now using mainly docs and mailing list, open for other ideas

Today - focus on packet marking

- Draft document
- Packet marking status:
  - Identify who is interested in participating - Done
    - Please sign yourself in the document if interested
  - Identify concrete technologies we'd like to look at - On-going
    - Core technologies and challenges identified - TBD
    - Still open for additional ideas
  - Perform feasibility study (for each technology) - Started
- Today - discuss the document in detail with the following goals:
  - Identify the main questions we want answered
  - Identify the list of appropriate tasks that one to a few people could undertake before our next meeting
  - Identify missing areas of work