Benchmarking Methodology WG (BMWG)
July 17, 2017 | 9:30-12:00 CEST (UTC-2)
Monday Morning session I   Room Karlin III   OPS

Remote Participation:
http://www.ietf.org/meeting/99/index/index.html
http://www.ietf.org/meeting/99/remote-participation.html

Note Taker: Marius Georgescu
A special thank you to our note taker, Marius, who kept very good minutes for us, reviewed the recording to shore them up, despite heavy participation in the room and at the mic as well. We have Marius to thank for the colour added to the minutes below ("Bradneresque" is going to have to be a new word! ☺)

-=-

0. Agenda
   - No agenda bashing

1. WG Status (Chairs)
   - The following drafts have all reached a state of completion:
     - IPv6 Neighbor Discovery
     - IPv6 Transition Benchmarking
     - VNF and Infrastructure Benchmarking Considerations
     - Benchmarking Virtual Switches in OPNFV
     - Data Center Benchmarking

2. Charter and Milestones (Chairs)
   - Larger rechartering discussion at the end of draft discussion; see end of this document for charter/rechartering notes from the discussion

3. Benchmarking Methodology for SDN Controllers
   Presenter: Sarah Banks
   Draft: https://tools.ietf.org/html/draft-ietf-bmwg-sdn-controller-benchmark-meth-04
   Draft: https://tools.ietf.org/html/draft-ietf-bmwg-sdn-controller-benchmark-term-04

   - Sarah:
     - Resolves comments from the last WGLC
     - This draft has been worked for quite some time and we would like to do another WGLC (Working Group Last Call).
   - Al: One of my comments is related to the VNF considerations draft. When I was looking at the 3x3 metric coverage matrix, the accuracy column was missing. I'm suggesting that we add the accuracy column and add as metric a measured loss-

ration, a two-dimensional plot with offered rates, loss rate and asynchronous achieved rates.

- Marius: I just wanted to say I'm satisfied with the way my feedback was covered.
- Sarah: I would prefer to make the changes and then go through WGLC.
- Al: We'll have a WGLC after the draft is revised.

4. Benchmarking Methodology for EVPN and PBB-EVPN
   Presenter: Sudhin Jacob
   Draft: https://tools.ietf.org/html/draft-kishjac-bmwg-evpntest-06
   Slides: https://datatracker.ietf.org/doc/slides-99-bmwg-bench-marking-of-evpnpbb-evpn/00/

- Marius: The setup needs clarifications. There are two traffic generators and no receiver.
- Sudhin: The CE acts as a receiver.
- Marius: It should be clarified which machines are part of the tester and which ones act as DUT.
- Sarah: When we see multiple routers on a diagram we think all of them are being tested. I see your point that not all are being tested, but I also see Marius' point that it needs to be clarified. I find it a little confusing too.
- Sudhin: You want to have a demarcation, it's noted. We will make it more explicit.
- Al: On the test diagram, can you explain how the packets would flow for those two cases (remote peer vs local peer)?
- Sudhin: The CE would act as a bridge.
- Al: that's the local case, can you also explain the remote?
- Sudhin: Traffic is sent directly to R1. The MACs in R1 will be advertised in BGP, then from the EVPN database it will be populated to the local path. R1 has the advertise and there is time difference there.
- Al: So, the fundamental answer to Marius's question is that all we really need to accomplish MAC learning is to generate packets with MACs that need to be learned and to measure the benchmarks we need to look at the control plane interactions between the devices in your setup in order to determine when a MAC has been learned and when it has been advertised in BGP. Is that correct?
- Sudhin: Yes, this is what we are doing. The BGP itself has serialization delay.
- James Uttaro: To echo what Al said, I was a little confused. So, bencharking is from DUT to R1, bidirectional advertisment of the type 2 route MAC/IP. Where I'm confused is going the other way CE1 generates traffic to DUT, DUT sends a type 2 advertisment to R1. What I was saying was the benchmark you're doing here is between DUT and R1 and R1 to MHPE2. Right?
- Sudhin: that's correct.
- James: My comment about this is in a multihoming active-active environment you may never get a MAC advertisment from MHPE2, just the ESI advertisment that is equal to the one on the DUT. When you benchmark and you say how quick, if you remove the ESI advertisment from MHPE2, you invalidate all the MACs at R1. So, it's possible that you won't get the type 2 advertisment.
- Sudhin: You are correct. We can make it a single-active.

- James: What I was saying is what you're measuring is the withdrwal of the ESI and the invalidation of all the MACs.
- Sudhin: It's a vanilla test not a trigger test.
- James: The ESI withdrawal is an optimization to make much faster than you would if you had to invalidate every MAC upon learing of the MAC withdrawal.
- Sudhin: We'll look into it.
- Sarah: What I'm hearing is that you should consider the withdrawal scenario. There is no learning without withdrawal. Learning might be its own discrete test, but withdrawal is very important to measure. The draft would be lacking without it. I really hope you will consider adding it.
- James: Just from a carrier perspective, a new customer comes on, you learn something. When the service goes down for a certain amount of time the customer will not be happy. So, when you benchmark that should be the primary thing.
- Sudhin: We added that as a link failure local and remote. How fast will the MAC be flushed.
- James: Going back to the diagram, in an active-backup scenario, as an operator I would like to know how fast will the service come back up. How long does the withdrawal take.
- Sudhin: If you break it here or the DUT, that is not covered. Flush and remote link failure is covered. The ESI cutting of is indirectly covered (snip).
- Sarah: From what I remember reading, you are covering MAC flush on the DUT. That's explicitly not what he was saying. We can take it to the list or discuss it after. I would like to convince you to consider adding that.
- Sudhin: Let me get back to you with that.
- James: I question for MAC ageing whether or not the timeout that comes from the withdrawal locally is as much a function of the EVPN as it is a function of the local implementation of reprogramming those boxes once you age out. Back to the ESI optimization, if a bunch of MAC timeout associate with a certain Ethernet segment, you're not going to get individual withdrawal for that. You're going to get the withdrwal for the ESI and you're going to invalidate everything.
- Sudhin: OK, thank you.
- Sarah: For the next meeting when you bring the slides back in, can you put the diagram on the side and report what's where? You're doing it with the highlighter here in the room and pointing to the screen, but for the remote folks that is not carrying through.
- Al: More than that, this is the kind of detail that needs to be added in the draft together with references from the RFC, so that everyone can understand these test setups. We all sort of understand the concepts here, but the unique details of EVPN are not my expertise. Obviously, it's James' expertise and your expertise and those are coming out in this discussion. Now we need to make sure that's going into the draft as well.
- Sarah: There's a lot of pushback to the feedback you're getting. If we are struggling with it, then other folks who read the RFC will struggle with it too.
- Sudhin: As Marius said, it has to be fine tuned. Fine tuning will be done.

- Sarah: I think it's a little more than fine tuning. A good chunk of the methodology: why you're doing the test, what's what and what's where, that is missing from the test cases. I think it's a good amount of surgery that needs to happen to make the document readable and the tests repeatable, so that if I were doing the tests and Al was doing the tests, we would both execute in the exact same way. This is what BMWG is about.
- Al: There are important background details missing.
- Sudhin: It was written in an environment where people are familiar with it.
- Sarah: Sometimes it's just not clear and I'm making assumptions. I shouldn't have to assume. The document should be very clear and straightforward.
- Sudhin: We will explain in such a way that a person who's not familiar with EVPN can understand.
- Sarah: Marius, could I ask. Can you take a look at the diagram and provide surgical feedback around that? Here's what does make sense to me, here is what doesn't. I think he's really struggling with where. Give me some examples. You'll take the first diagram, I'll take the first use case.
- Marius: Sure, I can come up with a solution for what I'm trying to correct. The thing is there's a lot of pushback to the sent feedback. I think it's important that you dig deeper in the feedback.
- Sarah: From three people now you've heard that the draft is not clear. It is not on us to do this. I see that you're not very receptive to the feedback. At the end of the day, if the draft is not readable, it is not going to get adopted. You're reading the overarching point that you have a readability problem.
- Al: If you want to type out things on the list that's fine, but you should really try to take advantage of the face-to-face time, to be sure that what we're writing down gets communicated. You have to understand that comments are a gift.
- James: Type 5 is essentially an IPv4 route with no associated MAC. Type 2 is MAC + IP. The IP routes have to be stored regardless if it's type 5 or type 2. I don't know what you're looking for here exactly with type 5 as supposed to a generic type 2.
- Sudhin: A type 5 will not be normally advertised as a type 2 because this is not part of EVPN.
- James: My point is both type 5 and type 2 contain an IPv4 prefix. You're trying to see how many of these prefixes you can store on a DUT. So, in many ways how is type 5 different than type 2? As a provider I would like to know what you're trying to tell me? Why do I need this specific test.
- Sudhin: This is about the scale of the type 5 the DUT can sustain. That was the feedback I received last IETF.
- Sarah: I think you should consider what James is saying now as well.
- James: One thing about this, there is a limit to the number of context that could be configured, regardless of how many things you configure in them. It might be interesting to know how many EVIs can be configured.
- Sudhin: That will give you the exact picture. In each context what is the capacity.
- Sarah: What you just said doesn't match your slide. Which is it?
- Sudhin: The DUT has to advertise and it should only be sent to the remote routers.

- Al: Thanks for being willing to accept lots of comments. You've gotten the feedback that you really needed to make this a better draft.

5. Benchmarking Methodology for Service Function Chain
   Presenter: Taekhee Kim
   Draft: https://tools.ietf.org/html/draft-kim-bmwg-sfc-benchmark-00
   Slides: https://datatracker.ietf.org/doc/slides-99-bmwg-considerations-for-benchmarking-service-function-chain/00/

- Al: A quick clarification, The speed and accuracy of the SFC creation. For all of these creations/deletions/modifications can have their own metrics.
- Sarah: Can you remind me, when you start to measure things like a TCAM usage, are you giving guidelines about what the switch should be running on.
- Taekhee: This time we tested on the white box switches, but did not expect the TCAM size to be affected.
- Sarah: So, is TCAM utilization something that we're going to do more on physical switches than virtual?
- Taekhee: No, I don't think so.
- Al: We'll find out as this work proceeds.
- Warren: is this being discussed in the SFC group at all?
- Taekhee: After this session, I will be going to the SFC WG. I need their opinion as well.
- Al: Very interesting work. As you mentioned the NSH (Network Service Header), which you weren't able to incorporate in your own work, we do have some people with expertise in this area. So, we could incorporate some of the work in this draft or a part 2 version.

6. Re-chartering BMWG
   Chairs and AD leading the discussion

- Marius: WLAN Benchmarking is something we are interested in and would like to see it a standard. I'm looking for other interested parties to start working on a draft. Al sent me an older draft (https://tools.ietf.org/html/draft-alexander-bmwg-wlan-switch-meth-01) and I think it's a very good draft and could be a food start.
- Al: Tom Alexander was working with IEEE 802.11 and he saw some gaps. We actually exchanged a liason with the testing group 802.11T and they said go ahead and do this. I was the only person to review the draft, with the exception of Scott Bradner, who became a co-author.
- Sarah: Let's check with the WiFi Alliance and see if they're doing anything like this. Separately there are a whole school of folks who do the protocol stuff here for WiFi. I already have a couple of contacts we can solicit to see if they can help. I'm certainly happy to look into a draft with you. If we were to take it on we would of course need people with expertise.
- Marius: I'm happy to contribute in any way I can, because I want to see an RFC written on this subject.

- Al: Tom raised this point when trying to get people to read the draft back at IETF71. The hospitals were beginning to use WIFi for the communication between their systems and if WiFi didn't work properly, lives might be at stake. Even with that call to arms, I was the only one who read the draft.
- Sarah: Maybe because we don't have expertise in the room. If we get the crosspollination working with the 802.11 experts in IETF, I firmly believe benchmarking for WiFi is such a good idea. We just need to talk to them to see if they are willing to participate.
- Al: I will volunteer to get in touch with him on this subject.

- Al: About the vswitch stuff, we had a draft on energy consumption. That was the kind of thing that didn't gather enough interest in the WG. If anyone is interested in that, I can share with you the previous work on that. However, we can't do that unless anyone champions it.

- Al: There's a couple of ways we can update RFC2544. I'll get to that in the NFVI presentation. As a spoiler, the back-to-back frames and latency need updates. The latency measurement in RFC2544 is based on single path. Of course, the manufacturers do more than that. I would like us to discuss between now and the next meeting the new items intensively on the list. If we are to have an interim meeting, we can do that as well. I hope we have text for the next charter by IETF100 and have a face-to-face discussion about what we're going to do.
- Warren: Is this new charter text or milestones?
- Al: It comes down to the way you want to see charters written. For the first 15 years we had a general charter. Dan Romascanu was the ops AD that suggested we have a more specific charter with bullet items in the charter to increase our visibility. That meant that every couple years we had to re-charter after completing the list. We always had proposal, like the IPv6 transition draft that fit our general charter, about which we asked our AD at the time, Joel. If your preference is to go back to the general charter, under which we evaluate as a leadership team what fits and what doesn't and not worry about the specific bulletin, then that's a lot shorter exercise.
- Sarah: If we choose one or the other, let's be consistent because 3 quarters of the charter are paragraphs about the specific work.
- Al: I'm happy to go back to the way we used to do things. We need the same discussion, but we don't need to write as much in the charter. That's the most awesome feedback for today. Makes our job a lot easier.

7. WG Discussion
Topic: Dataplane Performance: NFVI Benchmarking Measurements
Presenter: Al Morton
Slides: https://wiki.opnfv.org/download/attachments/10293193/VSPERF-Dataplane-Perf-Cap-Bench.pptx?api=v2
General Summit info: https://www.opnfv.org/opnfv-summit-2017-event-recap

(General Editor Note: The questions captured below make sense really once you've seen Al's presentation. I strongly suggest you review either the recording of the session, or at least the slides, to familiarize yourself with the content. For example, Sarah's comment about the results that changed when measuring 64 bytes versus 128 bytes don't make sense unless you're looking at the presentation, and even then, a bit is lost because the Slide number wasn't captured. The recording will give you that context, if you're interested.)

- Sarah: Did you draw anything from the fact that the OVS and the VPP numbers flip-flopped from 64 to the 128 measurement?
- Al: I think that's the instability of the maximum.
-
- Sarah: I feel like I'm channeling Scott. ☺ How many times was the test run?
- Al: A lot of times (in thick Bradneresque voice:). Basically, when we saw something inconsistent, we reran the test.
- Marius: Just to extend Sarah's question, how long was the test run?
- Al: Long enough :) They were at least 60 second duration.

- Venkatesh Palani: why not use the percentile instead of avg/min/max?
- Al: You're my best friend, because that's what I've been telling people to do around here for a long time. But today the test equipment doesn't support that.

- Sarah: You said a possible reason is test traffic is fixed size, but this is RFC2544.
- Al: Maybe we meant fixed duration of the test. I have to go back to my co-authors and ask why we said that.
- Sarah: Did you do any of the tests with an i-mix profile? (?)
- Al: It isn't a current capability of vsperf. We're still arguing how we're going to do that generically. It's not a generator limitation, it's more a vsperf limitation.

- Marius: Just to clarify, is that a 1Gbps linerate?
- Al: it was maximum of 25 million frames/sec.
- Marius: Thank you very much for the presentation, Al. It's nice to see every now and then some numbers to pick on them. Is there an article behind the presentation?
- Al: Everything is in the wiki: https://wiki.opnfv.org/display/vsperf/Traffic+Generator+Testing.

EOM

-=-