Benchmarking Methodology Working Group (BMWG)
IETF 95
Thursday, April 7, 2016
1000-1230 (Local Time) Morning Session I
Quebracho   OPS   bmwg

Remote Participation:
http://www.ietf.org/meeting/95/index.html
http://www.ietf.org/meeting/95/remote-participation.html

Minute Takers: Marius Georgescu, (A huge thanks to our note takers!)

-=

0. Agenda Bashing
    No agenda bashing.

1a. WG Status (Chairs)

- Need comments on the DC Benchmarking Draft.
- Please review the SDN Controller and IPv6 Drafts
- No new RFC's
- Al: Supplementary BMWG webpage is currently not reachable.
  Bill Cerveny (relayed by Joel Jaeggli): The tools wiki for BMWG page could probably be adapted for this purpose.
  Al: Good tentative solution. Thanks.

1b. Charter and Milestones (Chairs)
    No comments or questions.

2. Data Center Benchmarking Proposal
   Presenter: Jacob Rapp
   Presentation Link: https://www.ietf.org/proceedings/95/slides/slides-95-bmwg-1.pdf
   https://datatracker.ietf.org/doc/draft-ietf-bmwg-dcbench-terminology/
   https://datatracker.ietf.org/doc/draft-ietf-bmwg-dcbench-methodology/

- Trying to standardize on metrics for latency.
- Scott Bradner: The question of cut-through is certainly not a new one. It is in previous RFCs as well. We came to the conclusion that we should measure them differently. If you come up with a negative latency, that's certainly not ideal.
- Jacob: That's a good point. I think we should add more discussion around cut-through vs store-and-forward. What we are discovering now is that it's not one or the other. Some

devices act as cut-through at some packet sizes and then as store-and-forward for other packet sizes. That's why we called out in the draft that you should decide if it's doing cut-through or store and forward before starting the measurement. That's why we went back to first-in-last-out because it will give you a result be it cut-through or store-and-forward. If you do LIFO (last-in-first-out) you can miss it for cut-through.

- Al: That's when you get negative latency for cut-through.
- Scott: Which is great for advertising :). I looked at your updates and they look pretty good. I don't recall whether you specifically call out that when you're doing FILO (first-in-last-out) you're explicitly including the the line rate. I don't recall seeing it.
- Jacob: I'll double check and make sure we add it in if we didn't.
- Al: About including loss in the goodput measurement definition, I had a discussion with David Newman over a year ago. The simple solution is that it's unnecessary to mention loss. So maybe that's how we'll update the definition. I want to give that some thought and I'll put a proposal on the list as a participant.
- Al: It's more about the packets that were received twice than the ones that are lost.

- Al: Who has read the draft in the room?
- Scott has read the draft. Marius has read the draft as well.
- Al: We need more participation, especially since this is the last call. Otherwise, we will simply extend it.

3. IPv6 Transition Benchmarking
   Presenter: Marius Georgescu (remote)
   Presentation Link: https://www.ietf.org/proceedings/95/slides/slides-95-bmwg-2.pdf
   http://tools.ietf.org/html/draft-ietf-bmwg-ipv6-tran-tech-benchmarking-00

- Only 5% IPv6 world wide?

- Jacob: I have a question about the tags. Would you want to specify where to put the tags.
- Marius: I think that part should be left to the implementors. I don't think it's that important. What would you suggest.
- Jacob: You could put a tag at the very end vs a tag at the front and you could get completely different results.
- Marius: OK. I see your point. I guess we should specify that.
- Scott: I agree. Specifying is a good idea. Certainly when I was

testing I ran into different problems depending on the tag position. It really can make quite a difference.

- Marius: We will take care of this in the next version.
- Jacob: If you're measuring 500 frames, why not show the histogram instead of showing just the median.
- Al: On the topic of percentiles, there is no universal definition. There is a reference in RFC2330 (IPPM framework) that we could use. I think that's what we'll go with.
- Scott: This (Incremental performance Degradation) looks fine, except it doesn't say when to start the test. I would think you would want to have a latency after the last flow that you've added before you actually start the test.
- Marius: I think we neglected to mention that. We will amend that with the next version. Thank you.
- Al: I had a comment on this particular distribution, and it is good you showed this. This has some bimodal characteristics and gives the reason of why you would want to have the distribution shape occasionally and show the histogram. All the central measures get confused in this sort of world. We have to be careful going forward.
- Al: It looks as if this categories are enough for all the different technologies with nothing going amiss. Is that your general conclusion.
- Marius: I had a discussion with Fred (Baker) in IETF94 about any technology that might not fit in this categories, and we couldn't find any. If there are people who think this is not exhaustive, please challenge this categorization, so we can improve it.
- Al: I think your current path is fine. Let's proceed
- Tim Chown: In your tests do you cover the situation where on your client side you have IPv6 only and on your far side you have a destination that could be IPv4 or IPv6, and some of the traffic going through may be translated to IPv4 and some may be native IPv6. I think this would cover a typical scenario where your going to have more and more traffic over time.
- Marius: Thank you very much for the comment. I haven't given it much thought to be honest. I think you are suggesting mixed traffic, is that right?
- Tim: It may just be that there might be a negative impact, one way or the other, when having both translation and native. The chair is nodding, I think he's following what I'm saying.
- Al: The idea is to include characterization of forwarding of native IPv6 which would be potentially passing through a device that's also doing translation on a fraction of the traffic. This may deserve a paragraph.
- Marius: RFC5180 suggested something similar for dual stack devices and this is a good suggestion. I'll come up with at least some text about it. There was a question from Bill Cerveny about this as well. Maybe we can continue the

discussion on the mailing list.

- Marius: Should we care about NAT44?
- Al (as participant): There's a lot of work to specify these methodologies and if it it's not going to be a real competitive space, it may not be valuable to spend the time.
- Scott: How much effort would it be?
- Marius: It wouldn't be much effort. I don't like NAT44 much since it doesn't really push the IPv6 transition. I'm hoping it's going to go away soon enough.
- Scott: They're like cockroaches, they're not going to go away soon.
- Marius: I know, MAP-E/464XLAT use them. I'll give it some thought on how to include them specifically as well.
- Scott: Like our illustrious chair said, a lot of work goes into these methodologies, and if we can reuse some of them it's not a bad thing to do.
- Al: Marcelo (Bagnulo) was thinking to bring in a proposal in BMWG, for plain NAT but he hasn't done it, and maybe we'll his review and a little help with this.
- Marius: Thank you. Please try to get his feedback.
- Tim: A typical scenario for the transition would be IPv6 only on one side and dual stack on the other, IPv6 going native, IPv4 being translated. Another might be that you've got some dual-stack network on one side where NAT is being used along side native IPv6 with dual stack on the other side. I think it's worth considering it if it's not too much effort.
- Marius: Any opinions about including DN46 in the scope?
- Al: I don't see anyone rushing to the microphone to speak for it. Maybe that's a question to pose to the list as well. Let's also cross that on the DNSOP mailing list.
- Marius: We'll post that on the mailing lists and we'll decide after. Also, was wondering how far we are from the first WGLC?
- Al (as participant): I've seen this draft moving very quickly and with a lot of responsiveness to detail. I think that could be a good time-frame for a WGLC.
- Marius: Thank you very much.

4. VNF and Infrastructure Benchmarking Considerations
   Presenter: Al Morton
   Presentation Link:
   https://datatracker.ietf.org/doc/draft-ietf-bmwg-virtual-net/

   - Updates to draft
   - Ready for WGLC
   - Sarah Banks will lead the WGLC

- Scott: If you go up to Section 3, colossal may be overstated.
- Al: OK :) So, delete colossal. Thank you.
- Al: Any support for moving this draft forward to the ADs?
- 6-7 hands went up in the room. Three remote.
- Al: Not colossal support, but enough for BMWG.

5. Benchmarking SDN Controller
   Presenter: Bhuvaneswaran Vengainathan (remote)
   Presentation Link: https://www.ietf.org/proceedings/95/slides/slides-95-bmwg-7.pdf
   http://tools.ietf.org/html/draft-ietf-bmwg-sdn-controller-benchmark-term-01.txt
   http://tools.ietf.org/html/draft-ietf-bmwg-sdn-controller-benchmark-meth-01.txt

- Many updates.
- Goal: to compare various controller implementation across platforms.

- Al: Please read your hand if you've read either of the two draft.
- 5 people at the meeting have read this draft. 1 remote.
- Al: (Slide 4) Glad to hear you are adding some calibrations to understand what it is we're really measuring. Excellent work.
- Bhuvan: Thank you
- Bhuvan: considering Average vs. Median questions raised by Marius (Georgescu) at IETF 94.
- Al: That's discussion we had in many contexts, categorize and summarize these distributions carefully and completely.
- Marius: What I've shown is the distribution of the data. By analyzing the distribution of data produced following the methodology, we might get some hints on which summarizing function to use.
- Bhuvan: Consider for WGLC?
- Al: I have some comments. Some items in the matrix may need to be removed. Deactivation tests look more like functional tests. We might have at least one benchmark there still, and we need to talk about that. Also, we may want to expand throughput measurements. There may be some additional parameters to consider in this context. I've been talking to some folks running these types of tests, and when we'll have some conclusions on that, we'll bring them into this forum.

6. IPv6 Neighbor Discovery
   Presenter: Ron Bonica
   Presentation Links: https://www.ietf.org/proceedings/95/slides/

slides-95-bmwg-3.pdf

http://tools.ietf.org/html/draft-ietf-bmwg-sdn-controller-benchmark-term-01.txt

http://tools.ietf.org/html/draft-ietf-bmwg-sdn-controller-benchmark-meth-01.txt

- ND behavior could take down a router.
- How long does it take to go from Reachable to Stale states.

- Scott: Seems to me that maybe there's another question. When neighbor cache buffer fills up, what happens to router? Is buffering for neighbor cache the same as for data.   Common buffer pools?
- Ron: I'll describe the test very quickly. In the first test, when you're figuring out how long it takes to go stale, you send a packet, you see the NS/NA happen and it goes to reachable state. At that point, the test device doesn't send any more NAs. You keep sending packets and you measure the time between the first packet that got through and the last packet that got through and you get how long it takes to go stale. The next step is to discover when the neighbor cache is exhausted. You send a packet to the first possible address, the second the third... and you keep sending packets to all of these so that the neighbor cache never times-out. Then, you check to see how many destinations did you actually get traffic through to. That will tell you where the neighbor cache exhaustion is. It will also tell you something interesting, like "did the box crash at n+1?"
- Scott: Or is the buffer for the neighbor cache the same as the one for the packets? Assuming there is a common buffer pool, which I've seen in routers.
- Ron: That would be another interesting outcome. One of the things in the test is to record what kind of pathological behaviors you see at the end.
- Scott: Have you thought of maybe using SNMP queries to determine the time-out, rather than continuing to do traffic?
- Ron: We could do that, but isn't it against the rules to ever ask the box about itself in BMWG?
- Scott: Using a standard protocol in a standard way, doesn't make presumptions about what's in the box.
- Ron: In BMWG don't we assume that the box will always lie about its own behavior?
- Al: We believe it would if it could. That's why we treat these things as black-boxes, but I think if we're looking for a confirmation of externally observed behavior, I don't see anything wrong with that.
- Scott: Doing an SNMP query to see if a particular entry exists in the table should be OK.
- Joel: I think one of the observable phenomenons, not necessarily

related to benchmarking is that when it's full it's full, and essentially it has no performance characteristic. It holds a certain number and you count up to that number and it never goes above that, and that's a property of the system you can't measure by asking it, because it only holds a certain number and when the number is reached, that's it.

- Tim: I think the various mitigations that the vendors put in to prevent this, maybe vendor specific heuristics, so it might be that it isn't full, but you reserve a certain percentage of the cache for devices you've seen advertisements for internally.
- Ron: That brings us to the third test. You bring it to the point where you believe you've reached cache exhaustion, you let some of the cache entries go to the stale state. Once they've gone stale, you start send traffic to those and other destinations. If the ones you've seen before take preference than the one you've not seen before, it's behaving in accordance with Joel's spec.
- Joel: It wasn't so much as a specification, as a suggested mitigation, to the point the behavior you want to characterize in point 3 is that your expectation is this thing continue to work under duress, and the property of working is you can continue to add new neighbor entries. I've never cared about test 2 because I assume I can trivially achieve it on everything, which as it turned out, we could.
- Ron: You bring up an interesting point. Maybe we should rewrite all three tests. Maybe we can only have one test. You have n+1 senders from the test device. Any of them are trying to send legitimate traffic, one of them is doing a port scan. Do n-of-them get cache entries despite the fact that one is doing a port scan. If yes, the test succeeded, if not the test failed.
- Tim: It's good you've opened this can of worms, it's really interesting. So, part of it is that it might be the way the implementors have done it. It way be the way whether you're measuring on the internal interface of the router, or on the external. That would make a difference.
- Scott: You've used the magic word succeed. That's a no-no. BMWG treats the indication of performance not a value judgment. On the idea of having something which is doing a port scan, I was wondering how important the question of neighbor cache size is. Because the cache will always be bigger than the network actually needs. So, is it important to know what size the cache is? Maybe the key thing is: once the cache is exhausted by somebody doing a port scan, what happens?
- Ron: In fact that's what Joel's question is telling. The tests we've designed really aren't answering the question we set out to answer. The question is: do we survive in the face of a port scan? And the test we should do is do a port scan and try to get some real traffic through and see if we fall on

```
        our head.
   - Scott: There's the start up a flow and then do a port scan and
        see what happens to the flow. I'm not sure that how big the
        cache is makes any difference.
   - Ron: I agree. The spec basically need to be rewritten.
   - Joel: The original early tests are a verification of stated
        numbers on spec sheet. The properties of the cache can be
        described typically by x number of L2 entries and ARP entries
        and ND entries that they can store and those numbers tend to
        be documented pretty well.
   - Ron: A better way to design these would be:
        Test1: Start a port scan and let it run for half-an-hour,
                then start some legitimate traffic and see if it gets
                through
        Test2: Start the legitimate traffic first, then the port scan
        Test3: Start the legitimate traffic and the port scan, pause
                half of the legitimate traffic, then restart it and
                see how the 3 behave.
        Much simpler tests and it really answers the question. So,
        the bottom line is look for a new version of the draft.
   - Joel: We ask the right question. In the end game we care less
        about the fundamental properties that exist in each box.
   - Ron: We're looking less at the internals and more at the
        external properties of the box.
   - Joel: Turns out we're testing a black box :)
   - Al: By the way, I like your conclusions here, Ron.



PROPOSALS:

7. Benchmarking Virtual Switches in OPNFV
   Presenter: Maryam Tahhan (remote)
   Presentation Link: https://www.ietf.org/proceedings/95/slides/
slides-95-bmwg-4.pdf
   http://tools.ietf.org/html/draft-vsperf-bmwg-vswitch-opnfv-02.txt

   - Al: (Slide 4 - about vsperf) I want to emphasize to the group
        that this is a real test tool and many more tests have been
        defined. We're looking for additionla participation, if you'd
        like to join us.
   - SLIDE 5: Link to graph showing test results.

   - Ramki: Regarding more advanced features, such as Overylays. Do
        you capture those differences in your tests?
   - Al: This is the set of results which are quite stable OVS with
        DPDK. You can compare the results with the pure Vanilla OVS.
        With the Pure Vanilla OVS we're seeing less stability. This
        is one of the things that we have to investigate further.
   - Maryam: We just started support of overlays in release C. We
```

didn't have time to implement it. We hope to be able to publish some results for the next release cycle. We're trying to write the tests in a vSwitch agnostic way.

- Ramki: Another interesting thing would be if you're exercising anything with the flow tables, then there is perhaps a bigger variation in performance because Intel is cache based architecture.
- Al: There's an internal flow table and you can also have the controller flow table.
- Maryam: For those tests we are matching on the 5-tuple. That's not just matching in-port out-port, there is real matching and hashing going on.
- Ramki: Basically the OpenFlow rules.
- Al: That's the more realistic way to put it.
- Maryam: Full 150 page report is available. Please go through it and provide us with feedback.
- Al: If you had the chance to look at this and would care to support it, the call for adoption is open.
- 3 votes in favor of call for adoption in the room. Plus two messages on the list.

8. VNF Benchmarking Methodology
   Presenter: Raphael Rosa/Robert Szabo
   Presentation Link: https://www.ietf.org/proceedings/95/slides/
slides-95-bmwg-8.pdf
   http://www.ietf.org/id/draft-rosa-bmwg-vnfbench-00.txt

- Al: Are agents calibrated for tests?
- Raphael: Yes
- Jacob: You've mentioned Bare metal vs Hypervisor. Would you plan on running it side-by-side on all 3? Depending on the metric, would you allow tuning for different parameters, the hypervisor, for example.
- Raphael: It depends also on the specifications of the VNF, be it hypervisor or container.
- Pierre Louche: For VNFs it seems to be very concentrated on the data plane type of applications. Are you going to look at control plane type of VNFs, such as an MME diameter box, or something like that.
- Raphael: It would depend on the definition of the VNF itself. The VNF components, they could have data plane and control plane. In the perspective, we are treating all of these as a black-box. The are other metrics that could be defined, like the VNF instantiation time, which would consider the control plane.
- Pierre: How do you benchmark the performance of a VNF that's heavily control plane, data plane or both? I'm just wondering

if this is going to be part of this draft or not.

- Al: We're looking at the definitions of two things here. They're both called profiles, which is a little confusing. The benchmarking profile sounds like the test plan. The set of things that you're going to test when you're looking at this specific VNF. The subsequent result is the VNF profile, that specific set of tests for that specific VNF, the platform and other considerations. The overuse of the word profile is confusing.
- Ramki: Just to clarify the VNF profile, if you take two VNF examples, like CDN and firewall. The type of tests are going to be fundamentally different. So, are you really saying you're doing this as application benchmarking, or what is it exactly? What is really your goal?
- Raphael: The goal would be receiving this from the VNF developer as a black-box, not knowing if you're going for the control plane or data plane. We would like to see the resource allocated based on the performance metrics.
- Ramki: I'm seeing more of the latter, in the sense you're just trying to define this in a virtualized infrastructure. Is that it?
- Raphael: Yes.
- Al: To provide a helpful comment, when a VNF is delivered, it's going to be in the form of a package. It might be that this benchmarking profile can be collateral information, given as part of that package. It could be a script, it could be citing many RFCs asking to measure the specific metrics, suggesting configuration ranges. There could be a range of things, which could be delivered from the vendor, VNF specific. The foundation that we've created may not be complete, there's room for benchmarking all sorts of things.
- Ramki: If you start out as saying you are doing application benchmarking, then you have a fundamental problem that the case coverage is very very narrow. Practically speaking, if you take a real deployment like OpenStack there are tones of constraints.
- Robert: In a NFV environment, we would like to see this harnessing behavior, and that is the point of this benchmarking. So if you go for OpenStack, you have this small, medium and large footprint, that you can request. But, at the end of the day, in different OpenStack clouds you would like to see what is the performance delivered based on these small, medium and large allocations. The assumption is that even the OpenStack, NFVI both can change over time. So, you need a methodology to continuously on this with deterministic performance.
- Ramki: All I'm saying is you've got to consider how we're going to get to production. When we're going to go to production, you need to consider which part it is. Just doing this independently, you're going to run into trouble. You're

making very crafted assumptions on how these VNFs are going to be placed and you're loosing many real constraints. What you'll be producing would be a small subset of what happens in reality.
- Al: How many people have read the draft?
- Al: I see some interest. We thank you both for your efforts here and let's continue discussion on the list.

NEW/FIRST TIME:

9. Benchmarking Methodology for EVPN
   Presenter: Sudhin Jacob
   Presentation Link: https://www.ietf.org/proceedings/95/slides/slides-95-bmwg-5.pdf
   https://tools.ietf.org/html/draft-kishjac-bmwg-evpntest-00

   - Al: Has anyone in the room have read the draft or have any comments?
   - Al: I have read the draft and I have comments. First, just a simple one, You're missing one zero in your 10 million (very consistent thing). For many of the performance metrics, you have better names in slides than in the draft. You also need to define it a little better and separate form the methodology. We need a good definition of what we're going to measure. Let's try to separate these things. You've got them combined, and it makes for very long names.
   - Ramki: Perhaps you can leverage the work which was done previously. If there is some work already done you can reference it, or leverage it.

10. Benchmarking Performance Monitoring on a DUT
    Presenters: Sudhin Jacob
    Presentation Link: https://www.ietf.org/proceedings/95/slides/slides-95-bmwg-6.pdf
    https://tools.ietf.org/html/draft-jacpra-bmwg-pmtest-00

    - Al: Benchmarking loss, which of the Y.1731 methods of measuring loss were utilized? There's a method for which you insert a frame with frame-counts periodically. Also, there's a loss measurement, where you create synthetic packets or frames.
    - Sudhin: This is Y.1731 one-way-loss measurements, based on loss counts.
    - Praveen: Using the Single end loss measurements.
    - Al: Are there any performance diagnostics from Y.1731 that you're not covering?
    - Praveen: This only covers loss and delay.
    - Al: This might be valuable. If we decide to take-up work on this

topic, we will need a liaison with related ITU study group 15, the responsible for Y1731, and be sure they're OK with our work.

- Sudhin: The benchmarking is outside the scope of Y1731. We are trying to implement it, and reach a consensus between the different people.
- Al: This may be a very useful thing to do.
- Al: This is probably the last time we're going to see Scott face-to-face at one of our meetings. So, I want to add: thank you a million times and happy trails on all your future endeavors, Scott.

LAST. AOB