OZAN SAHIN: Good morning, good afternoon, and good evening. This is the RSS Metrics Work Party Teleconference held on the 30th of May 2019 at 15:00 UTC. On the call today, we have Duane Wessels, Kevin Right, Ryan Stephenson, Abdulkarim Oloyede, Fred Baker, Ihtisham Khalid, Jaap Akkerhuis, Jeff Osborn, Karl Reuss, Kazunori Fujiwara, Matt Larson, Paul Hoffman, Russ Mundy, Dave Lawrence, and Tom Miglin.

We have received apologies from [inaudible], and from staff, we have Steve Sheng and myself, Ozan Sahin.

I'd like to remind you all to please state your names before speaking for transcription purposes. Thank you, and over to you, Duane.

DUANE WESSELS: Alright, thanks very much. So, welcome everyone. What I think we're going to do today is go through the latest innovation of our document. You all should've seen this Google Document link in your email. A number of people have been inputting comments already. The primary, really the bulk of the work's been done since our last call, our last meeting, really has been to start to put, you know, our thoughts on the metrics into the format that's sort of the final format for an RSSAC document. So, there's kind of not really a lot of new content here. It's just sort of been reformatted and some of the metrics have been rewritten in a style that's more, sort of I guess, descriptive in the way that someone who is going to implement these metrics would need to read them.

So, I think let's sort of, I'll briefly sort of walk through the structure of the document and then we can take a look at some of the individual comments that have been made. If anyone has additional comments or thoughts, you know, as we go through it you can either add them or I guess raise your hand and we'll talk them through, as well. So, sound good Russ?

RUSS MUNDY: Yeah, that sounds good. One other thing I would like to just add is some of the discussions that were done as part of the Work Party meeting that was integrated into the RSSAC workshop, some of the values and missions we worked through at that meeting have been incorporated. So, as people look at the document they can see if it's, they think it's accurate with respect to what's been agreed to previously as you do your review. Thanks, Duane.

DUANE WESSELS: Yeah, that's a good point Russ. And I will confess that in the places where I have been doing writing in this document, I have been going from my memory of the meeting and my notes, which were not always complete. I have not listened through the transcripts of the previous meeting, so it's likely that I have misremembered something and please point that out if it seems like I misremembered something or forgotten something. That's quite likely actually.

So, the document is being shared here in Google Zoom. I don't know if, I think… Ozan, are you doing the sharing? I probably don't have the ability to scroll it, right?

OZAN SAHIN: I'm doing the sharing. I can scroll, but if you would like to, you can share it and you can do the scroll, too.

DUANE WESSELS: No, we'll just try it like this. So, as you see here on the screen, it's the standard RSSAC document format. We have an introduction, which has a little bit of text from the [inaudible] work, and then sort of how the report is organized. The Section 2 is background and scope, and one thing that's new here is we've added some terminology; we've added some definitions. These first few definitions are all essentially copied from the RSSAC terminology document, which is RSSAC026. So, these should be verbatim and just as a reminder to ourselves and to the reader of what these terms mean in this document.

Alright, let's scroll down a little bit more. We have some new terms here. These were suggested actually by Paul. We have measurement, and it says, "A measurement is an individual datapoint taken at a single point in time from a single location.", and it provides an example. And then metric is, "An aggregation of measurements over a range of time and locations using well-defined mathematical processes.", and such. So, that has been helpful as we sort of go through this and we try to be very clear about the difference between measurements and metrics.

I see some comments here from Kazunori about the way that measurement can be used both as a noun and a verb, I guess. I haven't really, I think that's a pretty new comment, so I haven't really considered if that's an issue in this document at this point, but it's

certainly something to be aware of. Paul responds that it's almost always data in this, almost always, I guess, the noun in this document, which would make sense to me. And then…

FRED BAKER: I think I would respond to that that a metric is something that one measures. A measurement is an instance of that being tested, and the verb form would be to measure; one goes and measures something. I'm not sure measurement would be directly a verb.

DUANE WESSELS: Yeah, okay. Thanks, Fred. We'll keep that in mind and you're, of course, welcome to type that in to keep us honest so we don't forget that.

So, there's also a definition down here for threshold, which is, "upper or lower limits on metrics beyond which action should be taken." At this point in our document, we haven't used thresholds a lot, you know. We haven't defined actual thresholds for any of the metrics at this point, but I think that is the goal. We will eventually, so that definition is here.

Alright, I'm going to try to just, again, run through the structure and then we'll come back and look at some of these individual comments. So, next in this background section we have a little bit about relationship with prior work. It references the other RSSAC documents and RC's and things. There's a section… Oh, and then there's, I guess that's the end of Section 2, so… Can you just scroll down to Section 3, Ozan?

This is just some general comments about the metrics and measurements. The order of these may not be ideal but these were sort of just written as they were thought of. There's a section on elapse time and time-outs, you know; what does a time-out mean. There's a part about how to calculate elapsed time for a DNS request over a TCP. There's a note that timestamps are always to be given or to be assumed to be in UTC time zone. We have this Section 3.3 which is really unchanged from the previous document. This is a little bit, it's just in outline form right now but this is our previous thoughts about what requirements there may be for a measurement platform.

Spoofing protections comments. We can talk about that point later. And then we have this Anycast note which again I think was in a previous document, just to say that the goal of this measurement, or this document, is not to try to measure all the Anycast sites of Open Root Server Operator.

RUSS MUNDY: Duane, one comment if I could add, just a little bit; 3 was I think trying to cover items that are, that would be, used in the later part of the document itself, and without having to repeat a lot of the things that might be included, so just sort of by the description of the metrics wouldn't have to be repeated in all of the things in Section 3 is kind of the idea.

DUANE WESSELS: Yeah, yeah. That's exactly right. Yes. Okay, so our Section 4 is sort of the definitions of metrics related to Individual Root Servers, or Root Server

Operators. We have four metrics that we've been talking about; availability, latency, correctness, and staleness. And then Section 5 is essentially, you know, the same sort of metrics but applied to the Roots of our system as a whole. And you'll notice that in this section, the text is really still in this previous style. It's probably just copied and pasted from the last document that we had been circulating, so this hasn't really been rewritten. I think it needs some attention.

Section 6 is titled… Oh, actually there's an unnumbered section for the BPQ stuff, and this is partially written. There's still a lot work to be done here. And then Section 6 is recommendations. This is where our specific thresholds might go. And then the rest of the document is the standard RSSAC document boilerplate stuff.

Alright. So, let's go back through and address some of the, or talk through some of these comments in particular. Let's, I guess, start in the terminology section again.

PAUL HOFFMAN: Duane, can I hop in just for a second? This is Paul Hoffman. Just on an organizational thing, Ken Renard, can you mute your phone? A bunch of us are getting some crosstalk.

DUANE WESSELS: Alright, thanks Paul. I wasn't sure who that was. I heard it as well, but I wasn't sure who it was.

So, we have some comments here. We already sort of touched on these a little bit. Metric versus measurement. People have made comments.

Would anyone like to say something that, you know, hasn't already been said about these definitions, or that is not in the comments of the document, or to reiterate something? Please feel free. Raise your hand, I guess. I'll try to find the place where I see the hands.

Russ, is your hand up?

RUSS MUNDY: Wasn't intended to be. Thanks.

DUANE WESSELS: Alright. Steve?

STEVE SHENG: Thanks, Duane. In the following section, we talk about RSO related metrics, and there was a discussion that the word RSO is not really the right word because we're not really measuring a organization, the availability. But really, the server, you know, the kind of entity they operate. The availability of that. So, you know, what is the right term for that, and I think once the Working Group decides on that, there probably needs to go into the terminology. Thanks.

DUANE WESSELS: Yeah, thanks Steve. I agree that this needs some attention. Right now, the document is sort of written, you know, it talks about like, say, RSO availability, and my point of making my comment was we need to be careful to be clear that we're not, you know, we're not measuring the

Operator. We're not measuring the O in RSO. We're measuring the service and it seems that we don't have a good, sort of, three-letter acronym like we have RSO, RSS, but we don't have an acronym for just the Root Service itself. Maybe we want to use RS, I don't know. I think Paul had a suggestion to just keep using RSO but keep in mind that it applies to the service but not necessarily the Operator, is that sort of what you were saying, Paul?

PAUL HOFFMAN: Yeah, so I don't mind us coming up with a new acronym either. Because at that point, quite frankly, then like I don't know that we would actually need RSO at that point. I was thinking of RSO as the equivalent of letter, and you make a good point that unfortunately Verisign has multiple letters, and not unfortunately, but I mean that that causes some confusion. So, either we can possibly redefine RSO as to mean, you know, the letter, or we can come up, and I know we can't really just say "letters", but you know, we could come up with like you say another acronym that means, you know, the service that you would find.

Steve had said in his comments in the document, he had suggested identity, and I would tend to stay away from that for two reasons. One is when you say identity, everyone in the internet thinks they know what you mean and it's certainly not what we're talking about, but the other is that to me an identity is a single point, and yet for all the services we have two points; an IPv4 and IPv6. And in the future, a letter might have multiple IPv4 or IPv6 addresses, so I don't think identity would be a good word here.

FRED BAKER: I would really want to stay away from the word “letter”, because, you know, if we were to add future RSO’s I don’t know that they would automatically be designated with a letter. That’s a convenience in the way the naming is done, but there’s nothing intrinsic about the letters.

RYAN STEPHENSON: Hi, sorry to cut you off Fred. While listening to this conversation -- this is Ryan Stevenson by the way -- I don’t know if maybe Root Server Instance would be better, and the acronym would RSI obviously. Just for something to think about.

DUANE WESSELS: Russ, you have your hand up?

RUSS MUNDY: Yes. I was… I pulled up RSSAC026 and looked at what we have in there as the definition of Root Server, because that is a defined term, and in there it is defined as an Anycast, or an instance of the Anycast configuration for particular server. So, I think that we have, unless we’re going to also revise 026 or suggest revisions to it, we need to be conscious about the terminology that we do land on.

One possibility that came to my mind is that we could, in the terminology section, come up with an acronym that basically represented the Root Server machinery operated by any individual or single organization. So that would I think cover the contingency Fred

was talking about, and what Paul had mentioned. You know, more than one letter being operated by a given organization. So, I think it would be, maybe, something good to be handled on the list so people should write down what they think is a good idea, or put it in the document itself, because just discussing it, it'll be hard to come to a conclusion without having words to look at. Thanks.

DUANE WESSELS: Okay, Russ. Thanks. Yeah, that's a good idea to sort of follow this up on the mailing list. I'll make a note of that. Alright, Ozan, can you scroll down a little bit to the later sections. More, more, more, more. Like, Section 3. Was there a, let's see… Oh, Andrew had a comment in Section 2. Okay. That seems like a minor comment. I see some comments from, I guess… I'm not sure where these are applied to, these comments from Robert.

PAUL HOFFMAN: Duane, those were him removing the word DNS for DNS response. Basically, his question was aren't these all going to be DNS?

DUANE WESSELS: Okay. Alright, that seems reasonable to me. Let's scroll down further. Alright, so, I don't think I heard Wes on the call, right? But Wes had this comment that… So, I added this paragraph about spoofing protections when I was writing some of the text about how to do correctness measurements, and, you know, I felt it was important to state that any software that's implementing these metrics needs to take all of the

precautions that, you know, all good DNS software takes to avoid being spoofed so you have to do, you know, choose random source ports and random Core ID's and things like that. Wes's comment was that this seems, I don't know, out of scope. Paul, I think you're sort of supporting that it does belong here. Does anyone else have opinions on this unwritten section, or partially written section?

RUSS MUNDY: This is Russ. It seems that perhaps spoofing protections isn't the best title, but I think the idea is a sound one that, you know, in terms of talking about the Work Party's view of what the test structure needs to be, expressing the fact that it should follow good operational practices seems like a reasonable thing to include. But the title I think is probably the hot button here, I think.

DUANE WESSELS: Yeah, is there an RC that we can refer to that covers, you know, this in other things?

PAUL HOFFMAN: There are many, and so I would say let's not try to boil the ocean on it. But I like Russ's idea of saying, "The reason why we're doing it is x and we suggest doing at least this", and then we can add them later.

DUANE WESSELS: What do you mean by add them later, Paul?

PAUL HOFFMAN: As we create the testbed, someone says, “Ah, but you didn’t think about spoofing this way!”, and then we can say, “Great. We’ll add that in as well.”

DUANE WESSELS: So, you mean as we get experience with implementing these, we would go back and update the document?

PAUL HOFFMAN: Yeah, or even just as other people come in, they might say, “Please add this to the list.” That’s why I had actually added to Wes’s thing of, at some point we can pull all of this out of the base metrics document and put it into an operational document for the test harness.

DUANE WESSELS: Okay, I see. Okay, thanks. Alright, let’s scroll down some more. So, this is the section where, or these comments are about what we’ve already talked about. The thing that Steve raised where we used the phrase RSO here, and we’ve had some good discussion on that. Paul had a… So, there’s a table here in this section that says, you know, if you’re going to, when you’re doing these sorts of measurements, these are how you would do these queries. The query name is Dot. The query type is SOA. Paul had a comment that this could possibly pollute measurements that other people are doing and maybe it would make sense to use a… Are you proposing like a metric specific RR type or something like that, Paul?

PAUL HOFFMAN: Yes, yeah. RR types are basically free. We just need to throw together a short document on it and we'll get a new one.

DUANE WESSELS: Okay. So, I guess the advantage to that is that these measurements could be done with an RR type that doesn't occur naturally, right? It doesn't occur in normal traffic, but it also means, you know, that nothing prevents others from abusing this same RR type and sort of, I guess, I mean it could still pollute DITL, I guess, right?

PAUL HOFFMAN: Correct, it could. This is not to prevent that. This is to prevent us from polluting metrics that other people are doing now. That there are Individual Root Server Operators who look through their data for this and that. And we don't know. They might already be looking for Dot SOA queries, so why all of a sudden make their data worse? We can create something that is completely new.

DUANE WESSELS: I see. Personally, I think I wrote here, I don't… Specific to DITL, that doesn't seem like a big concern to me. I don't know of anyone that goes through DITL data looking for SOA queries. Maybe there are some other

examples that I'm not aware of, though. Anyone else have thoughts on this topic?

Okay. Let's scroll down. So, the next section is about latency, and Paul, your comment here was, I think that, you know, you could reuse this measurement, you could reuse the queries from this one and the previous one to accomplish the same thing, which I agree with. That seems like a logical thing to do. Yeah. Any comments on that?

Alright, let's scroll on down. So, at the end of these metrics we've tried to give some examples of how they might look or be presented and Paul's comment here is that for these latency metrics, which are essentially medians, that there should be some indication of statistical precision for the derived metric. Which I agree with.

One of the standard ways that you might do statistical precision is something like variance or standard deviation. Based on, you know, my understanding of statistics, standard deviations are generally useful for distributions that are normally distributed, which latency usually isn't. It can have a, it has a, sort of a decaying heavy-tailed distribution, and so something like interquartile range might make more sense than a standard deviation, but really, I think it's up to us to define, you know, what sort of statistical precision do we want to see represented on these measurements.

Anyone else have opinions on that? Do we have any statisticians on the call?

WES HARDAKER: I know enough to be dangerous.

DUANE WESSELS: Just like me, then.

WES HARDAKER: I mean, I think Suzanne's point at the workshop was valid, that if we know, you know, what percentage of, you know, probability that we want to make sure that we are able to detect, I can do the math to reverse it. That I can tell you. Coming up with what's the right probability that we want to be able to catch something is the hard part because that's just opinion based.

DUANE WESSELS: Yeah. Ozan, can you scroll up just a little bit? Just a few lines. So, along the lines of what Suzanne was suggesting, Wes, these, you can see that the last sentence above that example says here that the minimum number of measurements for aggregation shall be 30, and that's just sort of a made-up number at this point. There's no math behind that, but that is designed to address sort of I think what Suzanne was getting at; that, you know, the metric is a little bit more useful, trustworthy, valid, the more measurement points you have and that there should be some lower limit on how many measurements you can aggregate in order to get a metric. And so, for this one, it says 30. You know, we can change that based on math if we want to. I think on the, there was also this text occurred in the previous metric which was about availability,

and I don't remember what the number was. Was it also 30? Oh, it's 60 there for whatever reason. Paul?

WES HARDAKER: Well, they probably should be consistent because otherwise that math gets worse if you're not consistent.

PAUL HOFFMAN: Yeah, so I agree with Wes, and I would also say since we're talking about measuring every minute, I think a, and we know that the internet routing changes over time and stuff happens, I think a more reasonable thing to say is that the minimum metric that we would use for, again, this would be for reporting plus possibly inclusion in thresholds, should be at least an hour's worth. And, it would not surprise me if we actually upped that to four hours' worth because sometimes things last for an hour, you know. Hopefully, they'd… I'm sorry. Sometimes negative things happen for an hour and they're much less likely to happen for four hours just because of the number of beepers that have gone off. So, upping the minimum number to the equivalent of four hours, or an hour, I think would solve most of the concerns.

DUANE WESSELS: Okay. Thanks. Yeah, in some of the later ones the text does read in a way that it has minimums on both. For the availability one, it does not have a minimum on both, although it certainly could. But my rationale there was that the way availability is defined, it's really the number of queries sent that matters, not the number of responses received,

because that's the whole goal of measuring availability. But I agree that for something like latency you could say, in order, a minimum for aggregation is x time and n data points. So, we'll work on making those the same. Russ?

RUSS MUNDY: Yeah, thanks Duane. I think because of the nature and the difference of the metrics themselves, having the, trying to achieve the same might actually require us to use different numbers, different timings. And I think it's great for us to be discussing these now, and what the implications are, but I don't know if in the end in terms of, like you say, by how they are defined we might need in fact to have variation based upon what the metric itself is, so I think the values that are used is something we need to continue to work on and discuss throughout the development of the document.

DUANE WESSELS: Yeah, okay. Thanks. Paul, is your hand up to speak again? No, okay. Okay, let's continue on through this a little bit farther down. So, we talked about latency and availability. The next one is correctness. So, I apologize. This is something that was written sort of… This part of the document changed after the link went out, the list. So, this may be new to some of you if you weren't paying close attention.

The idea in this metric is to use dataset validation as a way to sort of determine correctness over some period of time. So, there's a few things different from some of the other previously defined metrics. One is, here, that, it says measurements shall be made at five-minute

intervals, whereas previously we were doing one-minute intervals. That is based on my recollection of our meeting in April with RSSAC, but again I could be wrong about that. If there is agreement that they should be all at one-minute intervals, that's fine with me.

There's this table here that sort of defines, you know, the name and type and the, that the data [inaudible] bid must be set and so on. You see that in the table there is different names with different probabilities. This is, the idea here is, to sort of get different coverage of the contents of the zone. So, with the 20 percent probability you ask for the SOA record. With a 20 percent probability you ask for the root DNS key. With 50 percent probability you would ask for the DS record of a TLD, chosen at random. And with 10 percent probability you would ask for basically a name that you know will be NXDOMAIN response, so that is to test the correctness of NXDOMAIN responses.

And then I also defined here some buffer sizes for queries that are sent over UDP. So, I see Andrew has made some comments here. I haven't had a chance to read these before. Yeah, so I think Andrew's point is that there's an updated list of Special-Use Domains that might be appropriate instead of RFC 2606 which list just these four reserved TLD's.

Paul suggests just use the, just be very specific and use the ones for this test. That's fine with me, as well. I was, you know, trying to refer to RFC's as much as possible, but we don't have to.

Can you scroll down a little bit more so I can see Paul's comment to Andrew? Okay. Alright, so yeah, Paul, your comment is to maybe have a

longer list of, sort of, specially chosen names that are not necessarily reserved but that we think will be NXDOMAIN for some, you know, long future amount of time that we would check periodically to make sure of that. And then, I'm fine with that as well if that's the way we want to go.

So, there's an example results table here which is similar to the others. Since this metric is basically just simple division, you know, it's number of queries sent divided by number of… Well actually that's not true. It's a little bit more complicated because what we agreed on, based on my recollection of the April workshop, is that the way that you aggregate is that if any one measurement within the aggregation interval results in an incorrect response, then you say that the metric was incorrect for that whole interval. And that is one reason why here it says that for this metric the aggregation interval shall be exactly one day. So, it doesn't make sense to aggregate at a shorter time period or at a longer time period. And it also specifies a minimum number of aggregated measurements which is, here it's set to half of the expected measurements that you would get over that one-day period.

And then in the example results table, so there's no, I realized, there's no text for this, but in the example table it shows that this last column, Response Count, shows how many responses were used to derive this measurement. So that is the indication of precision for this metric.

So, I've been talking a lot. Any questions or discussions about this particular metric right now? I'm not sure I'm able to see everyone's hands if their raised. Let's see. Alright. I guess not. Not seeing any hands.

So, one thing still, I think, to be addressed in this metric is to document ways that these measurements can be falsified, as we talked about by spoofing. I think that a lot of us may be aware of these sorts of things already but other readers of this document may not be aware and if in the future we come to find that, you know, there are a lot of cases where it looks like some third-party is interfering with these measurements, then, you know, we can say, "Yeah, that's something that we were aware of at the time. Maybe we need to take these certain steps to address that." Or maybe we say that, you know, this metric is always susceptible to interference and is not something that can be reliably measured. Comments?

RUSS MUNDY: Hi Duane, this is Russ. I think one of the concerns about all of the metrics that get defined is exactly that area, and perhaps the place to treat it would be up in the Section 3 area where it is pointed out that, some words to the effect that these aren't, there's not any, other than current good operational practices, there's nothing to prevent attacks or manipulation rather than trying to deal with them on a one-on-one basis. That seems like it might be a better way to try to handle the how easy or hard it is to muck with the results.

DUANE WESSELS: Okay, yeah. I think that's fair. Paul?

PAUL HOFFMAN: So, I would like to strongly disagree with what Russ just said. At the Prague meeting, I thought that the sense of the table was that a metric that says that a Root Server Operator sending out incorrect information was much, much more serious than even an availability issue. And therefore, the mucking, someone mucking with availability would have a lower effect than someone mucking with the correctness. I think maybe we can talk about all of those together, but collapsing them, saying, "Oh, someone might muck with these", and not differentiating I think goes against the sense of what I heard in Prague.

RUSS MUNDY: If I could respond to that. I think that either I didn't say it very well or it didn't, you didn't hear what I was trying to infer, Paul. I was trying to, I was thinking of it from the perspective of a third-party, if you will, an independent entity trying to affect or impact the results that would be going on between the testing harness, whatever it is, and the Root Server. Not the Root Server themselves doing something bad or wrong. That was what I was trying to talk about for putting in to a single statement.

PAUL HOFFMAN: So, I agree that that is true, but again, the effect of a third-party diddling with your availability is significantly less, at least from what I heard people saying in the meeting in Prague, than the third-party, the effect of a third-party changing your correctness. Correctness, as I understood it, everyone was like, "Oh, yeah, my God, this is the big one." And so, maybe we can stop having this discussion here and wait until we have

some words to put up there, but I just don't want it to seem like all four of these measurements, that somebody twiddling with them, a third-party twiddling with them, would have the same affect. I think it would be much more devastating for, you know, Y server to be seen as giving incorrect data, you know, once a day than it would be for them being offline once a day.

DUANE WESSELS: Well, Paul, you're right because of the way this metric is defined. It only takes one, you know, incorrect measurement per day here to sort of turn the whole day incorrect, right, or the whole interval incorrect. That's just the way this metric is defined. Whereas for some of the other ones, you know, twiddling with one individual measurement probably doesn't have a big impact on the value.

PAUL HOFFMAN: And I think the other thing that people were saying in Prague was that the, you know, since these metrics are meant to be used by people outside looking at them, that some it is much more easy to misunderstand this correctness one being anything other than 100 percent correct than it is the others. It would, I think a normal person would assume that Root Operator Y lied on purpose if they saw an incorrect thing, whereas we of course know, "Oh, that might've been, you know, pushed in" or whatever. So, again, I just want to say that what I heard was that people said this is a very different kind of metric, so I don't think we should necessarily even gang in the discussion of the

results of how things could change, but let's look at words and then maybe work from there.

DUANE WESSELS: So, Paul, if I can ask if, you know, if we were to have some words in this document saying, you know, that here be dragons, be careful with spoofing and stuff, would you want that to be only in this section or would you be okay if it was in the general section up above?

PAUL HOFFMAN: I'm not sure. That's a very good question. When you put it in that section above, I didn't think, "Oh, it's really important to put below." So, let me think about that.

DUANE WESSELS: Okay, sure. Any other comments about this metric? Okay, let's scroll down then to the next one. So, this is the staleness metric and this one is a little bit complicated because… So, here the idea is to have a metric that describes the staleness of data served by a Root Server, and for the staleness we're basically looking at the SOA serial number. Since the root zone is published and pushed out at, sort of, unpredictable times, these measurements are complicated by that fact because you may be doing a staleness measurement at the same time as the root zone is being updated and that could lead to confusing results. The previous… Oh, and so another, sort of, factoid about the way this metric works is the way that staleness is determined is essentially you query all of the

Root Servers and sort of see the extent to which they have the same SOA serial number, and so you compare them to each other.

And the assumption, or I guess what seems logical to me, is that you would do that comparison at some central location rather than on the individual highly distributed probes. So, the previous definition for this metric was something like you query all the Root Servers from all the probes; you get their serial numbers. If they're all the same then good, none of them are stale. If there's a difference, you take the maximum value and then you remember that maximum serial number value, you wait ten minutes, and then you query them all again and compare them to the previously determined maximum serial number value.

So, in order for that to actually work, you would have to have sort of a lot of communication and coordination between some central point in the probes. You know, all of those measurements would have to go back to the central location within this, sort of, this ten-minute time period, and then the central location would have to direct them all to, you know, do the re-querying. And so, to me, that felt a little bit, I don't know, a little bit difficult and a lot of work, so the way this is currently written is slightly different.

It says that basically every hour, these staleness measurements start, and the probes do a pair of measurements. The pair of measurements, the first one is named Alpha, the second one is named Beta, and the two measurements are separated by this ten minutes of time. And then, so each probe is always doing two measurements separated at the time and those results are going back to the central processing location, sort of, as they can. And then, the system that's doing all that central

processing has both sets of measurements to use if it needs to. So that's why this is a little bit complicated here.

One thing that we need to discuss and sort of talk about is whether these measurements should be made at, you know, very predictable times or not. So, for example, should the measurement always start at the top of the hour, and then, you know, exactly ten minutes later, or should there be some variation there? I think there's tradeoffs to both. You know, there's a complexity in doing that, but also probably some advantages to, you know, distributing things up a little bit.

Can you scroll down for me a little bit, Ozan? So, the way, the text that's here is, I think, well written up to the point where the measurements are, you know, it describes how to perform the measurements and how to do the thing where you check the first set and then decide if you need to check the next set. But there's nothing here that describes how the measurements are aggregated in to the metric. I think there are… That's plenty far, Ozan, can you scroll back a little bit? You can put the example table at the bottom of the screen.

I think that this metric has some, has the potential that the interval of measurements, so right now it says one hour, if we change that, that sort of has the potential to impact the results, so I think we need to think carefully about how that measurement interval will affect the actual results, potentially affect the results in cases where there is staleness occurring. I'd like to ask for volunteers to take a close look at this and maybe help write the remainder of this section. Is anyone feel up to that task? Paul, I see your hand is up, or it was until I asked for volunteers.

PAUL HOFFMAN: So, my hand was up because I think, and again, always terrible to do on a phone call especially with only five minutes left, I think I have a radically different design that would be a lot simpler. If the central, you described, you know, a central test place. If that knew when the root zone was being changed, if it somehow got a notification either from Verisign, or just in some other way, it could then signal to all the test things saying, "The root zone has been updated. Please check in ten minutes." Or maybe it just sends the message in ten minutes. Then that's a one-shot thing. How many, since what we care about is, are the instances updated within, and we picked ten minutes, just only test after you know that it is changed at that interval, and you get a yes no.

DUANE WESSELS: Yeah, I could see that working. I can think of ways where it could fail as well but I mean it's probably as reasonable as any other.

PAUL HOFFMAN: So, I raised my hand for the I'd be willing to work on this, and so I'd be happy to write that up but also, I'd be willing to work with anyone who wanted to work with your current suggestion as well.

DUANE WESSELS: Okay, thanks Paul. Alright, so as Paul said we're almost out of time here. Ozan, let's scroll down to the… Okay, so this part where it's the sub-metrics. This is old text that essentially needs to be deleted because it got rewritten above. Let's scroll down some more to the RSS metric. So,

like I said before, this also need to be rewritten. I would like to ask for volunteers to take on some of these sections and sort of take a stab at writing them in the style of, you know, for someone who has to implement these, right? This document will be essentially implementation advise on how to do these measurements. Paul, I don't know if your hand is still up or if you're volunteering for more work? No? Can I get any other volunteers to do some writing in this document please?

WES HARDAKER: I can probably help in a couple of weeks, but I'm swamped for the next few weeks.

DUANE WESSELS: Alright. Speaking of the next couple of weeks, in not too many weeks we have our ICANN meeting in Marrakech where we had hoped to talk more about this work with the RSSAC. Fred, do you have any information about the schedule for that meeting?

FRED BAKER: Well, yeah. We've been discussing that, Brad and myself and the ICANN staff, and planning that meeting. Right now, we are planning to have two ninety-minute slots, so basically all morning. I think that's on the 27th. Ozan, correct me if I've got that wrong, during the meeting. And, you know, my deepest apologies to people in North America, that'll be the middle of the night for you. I'm sorry. It's daytime in Morocco. So, those two sessions will be, kind of like in the workshop a few weeks ago,

will be a Metrics Working Party discussion that, you know, RSSAC is all members of this so I don't know if it makes sense to say that the RSSAC is present for it, but something like that. We have scheduled two ninety-minute sessions for that.

One of the things we really need to, that Brad and I really want to see come out of that, is a discussion of thresholds. Now, Steve Sheng has an intern that is doing a study this month on using, basically, prototype metrics using ripe Atlas data and they'll make a report on what they've observed. That hopefully will give us some feedback on whether the metrics makes sense in implementation. And then what values we observed. And from that we would have some basis for a discussion of thresholds.

If we can't come up with thresholds that we agree to during the Marrakesh meeting, Brad and I would really like to see a workshop, probably in the August or the October timeframe, which would be primarily a metrics workshop with the target of coming up with thresholds. And of course, when we write thresholds into the document, that's an initial stab at things. People will come back and say, "Oh, I think that should be different.", and will give their reasons and we can discuss that. But we want to get at least an initial stab at thresholds during the Marrakesh meeting or a workshop following. This will be discussed during the RSSAC call next week, on Tuesday, which is an open meeting. As far as I know you're all welcome to be there. We'll have a brief report and a discussion of the Marrakesh and the possible workshop meetings.

DUANE WESSELS: Alright. Thanks, Fred. So, I'm going to sort of wrap this up since we're out of time, but I do want to strongly encourage people to do some writing on this document. I mean, that's what it means to be on a Work Party. You have to do some work; you can't just lurk, and we need people to do some writing on this document before the Marrakesh meeting so if you're willing to do that please contact me or Steve or one of the staff via email and we'll get you set up. And also, we'll go to the email list to schedule our next Metrics Work Party call. Anything else before we disconnect, Steve?

STEVE SHENG: No, sounds good.

DUANE WESSELS: Okay, thank you everyone.

RUSS MUNDY: Thanks everybody.

[END OF TRANSCRIPTION]