Internet Engineering Task Force	R. Geib, Ed.	
Internet-Draft	Deutsche Telekom	
Intended status: Informational	A. Morton	
Expires: August 15, 2010	AT&T Labs	
	R. Fardid	
	Covad Communications	
	February 11, 2010	

TOC

IPPM standard advancement testing
draft-geib-ippm-metrictest-02

Abstract

This document specifies tests to determine if multiple independent instantiations of a performance metric RFC have implemented the specifications in the same way. This is the performance metric equivalent of interoperability, required to advance RFCs along the standards track. Results from different implementations of metric RFCs will be collected under the same underlying network conditions and compared using state of the art statistical methods. The goal is an evaluation of the metric RFC itself, whether its definitions are clear and unambiguous to implementors and therefore a candidate for advancement on the IETF standards track.

Status of this Memo

Table of Contents

1. Introduction
    1.1. Requirements Language
2. Basic idea
3. Verification of conformance to a metric specification
    3.1. Tests of an individual implementation against a metric specification
    3.2. Test setup resulting in identical live network testing conditions
    3.3. Tests of two or more different implementations against a metric specification
    3.4. Clock synchronisation
    3.5. Recommended Metric Verification Measurement Process
    3.6. Miscellaneous
4. Acknowledgements
5. Contributors
6. IANA Considerations
7. Security Considerations
8. References
    8.1. Normative References
    8.2. Informative References
Appendix A. An example on a One-way Delay metric validation
    A.1. Compliance to Metric specification requirements
    A.2. Examples related to statistical tests for One-way Delay
Appendix B. Glossary
§ Authors' Addresses

TOC

1. Introduction

The Internet Standards Process RFC2026 (Bradner, S., "The Internet Standards Process -- Revision 3," October 1996.) [RFC2026] requires that for a IETF specification to advance beyond the Proposed Standard level, at least two genetically unrelated implementations must be shown to interoperate correctly with all features and options. There are two distinct reasons for this requirement.

In the case of a protocol specification, the notion of "interoperability" is reasonably intuitive - the implementations must successfully "talk to each other", while exercising all features and options. To achieve interoperability, two implementors need to interpret the protocol specifications in equivalent ways. In the case of IP Performance Metrics (IPPM), this definition of interoperability is only useful for test and control protocols like OWAMP [RFC4656] (Shalunov, S., Teitelbaum, B., Karp, A., Boote, J., and M. Zekauskas, "A One-way Active Measurement Protocol (OWAMP)," September 2006.) and TWAMP [RFC5357] (Hedayat, K., Krzanowski, R., Morton, A., Yum, K., and J. Babiarz, "A Two-Way Active Measurement Protocol (TWAMP)," October 2008.).

A metric specification RFC describes one or more metric definitions, methods of measurement and a way to report the results of measurement. One example would be a way to test and report the One-way Delay that data packets incur while being sent from one network location to another, One-way Delay Metric.

In the case of metric specifications, exactly what constitutes "interoperation" is less obvious. The IETF has not yet agreed on how to judge metric specification "interoperability" in the context of the IETF Standards Process. This draft provides methods which should be suitable to evaluate metric specifications for standards track advancement. The methods proposed here MAY be generally applicable to metric specification RFCs beyond those developed under the IPPM Framework [RFC2330] (Paxson, V., Almes, G., Mahdavi, J., and M. Mathis, "Framework for IP Performance Metrics," May 1998.).

Since many implementations of IP metrics are embedded in measurement systems that do not interact with one another (they were built before OWAMP and TWAMP), the interoperability evaluation called for in the IETF standards process cannot be determined by observing that independent implementations interact properly for various protocol exchanges. Instead, verifying that different implementations give statistically equivalent results under controlled measurement conditions takes the place of interoperability observations. Even when evaluating OWAMP and TWAMP RFCs for standards track advancement, the methods described here are useful to evaluate the measurement results because they have little value otherwise.

The standards advancement process aims at producing confidence that the metric definitions and supporting material are clearly worded and unambiguous, or reveals ways in which the metric definitions can be revised to achieve clarity. The process also permits identification of

options that were not implemented, so that they can be removed from the advancing specification. Thus, the product of this process is information about the metric specification RFC itself: determination of the specifications or definitions that are clear and unambiguous and those that are not (as opposed to an evaluation of the implementations which assist in the process).

This document defines a process to verify that implementations (or practically, measurement systems) have interpreted the metric specifications in equivalent ways, and produce equivalent results. Testing for statistical equivalence requires ensuring identical test setups (or awareness of differences) to the best possible extent. Thus, producing identical test conditions is a core requirement. Another important aspect of this process is to test individual implementations against specific requirements in the metric specifications using customized tests for each requirement. These tests can distinguish equivalent interpretations of each specific requirement.

Conclusions on equivalence are reached by two measures.

First, implementations are compared against individual metric specifications to make sure that differences in implementation are minimised or at least known.

Second, a test setup is proposed ensuring identical networking conditions so that unknowns are minimized and comparisons are simplified. The resulting separate data sets may be seen as samples taken from the same underlying distribution. Using state of the art statistical methods, the equivalence of the results is verified. To illustrate application of the process and methods defined here, evaluation of the One-way Delay Metric (Almes, G., Kalidindi, S., and M. Zekauskas, “A One-way Delay Metric for IPPM,” September 1999.) [RFC2679] is provided in an Appendix. While test setups will vary with the metrics to be validated, the general methodology of determining equivalent results will not. Documents defining test setups to evaluate other metrics should be developed once the process proposed here has been agreed and approved.

Changes from -01 to -02 version

*Major editorial review, rewording and clarifications on all contents.

*Additional text on parrallel testing using VLANs and GRE or Pseudowire tunnels.

*Additional examples and a glossary.

Changes from -00 to -01 version

*Addition of a comparison of individual metric implementations against the metric specification (trying to pick up problems and solutions for metric advancement (Morton, A., “Problems and

Possible Solutions for Advancing Metrics on the Standards Track," July 2009.) [morton-advance-metrics]).

*More emphasis on the requirement to carefully design and document the measurement setup of the metric comparison.

*Proposal of testing conditions under identical WAN network conditions using IP in IP tunneling or Pseudo Wires and parallel measurement streams.

*Proposing the requirement to document the smallest resolution at which an ADK test was passed by 95%. As no minimum resolution is specified, IPPM metric compliance is not linked to a particular performance of an implementation.

*Reference to RFC 2330 and RFC 2679 for the 95% confidence interval as preferred criterion to decide on statistical equivalence

*Reducing the proposed statistical test to ADK with 95% confidence.

1.1. Requirements Language

TOC

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 (Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels," March 1997.) [RFC2119].

---

2. Basic idea

TOC

The implementation of a standard compliant metric is expected to meet the requirements of the related metric specification. So before comparing two metric implementations, each metric implementation is individually compared against the metric specification.
Most metric specifications leave freedom to implementors on non-fundamental aspects of an individual metric (or options). Comparing different measurement results using a statistical test with the assumption of identical test path and testing conditions requires knowledge of all differences in the overall test setup. Metric specification options chosen by implementors have to be documented. It is REQUIRED to use identical implementation options wherever possible

for any test proposed here. Calibrations proposed by metric standards should be performed to further identify (and possibly reduce) potential sources of errors in the test setup.

The Framework for IP Performance Metrics (Paxson, V., Almes, G., Mahdavi, J., and M. Mathis, “Framework for IP Performance Metrics,” May 1998.) [RFC2330] expects that a "methodology for a metric should have the property that it is repeatable: if the methodology is used multiple times under identical conditions, it should result in consistent measurements." This means an implementation is expected to repeatedly measure a metric with consistent results (repeatability with the same result). Small deviations in the test setup are expected to lead to small deviations in results only. To characterise statistical equivalence in the case of small deviations, RFC 2330 and (Almes, G., Kalidindi, S., and M. Zekauskas, “A One-way Delay Metric for IPPM,” September 1999.) [RFC2679] suggest to apply a 95% confidence interval. Quoting RFC 2679, "95 percent was chosen because ... a particular confidence level should be specified so that the results of independent implementations can be compared."

Two different implementations are expected to produce statistically equivalent results if they both measure a metric under the same networking conditions. Formulating in statistical terms: separate metric implementations collect separate samples from the same underlying statistical process (the same network conditions). The statistical hypothesis to be tested is the expectation that both samples do not expose statistically different properties. This requires careful test design:

* The measurement test setup must be self-consistent to the largest possible extent. To minimize the influence of the test and measurement setup on the result, network conditions and paths MUST be identical for the compared implementations to the largest possible degree. This includes both the stability and non-ambiguity of routes taken by the measurement packets. See RFC 2330 for a discussion on self-consistency.

* The error induced by the sample size must be small enough to minimize its influence on the test result. This may have to be respected, especially if two implementations measure with different average probing rates.

* Every comparison must be repeated several times based on different measurement data to avoid random indications of compatibility (or the lack of it).

* To minimize the influence of implementation options on the result, metric implementations SHOULD use identical options and parameters for the metric under evaluation.

*The implementation with the lowest probing frequency determines the smallest temporal interval for which samples can be compared.

The metric specifications themselves are the primary focus of evaluation, rather than the implementations of metrics. The documentation produced by the advancement process should identify which metric definitions and supporting material were found to be clearly worded and unambiguous, OR, it should identify ways in which the metric specification text should be revised to achieve clarity and unified interpretation.
The process should also permit identification of options that were not implemented, so that they can be removed from the advancing specification (this is an aspect more typical of protocol advancement along the standards track).
Note that this document does not propose to base interoperability indications of performance metric implementations on comparisons of individual singletons. Individual singletons may be impacted by many statistical effects while they are measured. Comparing two singletons of different implementations may result in failures with higher probability than comparing samples.

---

3. Verification of conformance to a metric specification

TOC

This section specifies how to verify compliance of two or more IPPM implementations against a metric specification. This document only proposes a general methodology. Compliance criteria to a specific metric implementation need to be defined for each individual metric specification. The only exception is the statistical test comparing two metric implementations which are simultaneously tested. This test is applicable without metric specific decision criteria.

3.1. Tests of an individual implementation against a metric specification

TOC

A metric implementation MUST support the requirements classified as "MUST" and "REQUIRED" of the related metric specification to be compliant to the latter.
Further, supported options of a metric implementation SHOULD be documented in sufficient detail. The documentation of chosen options is RECOMMENDED to minimise (and recognise) differences in the test setup if two metric implementations are compared. Further, this documentation is used to validate and improve the underlying metric specification option, to remove options which saw no implementation or which are badly specified from the metric specification to be promoted to a

standard. This documentation SHOULD be made for all implementation relevant specifications of a metric picked for a comparison, which aren't explicitly marked as "MUST" or "REQUIRED" in the metric specification. This applies for the following sections of all metric specifications:

- *Singleton Definition of the Metric.
- *Sample Definition of the Metric.
- *Statistics Definition of the Metric. As statistics are compared by the test specified here, this documentation is required even in the case, that the metric specification does not contain a Statistics Definition.
- *Timing and Synchronisation related specification (if relevant for the Metric).
- *Any other technical part present or missing in the metric specification, which is relevant for the implementation of the Metric.

RFC2330 and RFC2679 emphasise precision as an aim of IPPM metric implementations. A single IPPM conformant implementation MUST under otherwise identical network conditions produce precise results for repeated measurements of the same metric.
RFC 2330 prefers the "empirical distribution function" EDF to describe collections of measurements. RFC 2330 determines, that "unless otherwise stated, IPPM goodness-of-fit tests are done using 5% significance." The goodness of fit test determines by which precision two or more samples of a metric implementation belong to the same underlying distribution (of measured network performance events). The goodness of fit test to be applied is the Anderson-Darling K sample test (ADK sample test, K stands for the number of samples to be compared) (Scholz, F. and M. Stephens, “K-sample Anderson-Darling Tests of fit, for continuous and discrete cases,” May 1986.) [ADK]. Please note that RFC 2330 and RFC 2679 apply an Anderson Darling goodness of fit test too.
The results of a repeated test with a single implementation MUST pass an ADK sample test with confidence level of 95%. The resolution for which the ADK test has been passed with the specified confidence level MUST be documented. To formulate this differently: The requirement is to document the smallest resolution, at which the results of the tested metric implementation pass an ADK test with a confidence level of 95%. The minimum resolution available in the reported results from each implementation MUST be taken into account in the ADK test.

TOC

3.2. Test setup resulting in identical live network testing conditions

Two major issues complicate tests for metric compliance across live networks under identical testing conditions. One is the general point that metric definition implementations cannot be conveniently examined in field measurement scenarios. The other one is more broadly described as "parallelism in devices and networks", including mechanisms like those that achieve load balancing (see (Salowey, J. and R. Droms, “RADIUS Delegated-IPv6-Prefix Attribute,” April 2007.) [RFC4818]).
This section proposes two measures to deal with both issues. Tunneling mechanisms can be used to avoid parallel processing of different flows in the network. Measuring by separate parallel probe flows results in repeated collection of data. If both measures are combined, WAN network conditions are identical for a number of independent measurement flows, no matter what the network conditions are in detail.
Any measurement setup MUST be made to avoid the probing traffic itself to impede the metric measurement. The created measurement load MUST NOT result in congestion at the access link connecting the measurement implementation to the WAN. The created measurement load MUST NOT overload the measurement implementation itself, eg. by causing a high CPU load or by creating imprecisions due to internal transmit (receive respectively) probe packet collisions.
IP in IP tunnels can be used to avoid Equal-Cost Multi-Path (ECMP) routing of different measurement streams if they carry inner IP packets from different senders in a single tunnel with the same outer origin and destination address as well as the same port numbers.
>>>> Comment: The author is not an expert on tunneling and appreciates guidance on the applicability of one or more of the following protocols: IP in IP (Perkins, C., “IP Encapsulation within IP,” October 1996.) [RFC2003] or L2TP (Townsley, W., Valencia, A., Rubens, A., Pall, G., Zorn, G., and B. Palter, “Layer Two Tunneling Protocol "L2TP",” August 1999.) [RFC2661] or [RFC3931] (Lau, J., Townsley, M., and I. Goyret, “Layer Two Tunneling Protocol - Version 3 (L2TPv3),” March 2005.).RFC 4928 (Swallow, G., Bryant, S., and L. Andersson, “Avoiding Equal Cost Multipath Treatment in MPLS Networks,” June 2007.) [RFC4928] proposes measures how to avoid ECMP treatment in MPLS networks.
By applying 802.1Q VLANs combined with an Ethernet port based tunnel mechanism likeGeneric Routing Encapsulation (GRE) (Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, “Generic Routing Encapsulation (GRE),” March 2000.) [RFC2784] or Ethernet Pseudo Wires (Martini, L., Rosen, E., El-Aawar, N., and G. Heron, “Encapsulation Methods for Transport of Ethernet over MPLS Networks,” April 2006.) [RFC4448] the desired test environment can be set up (see Figures 1 and 2). By this solution, Ethernet frames are transmitted, containing the measurement packets. The IP packet size of the metric implementation SHOULD be chosen small enough to avoid fragmentation due to tghe added Ethernet and tunnel headers.

If tunneling is applied, two tunnels MUST carry all test traffic in between the test site and the remote site. For example, if 802.1Q Ethernet Virtual LANs (VLAN) are applied and the measurement streams are carried in different VLANs, the GRE tunnel or Pseudo Wires respectively MUST be set up in physical port mode to avoid set up of Pseudo Wires per VLAN (which may see different paths due to ECMP routing), see RFC 4448. The remote router and the Ethernet switch shown in figure 2 must support 802.1Q in this set up.
The tunneled packets carry an overhead. To avoid fragmentation in the Internet, it is suggested to limit the size of the test packets. The following headers are added if VLANs and GRE tunnels are applied:

*Ethernet 802.1Q: 22 Byte.

*GRE Header: 8 Byte.

*IPv4 Header (outer IP header): 20 Byte.

*MPLS Labels may be added by a carrier. Each MPLS Label has a length of 4 Bytes. By the time of writing, between 1 and 4 Labels seems to be a fair guess of what's expectable.

Each test is repeated several times. WAN conditions may change over time. Sequential testing is desirable, but may not be a useful metric test option. However tests can be carried out by establishing n different parallel measurement flows. Two or three linecards per implementation serving to send or receive measurement flows should be sufficient to create 5 or more parallel measurement flows. If three linecards are used, each card sends and receives 2 flows. Other options are to separate flows by DiffServ marks (without deploying any QoS in the inner or outer tunnel) or using a single CBR flow and evaluating every n-th singleton to belong to a specific measurement flow.
Tunneling setups like the one proposed by GRE encapsulated multicast probing (Gu, Y., Duffield, N., Breslau, L., and S. Sen, “GRE Encapsulated Multicast Probing: A Scalable Technique for Measuring One-Way Loss,” June 2007.) [GU+Duffield] should be applied (note that one or more remote tunnel end points and the same number of additional routers are required).
An illustration of a test setup with two tunnels and two flows between two linecards of one implementation is given in Figure 1.

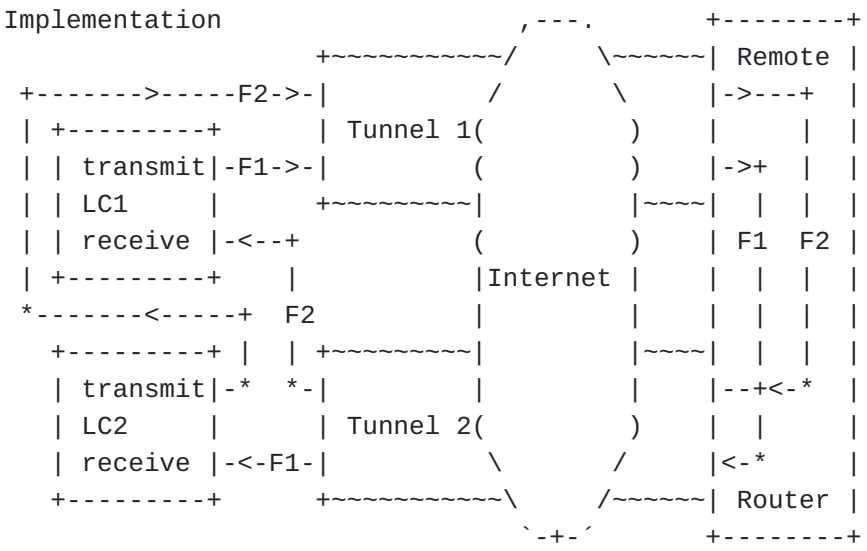

Illustration of a test setup with two tunnels and two flows F between two linecards LC of one implementation.

Figure 1

Figure 2 shows the network elements required to set up GRE tunnels or as shown by figure 1.

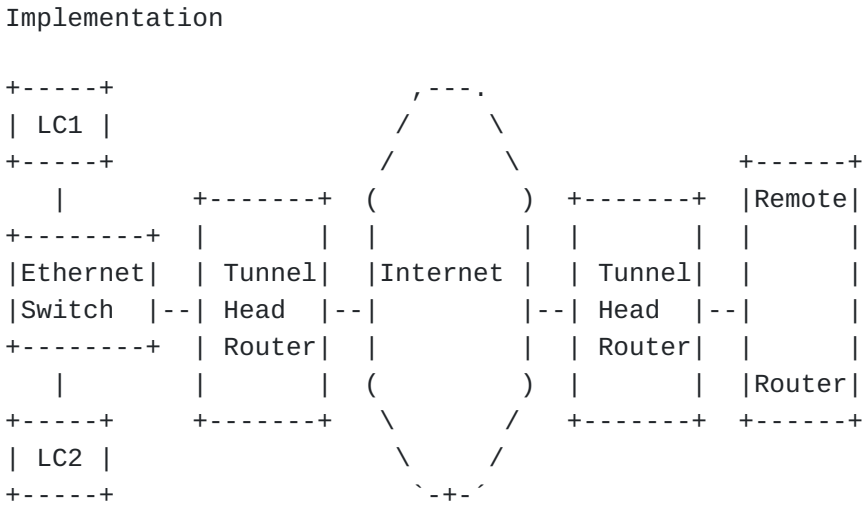

Illustration of a hardware setup to realise the test setup illustrated by figure 1 with GRE tunnels or Pseudowires.

Figure 2

Some additional rules to calculate and compare samples have to be respected to perform a metric test:

- To compare different probes of a common underlying distribution in terms of metrics characterising a communication network requires to respect the temporal nature for which the assumption of common underlying distribution may hold. Any singletons or samples to be compared MUST be captured within the same time interval.

- Whenever statistical events like singletons or rates are used to characterise measured metrics of a time-interval, at least 5 singletons of a relevant metric SHOULD be present to ensure a minimum confidence into the reported value (see Wikipedia on confidence (N., N., “Confidence interval,” October 2008.) [Rule of thumb]). Note that this criterion also is to be respected e.g. when comparing packet loss metrics. Any packet loss measurement interval to be compared with the results of another implementation SHOULD contain at least five lost packets to have a minimum confidence that the observed loss rate wasn't caused by a small number of random packet drops.

- The minimum number of singletons or samples to be compared by an Anderson-Darling test SHOULD be 100 per tested metric implementation. Note that the Anderson-Darling test detects small differences in distributions fairly well and will fail for high number of compared results (RFC2330 mentions an example with 8192 measurements where an Anderson-Darling test always failed).

- Generally, the Anderson-Darling test is sensitive to differences in the accuracy or bias associated with varying implementations or test conditions. These dissimilarities may result in differing averages of samples to be compared. An example may be different packet sizes, resulting in a constant delay difference between compared samples. Therefore samples to be compared by an Anderson Darling test MAY be calibrated by the difference of the average values of the samples. Any calibration of this kind MUST be documented in the test result.

3.3. Tests of two or more different implementations against a metric specification

TOC

RFC2330 expects "a methodology for a given metric [to] exhibit continuity if, for small variations in conditions, it results in small variations in the resulting measurements. Slightly more precisely, for

every positive epsilon, there exists a positive delta, such that if two sets of conditions are within delta of each other, then the resulting measurements will be within epsilon of each other." A small variation in conditions in the context of the metric test proposed here can be seen as different implementations measuring the same metric along the same path.
IPPM metric specification however allow for implementor options to the largest possible degree. It can't be expected that two implementors pick identical options for the implementations. Implementors SHOULD to the highest degree possible pick the same configurations for their systems when comparing their implementations by a metric test.
In some cases, a goodness of fit test may not be possible or show disappointing results. To clarify the difficulties arising from different implementation options, the individual options picked for every compared implementation SHOULD be documented in sufficient detail. Based on this documentation, the underlying metric specification should be improved before it is promoted to a standard.
The same statistical test as applicable to quantify precision of a single metric implementation MUST be passed to compare metric conformance of different implementations. To document compatibility, the smallest measurement resolution at which the compared implementations passed the ADK sample test MUST be documented.
For different implementations of the same metric, "variations in conditions" are reasonably expected. The ADK test comparing samples of the different implementations may result in a lower precision than the test for precision of each implementation individually.

3.4. Clock synchronisation

TOC

Clock synchronization effects require special attention. Accuracy of one-way active delay measurements for any metrics implementation depends on clock synchronization between the source and destination of tests. Ideally, one-way active delay measurement (RFC 2679, (Almes, G., Kalidindi, S., and M. Zekauskas, “A One-way Delay Metric for IPPM,” September 1999.) [RFC2679]) test endpoints either have direct access to independent GPS or CDMA-based time sources or indirect access to nearby NTP primary (stratum 1) time sources, equipped with GPS receivers. Access to these time sources may not be available at all test locations associated with different Internet paths, for a variety of reasons out of scope of this document.
When secondary (stratum 2 and above) time sources are used with NTP running across the same network, whose metrics are subject to comparative implementation tests, network impairments can affect clock synchronization, distort sample one-way values and their interval statistics. It is RECOMMENDED to discard sample one-way delay values

for any implementation, when one of the following reliability conditions is met:

- *Delay is measured and is finite in one direction, but not the other.

- *Absolute value of the difference between the sum of one-way measurements in both directions and round-trip measurement is greater than X% of the latter value.

Examination of the second condition requires RTT measurement for reference, e.g., based on TWAMP (RFC5357, RFC 5357 (Hedayat, K., Krzanowski, R., Morton, A., Yum, K., and J. Babiarz, "A Two-Way Active Measurement Protocol (TWAMP)," October 2008.) [RFC5357]), in conjunction with one-way delay measurement.
Specification of X% to strike a balance between identification of unreliable one-way delay samples and misidentification of reliable samples under a wide range of Internet path RTTs probably requires further study.
An IPPM compliant metric implementation whose measurement requires synchronized clocks is however expected to provide precise measurement results. Any IPPM metric implementation MUST be of a precision of 1 ms (+/- 500 us) with a confidence of 95% if the metric is captured along an Internet path which is stable and not congested during a measurement duration of an hour or more. [Editor: this latter definition may avoid NTP (stratum 2 or worse) synchronized IPPM implementations from becoming IPPM compliant. However internal PC clock synched implementations can't be rejected that way.
>>>> Comment: Ideas on criteria to deal with the latter are welcome. May drift be one, as GPS synched implementations shouldn't have one or the same on origin and destination, respectively.]

3.5. Recommended Metric Verification Measurement Process

TOC

In order to meet their obligations under the IETF Standards Process the IESG must be convinced that each metric specification advanced to Draft Standard or Internet Standard status is clearly written, that there are the required multiple verifiably equivalent implementations, and that all options have been implemented.
In the context of this document, metrics are designed to measure some characteristic of a data network. An aim of any metric definition should be that it should be specified in a way that can reliably measure the specific characteristic in a repeatable way.
Each metric, statistic or option of those to be validated MUST be compared against a reference measurement or another implementation by at least 5 different basic data sets, each one with sufficient size to reach the specified level of confidence, as specified by this document.

Finally, the metric definitions, embodied in the text of the RFCs, are the objects that require evaluation and possible revision in order to advance to the next step on the standards track.
IF two (or more) implementations do not measure an equivalent metric as specified by this document,
AND sources of measurement error do not adequately explain the lack of agreement,
THEN the details of each implementation should be audited along with the exact definition text, to determine if there is a lack of clarity that has caused the implementations to vary in a way that affects the correspondence of the results.
IF there was a lack of clarity or multiple legitimate interpretations of the definition text,
THEN the text should be modified and the resulting memo proposed for consensus and advancement along the standards track.
The complete process of advancing a metric specification to a standard as defined by this document is illustrated in Figure 3.

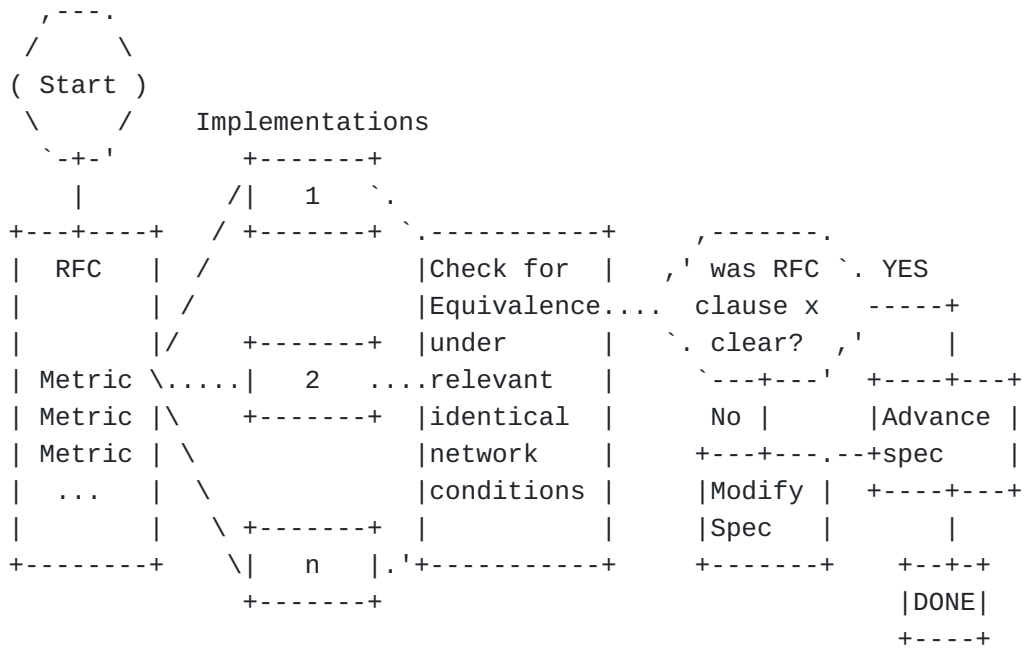

Illustration of the metric standardisation process

Figure 3

Any recommendation for the advancement of a metric specification MUST be accompanied by an implementation report, as is the case with all

requests for the advancement of IETF specifications. The implementation report needs to include the tests performed, the applied test setup, the specific metrics in the RFC and reports of the tests performed with two or more implementations. The test plan needs to specify the precision reached for each measured metric and thus define the meaning of "statistically equivalent" for the specific metrics being tested. Ideally, the test plan would co-evolve with the development of the metric, since that's when people have the most context in their thinking regarding the different subtleties that can arise.
In particular, the implementation report MUST as a minimum document:

- The metric compared and the RFC specifying it. This includes statements as required by the section "Tests of an individual implementation against a metric specification" of this document.

- The measurement configuration and setup.

- A complete specification of the measurement stream (mean rate, statistical distribution of packets, packet size or mean packet size and their distribution), DSCP and any other measurement stream properties which could result in deviating results. Deviations in results can be caused also if chosen IP addresses and ports of different implementations can result in different layer 2 or layer 3 paths due to operation of Equal Cost Multi-Path routing in an operational network.

- The duration of each measurement to be used for a metric validation, the number of measurement points collected for each metric during each measurement interval (i.e. the probe size) and the level of confidence derived from this probe size for each measurement interval.

- The result of the statistical tests performed for each metric validation as required by the section "Tests of two or more different implementations against a metric specification" of this document.

- A parameterization of laboratory conditions and applied traffic and network conditions allowing reproduction of these laboratory conditions for readers of the implementation report.

- The documentation helping to improve metric specifications defined by this section.

All of the tests for each set SHOULD be run in a test setup as specified in the section "Test setup resulting in identical live network testing conditions."
If a different test set up is chosen, it is RECOMMENDED to avoid effects falsifying results of validation measurements caused by real

data networks (like parallelism in devices and networks). Data networks may forward packets differently in the case of:

*Different packet sizes chosen for different metric implementations. A proposed countermeasure is selecting the same packet size when validating results of two samples or a sample against an original distribution.

*Selection of differing IP addresses and ports used by different metric implementations during metric validation tests. If ECMP is applied on IP or MPLS level, different paths can result (note that it may be impossible to detect an MPLS ECMP path from an IP endpoint). A proposed counter measure is to connect the measurement equipment to be compared by a NAT device, or establishing a single tunnel to transport all measurement traffic The aim is to have the same IP addresses and port for all measurement packets or to avoid ECMP based local routing diversion by using a layer 2 tunnel.

*Different IP options.

*Different DSCP.

*If the five measurements are captured by repeated measurements instead of simultaneous ones: Changing pathes and load conditions time.

3.6. Miscellaneous

TOC

In the case that a metric validation requires capturing rare events, an impairment generator may have to be added to the test set up. Inclusion of an impairment generator and the parameterisation of the impairments generated MUST be documented. Rare events could be packet duplications, packet loss rates above one digit percentages, loss patterns or packet re-ordering and so on.
As specified above, 5 singletons are the recommended basis to minimise interference of random events with the statistical test proposed by this document. In the case of ratio measurements (like packet loss), the underlying sum of basic events, against the which the metric's monitored singletons are "rated", determines the resolution of the test. A packet loss statistic with a resolution of 1% requires one packet loss statistic-datapoint to consist of 500 delay singletons (of which at least 5 were lost). To compare EDFs on packet loss requires one hundred such statistics per flow. That means, all in all at least 50 000 delay singletons are required per single measurement flow. Live network packet loss is assumed to be present during main traffic hours

only. Let this interval be 5 hours. The required minimum rate of a single measurement flow in that case is 2.8 packets/sec (assuming a loss of 1% during 5 hours). If this measurement is too demanding under live network conditions, an impairment generator should be used.

---

4. Acknowledgements

TOC

Gerhard Hasslinger commented a first version of this document, suggested statistical tests and the evaluation of time series information. Henk Uijterwaal pushed this work and Mike Hamilton, Scott Bradner and Emile Stephan commented on versions of this draft before initial publication. Carol Davids reviewed the 01 version of this draft.

---

5. Contributors

TOC

Scott Bradner, Vern Paxson and Allison Mankin drafted bradner-metrictest (Bradner, S., Mankin, A., and V. Paxson, “Advancement of metrics specifications on the IETF Standards Track,” July 2007.) [bradner-metrictest], and major parts of it are included in this document.

---

6. IANA Considerations

TOC

This memo includes no request to IANA.

---

7. Security Considerations

TOC

This draft does not raise any specific security issues.

---

8. References

TOC

8.1. Normative References

TOC

[RFC2003]	Perkins, C., "IP Encapsulation within IP," RFC 2003, October 1996 (TXT, HTML, XML).
[RFC2026]	Bradner, S., "The Internet Standards Process -- Revision 3," BCP 9, RFC 2026, October 1996 (TXT).
[RFC2119]	Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels," BCP 14, RFC 2119, March 1997 (TXT, HTML, XML).
[RFC2330]	Paxson, V., Almes, G., Mahdavi, J., and M. Mathis, "Framework for IP Performance Metrics," RFC 2330, May 1998 (TXT, HTML, XML).
[RFC2661]	Townsley, W., Valencia, A., Rubens, A., Pall, G., Zorn, G., and B. Palter, "Layer Two Tunneling Protocol "L2TP"," RFC 2661, August 1999 (TXT).
[RFC2679]	Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Delay Metric for IPPM," RFC 2679, September 1999 (TXT).
[RFC2680]	Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Packet Loss Metric for IPPM," RFC 2680, September 1999 (TXT).
[RFC2681]	Almes, G., Kalidindi, S., and M. Zekauskas, "A Round-trip Delay Metric for IPPM," RFC 2681, September 1999 (TXT).
[RFC2784]	Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, "Generic Routing Encapsulation (GRE)," RFC 2784, March 2000 (TXT).
[RFC3931]	Lau, J., Townsley, M., and I. Goyret, "Layer Two Tunneling Protocol - Version 3 (L2TPv3)," RFC 3931, March 2005 (TXT).
[RFC4448]	Martini, L., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks," RFC 4448, April 2006 (TXT).
[RFC4656]	Shalunov, S., Teitelbaum, B., Karp, A., Boote, J., and M. Zekauskas, "A One-way Active Measurement Protocol (OWAMP)," RFC 4656, September 2006 (TXT).
[RFC4818]	Salowey, J. and R. Droms, "RADIUS Delegated-IPv6-Prefix Attribute," RFC 4818, April 2007 (TXT).
[RFC4928]	Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks," BCP 128, RFC 4928, June 2007 (TXT).

8.2. Informative References

TOC

[ADK]	Scholz, F. and M. Stephens, “K-sample Anderson-Darling Tests of fit, for continuous and discrete cases,” University of Washington, Technical Report No. 81, May 1986.
[GU+Duffield]	Gu, Y., Duffield, N., Breslau, L., and S. Sen, “GRE Encapsulated Multicast Probing: A Scalable Technique for Measuring One-Way Loss,” SIGMETRICS’07 San Diego, California, USA, June 2007.
[RFC5357]	Hedayat, K., Krzanowski, R., Morton, A., Yum, K., and J. Babiarz, “A Two-Way Active Measurement Protocol (TWAMP),” RFC 5357, October 2008 (TXT).
[Rule of thumb]	N., N., “Confidence interval,” October 2008.
[bradner-metrictest]	Bradner, S., Mankin, A., and V. Paxson, “Advancement of metrics specifications on the IETF Standards Track,” draft -morton-ippm-advance-metrics-00, (work in progress), July 2007.
[morton-advance-metrics]	Morton, A., “Problems and Possible Solutions for Advancing Metrics on the Standards Track,” draft -bradner-metricstest-03, (work in progress), July 2009.

Appendix A. An example on a One-way Delay metric validation

TOC

The text of this appendix is not binding. It is an example how parts of a One-way Delay metric test could look like. http://xml.resource.org/public/rfc/bibxml/

A.1. Compliance to Metric specification requirements

TOC

One-way Delay, Loss threshold, RFC 2679
This test determines if implementations use the same configured maximum waiting time delay from one measurement to another under different delay conditions, and correctly declare packets arriving in excess of the waiting time threshold as lost. See Section 3.5 of RFC2679, 3rd bullet point and also Section 3.8.2 of RFC2679.

(1) Configure a path with 1 sec one-way constant delay.

(2) Measure one-way delay with 2 or more implementations, using identical waiting time thresholds for loss set at 2 seconds.

(3) Configure the path with 3 sec one-way delay.

(4) Repeat measurements.

(5) Observe that the increase measured in step 4 caused all packets to be declared lost, and that all packets that arrive successfully in step 2 are assigned a valid one-way delay.

One-way Delay, First-bit to Last bit, RFC 2679
This test determines if implementations register the same relative increase in delay from one measurement to another under different delay conditions. This test tends to cancel the sources of error which may be present in an implementation. See Section 3.7.2 of RFC2679, and Section 10.2 of RFC2330.

(1) Configure a path with X ms one-way constant delay, and ideally including a low-speed link.

(2) Measure one-way delay with 2 or more implementations, using identical options and equal size small packets (e.g., 100 octet IP payload).

(3) Maintain the same path with X ms one-way delay.

(4) Measure one-way delay with 2 or more implementations, using identical options and equal size large packets (e.g., 1500 octet IP payload).

(5) Observe that the increase measured in steps 2 and 4 is equivalent to the increase in ms expected due to the larger serialization time for each implementation. Most of the measurement errors in each system should cancel, if they are stationary.

One-way Delay, RFC 2679
This test determines if implementations register the same relative increase in delay from one measurement to another under different delay conditions. This test tends to cancel the sources of error which may be present in an implementation. This test is intended to evaluate measurments in sections 3 and 4 of RFC2679.

(1) Configure a path with X ms one-way constant delay.

(2) Measure one-way delay with 2 or more implementations, using identical options.

(3) Configure the path with X+Y ms one-way delay.

(4) Repeat measurements.

(5)

> Observe that the increase measured in steps 2 and 4 is ~Y ms for each implementation. Most of the measurement errors in each system should cancel, if they are stationary.

Error Calibration, RFC 2679
This is a simple check to determine if an implementation reports the error calibration as required in Section 4.8 of RFC2679. Note that the context (Type-P) must also be reported.

A.2. Examples related to statistical tests for One-way Delay

TOC

A one way delay measurement may pass an ADK test with a timestamp resultion of 1 ms. The same test may fail, if timestamps with a resolution of 100 microseconds are eavluated. The implementation then is then conforming to the metric specification up to a timestamp resolution of 1 ms.
Let's assume another one way delay measurement comparison between implementation 1, probing with a frequency of 2 probes per second and implementation 2 probing at a rate of 2 probes every 3 minutes. To ensure reasonable confidence in results, sample metrics are calculated from at least 5 singletons per compared time interval. This means, sample delay values are calculated for each system for identical 6 minute intervals for the whole test duration. Per 6 minute interval, the sample metric is calculated from 720 singletons for implementation 1 and from 6 singletons for implementation 2. Note, that if outliers are not filtered, moving averages are an option for an evaluation too. The minimum move of an averaging interval is three minutes in this example.
The data in table 1 may result from measuring One-Way Delay with implementation 1 (see column Implemnt_1) and implementation 2 (see column implemnt_2). Each data point in the table represents a (rounded) average of the sampled delay values per interval. The resolution of the clock is one micro-second. The difference in the delay values may result eg. from different probe packet sizes.

Implemnt_1	Implemnt_2	Implemnt_2 - Delta_Averages
5000	6549	4997
5008	6555	5003
5012	6564	5012
5015	6565	5013
5019	6568	5016

5022	6570	5018
5024	6573	5021
5026	6575	5023
5027	6577	5025
5029	6580	5028
5030	6585	5033
5032	6586	5034
5034	6587	5035
5036	6588	5036
5038	6589	5037
5039	6591	5039
5041	6592	5040
5043	6599	5047
5046	6606	5054
5054	6612	5060

Table 1

Average values of sample metrics captured during identical time intervals are compared. This excludes random differences caused by differing probing intervals or differing temporal distance of singletons resulting from their Poisson distributed sending times.
In the example, 20 values have been picked (note that at least 100 values are recommended for a single run of a real test). Data must be ordered by ascending rank. The data of Implemnt_1 and Implemnt_2 as shown in the first two columns of table 1 clearly fails an ADK test with 95% confidence.
The results of Implemnt_2 are now reduced by difference of the averages of column 2 (rounded to 6581 us) and column 1 (rounded to 5029 us), which is 1552 us. The result may be found in column 3 of table 1. Comparing column 1 and column 3 of the table by an ADK test shows, that the data contained in these columns passes an ADK tests with 95% confidence.

Appendix B. Glossary

TOC

ADK	

	Anderson-Darling K-Sample test, a test used to check whether two samples have the same statistical distribution.
ECMP	Equal Cost Multipath, a load balancing mechanism evaluating MPLS labels stacks, IP addresses and ports.
EDF	The "Empirical Distribution Function" of a set of scalar measurements is a function F(x) which for any x gives the fractional proportion of the total measurements that were smaller than or equal as x.
Metric	A measured quantity related to the performance and reliability of the Internet, expressed by a value. This could be a singleton (single value), a sample of single values or a statistic based on a sample of singletons.
OWAMP	One-way Active Measurement Protocol, a protocol for communication between IPPM measurement systems specified by IPPM.
OWD	One-Way Delay, a performance metric specified by IPPM.
Sample metric	A sample metric is derived from a given singleton metric by evaluating a number of distinct instances together.
Singleton metric	A singleton metric is, in a sense, one atomic measurement of this metric.
Statistical metric	A 'statistical' metric is derived from a given sample metric by computing some statistic of the values defined by the singleton metric on the sample.
TWAMP	Two-way Active Measurement Protocol, a protocol for communication between IPPM measurement systems specified by IPPM.

Table 2

Authors' Addresses

TOC

	Ruediger Geib (editor)
	Deutsche Telekom
	Heinrich Hertz Str. 3-7
	Darmstadt, 64295
	Germany
Phone:	+49 6151 628 2747
Email:	Ruediger.Geib@telekom.de
	Al Morton

	AT&T Labs
	200 Laurel Avenue South
	Middletown, NJ 07748
	USA
Phone:	+1 732 420 1571
Fax:	+1 732 368 1192
Email:	acmorton@att.com
URI:	http://home.comcast.net/~acmacm/
	Reza Fardid
	Covad Communications
	2510 Zanker Road
	San Jose, CA 95131
	USA
Phone:	+1 408 434-2042
Email:	RFardid@covad.com