Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: April 29, 2010

R. Geib, Ed.
Deutsche Telekom
A. Morton
AT&T Labs
R. Fardid
Covad Communications

October 26, 2009

IPPM standard compliance testing
draft-geib-ippm-metrictest-01

Status of this Memo

Copyright Notice

Abstract

This document specifies tests to determine if multiple, independent, and interoperable implementations of a metrics specification document are at hand so that the metrics specification can be advanced to an Internet standard. Results of different IPPM implementations can be compared if they measure under the same underlying network conditions. Results are compared using state of the art statistical methods.

Table of Contents

1. Introduction

Draft bradner-metrictest [bradner-metrictest] states:

The Internet Standards Process RFC2026 [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.

In the case of a specification for a performance metric, network latency for example, exactly what constitutes "interoperation" is less obvious. The IESG didn't yet decide how to judge "metric specification interoperability" in the context of the IETF Standards Process and this new draft suggests a methodology which (hopefully) is suitable for IPPM metrics. General applicability of the methods proposed in the following should however not be excluded.

A metric specification describes a method of testing and a way to report the results of this testing. One example of such a metric would be a way to test and report the latency that data packets would incur while being sent from one network location to another.

Since implementations of testing metrics are by their nature stand-alone and do not interact with each other, the level of the interoperability called for in the IETF standards process cannot be simply determined by seeing that the implementations interact properly. Instead, verifying equivalence by proofing that different implementations verifiably give statistically equivalent results Verifiable equivalence may take the place of interoperability.

This document defines the process of verifying equivalence by using a specified test set up to create the required separate data sets (which may be seen as samples taken from the same underlying distribution) and then apply state of the art statistical methods to verify equivalence of the results. To illustrate application of the process defined her, validating compliance with RFC2679 [RFC2679] is picked as an example. While test set ups may vary with the metrics to be validated, the statistical methods will not. Documents defining test setups to validate other metrics should be created by the IPPM WG, once the process proposed here has been agreed upon.

This document defines the process of verifying equivalence by using a

specified test set up to create the required separate data sets (which may be seen as samples taken from the same underlying distribution) and then apply state of the art statistical methods to verify equivalence of the results. To illustrate application of the process defined her, validating compliance with RFC2679 [RFC2679] is picked as an example. While test set ups may vary with the metrics to be validated, the statistical methods will not. Documents defining test setups to validate other metrics should be created by the IPPM WG, once the process proposed here has been agreed upon.

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-advance-metrics]).

- More emphasis on the requirement to carefully design and document the measurement set up of the metric comparison.

- Proposal of testing conditions under identical WAN netwrok 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

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 [RFC2119].

2. Basic idea

The Framework for IP Performance Metrics (RFC 2330, [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 IPPM implementation is expected to measure a metric with high precision. The metric compliance test

specified in the following emphasises precision over accuracy. Further the methodology and test methods proposed by RFC 2330 are used by this document too.

The implementation of a standard compliant metric is expected to meet the requrirements of the related a metric specification. So before comparing two metrice implementations, each metric implementation is individually compared against the metric specification. As an example, an implementation of the OWD metric must be calibrated. Calibration results of a standard conformant metric implementation must be published then.

Most metric specificatios leave freedom to implementors on those aspects, which aren't fundamental for an individual metric implementation. Calibration of individual metric implementations and comparing different ones requires a careful design and documentation of the metric implementation and of the testing conditions.

The IPPM framework expects repeating measurements to lead to the same results, if the conditions under which these measurements have been collected are identical. Small deviations are expected to lead to small deviations in results only. To charaterise statistical

equivalence in the case of small deviations, RFC 2330 and RFC 2679 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 IPPM implementations are expected to measure statistically equivalent results, if they both measure a metric under the same networking conditions. Formulating the measurement in statistical terms: separate samples are collected (by separate metric implementations) 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 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.

- If statistics of time series are compared, the implementation with the lowest probing frequency determines the smallest temporal interval for which results can be compared.

- Every comparison must be repeated several times based on different measurement data to avoid random indications of compatibility (or

the lack of it).

- The measurement test set up must be self-consistent to the largest possible extent. This means, network conditions, paths and IPPM metric implementations SHOULD be identical for the compared implementations to the largest possible degree to minimize the influence of the test and measurement set up on the result. This includes e.g. aspects of the stability and non-ambiguity of routes taken by the measurement packets. See RFC 2330 for a discussion on self-consistency RFC 2330 [RFC2330].

As addressed by "problems and solutions for metric advancement" [morton-advance-metrics], documentation of the metric test will indicate which requirements and options of a metric specification are specified clear enough for an implementation or uncover gaps in the

metric specification. The final step in advancing a metric specification to standard is by improving unclear specifications and by cleaning it from not supported options.

3. Verification of conformance to a metric specification

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 are expected to be drafted 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

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 to validate and improve the underlying metric specification option or remove options which saw no implementation or which are badly specified from the metric specification to be promoted to a standard.

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 required to determine the preciusion of a metric implementation consists of testing, whether two or more samples 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 test, K stands for

the number of samples to be compared). Please note that RFC 2330 and RFC 2679 apply an Anderson Darling goodness of fit test too.

The results of a repeated tests 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 different: The requirement is to document the smalles resolution, at which the results of the tested metric implementation pass an ADK test with a confidence level of 95%.

As an example, 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.

3.2. Test set up resulting in identical live network testing conditions

Two major issues complicate tests for metric compliance across live networks under identical testing conditions. One of these is the general posit, "metric definition implementations cannot be conveniently examined in field measurement scenarios". The other is more more specificcally addressing "parallelism in devices and networks", by which mechanisms like load balancing are meant. As a reference for the latter, [RFC 4814] is given.

This section proposes two measures how to deal with both. Tunneling mechanisms can be used to avoid pallalel processing of different flows in the network. Measuring by separate parallel probe flows results in repeated collection of data. In both cases, WAN network conditions are identical, no matter what they are in detail.

Any measurement set up 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 send/receive probe packet collisions.

IP in IP tunnels can be used to avoid ECMP routing of different

measurement streams if they allow to 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. 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 [RFC2003], GRE [RFC2784] or L2TP [RFC2661] or [RFC3931]. RFC 4928 [RFC4928] proposes measures how to avoid ECMP treatment in MPLS networks. Applying Pseudo-Wires for a metric implementation test is one way to avoid MPLS based ECMP treatment. If tuneling is applied, a single tunnel MUST carry all test traffic in one direction. If eg. Ethernet Pseudo Wires are applied and the measurement streams are carried in different VLANs, the Pseudo Wires 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 [RFC4448].

To have statsitical significance, a test MUST be repeated 5 times at least (see below). WAN conditions may change over time. Sequential testing is no useful metric test option. However tests can be carried out by applying 5 or more different parallel measuremet flows. The author takes no position, whether such a test is carried out by sending eg a single CBR flow and defining avery n-th (n = 1..5) packet to belong to a specific measurement flow, or whether multiple network cards are applied to create several distinct flows of a single implementation. In the latter case, three different cards of one implementation at a single test site will do, if tunneling set ups like the one proposed by GRE encapsulated multicast probing [GU&Duffield] are applied (note that one or more remote tunnel end points and the same number of routers are required).

Some additional rules to calculate and compare samples have to be respected. The following rules are of importance for the IPPM 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 events of a relevant metric MUST be present to ensure a minimum confidence into the reported value (see Wikipedia on confidence [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 needs to contain at least five lost packets

to have a minimum confidence that the observed loss rate wasn't caused by a samll number of random packet drops.

- The minimum number of singletons or samples to be compared by an Anderson-Darling test is 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 to guarantee a failure of an Anderson-Darling test).

- The Anderson-Darling test is sensible against differing accuracy or bias of different implementations. These differences result in differing averages of compared samples. In general, differences in averages of samples may result from differing test conditions. 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.

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

RFC2330 expects that a "a methodology for a given metric exhibits 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 a metric comparison can be seen as different implementations measuring the same metric along the same path.

RFC2679 comments that a "95 percent [confidence level for an Anderson-Darling goodness of fit test] was chosen because....a particular confidence level should be specified so that the results of independent implementations can be compared." While the RFC 2679 statement refers to calibration, it expresses the expectation that the methodology allows for comparisons between different implementations.

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

dissapointing results. To clarify the difficulties arising from different implemenation 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 implemenations. 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 implemenations may result in a lower precision than the test for precision of each implementation individually.

3.4. Clock synchronisation

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, [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 acrossthe 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:

o Delay is measured and is finite in one direction, but not the

other.

- o 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 [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 synchonized 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) synchonized IPPM implementations from becoming IPPM compliant. However internal PC clock synched implementations can't be rejected that way. 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

The proposal made by the authors of bradner-metrictest [bradner-metrictest] is picked up and slightly enhanced:

"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 memo, 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 on with sufficient size to reach the specified level of confidence.

"In the same way, sequentially running different implementations of software that perform the tests described in the metric document on a stable network, or simultaneously on a network that may or may not be stable should produce essentially the same results."

Following these assumptions any recommendation for the advancement of a metric specification needs to 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 a

specific plan to test the specific metrics in the RFC in lab or real-world networks and reports of the tests performed with two or more implementations of the software. The test plan should cover key parts of the specification, 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, including the chosen options (like e.g. the implemented selection function in the case of IPDV).

- 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 property 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.

- The measurement configuration and set up.

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

All of the tests for each set MUST be run in a test set up as specified in the section "Test set up resulting in identical live network testing conditions."

It is RECOMMENDED to avoid effects falsifying results of real data networks, if validation measurements are taken over them. Obviously, the conditions met there can't be reproduced. As the measurement equipment compared is designed to reliable quantify real network

performance, validating metrics under real network conditions is desirable of course.

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.

4. Acknowledgements

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 reviewed the document before publication.

5. Contributors

Scott Bradner, Vern Paxson and Allison Manking drafted bradner-metrictest [bradner-metrictest], and major parts of it are quoted in this document. Scott Bradner and Emile Stephan commented this draft before publication.

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

This draft does not raise any specific security issues.

8. References

8.1. Normative References

[RFC2003] Perkins, C., "IP Encapsulation within IP", RFC 2003, October 1996.

[RFC2026] Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2330] Paxson, V., Almes, G., Mahdavi, J., and M. Mathis, "Framework for IP Performance Metrics", RFC 2330, May 1998.

[RFC2661] Townsley, W., Valencia, A., Rubens, A., Pall, G., Zorn, G., and B. Palter, "Layer Two Tunneling Protocol "L2TP"", RFC 2661, August 1999.

[RFC2679] Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Delay Metric for IPPM", RFC 2679, September 1999.

[RFC2784] Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 2784, March 2000.

[RFC3931] Lau, J., Townsley, M., and I. Goyret, "Layer Two Tunneling Protocol - Version 3 (L2TPv3)", RFC 3931, March 2005.

[RFC4448] Martini, L., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, April 2006.

[RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, June 2007.

8.2. Informative References

[Autocorrelation]
N., N., "Autocorrelation", December 2008.

[Correlation]
N., N., "Correlation", June 2009.

[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.

[Precision]
N., N., "Accuracy and precision", June 2009.

[RFC5357] Hedayat, K., Krzanowski, R., Morton, A., Yum, K., and J. Babiarz, "A Two-Way Active Measurement Protocol (TWAMP)", RFC 5357, October 2008.

[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. Further ideas on statistical tests

IPPM metrics are captured by time series. Time series can be checked for correlation. There are two expectations on statistical time series properties which should be met by separate measurements probing the same underlying network performance distribution:

- The Autocorrelation indicates, whether there are any repeating patterns within a time series. For the purpose of this document, it does not matter whether there is autocorrelation in a measurement. It is however expected, that two measurements expose the same autocorrelation on identical "lag" intervals. If calculable, the autocorrelation lies within an interval [-1;1], (see Wikipedia on autocorrelation [Autocorrelation]).

- The correlation coefficient "indicates the strength of a linear relationship between two random variables." The two random

variables in the case of this document are the measurement time series of the IPPM implementations to be compared. The expectation is, that both are strongly correlated and the resulting correlation coefficient is close to 1, (see Wikipedia on correlation [Correlation]).

A metric test can derive additional statistics from time series analysis. Further, formulation of a test hypothesis is possible for autocorrelation and the correlation coefficient. It is however not clear, whether an appropriate statistical test to validate the hypothesis by 95% significance exists. Applicability of time series analysis for a metric test requires further input from statisticians.

In the absence of any metric test on time series, any test result SHOULD provide the autocorrelation of the compared metrics time series by lags from 1 to 10. In addition, the value of the correlation coefficient SHOULD be provided. Autocorrelation and Correlation coefficient are expected to be rather close to the value 1.

As mentioned earlier, the time series analysis requires application of identical time intervals to allow a comparison. In our delay example, single sample delay metric values are calculated for 9 minute intervals. If 200 consecutive sample delay metrics with the same start and end interval are available for each implementation, autocorrelation can be calculated for different n * 9 minute lags. The autocorrelation calculated for the time series of each implementation should be very close to the autocorrelation of the other implementation for the same time lag. Further, the correlation coefficient for both time series should be close to 1.

The way to prove that two IPPM metric measurements provide compatible results then could be performed stepwise:

- First prove that the two compared implementations have the same precision by comparing statistics of the distribution of singletons (or samples) of a metric by comparing the EDF of the samples captured by the two implementations.

- Second indicate that two compared implementations produce strongly correlated time series of which each one individually has the same autocorrelation as the other one.

Comparing "Accuracy" of IPPM implementations based on averages and variations may require prior checks for the absence of long range dependency within the compared measurements. Large outliers as typically occurring in the case of long range dependency, can have a

serious impact on mean values. The median or percentiles may be more

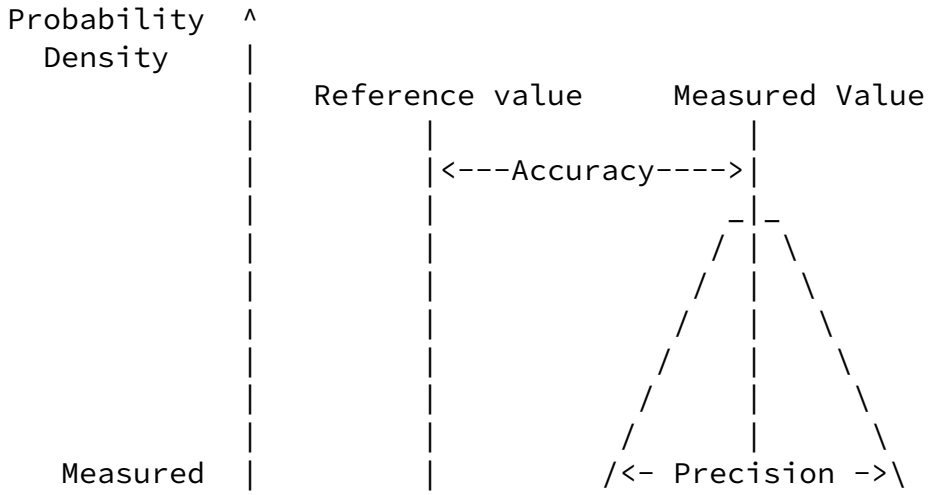
robust measures on which to compare the accuracy of different IPPM implementations. An idea may be to consider data up to a certain percentile, calculate the mean for data up to this percentile and then compare the means of the two implementations. This could be repeated for different percentiles. If long range dependencies impact is limited to large outliers, the method may work for lower percentiles. Whether this makes sense must be confirmed by a statistician, so this attempt requires further study.

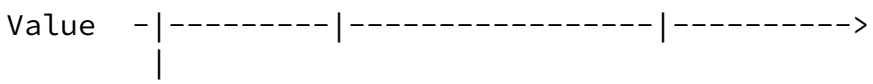
Appendix B. Verification of measurement precision by statistical methods

Following the definition of statistical precision [Precision], a measurement process can be characterised by two properties:

- Accuracy, which is the degree of conformity of a measured quantity to its actual (true) value.
- Precision, also called reproducibility or repeatability, the degree to which repeated measurements show the same or similar results.

Figure 1 further clarifies the difference between accuracy and precision of a measurement.





Measurement accuracy and precision [Precision].

Figure 1

The Framework for IP Performance Metrics (RFC 2330, [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 IPPM implementation is expected to measure a metric with high precision.

A guideline for an IPPM conformant metric implementation can be taken from these principles:

Two different implementations measuring the same IPPM metric must produce results with a limited difference if measuring under to the largest extent possible identical network conditions.

In a metric test, both conditions are expected to hold, meaning that repeated tests of two implementations MUST produce precise results for all repetition intervals.

A suitable statistical test and and a level of confidence to define whether differences are rather limited and whether a measurement is highly precise are specified below.

Let's assume a one way delay measurement comparison between system A, probing with a frequency of 2 probes per second and system B 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 system A and from 6 singletons for system B). 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 our example.

The test set up for the delay measurement is chosen to minimize errors by locating one system of each implementation at the same end of two separate sites, between which delay is measured for the metric test. Both measurement sites are connected by one IPSEC tunnel, so that all measurement packets cross the Internet with the same IP addresses. Both measurement systems measure simultaneously and the local links are dimensioned to avoid congestion caused by the probing traffic itself.

The measured delay values are reported with a resolution above the measurement error and above the synchronisation error. This is done to avoid comparing these errors between two different metric implementations instead of comparing the IPPM metric implementation itself.

The overall duration of the test is chosen so that more than 1000 six minute measurement intervals are collected. The amount of data collected allows separate comparisons for e.g. 200 consecutive 6 minute intervals. intervals, during which routes were instable, are discarded prior to evaluation.

The captured delays may have been captured singletons ranging from an absolute minimum Delay Dmin to values Dmin + 5 ms. To compare distributions, the set of singletons of a chosen evaluation interval (e.g. the data of one of the five 1800 minute capture sequences, see above) is sorted for the frequency of singletons per Dmin + N * 0.5 ms (n = 1, 2, ...). After that, a comparison of the two probe sets with any of the mentioned tests may be applied.

Authors' Addresses

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