Authors: G. Mirsky   J. Halpern   X. Min   A. Clemm
         Ericsson   Ericsson   ZTE Corp.   Futurewei
         J. Strassner J. Francois
         Futurewei   Inria

Precision Availability Metrics for SLO-Governed End-to-End Services

Abstract

This document defines a set of metrics for networking services with
performance requirements expressed as Service Level Objectives
(SLO). These metrics, referred to as Precision Availability Metrics
(PAM), can be used to assess the service levels that are being
delivered. Specifically, PAM can be used to assess whether a service
is provided in compliance with its specified quality, i.e., in
accordance with its defined SLOs.

Status of This Memo

This Internet-Draft is submitted in full conformance with the
provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering
Task Force (IETF). Note that other groups may also distribute
working documents as Internet-Drafts. The list of current Internet-
Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six
months and may be updated, replaced, or obsoleted by other documents
at any time. It is inappropriate to use Internet-Drafts as reference
material or to cite them other than as "work in progress."

This Internet-Draft will expire on 20 December 2022.

Copyright Notice

Table of Contents

1. Introduction

Network operators and network users often need to assess the quality
with which network services are being provided and delivered. In
particular in cases where service level guarantees are given and
service level objectives (SLOs) are defined, it is essential to
provide a measure of the degree with which actual service levels
that are delivered comply with SLOs that were agreed, typically in a
contract or agreement. Examples of service levels include service
latency and packet loss. Simple examples of SLOs associated with
such service levels would be target values for the maximum packet
delay (one-way and/or round trip) or maximum packet loss ratio that
would be deemed acceptable.

An example of an SLO is one that characterizes the continued ability
of a particular set of nodes to communicate. Essentially, the
absence of what is, in other contexts, is called a defect. The SLO
would include the various time and measurement aspects that would be
interpreted as a defect or failure to communicate. It is important
to note that it is being defined as a state, and thus, it has
conditions that define entry into it and exit out of it. It is
expected that an SLA includes a defect-related SLO, possibly in
addition to other SLOs.

To express the perceived quality of delivered networking services versus their SLOs, a set of metrics are needed to characterize the quality of the service being provided. Of concern is not so much the absolute service level (for example, actual latency experienced), but whether the service is provided in accordance with the negotiated, and eventually contracted, service levels. For instance, this may include whether the packet delay that is experienced falls within an acceptable range that has been contracted for the service. The specific quality of service depends on the SLO that is in effect. A non-conformance to an SLO might result in degradation of the quality of experience for gamers or even jeopardize the safety of a large geographical area. However, as those applications represent clear business opportunities, they demand dependable technical solutions.

The same service level may be deemed acceptable for one application, while unacceptable for another, depending on the needs of the application. Hence it is not sufficient to simply measure service levels per se over time, but to assess the quality of the service being provided with the applicable SLO in mind. However, at this point, there are no standard metrics in place that can be used to account for the quality with which services are delivered relative to their SLOs, and whether their SLOs are being met at all times. Such metrics and the instrumentation to support them are essential for a number of purposes, including monitoring (to ensure that networking services are performing according to their objectives) as well as accounting (to maintain a record of service levels delivered, important for monetization of such services as well as for triaging of problems).

The current state-of-the-art of metrics available today includes, for example, interface metrics, useful to obtain data on traffic volume and behavior that can be observed at an interface [RFC2863] and [RFC8343], but agnostic of actual service levels and not specific to distinct flows. Flow records [RFC7011] and [RFC7012] maintain statistics about flows, including flow volume and flow duration, but again, contain very little information about end-to-end service levels, let alone whether the service levels delivered to meet their targets, i.e., their associated SLOs.

This specification introduces a new set of metrics, Precision Availability Metrics (PAM), aimed at capturing end-to-end service levels for a flow, specifically the degree to which flows comply with the SLOs that are in effect. PAM can be used to assess whether a service is provided in compliance with its specified quality, i.e., in accordance with its defined SLOs. This information can be used in multiple ways, for example, to optimize service delivery, take timely counteractions in the event of service degradation, or account for the quality of services being delivered.

Availability is discussed in Section 3.4 of [RFC7297]. In this document, the term "availability" reflects that a service which is characterized by its SLOs is considered unavailable whenever those SLOs are violated, even if basic connectivity is still working. "Precision" refers to the fact that services whose end-to-end service levels are governed by SLOs, and which must therefore be precisely delivered according to the associated quality and performance requirements. It should be noted that "precision" refers to what is being assessed, not to the mechanism used to measure it; in other words, it does not refer to the precision of the mechanism with which actual service levels are measured. The specification and implementation of methods that provide for accurate measurements is a separate topic independent of the definition of the metrics in which the results of such measurements would be expressed.

[Ed.note: It should be noted that at this point, the set of metrics proposed here is intended as a "starter set" that is intended to spark further discussion. Other metrics are certainly conceivable; we expect that the list of metrics will evolve as part of the Working Group discussions.]

2. Conventions and Terminology

2.1. Terminology

In this document, SLA and SLO are used as defined in Section 4.1 [I-D.ietf-teas-ietf-network-slices].

2.2. Acronyms

[Ed.Note: needs updating.]

PAM Precision Availability Metric

OAM Operations, Administration, and Maintenance

VI Violated Interval

VIR Violated Interval Ratio

SVI Severely Violated Interval

SVIR Severely Violated Interval Ratio

VFI Violation-Free Interval

3. Performance Availability Metrics

3.1. Introducing Violated Intervals

When analyzing the availability metrics of a service flow between
two nodes, we need to select a time interval as the unit of PAM. In
[ITU.G.826], a time interval of one second is used. That is
reasonable, but some services may require different granularity. For
that reason, the time interval in PAM is viewed as a variable
parameter though constant for a particular measurement session.
Further, for the purpose of PAM, each time interval, e.g., second or
decamillisecond, is classified either as Violated Interval (VI),
Severely Violated Interval (SVI), or Violation-Free Interval (VFI).
These are defined as follows:

  *VI is a time interval during which at least one of the
   performance parameters degraded compared to its pre-defined
   optimal level threshold.

  *SVI is a time interval during which at least one the performance
   parameters degraded compared to its pre-defined critical
   threshold.

  *Consequently, VFI is a time interval during which all performance
   objectives are at or better than their respective pre-defined
   optimal levels. In such a case, the service is in compliance with
   its specification.

Mechanisms of setting levels of threshold of an SLO are outside the
scope for this document.

From these defitions, a set of basic metrics can be defined that
count the numbers of time intervals that fall into each category:

  *VI count.

  *SVI count.

  *VFI count.

These count metrics are essential in calculating respective ratios
that can be used to assess the instability of the service.

3.2. Derived Precision Availability Metrics

A set of metrics can be created based on PAM introduced in Section
3. In this document, these metrics are referred to as derived PAM.
Some of these metrics are modeled after Mean Time Between Failure

(MTBF) metrics - a "failure" in this context referring to a failure
to deliver a packet according to its SLO.

  *Time since the last violated interval (e.g., since last violated
   ms, since last violated second). (This parameter is suitable for
   monitoring the current compliance status of the service, e.g.,
   for trending analysis.)

  *Packets since the last violated packet. (This parameter is
   suitable for the monitoring of the current compliance status of
   the service.)

  *Mean time between EIs (e.g., between violated milliseconds,
   violated seconds) is the arithmetic mean of time between
   consecutive EIs.

  *Mean packets between EIs is the arithmetic mean of the number of
   SLO-compliant packets between consecutive EIs. (Another variation
   of "MTBF" in a service setting.)

An analogous set of metrics can be produced for SVI:

  *Time since the last SVI (e.g., since last violated ms, since last
   violated second). (This parameter is suitable for the monitoring
   of the current compliance status of the service.)

  *Packets since the last severely violated packet. (This parameter
   is suitable for the monitoring of the current compliance status
   of the service.)

  *Mean time between SVIs (e.g., between severely violated
   milliseconds, severely violated seconds) is the arithmetic mean
   of time between consecutive SVIs.

  *Mean packets between SVIs is the arithmetic mean of the number of
   SLO-compliant packets between consecutive SVIs. (Another
   variation of "MTBF" in a service setting.)

Determining the condition in which the path is currently with
respect to availability/unavailability is helpful. But because
switching between periods requires ten consecutive intervals,
shorter conditions may not be adequately reflected. Two additional
PAMs can be used, and they are defined as follows:

  *violated interval ratio (VIR) is the ratio of VI to the total
   number of time unit intervals in a time of the availability
   periods during a fixed measurement interval.

*severely violated interval ratio (SVIR) - is the ratio of SVIs to
 the total number of time unit intervals in a time of the
 availability periods during a fixed measurement interval.

3.3. Service Availability in PAMs

VI, SVI, and VFI characterize the communication between two nodes
relative to the level of required and acceptable performance and
when the performance level degrades below an acceptable level. The
former condition in this document defined to as service
availability. The latter is defined as service unavailability. Based
on the definitions in [Section 3.1](#), SVI is the one time interval of
service unavailability while VI and VFI present an interval of
service availability. Since the conditions of the service are are
continually changing, periods of availability and unavailability
need to be defined with duration larger than one time interval to
reduce the number of state changes while correctly reflecting the
service condition. The method to determine the state of the service
in terms of PAM is described below:

  *If ten consecutive SVIs been detected, then the PAM state of the
   service is defined as unavailability, and the beginning of that
   period of unavailability state is at the start of the first SVI
   in the sequence of the consecutive SVIs.

  *Similarly, for ten consecutive non-SVIs (i.e., either VIs or
   VFIs), the service is defined to be available. The start of that
   period is at the beginning of the first non-SVI.

  *Resulting from these two definitions, a sequence of less than ten
   consecutive SVIs or non-SVIs does not change the PAM state of the
   service. For example, if the PAM state is determined as
   unavailable, a sequence of seven VFI s is not viewed as an
   availability period.

4. Statistical SLO

It should be noted that certain Service Level Agreements (SLA) may
be statistical, requiring the service levels of packets in a flow to
adhere to specific distributions. For example, an SLA might state
that any given SLO applies only to a certain percentage of packets,
allowing for a certain level of, for example, packet loss and/or
exceeding packet delay threshold to take place. Each such event, in
that case, does not necessarily constitute an SLO violation.
However, it is still useful to maintain those statistics, as the
number of out-of-SLO packets still matters when looked at in
proportion to the total number of packets.

Along that vein, an SLA might establish an SLO of, say, end-to-end
latency to not exceed 20 ms for 99% of packets, to not exceed 25ms

for 99.999% of packets, and to never exceed 30ms for any packet. In
that case, any individual packet with latency larger than 20 ms
latency and lower than 30 ms cannot be considered an SLO violation
in itself, but compliance with the SLO may need to be assessed after
the fact.

To support statistical services more directly requires additional
metrics, such as metrics that represent histograms for service level
parameters with buckets corresponding to individual service level
objectives. For the example just given, a histogram for a given flow
could be maintained with three buckets: one containing the count of
packets within 20ms, a second with a count of packets between 20 and
25ms (or simply all within 25ms), a third with a count of packets
between 25 and 30ms (or merely all packets within 30ms, and a fourth
with a count of anything beyond (or simply a total count). Of
course, the number of buckets and the boundaries between those
buckets should correspond to the needs of the SLA associated with
the application, i.e., to the specific guarantees and SLOs that were
provided. The definition of histogram metrics is for further study.

5. Other PAM Benefits

PAM provides a number of benefits with other, more conventional
performance metrics. Without PAM, it would be possible to conduct
ongoing measurements of service levels and maintain a time-series of
service level records, then assess compliance with specific SLOs
after the fact. However, doing so would require the collection of
vast amounts of data that would need to be generated, exported,
transmitted, collected, and stored. In addition, extensive
postprocessing would be required to compare that data against SLOs
and analyze its compliance. Being able to perform these tasks at
scale and in real-time would present significant additional
challenges.

Adding PAM allows for a more compact expression of service level
compliance. In that sense, PAM does not simply represent raw data
but expresses actionable information. In conjunction with proper
instrumentation, PAM can thus help avoid expensive postprocessing.

6. Discussion Items

The following items require further discussion:

  *Metrics. The foundational metrics defined in this draft refer to
   violated intervals. In addition, counts of violations related to
   individual packets may also need to be maintained. Metrics
   referring to violated packets (i.e., packets that on an
   individual basis miss a performance objective) may be added in a
   later revision of this document.

The following is a list of items for which further discussion is
needed as to whether they should be included in the scope of this
specification:

   *A YANG data model.

   *A set of IPFIX Information Elements.

   *Statistical metrics: e.g., histograms/buckets.

   *Policies regarding the definition of "violated" and "severely
    violated" time interval.

   *Additional second-order metrics, such as "longest disruption of
    service time" (measuring consecutive time units with SVIs).

7. IANA Considerations

   This document has no IANA actions.

8. Security Considerations

   Instrumentation for metrics that are used to assess compliance with
   SLOs constitute an attractive target for an attacker. By interfering
   with the maintaining of such metrics, services could be falsely
   identified as complying (when they are not) or vice-versa (i.e.,
   flagged as being non-compliant when indeed they are). While this
   document does not specify how networks should be instrumented to
   maintain the identified metrics, such instrumentation needs to be
   adequately secured to ensure accurate measurements and prohibit
   tampering with metrics being kept.

   Where metrics are being defined relative to an SLO, the
   configuration of those SLOs needs to be adequately secured.
   Likewise, where SLOs can be adjusted, the correlation between any
   metrics instance and a particular SLO must be clear. The same
   service levels that constitute SLO violations for one flow that
   should be maintained as part of the "violated time units" and
   related metrics, may be perfectly compliant for another flow. In
   cases when it is impossible to tie together SLOs and PAM properly,
   it will be preferable to merely maintain statistics about service
   levels delivered (for example, overall histograms of end-to-end
   latency) without assessing which constitutes violations.

   By the same token, where the definition of what constitutes a
   "severe" or a "significant" violation depends on policy or context.
   The configuration of such policy or context needs to be specially
   secured. Also, the configuration of this policy must be bound to the
   metrics being maintained. This way, it will be clear which policy
   was in effect when those metrics were being assessed. An attacker

that can tamper with such policies will render the corresponding metrics useless (in the best case) or misleading (in the worst case).

9. Acknowledgments

TBA

10. References

10.1. Informative References

[I-D.ietf-teas-ietf-network-slices]
          Farrel, A., Drake, J., Rokui, R., Homma, S., Makhijani, K., Contreras, L. M., and J. Tantsura, "Framework for IETF Network Slices", Work in Progress, Internet-Draft, draft-ietf-teas-ietf-network-slices-10, 27 March 2022, <https://datatracker.ietf.org/doc/html/draft-ietf-teas-ietf-network-slices-10>.

[ITU.G.826] ITU-T, "End-to-end error performance parameters and objectives for international, constant bit-rate digital paths and connections", ITU-T G.826, December 2002.

[RFC2863] McCloghrie, K. and F. Kastenholz, "The Interfaces Group MIB", RFC 2863, DOI 10.17487/RFC2863, June 2000, <https://www.rfc-editor.org/info/rfc2863>.

[RFC7011] Claise, B., Ed., Trammell, B., Ed., and P. Aitken, "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information", STD 77, RFC 7011, DOI 10.17487/RFC7011, September 2013, <https://www.rfc-editor.org/info/rfc7011>.

[RFC7012] Claise, B., Ed. and B. Trammell, Ed., "Information Model for IP Flow Information Export (IPFIX)", RFC 7012, DOI 10.17487/RFC7012, September 2013, <https://www.rfc-editor.org/info/rfc7012>.

[RFC7297] Boucadair, M., Jacquenet, C., and N. Wang, "IP Connectivity Provisioning Profile (CPP)", RFC 7297, DOI 10.17487/RFC7297, July 2014, <https://www.rfc-editor.org/info/rfc7297>.

[RFC8343] Bjorklund, M., "A YANG Data Model for Interface Management", RFC 8343, DOI 10.17487/RFC8343, March 2018, <https://www.rfc-editor.org/info/rfc8343>.

Contributors' Addresses

   Liuyan Han
   China Mobile
   32 XuanWuMenXi Street
   Beijing
   100053
   China

   Email: hanliuyan@chinamobile.com

   Mohamed Boucadair
   Orange
   35000 Rennes
   France

   Email: mohamed.boucadair@orange.com

Authors' Addresses

   Greg Mirsky
   Ericsson

   Email: gregimirsky@gmail.com

   Joel Halpern
   Ericsson

   Email: joel.halpern@ericsson.com

   Xiao Min
   ZTE Corp.

   Email: xiao.min2@zte.com.cn

   Alexander Clemm
   Futurewei
   2330 Central Expressway
   Santa Clara, CA 95050
   United States of America

   Email: ludwig@clemm.org

   John Strassner
   Futurewei
   2330 Central Expressway
   Santa Clara, CA 95050
   United States of America

   Email: strazpdj@gmail.com

Jerome Francois
Inria
615 Rue du Jardin Botanique
54600 Villers-les-Nancy
France

Email: jerome.francois@inria.fr