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

draft-mhmcsfh-ippm-pam

Greg Mirsky
Joel Halpern
Xiao Min
Liuyan Han

Alexander Clemm
John Strassner
Jérôme François

IETF-113, March 2022

Something new, something old

draft-mirsky-ippm-epm

+ draft-csfx-ippm-hipmetrics

---

draft-mhmcsfh-ippm-pam

What is PAM?

- Precision Availability Metrics express the availability of a service in accordance with the performance requirements reflected in a contract and expressed using SLOs
  - Example: a service with the requirement for not-to-exceed end-to-end latency
- Performance requirements for various networking services can be expressed through a combination of Service Level Objectives (SLO). An SLO usually sets a threshold of one measurable metric that a service provider accepts as part of a service contract.
- Precision Availability Metrics (PAM) can be used:
  - To determine the degree of compliance with which service levels are delivered relative to pre-defined SLOs.
  - To provide service according to its SLO as part of accounting records; to account for the actual quality with which services were delivered and whether or not any SLO violations had occurred.
  - To continuously monitor the quality with which the service is delivered.

Elements of PAM

- A PAM time unit, a.k.a. PAM interval, can be characterized as:
  - Errored Interval (EI) – an interval during which at least one of service level degraded below the pre-defined threshold
  - Error-Free Interval (EFI) – all performance parameters are at or above their respective pre-defined optimal levels, and no defects have been detected
- Time interval: e.g., 1 second, or 1 msec
- Extensions possible, e.g., to differentiate “slight” and “severe” violations
  - Severely Errored Interval (SEI) – at least one of performance parameters degraded below the pre-defined critical threshold or a defect was detected
- Based on these definitions, a set of basic metrics that count respective intervals is defined:
  - EI count, EFI count, and SEI count
- Violated packets can also be counted, but intervals are often more meaningful
  - Violations can occur in bursts: e.g., temporary overload conditions, route reconvergence
  - Differentiate “on rare occasions, sucks a lot” vs. “frequently, sucks just a little”
  - Compare Errored Seconds for transmissions

Derived PAM Metrics

- Based on basic PAM metrics a set of derived metrics is introduced for an EI:
  - Time since the last EI
  - Mean time between EIs
  - # Packets since the last EI
  - Mean # packets between EIs
- Analogous metrics introduced for SEI:
  - Time since the last SEI
  - Mean time between SEIs
  - # Packets since the last SEI
  - Mean # packets between SEIs

PAM extensions

- Account for lengthy disruptions, e.g.
  - Define significant duration threshold, e.g. ,10
  - Extended unavailability metrics measure occurrence of consecutive EIs/SEIs beyond that threshold
- Complement with state model: service is deemed unavailable when the most recent intervals were all violated (or severely violated)
  - E.g., 10 consecutive SEIs constitute service unavailability state that begins at the start of the first SEI
  - E.g., 10 consecutive non-SEIs constitute service availability state that begins at the start of the first non-SEI
- Complement with additional derived metrics:
  - EI ratio – ratio of EIs to the total number of PAM intervals
  - SEI ratio – ratio of SEIs to the total number of PAM intervals

Discussion items

- Terminology: “Errored” vs. “Violated”. Is a singleton of non-conformance to an SLO an error or violation of a contract?
- Metrics: individual packets that breach SLO(s)?

Future work (beyond this draft)

- YANG data model
- IPFIX Informational Elements
- Support for statistical SLOs, e.g., histogram and/or bucket
- Policies to define error/violated time unit, configure metrics
- Additional second-order metrics, e.g., “longest disruption of service time”

Next steps

- Welcome comments, questions
- Contributions, cooperation are most appreciated
- WG adoption?

Thank you