CATS Metrics Definition

draft-ietf-cats-metric-definition-02

Kehan Yao, China Mobile
Hang Shi, Huawei Technologies
Cheng Li, Huawei Technologies
L. M. Contreras, Telefonica
Jordi Ros-Giralt, Qualcomm Europe, Inc

Updates Since the Interim Meeting

- Focused on addressing comments from Adrian and Joel.
- Polished Abstract and Introduction (added more references, clarified the scope of CATS metrics).
- Updated Section 3 CATS metrics definition (updated the classifications of L1 metrics, updated Figure 1).
- Updated Section 4, CATS metrics representation (clearly defined CATS metric fields in section 4.1, modified the examples of three metric levels and the basic encoding format of the example metric)
- Added a new section, section 6 “Implementation Guidance on Using CATS Metrics” to propose some options for service providers and vendors.

The Scope of CATS Metrics

- About the scope of CATS metrics:

"This document defines a set of metrics from the computing domain used for CATS."

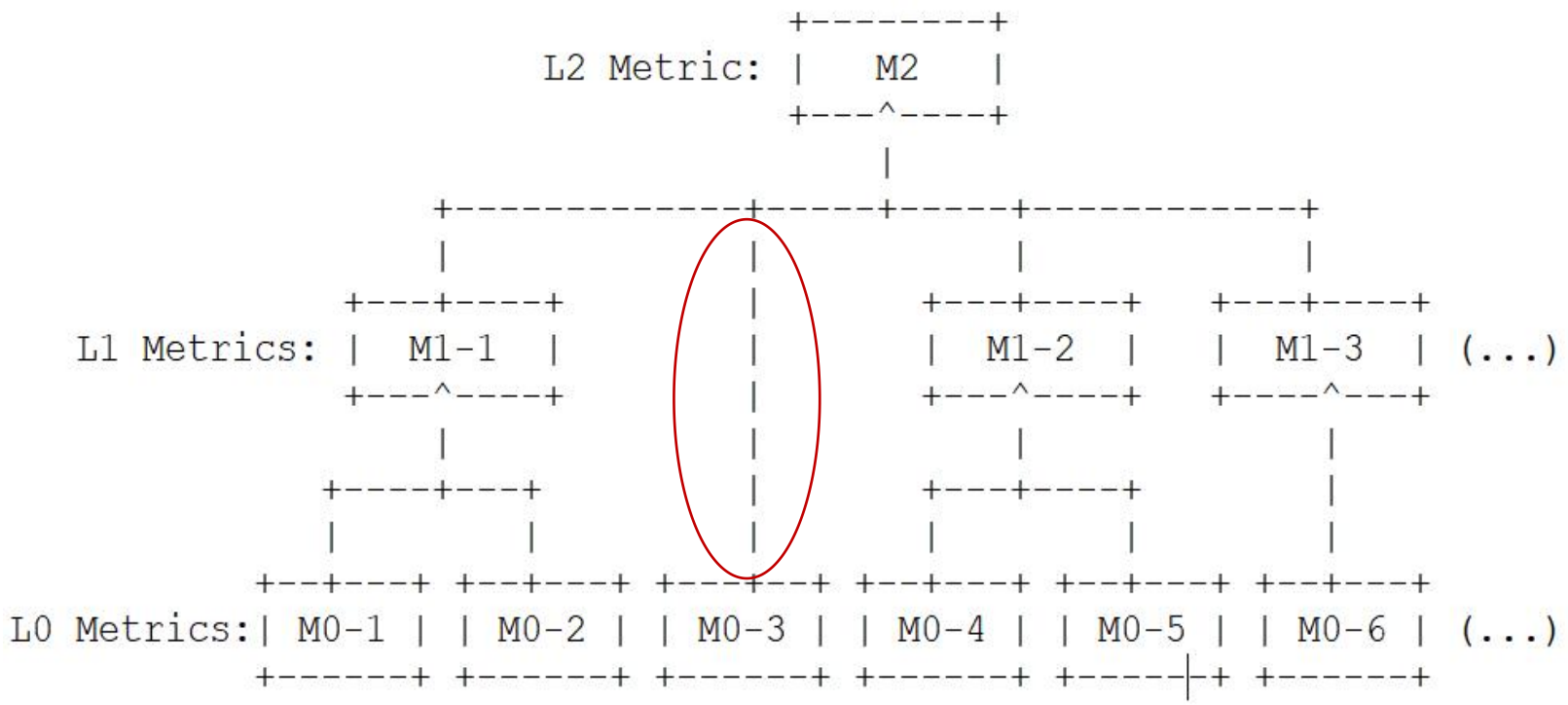
- **"computing domain"** is a term analogous to the network domain. It means that CATS metrics not only covers compute metrics, but also other types of metrics (including communication metrics) generated from the computing domain.
- **"a set of "** means not all metrics will be covered. In particular, focus is on metrics that are useful for traffic steering.

CATS Metrics Definition

- L0 metrics: clarify its position that "*L0 metrics serve as foundational data and do not require classification.* "
- L1 metrics: Re-organized them into three classes:
  - **Computing:** A normalized value derived from computing-related L0 metrics, such as CPU, GPU, and NPU metrics. Added storage metrics inside this class, since storage metrics have different levels, some can be seen as part of the compute component, such as memory or cache.
  - **Communication:** A normalized value derived from communication-related L0 metrics. Replaced the class "networking"
  - **Composed:** A normalized value derived from an aggregation function gathering compute and communication metrics. For example, delay is the sum of all delays along the path. Merged "Delay" into this new class.

CATS Metrics Definition -- Cont'd

- Modified figure 1. L0 metrics can be directly connected with L2 metrics, which means L2 metrics can be either normalized by L1 metrics only, L0 metrics only, or both L1 and L0 metrics.



Representation of CATS Metrics

- CATS metrics fields, currently including 9 types of fields:
  - **Metric_Type** (type), **Format standard** (format_std, optional), **Format** (format), **Length** (length), **Unit** (unit), **Source** (source, optional), **Statistics** (statistics, optional), **Level** (level), **Value** (value).
  - Format standard (optional), only formal standard metrics, like ieee_754 and ascii.
  - Format, this document only covers **unsigned integer** and **floating point**.
  - Source (optional), the origin of the information, including 5 subtypes.
    - '**nominal**', '**estimation**', '**directly measured**', '**normalization**', '**aggregation**'
  - Statistics (optional), provides additional details when there are some pre-computation performed on the metrics, including 4 subtypes.
    - '**max**', '**min**', '**mean**', '**cur**'.
    - 'cur' stands for current value of the metrics collected.

Representation of CATS Metrics -- cont'd

- Examples of L0 and L1 metrics.
- The representation of L2 metric is similar to L1 metrics, with the only differences in "Metric Type" and "Level"
- Need further discussion on the encoding, are the bits defined for each type appropriate?

```
Basic fields:
      Metric Type: "compute type_CPU_frequency"
      Level: L0
      Format: unsigned integer, floating point
      Unit: GHZ
      Length: four octets
      Value: 2.2
Source:
      nominal

|Metric Type|Level|Format| Unit|Length| Value|Source|
    8bits    2bits  1bit  4bits  3bits 32bits  3bits
```

L0 Metric, CPU frequency

```
Basic fields:
      Metric type: "communication_norm"
      Level: L1
      Format: unsigned integer
      Length: one octet
      Value: 1
Source:
      normalization

|Metric Type|Level|Format|Length|Value|Source|
    8bits    2bits  1bit   3bits 8bits  3bits
```

L1 Metric, communication type

Implementation Guidance for Using CATS Metrics

- To propose some options for service providers and vendors.
- To better align with the CATS framework.
- **Key question: How to define normalization and aggregation functions and where to put them?**
  - Option 1: **CATS Service Contact Instance**.
    - pros: don't need to worry about resources and performance.
    - cons: only local knowledge
  - Option 2: **CATS Service Metric Agent(C-SMA)**.
    - pros: can synthesize CATS metrics as well as metrics from network domain.
    - cons: limited resource in C-SMA, depending on how it's located with CATS-forwarders
  - Option 3: **Both but split functions**. This document suggest that the aggregation functions be placed at C-SMA while normalization functions as well as aggregation functions can be both placed at service contact instances.
  - Other Options?

Next Steps

- Polish and solidify CATS metrics representation and encoding.
- Solicit for comments on implementation options and agree on at least one option.
- Authors will set the **submission to IESG at March 2026**, but also consider the progress of CATS framework. The potential connection between these documents lies in the necessity of defining a logical component to serve as the metric aggregator within the CATS framework.

THANKS!

Supplementary Considerations on CATS Metrics

Default Policy and Related Metrics

draft-du-cats-default-policy-and-metrics-00, Zongpeng

Introduction of draft: Default Policy and Related Metrics

- A new draft is proposed instead of the old one
  - **New:** Default Policy and Related Metrics in CATS, draft-du-cats-default-policy-and-metrics-00
  - **Old:** Compute Modeling and Metrics, draft-du-cats-computing-modeling-description-03
- It is suggested that some work in the draft can be considered in the working group draft

- Motivation: the scenarios of task allocation and scheduling in cloud computing is similar to the traffic scheduling in CATS
  - In task allocation strategies of cloud computing systems, the main optimization objectives considered are **task execution time**, execution cost, and **load balancing;** Execution time and execution cost are used to evaluate the user satisfaction, and load balancing is used to measure the reliability and availability of the system
  - In CATS, we also can consider similar metrics in the traffic steering, and the proposed ones are the **computing delay**, and the **service capability** (or called the load balancing value, **LBValue**)

Introduction of draft: Default Policy and Related Metrics

- The Ingress, which is the traffic steering point in CATS, it have two methods of traffic steering for CATS traffic with an anycast IP as the Destination IP Address
  - Method one: One primary path + one or several backup path
  - Method two: Several candidate paths with different weights
- If the method one is the default forwarding policy of the CATS service, we can use the **computing delay** as the related metrics, which can achieve a minimum total delay
- If the method two is the default forwarding policy of the CATS service, we can use the **service capability** as the related metrics, which can achieve a better LB result

Method one:

path1: primary path (active)
path2: backup path (inactive)

Method Two:

path1: active (weight 1)
path2: active (weight 3)