CATS Metrics

yaokehan@chinamobile.com, shihang9@huawei.com, c.l@huawei.com, sabine.randriamasy@nokia-bell-labs.com, luismiguel.contrerasmurillo@telefonica.com, jros@qti.qualcomm.com, Roland.Schott@telekom.de

IETF 121, Dublin, November 2024

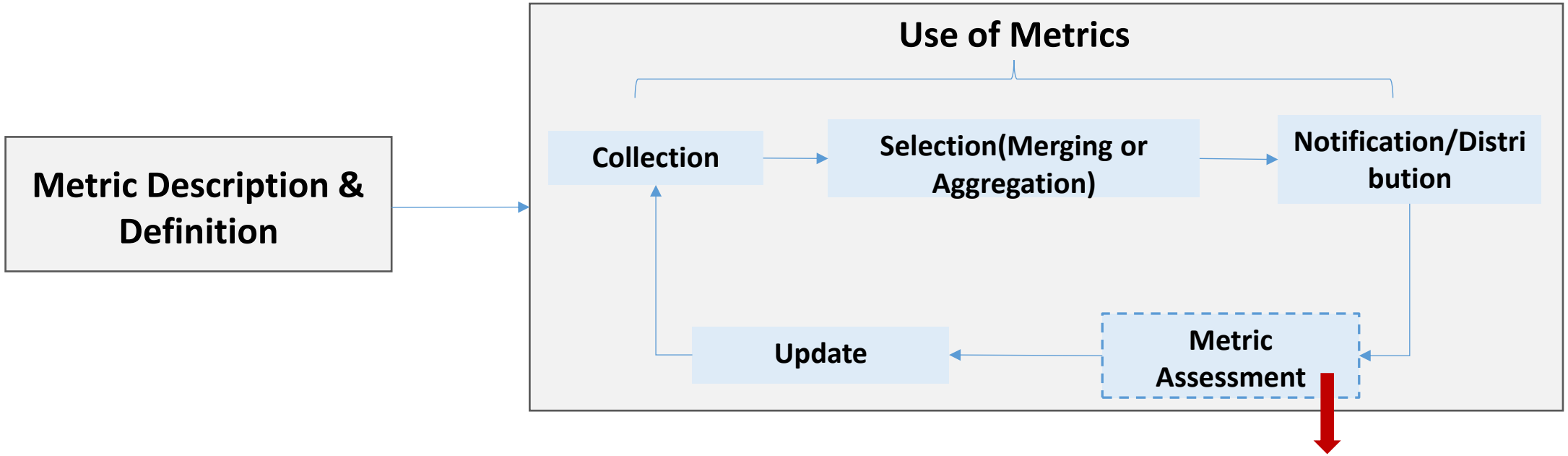
Why we need to define CATS Metrics

- All five use cases need network & compute metrics for path and instance selection.
- But they cover different metrics categories.
- Need an agreement on metrics definition to guide vendor implementations.

Use case	Metrics needed in each use cases
AR/VR	Delay, compute metrics (memory IO, GPU, etc.)
Intelligent Transportation(Internet of Vehicles)	Delay, compute metrics (memory IO, GPU, etc.)
Digital Twin	Delay, storage, compute metrics
SD-WAN	Delay, compute metrics (virtual CPU, etc.)
AI Large Model Inference	Delay, network (bandwidth, etc.), compute metrics (memory IO, memory size, GPU, etc.)

The Basis for CATS Metrics definition: Requirements on Metrics

- Understand how the metrics will be used before defining them



This module belongs to service assurance, it might not be a "must" choice

- **Metric collection:** Raw metrics or normalized metrics?
- **Metric selection:** Aggregated metrics or all metrics?
- **Metric distribution:** How will the metrics be distributed? Will they incur much overhead?
- **Metric assessment:** If the selected metrics are not useful, how to tune and update?
- **Metric update:** Can the metrics be updated? If so, what's the duration for metrics to be updated?

How do We Define CATS Metrics? ALTO RFC 9439 as an Example

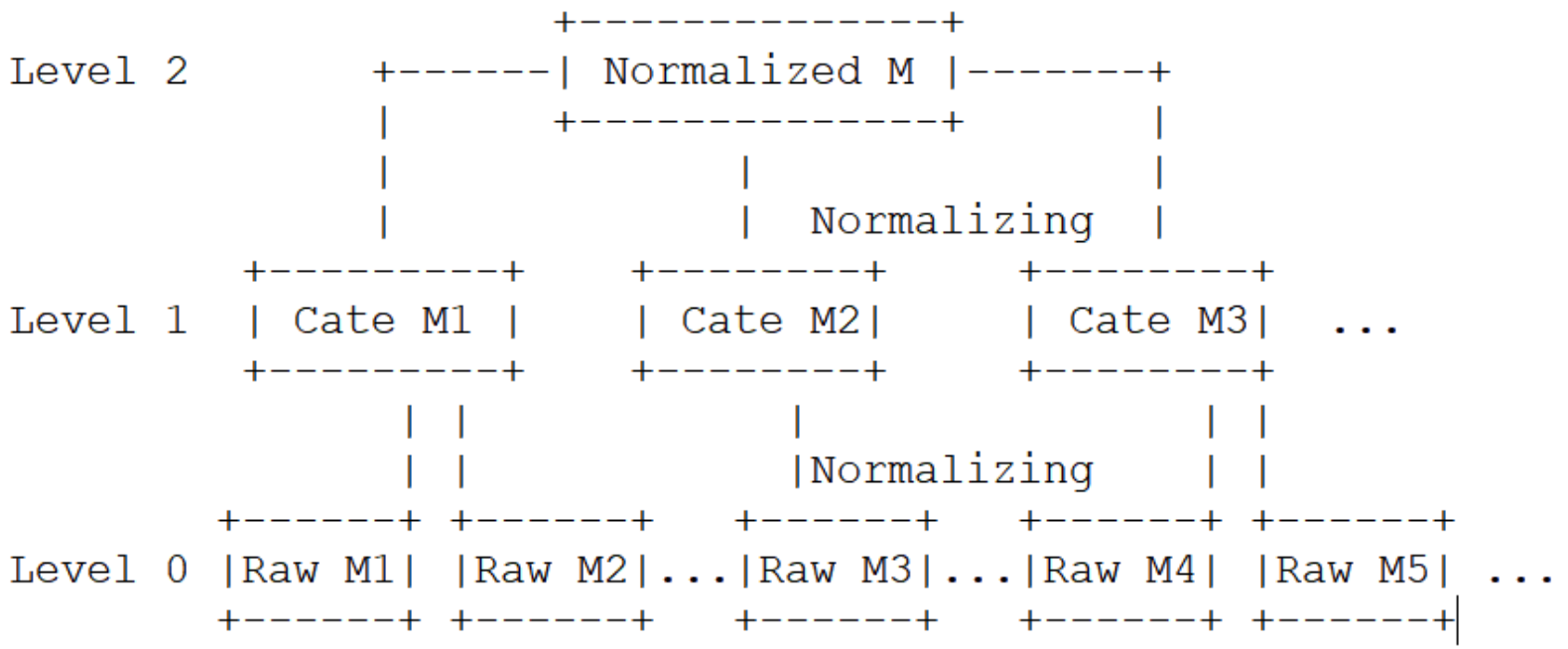
- Defining network performance metrics (delay, jitter, packet loss, etc.) for network capability exposure.
- For each metric, it defines the following fields:
  - Base identifier
  - Value Representation
  - Semantics and Use
  - Cost-context (source of the metrics and statistics)
- Source of metrics:
  - nominal: the metric is statically configured by the underlying devices
  - sla: metric is derived from some commitment
  - estimation: metric is computed through an estimation process.
- Metric statistics:
  - min, max, median, mean, stddev, stdvar, etc.

CATS Metrics Description and Definition

- Basic fields for CATS metrics:
  - **Metric type:** describe what the metric is and how network devices can recognize it**.**
  - **Unit:** for quantitatively counting the metric
  - **Format:** for accurately describing a metric.
  - **Bits occupation:** number of bits needed to carry the metric in a packet.
  - **Score**: a unitless value representing the score of the metric (for L1 and L2 metrics).
- Source of CATS metrics: nominal, sla, estimation:
  - further divide **estimation** into three sub-types:
    - **Directly measured metrics:** have physical meanings and units without any processing.
    - **Aggregated metrics:** can be either physically meaningful or not, and they maintain their meanings compared to the directly measured metrics
    - **Normalized metrics:** can have physical meanings or not, but they don’t have units.
- Metric statistics:
  - min, max, median, mean, stddev, stdvar, etc.

3-level CATS Metric Definition

- **Level 0. Raw metrics:** CPU/GPU Frequency, Memory BW, etc.
- **Level 1. Normalized metrics in Categories:** Networking, computing, storage, or delay.
- **Level 2. One fully normalized value:** A single score.



Hierarchical view of 3-level CATS metrics (to be updated according to groups comments)

CATS Metric Representation

Level 0 CATS metric representation:

Note: There are many Level 0 raw metrics defined in other SDOs, (e.g., DMTF). This document doesn't want to define detailed level 0 metrics, but rather show examples and how they can be used for derivation of level 1 and level 2 CATS metrics.

Compute raw metrics example:

{
Basic fields:
Metric type: "compute type_CPU"
Format: integer, FP8
Bits occupation: 4 octets
Score: integer from 1 to 5
Special fields:
Frequency unit: GHZ
Compute capabilities unit: FLOPs
Source:
Direct measurement
Statistics:
mean
}

Network raw metrics example:

{
Basic fields:
Metric type: "network type_Bandwidth"
Format: integer
Unit: Gb/s
Bits occupation: 2 octets
Score: integer from 1 to 5
Source:
nominal
Statistics:
cur
}

Delay raw metrics example:

{
Basic fields:
Metric type: "delay_raw"
Format: integer, FP8
Unit: Microsecond (us)
Bits occupation: 4 octets
Score: integer from 1 to 5
Source:
Aggregation
Statistics:
max
}

CATS Metric Representation

- **Level 1 CATS metric representation:**
  - Normalized compute metric: denoted as **“compute_norm”**, integer, no unit, one octet, score
  - Normalized network metric: denoted as **“network_norm”**, integer, no unit, one octet, score
  - Normalized storage metric: denoted as **“storage_norm”**, integer, no unit, one octet, score
  - Delay: denoted as **“delay_norm”**, integer, no unit, one octet , score
  - Source of L1 metrics: normalization, aggregation
- **Level 2 CATS metric representation:**
  - A single value without any physical meaning or unit
  - Each provider may have its own methods to derive it, but all providers MUST follow the definition:

{
Basic fields:
Metric_type: “Norm_fi”
Format: non-negative interger
Unit: null
Bits occupation: one octet
Score: integer
Source:
Normalization
}

CATS Level Metrics Comparison

Level	Encoding Complexity	Extensibility	Stability	Accuracy
Level 0	Complicated	Bad	Bad	Good
Level 1	Medium	Medium	Medium	Medium
Level 2	Simple	Good	Good	Medium

- Make comparisons by following the design principles.
- Intuitively, L2 metrics are recommended because of its simplicity, extensibility and stability.
- Its accuracy can also be improved by using proper decision making algorithms.

Recap on Current Status for Metrics Work in IETF CATS

I-Drafts:

- Draft 1: CATS Metric Description and Definition. https://datatracker.ietf.org/doc/draft-ysl-cats-metric-definition/
- Draft 2: Joint Exposure of Network and Compute Information for Infrastructure-Aware Service Deployment. https://datatracker.ietf.org/doc/draft-rcr-opsawg-operational-compute-metrics/

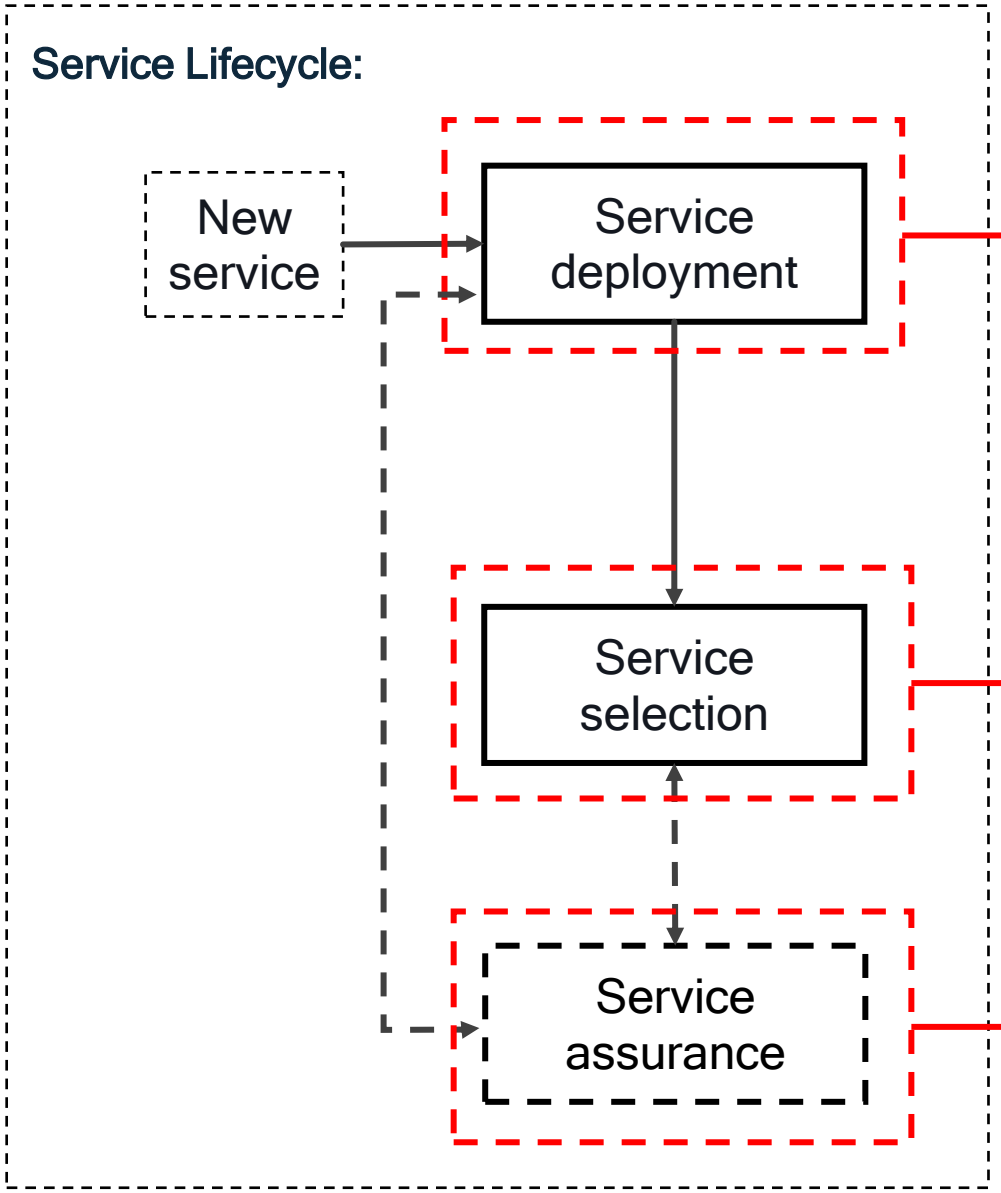
CATS Milestone:

- Mar 2025: Adopt document describing CATS metrics
- Mar 2026: Submit document describing CATS metrics to the IESG for publication as Informational

• General Approach:

- Unifying work towards delivering the March 2025 milestone. (Non-spoiler: Concrete proposal in a few slides.)

General Problem Space: Service Lifecycle and Information Exposure



Action to take	Information needed	Who needs it
(1) Service placement	Compute and communication	Service provider
(2.a) Service selection: compute node selection	Compute and communication	Network provider, service provider or application
(2.b) Service selection: path selection	Communication	Network provider or application
(3) Service assurance	Compute and communication	Network provider, service provider or application

Table 1: Problem space, needs, and stakeholders.

CATS problem space

Problem Space: Metric Definition and Exposure Mechanism

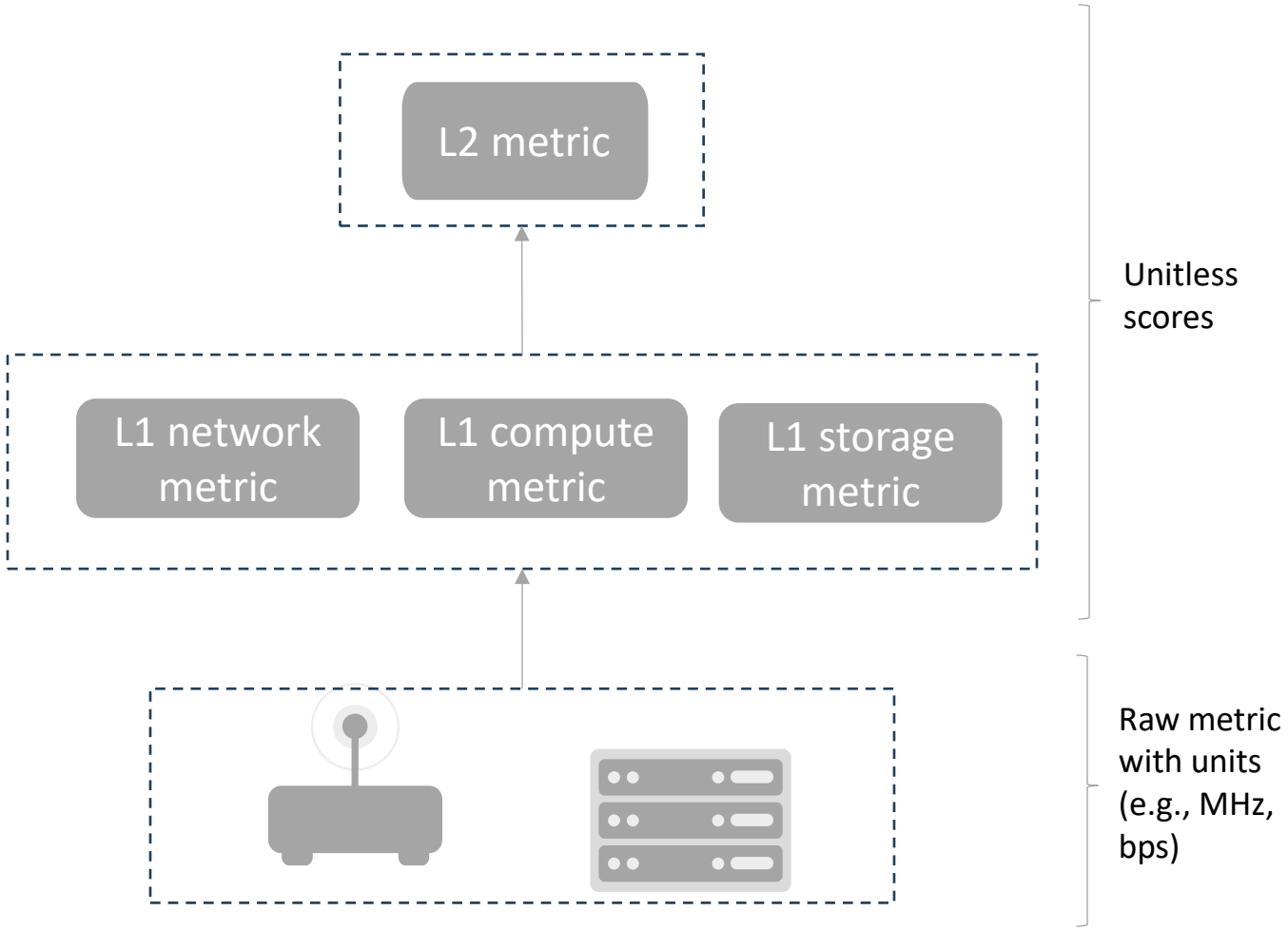
Two main problems to work on:

(1) Definition of the metrics

(2) Definition of the protocol interface to expose the metrics to the consumer

Definition of the Metrics: Summary of Approach

CATS Metric Model: A 3-level framework to meet the trade-off interoperability vs scalability vs usefulness.



Analogy: Works similarly to the University Grade Point Average system. Every university abstracts out a single score for each student that is specific to its country (e.g., country A score goes from 1 to 10, country B score goes from 1 to 5). When a student travels to another country, the score can be easily translated to that country's metric. This allows for each country to independently implement their own metrics without global coordination, while achieving global interoperability.

Definition of the Interface to Expose the Metrics

- I-D.ldbc-cats-framework presents three CATS models: distributed, centralized and hybrid. Their corresponding distribution mechanism are:
  - Distributed: Directly distributed to the network devices.
  - Centralized: Collected by a centralized control plane.
  - Hybrid: Some directly, some centralized.
- Optimal choice depends on dynamicity: higher-frequency metric updates tend to favor a centralized collection approach, and vice versa.
- For decentralized approach, draft-ll-idr-cats-bgp-extension and draft-ietf-idr-5g-edge-service-metadata propose using BGP.
- For centralized approach, a potential candidate solution is to leverage ALTO (e.g., RFC7285, RFC9240)

Proposed Roadmap

- **Overall document structure. Three main documents:**
  - A document focusing on metrics definition (Deadline / Milestone March 2025 per CATS charter)
  - Documents focusing on metrics exposure depending on consumer: centralized vs decentralized, on-path vs off-path (is there a need to define a milestone?)
  - An informational document describing the overall global picture for compute metrics, focusing on: (1) the broader use cases (deployment, selection, assurance) and (2) positioning the current CATS use case (selection) within the overall global picture. (This doc could be done within OPSAWG).
- **Technical approach:**
  - Metrics definition document to be based on draft-ysl-cats-metric-definition.
    - Approach: metric levels abstraction
  - Metrics exposure documents:
    - Approach: on-path vs off-path / YANG Model.
  - Informational document to be based on draft-rcr-opsawg-operational-compute-metrics.
    - Approach: service life cycle (deployment, selection, assurance)