Computing-Aware Traffic Steering
Internet-Draft
Intended status: Informational
Expires: 26 October 2026

M. Zhu
China mobile

24 April 2026

Operational Semantics for CATS Metric Consumption
draft-zhu-cats-metric-semantics-00

Abstract

The CATS framework introduces computing-related information into traffic steering decisions. Existing work defines how such metrics are represented, distributed, and used within the CATS architecture. However, it does not fully address whether a metric remains suitable for use at the point of consumption.

This document introduces a set of operational semantics for CATS metrics, including Freshness, Operational acceptability, and Assurance exposure. These semantics describe whether a metric remains temporally aligned with the underlying condition, whether it remains suitable for operational use in steering, and whether degraded consumption is externally visible to management or OAM functions.

The document further explains how these semantics apply across centralized, distributed, and hybrid deployments, including cases where different metric sources contribute under different conditions. The goal is to provide a consistent basis for interpreting metric usability in CATS without introducing a new metric level or prescribing a single derivation method.

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 26 October 2026.

Table of Contents

1. Introduction

Computing-Aware Traffic Steering (CATS) extends traffic steering beyond traditional network reachability and path selection by incorporating computing-related inputs into forwarding and service-selection decisions. This change is not merely an incremental extension of traditional routing inputs. Many computing-related metrics vary more quickly, are aggregated and distributed through

more diverse paths, and lose operational meaning more rapidly. As a result, the difficulty in CATS is not only how to define more metrics, but also how to determine whether a received metric remains suitable for operational consumption as a steering input.

Existing CATS work explains how metrics are represented, distributed, and used [CATS-FRAMEWORK] [CATS-METRIC-DEFINITION]. Related requirements and OAM work identify update, stability, service-continuity, consistency, and black-holing concerns [CATS-REQUIREMENTS] [CATS-OAM]. However, such metrics cannot always be directly consumed by existing steering or routing protocols. Many computing-related metrics evolve at timescales that are shorter than those assumed by traditional control-plane mechanisms. Excessively frequent metric updates may introduce instability or oscillation into the steering process. Infrequent updates, by contrast, may cause decisions to rely on stale conditions that no longer reflect the current operational state.

This document addresses a related issue at the point of consumption: whether a metric remains operationally suitable when it is consumed for steering. A metric may remain visible and well-formed while no longer remaining suitable for normal steering use. This problem is more likely to arise when computing-related information changes quickly, is collected and redistributed before use, or is consumed under different deployment conditions.

In conventional routing, slightly outdated cost information often leads only to a suboptimal path. In CATS, a decision may rely on utilization, admission headroom, or service-state information that no longer reflects the current operational condition. In centralized deployments, this may result from control-loop delay. In distributed deployments, it may result from divergence across local observations. In hybrid deployments, it may result from the joint use of inputs that do not share the same temporal behavior or operational conditions. The result may be admission rejection, degraded service continuity, or steering behavior resembling black-holing.

This document defines an orthogonal set of operational semantics that can be associated with any CATS metric, regardless of abstraction level. These semantics are intended to express whether a metric remains sufficiently fresh, whether it remains operationally acceptable for steering use, and whether degraded consumption becomes externally visible to OAM or management functions.

2. Scope and positioning

The semantics are intended to complement the metric abstraction model. Metric abstraction explains how raw measurements are normalized or combined into higher-level indicators. This document addresses a different dimension: the operational condition of a metric at the point of consumption. More specifically, it defines three orthogonal semantics, i.e., Freshness, Operational acceptability, and Assurance exposure, to describe whether a metric remains temporally suitable for use, whether it remains acceptable for operational consumption in steering, and whether degraded consumption or fallback become externally visible.

These semantics are not tied to any single deployment model. They can be applied across existing abstraction levels and across centralized, distributed, and hybrid operation. This document also does not define a new transport, encoding, or control-plane protocol. Instead, it defines semantic information that may later be carried, derived, or exposed by future protocol elements, data models, management objects, or OAM procedures. A steering consumer may use these semantics to determine whether a metric can still participate in normal steering logic. A control, management, or OAM function may use them to distinguish normal consumption from degraded consumption, fallback behavior, or source-specific semantic degradation.

3. Terminology

Metric-consuming decision point: A functional point at which CATS metrics are consumed to derive or support steering, path-selection, or service-selection decisions. Depending on the deployment model, this function may be realized by a centralized CATS Path Selector (C-PS), by an Ingress CATS-Forwarder with embedded decision logic, or by a combination of both in hybrid deployments.

Freshness: The extent to which a metric remains temporally suitable for its intended operational use.

Operational acceptability: The extent to which a metric remains suitable for operational consumption at the current time.

Assurance exposure: The extent to which degraded metric consumption, inconsistency, or fallback behavior is visible to OAM or management systems.

4. Operational gap

The gap addressed in this document is the lack of an explicit description of metric usability at the point of consumption. A metric may remain visible and well-formed while no longer remaining suitable for normal steering use.

This missing layer appears in three ways. First, a metric may lose temporal alignment with the condition it is intended to describe while still remaining available to the consumer. For example, a controller-based deployment may continue to distribute a site-level utilization metric whose indicated admission headroom no longer reflects the current service state.

Second, a metric may remain present and syntactically valid while no longer remaining suitable for normal operational consumption in steering. For example, repeated delay, poor update continuity, or inconsistency with other observations may make a metric unsuitable for fine-grained steering even though it is still retained for reduced-trust or fallback use.

Third, degraded metric consumption may remain invisible to management or OAM even after steering shifted into fallback or reduced-trust behavior. In such a case, the problem is not only metric degradation itself, but also the lack of external visibility into the semantic condition under which steering is proceeding.

These gaps are amplified by deployment conditions. In centralized operation, semantic degradation may be introduced within the control loop before the metric is used. In distributed operation, different decision points may rely on different local versions of what is nominally the same condition. In hybrid operation, the problem is further complicated by the joint use of metric inputs that do not share the same temporal behavior or consumption assumptions.

5. Operational semantics in different deployment modes

This document introduces three operational semantics for CATS metrics: Freshness, Operational acceptability, and Assurance Exposure. They describe the operational condition of a metric at the point of consumption. These semantics can support consistent steering, path-selection, and service-selection decisions across centralized, distributed, and hybrid deployments.

A derivation method for these semantics may depend on observable factors such as metric age, update continuity, source consistency, and deployment-specific trust conditions. These factors may be combined differently depending on the dynamics of the metric and the operational objectives of the deployment. Appendix A provides one illustrative realization of such logic.

5.1. Freshness

Freshness captures whether a metric remains temporally aligned with the condition it represents, particularly when update frequency does not match the dynamics of the underlying system. In many deployments, Freshness depends at least in part on the elapsed age of the metric relative to the time sensitivity of the condition it represents. A metric that is only a few seconds old may remain operationally usable for relatively stable capability information, while the same age may be excessive for rapidly varying utilization or admission-related state. Freshness therefore concerns whether the temporal separation between metric generation and metric consumption remains consistent with the operational purpose for which the metric is used.

5.2. Operational acceptability

Operational acceptability captures whether the metric remains suitable for operational consumption in steering at the current time. A metric may remain visible, syntactically valid, and even partially informative while no longer remaining appropriate for normal fine-grained steering use. For clarity, this document uses a lightweight three-state interpretation: acceptable, degraded, and unacceptable. More detailed state distinctions are possible, but they are outside the scope of this document. An acceptable metric remains suitable for normal steering input under the assumptions of the deployment. A degraded metric no longer supports normal steering use, but may still be retained for fallback or reduced-trust behavior. An unacceptable metric is not suitable for steering input. A deployment may derive these states from one or more factors, including metric age, update continuity, source consistency, or other deployment-specific conditions.

5.3. Assurance exposure

Assurance exposure captures whether degraded usage, inconsistency, or fallback behavior is externally visible to management or OAM. A system may continue to forward traffic and may continue to retain metric values internally while no longer operating under the semantic conditions that would justify normal steering. Assurance exposure therefore concerns whether degraded consumption, semantic divergence,

or fallback-driven behavior can be distinguished from normal operation by external functions for diagnosis, monitoring, or operational control.

5.4. Deployment-specific considerations

The effect of these semantics depends on where metrics are consumed for decisions and how metric-related information is exchanged among CATS functional entities. In centralized deployments, decisions are made primarily in a centralized CATS Path Selector (C-PS) or equivalent control-side function. In distributed deployments, decisions are made at, or near, an Ingress CATS-Forwarder. In hybrid deployments, decision logic is split across centralized and ingress-side functions.

In this document, communication among network elements refers mainly to the exchange of metric information or decision-related information, such as metric reporting from computing or service nodes to a decision function, or decision distribution from a C-PS to an Ingress CATS-Forwarder. These exchanges are distinct from data-plane traffic, where user traffic is forwarded toward a selected service instance. Degraded semantic conditions may also need to be exposed through management or OAM functions.

5.4.1. Centralized deployments

In centralized deployments, metrics typically reach the decision point only after collection, transport, processing, and possible aggregation. Metric information may be reported from computing or service nodes, possibly through metric agents, to a centralized C-PS or equivalent control-side function. The resulting decision-related information may then be provided to Ingress CATS-Forwarders for steering execution. As a result, a metric may no longer accurately reflect the condition on which the centralized decision is intended to rely by the time it reaches the decision function.

In this setting, freshness helps determine whether the metric remains temporally aligned with the underlying operational condition. Operational acceptability helps determine whether the metric can still be used as normal input to centralized steering logic, or whether it should instead be treated as degraded or reduced-trust input. Assurance exposure helps determine whether such degradation in metric consumption is externally visible, even when the centralized system continues to steer traffic and continues to receive metrics from the underlying sources.

5.4.2. Distributed deployments

In distributed deployments, metrics are consumed at, or close to, the ingress-side decision point. Metric information may be distributed directly to an Ingress CATS-Forwarder or to a co-located decision function, and the resulting steering decision may be applied locally. The main issue is that different local decision points may consume different observations, update histories, or local versions of what is operationally treated as the same condition.

A locally available metric may remain fresh from the perspective of one ingress decision point, while another ingress decision point has shifted to a different view of the same service or resource condition. In this setting, freshness helps determine whether the locally available metric remains temporally suitable. Operational acceptability helps determine whether that local metric can still support normal steering at that decision point, or whether it should instead be treated as degraded or reduced-trust input. Assurance exposure helps determine whether divergence across local decision points is externally visible, rather than remaining only an internal difference among distributed observations.

5.4.3. Hybrid deployments

In hybrid deployments, metric-consuming decisions are split across centralized and ingress-side functions, and different metric sources may be consumed at different layers of the same steering process. Some metric information may be collected and interpreted by a centralized C-PS, while other metric information may be consumed directly by an Ingress CATS-Forwarder or local decision function. The main issue is that jointly consumed inputs may not share the same temporal behavior, trust conditions, or operational scope. A relatively stable local metric may remain suitable for normal steering use, while a centrally distributed dynamic metric may be suitable only for degraded or reduced-trust use.

In this setting, freshness helps distinguish inputs whose temporal validity differs across sources. Operational acceptability helps distinguish source-specific degradation, so that one input may remain acceptable while another is retained only for degraded use. Assurance exposure helps determine whether such partial semantic degradation is externally visible.

For this reason, a hybrid deployment should be able to distinguish metrics that arrive from different sources and that do not share the same consumption conditions. It should also support source-specific degradation, so that one degraded input does not force all other inputs into the same state, and one acceptable input does not hide degradation in another.

6. Operational implications

6.1. Relationship to service continuity

In CATS, service continuity depends not only on whether traffic can still be forwarded, but also on whether the selected service instance or computing target remains suitable after the steering decision is made. A steering outcome may therefore remain valid from a forwarding perspective while no longer remaining valid from a service perspective. At the same time, the steering decision itself may depend on traffic- and service-related conditions whose validity is highly sensitive to metric freshness. Excessively frequent metric updates may introduce instability or oscillation into the steering process. Infrequent updates, by contrast, may cause steering decisions to rely on stale conditions that no longer reflect the current operational state. For this reason, freshness is relevant not only to the suitability of the selected service target, but also to the continued validity of the steering decision that directs traffic toward it.

Freshness helps determine whether a metric reflects the service condition on which continuity-related steering depends. Operational acceptability helps determine whether that metric can support normal continuity-sensitive steering or should instead be treated as degraded or fallback input. Assurance exposure helps make continuity-relevant degradation externally visible once the system shifted away from normal semantic conditions.

These semantics do not themselves provide continuity procedures, migration behavior, or affinity handling. They indicate when a metric should no longer be treated as a normal input to continuity-sensitive steering.

6.2. Control, management, and OAM relevance

These semantics are relevant not only at the metric-consuming decision point, but also to control, management, and OAM functions around it.

A control function may use these semantics to distinguish normal metric use from degraded or fallback use in the steering process. A management function may use them to determine whether steering is operating under normal semantic conditions or shifted into reduced-confidence behavior. An OAM function may use them to observe whether degraded consumption, semantic divergence, or fallback handling is operationally visible even though forwarding succeeds.

6.3. Lightweight signaling considerations

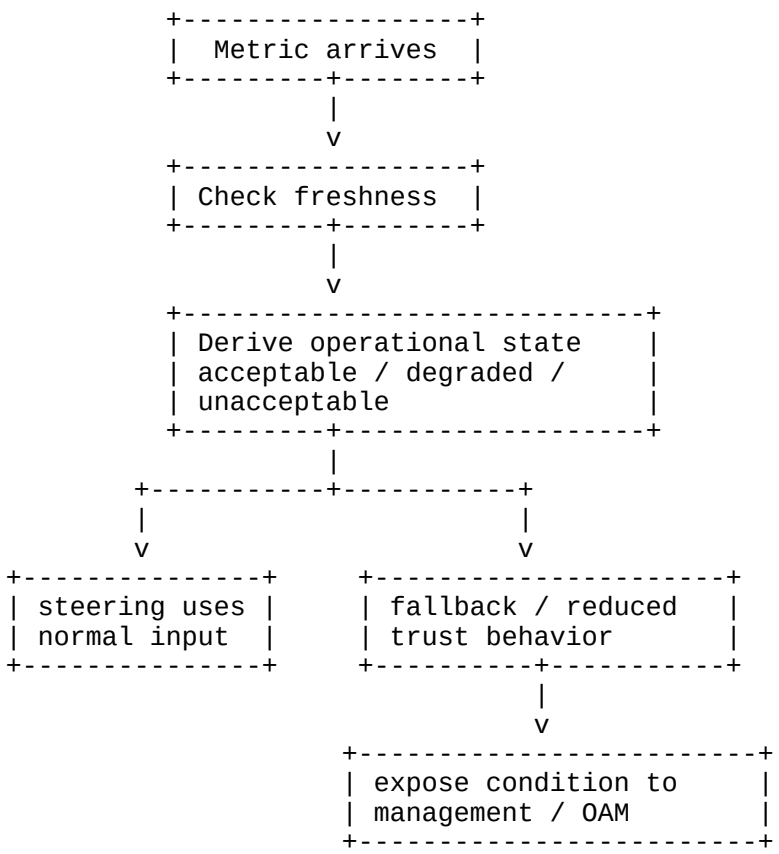
This document does not define protocol fields, but the semantics above are intended to be protocol-ready and lightweight.

Freshness could be reflected by a timestamp, an age value, or a validity window carried with the metric or its enclosing object, allowing the consumer to interpret the metric under different update frequencies. Operational acceptability could be represented as a compact three-state indication associated with the metric or with the result of consuming that metric. Assurance exposure could be realized by exposing degraded-consumption state to management or OAM systems, for example by attaching state to an OAM record, a management object, or a troubleshooting signal. Such signaling may occur on different information paths depending on deployment, such as metric reporting toward a decision function, decision distribution toward an ingress forwarder, or exposure toward management and OAM systems.

7. Illustrative example

Consider a hybrid deployment in which a consumer uses relatively stable site capability information learned through one path and fast-changing utilization information received through a centralized controller path. At time T1, both inputs are current enough that the consumer selects Site B for dynamic steering. At time T2, the capability information remains unchanged, but the utilization information distributed by the controller is several seconds old. If the consumer continues to treat both inputs as equally current, it may still steer new requests toward Site B even though Site B has lost the headroom assumed by the old utilization value.

The semantic decision flow can be illustrated as follows:



Under the semantics defined here, the capability information may remain acceptable, while the utilization information is only degradedly acceptable or even unacceptable. The consumer may therefore fall back to a coarser policy, and that fallback can be exposed to management or OAM.

8. Security Considerations

If an attacker can manipulate freshness-related metadata, acceptability state, or assurance visibility, traffic may be steered on the basis of information that appears valid but is not. This can amplify the impact of stale or falsified compute-related inputs and may lead to traffic mis-steering, localized resource exhaustion, or service disruption.

9. IANA Considerations

This document has no IANA actions.

10. Informative References

[CATS-FRAMEWORK]
IETF CATS Working Group, "A Framework for Computing-Aware Traffic Steering (CATS)", n.d., <https://datatracker.ietf.org/doc/draft-ietf-cats-framework/>.

[CATS-METRIC-DEFINITION]
IETF CATS Working Group, "Computing-Aware Traffic Steering (CATS) Metrics Definition", n.d., <https://datatracker.ietf.org/doc/draft-ietf-cats-metric-definition/>.

[CATS-OAM] IETF, "CATS OAM Framework", n.d., <https://datatracker.ietf.org/doc/draft-fu-cats-oam-fw/>.

[CATS-REQUIREMENTS]
IETF CATS Working Group, "Use Cases and Requirements for Computing-Aware Traffic Steering (CATS)", n.d., <https://datatracker.ietf.org/doc/draft-ietf-cats-usecases-requirements/>.

Acknowledgments

Illustrative Multi-Factor Derivation Model

This appendix provides one illustrative realization of the acceptable, degraded, and unacceptable states described in the main body of this document. It is included for illustration only.

For illustration, let T_age denote the elapsed time since metric generation. A deployment may compute T_age as the difference between the current time and the timestamp associated with the metric. Let T_validity denote a duration within which the metric is considered suitable for normal steering use. Let T_grace denote an additional duration during which the metric may still be retained for degraded or fallback use.

In addition, let U_gap denote the elapsed time since the last successful metric update, or more generally a measure of update continuity. This allows the example to capture not only whether a metric is old, but also whether the metric source is updating in a sufficiently continuous manner for operational use.

In this example, metric age provides the baseline timing model:

* acceptable: T_age <= T_validity

* degraded: T_validity < T_age <= T_validity + T_grace

* unacceptable: T_age > T_validity + T_grace

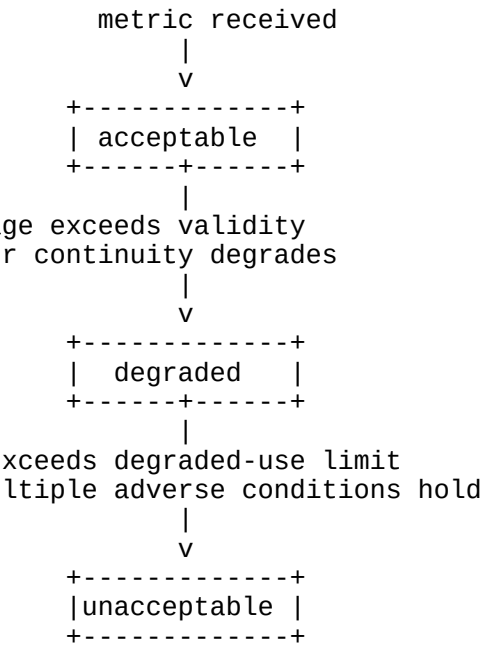
Update continuity then acts as an additional modifying condition. One simple interpretation is that poor update continuity may trigger an earlier transition to degraded or unacceptable states, even when the nominal age-based condition alone would not yet do so. For example:

* if U_gap > U_threshold, the metric may be treated as at least degraded, even if T_age <= T_validity

* if both T_age > T_validity + T_grace and U_gap > U_threshold, the metric may be treated as unacceptable

Under this example, T_validity represents a normal-use interval, while T_grace represents a degraded-use interval rather than an extension of full validity. The point of the example is not to define a universal formula, but to illustrate that a simple state space may still depend on more than one observable condition.

A simplified state transition model can be represented as:



A newer valid update may move the metric back to acceptable.

The same example may be summarized as:

Condition	Derived State	Interpretation
T_age <= T_validity and U_gap <= U_threshold	acceptable	usable for normal steering
T_validity < T_age <= T_validity + T_grace, or U_gap > U_threshold	degraded	usable only for reduced-trust or fallback behavior
T_age > T_validity + T_grace, or multiple adverse conditions persist	unacceptable	not suitable for steering input

Table 1

In a centralized deployment, T_age may dominate because the control loop of collection, processing, and redistribution can introduce significant delay before the metric reaches the decision point. In such a case, age-based degradation may become the primary reason that a metric transitions from acceptable to degraded.

In a distributed deployment, update continuity may become more significant because local decision points may rely on rapidly refreshed but independently observed inputs. In such a case, poor continuity or irregular local update behavior may cause a metric to lose normal steering utility even if its nominal age remains small.

In a hybrid deployment, different sources may be interpreted under different semantic conditions within the same decision process. A relatively stable local capability-related metric may remain acceptable, while a centrally distributed utilization-related metric may only remain suitable for degraded or reduced-trust use. This illustrates that state derivation may be both multi-factor and source-specific, rather than globally uniform across all inputs.

Author’s Address

Mengfei Zhu
China mobile
Email: zhumengfei@cmdi.chinamobile.com