opsawg
Internet-Draft
Intended status: Standards Track
Expires: 23 April 2026

K. Yao
China Mobile
L. M. Contreras
Telefonica
J. Ros-Giralt
Qualcomm Europe, Inc.

20 October 2025

Service Instance Deployment based on Integrated Network and Compute Metrics
draft-ymg-opsawg-service-deployment-with-alto-cats-00

Abstract

The deployment of service instances across distributed interconnected edge-cloud environments can be optimized in terms of performance expectations and Service Level Objectives (SLOs) satisfaction when performed taken into account both network and compute metrics. In order to do so, this document primarily concentrates on existing standardized mechanisms, namely ALTO and CATS, to facilitate such integration of metrics. The ALTO protocol can be extended to expose compute metrics from a cloud manager to a network orchestrator or as part of the network and cost maps, enabling improved deployment of compute service instances based on joint awareness of both network and computing information. This document proposes protocol extensions, workflows, and operational considerations for ALTO enhancements using CATS metrics.

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 23 April 2026.

Table of Contents

1. Introduction

Applications such as artificial intelligence (AI) inference and cloud rendering require performance optimization based on joint awareness of both network and computing information. [I-D.rcr-opsawg-operational-compute-metrics] introduces the service lifecycle, including service deployment, service selection, and service assurance. The discussion and documentation of the service selection problem are currently being undertaken by the Computing-Aware Traffic Steering (CATS) Working Group, while the service deployment problem still lacks a dedicated venue for resolution. This document primarily focuses on the service deployment problem and leverages the flexibility of ALTO protocol extensions to enable joint awareness of both network and computing information for improved compute service instance deployment.

[I-D.ietf-cats-metric-definition], adopted by the CATS Working Group, defines three metric levels for all CATS-related metrics from the computing domain (e.g., cloud). These metrics are initially intended to support service instance selection for traffic steering, but they can also be exposed for compute service deployment. [I-D.contreras-alto-service-edge] introduces a method for using ALTO protocol extensions for service deployment, but it does not cover CATS metrics or their encodings. This document provides supplementary information to [I-D.contreras-alto-service-edge], including protocol extensions for encoding CATS metrics, workflows, and operational considerations for compute service instance deployment.

2. Workflows

The main entities involved in this solution are the ALTO server and the ALTO client.

*ALTO Server: Deployed in the cloud manager and network controller, it is responsible for collecting various CATS metrics.

*ALTO Client: Located in the service orchestrator, it is responsible for requesting and receiving information from the ALTO server, and for formulating compute instance deployment strategies based on the collected data.

There are two basic implementation schemes. The interactions between the ALTO client and servers may vary depending on the chosen mode.

2.1. Request and Response

Figure 1 shows the workflow under request and response mode.

The ALTO client in the service orchestrator within the network domain first requests computing domain information from the ALTO server located in the cloud manager, and then requests network domain information from the ALTO server in the network controller.

The ALTO server in the cloud manager collects CATS metrics, including various L0 metrics, calculates L1 and L2 metrics, determines and selects all or part of the metrics from L0, L1, and L2 to pass to the ALTO client, encapsulates the message, and sends the ALTO message to the ALTO client. The basis for the ALTO server in the cloud manager to determine which level of computing metrics to send is the request information from the ALTO client. The initial request from the ALTO client will clearly specify the requested level of CATS metrics (such as L2, L1, and/or L0), and for L1 and/or L0 metrics, it will specify whether to request all metrics at that level or specific ones (e.g., only the "compute type" L1 metric for L1, or only CPU utilization for L0).

The ALTO server in the network controller obtains network domain information (such as link bandwidth, latency, etc.), encapsulates all the obtained information in an ALTO message, and sends it to the ALTO client.

The ALTO client sends confirmation messages to both ALTO servers.

The ALTO client in the service orchestrator calculates the compute instance deployment method based on the obtained computing domain and network domain information, and then sends the deployment strategy information to the ALTO server in the cloud manager, which notifies the cloud manager to deploy the corresponding computing instances.

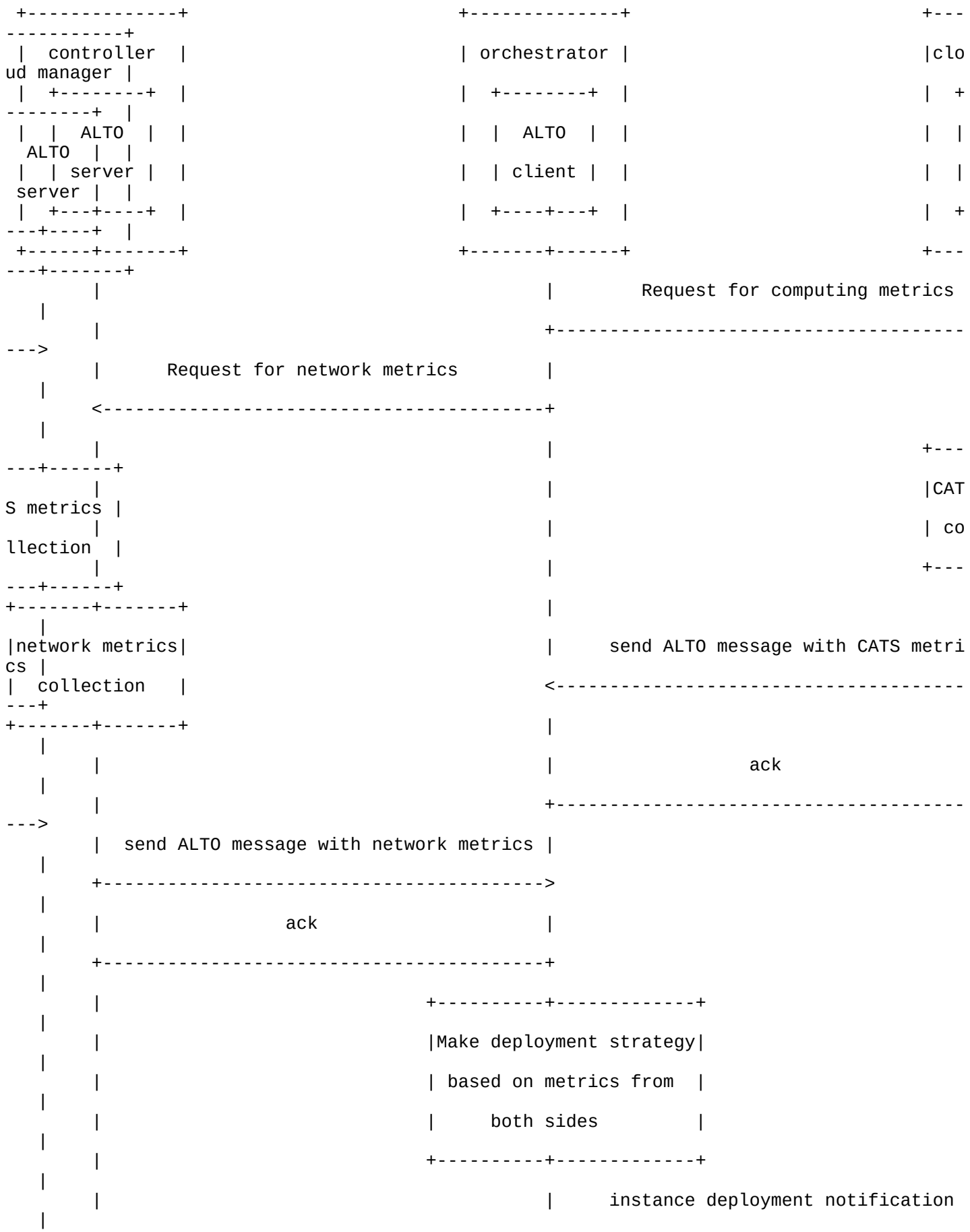

controller
orchestrator
cloud manager
ALTO server
ALTO client
ALTO
Request for computing metrics
Request for network metrics
CATS metrics collection
network metrics collection
send ALTO message with CATS metrics
ack
send ALTO message with network metrics
ack
Make deployment strategy based on metrics from both sides
instance deployment notification

Figure 1: Request-Response Mode

2.2. Active Push

Figure 2 shows the workflow under active push mode.

The ALTO client establishes Server-Sent Events(SSE) long connections with the ALTO server in the cloud manager and the ALTO server in the network controller respectively. During connection maintenance, ALTO servers actively push ALTO messages to the ALTO client.

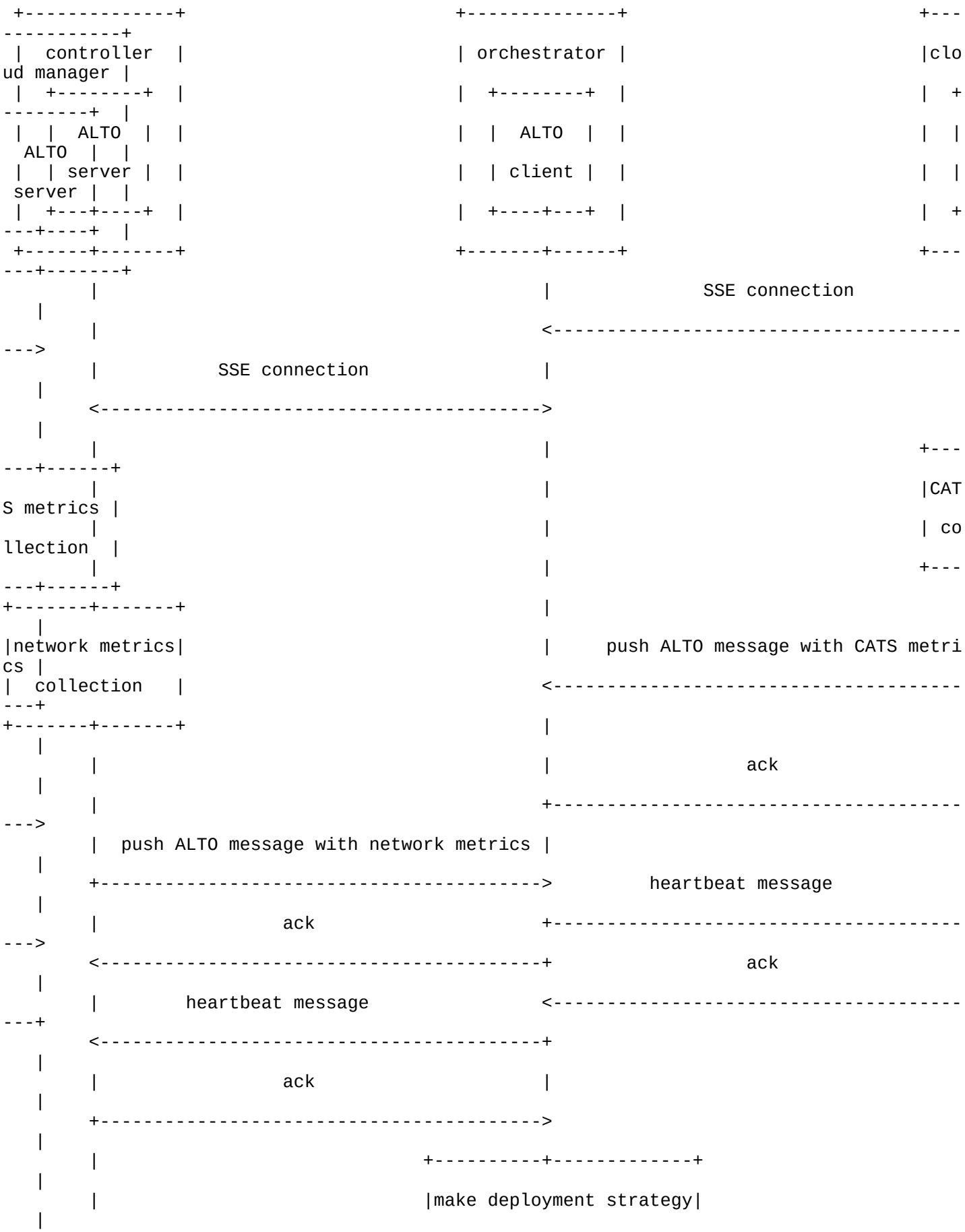
The ALTO server in the cloud manager obtains CATS metrics, including various L0 metrics, calculates L1 and L2 metrics, determines and selects all or part of the metrics from L0, L1, and L2 to pass to the ALTO client, encapsulates the message, and actively pushes the ALTO message containing the selected computing domain information.

The ALTO server in the network controller obtains network domain information, encapsulates it in an ALTO message, and pushes it to the ALTO client.

The ALTO client sends confirmation messages and periodic heartbeat messages to maintain connection status. ALTO servers then reply with acknowledgment (ACK) messages if the connections are valid.

The ALTO client formulates the compute instance deployment strategy based on the obtained information and sends the deployment strategy information.

Note that in the active push mode, the ALTO server initially sends L2 level computing metrics by default. The ALTO client can carry requests for L1 and/or L0 level metrics (all or part) in the confirmation message, and the ALTO server will start pushing the requested metrics from the next cycle.


controller
orchestrator
cloud manager
ALTO server
ALTO client
SSE connection
SSE connection
CATS metrics collection
network metrics collection
push ALTO message with CATS metrics
ack
push ALTO message with network metrics
heartbeat message
ack
ack
heartbeat message
ack
make deployment strategy

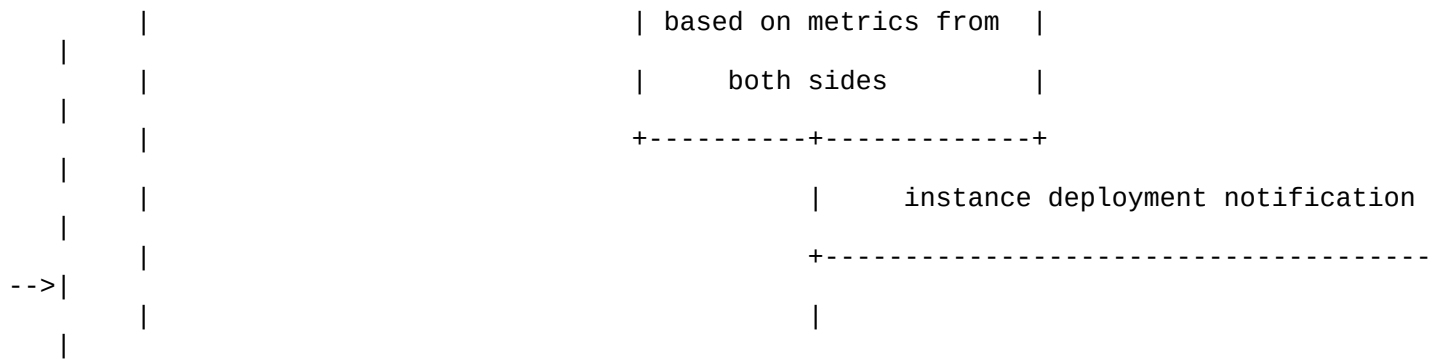


Figure 2: Active Push Mode

2.3. Protocol Extention Examples

The three-layer metric information can be defined by extending the ALTO endpoint cost service in RFC 7285, through extending the "cost-type" field.

Figure 3 shows the example in JSON format:

```
{
  "meta": {
    "cost-types": {
      "L2-metric": {
        "metric-type": "Fully normalized metric",
        "level": "L2",
        "cost-mode": "numerical",
        "cost-metric": "normalized-value"
      },
      "L1-metric": {
        "metric-type": "Compute metric",
        "level": "L1",
        "cost-mode": "numerical",
        "cost-metric": "normalized-value"
      },
      "L0-metric": {
        "metric-type": "CPU frequency",
        "level": "L0",
        "cost-mode": "numerical",
        "cost-metric": "GigaHertz"
      }
    }
  }
}
```

Figure 3: cost-type extension

Figure 4 describes the encapsulation of CATS metrics as well as site information based on the ALTO protocol extensions in JSON format:

```
{
  "meta": {
    "dependent-vtags": [
      {
        "resource-id": "my-default-networkmap",
        "tag": "3ee2cb7e8d63d9fab71b9b34cbf764436315542e"
      }
    ]
  },
  "endpoint-properties": {
    "ipv4:192.0.2.34": {
      "Fully normalized metric": {
        "level": "L2",
        "value": 5
      }
    },
    "ipv4:203.0.113.56": {
      "Compute metric": {
        "level": "L1",
        "value": 3
      },
      "CPU frequency": {
        "level": "L0",
        "value": "2.2 GigaHertz"
      }
    }
  }
}
```


Figure 4: alto encodings

In the example above, the ALTO message carries metric information of two service endpoints. The ALTO server can define and set the content to be sent, choosing to send all levels of metric information for the corresponding endpoint, or select one level of metrics or a specific metric within a level.

3. Operational Considerations

3.1. Compute Instance Deployment Strategy

The ALTO client in the service orchestrator formulates various computing service instance deployment methods based on computing and network information. The procedure of how to generate and deploy the strategy should also be considered.

Firstly, determine the availability for instance deployment. The ALTO client must first verify whether a specific site is capable of hosting the service. Therefore, supplementary procedures may be required at the beginning of both workflows described in the previous sections. Upon receiving a service request, the ALTO client needs to notify the ALTO server in the cloud manager of the specific resources required for deployment (e.g., X CPU cores or Y GB of GPU memory).

Secondly, determine the priority for instance deployment. If the ALTO client receives Level 2 (L2) metrics, it may perform a direct summation. If it receives Level 1 (L1) metrics, it may apply a weighted summation, for example:

(Score of computing class L1 metric of a node from the cloud manager’s ALTO server * weight1) + (Score of normalized network class L1 metric of a link from the network controller’s ALTO server * weight2)

If the ALTO client receives Level 0 (L0) metrics, the algorithm may involve applying a polynomial function over multiple metrics. After computation, the ALTO client sorts the results to determine the priority of instance deployment, with higher scores indicating higher priority.

Thirdly, determine remaining resources after instance deployment. Once the compute service node for deployment is selected, the ALTO server completes the instance deployment, calculates the remaining resource availability, and notifies the ALTO client.

3.2. Deployment Considerations of ALTO Client and Server

The ALTO server can be co-located with the network controller or cloud manager, or deployed separately. Similarly, the ALTO client can be co-located with the service orchestrator or deployed separately.

The three-level metric framework provides flexibility in information exposure, allowing adaptation to different scenarios where the computing and network domains may belong to the same or different service entities.

Dynamic updates of metrics should be considered to ensure the timeliness and accuracy of information for effective deployment decisions.

4. Security Considerations

This document does not introduce new security risks beyond those inherent in the ALTO protocol. Security mechanisms specified in [RFC7285] and related ALTO extensions (such as access control in [RFC7971]) should be applied to protect sensitive computing and network information, especially when computing and network domains belong to different service entities.

5. IANA Considerations

This document has no IANA actions.

6. Acknowledgements

7. References

7.1. Normative References

[RFC7285] Alimi, R., Ed., Penno, R., Ed., Yang, Y., Ed., Kiesel, S., Previdi, S., Roome, W., Shalunov, S., and R. Woundy, "Application-Layer Traffic Optimization (ALTO) Protocol", RFC 7285, DOI 10.17487/RFC7285, September 2014, <https://www.rfc-editor.org/rfc/rfc7285>.

[RFC7971] Stiemerling, M., Kiesel, S., Scharf, M., Seidel, H., and S. Previdi, "Application-Layer Traffic Optimization (ALTO) Deployment Considerations", RFC 7971, DOI 10.17487/RFC7971, October 2016, <https://www.rfc-editor.org/rfc/rfc7971>.

7.2. Informative References

[I-D.contreras-alto-service-edge] Contreras, L. M., Randriamasy, S., Ros-Giralt, J., Perez, D. A. L., and C. E. Rothenberg, "Use of ALTO for Determining Service Edge", Work in Progress, Internet-Draft, draft-contreras-alto-service-edge-10, 13 October 2023, <https://datatracker.ietf.org/doc/html/draft-contreras-alto-service-edge-10>.

[I-D.ietf-cats-metric-definition] Yao, K., Li, C., Contreras, L. M., Ros-Giralt, J., and H. Shi, "CATS Metrics Definition", Work in Progress, Internet-Draft, draft-ietf-cats-metric-definition-04, 20 October 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-cats-metric-definition-04>.

[I-D.rcr-opsawg-operational-compute-metrics]
Randriamasy, S., Contreras, L. M., Ros-Giralt, J., and R. Schott, "Joint Exposure of Network and Compute Information for Infrastructure-Aware Service Deployment", Work in Progress, Internet-Draft, draft-rcr-opsawg-operational-compute-metrics-08, 21 October 2024, <https://datatracker.ietf.org/doc/html/draft-rcr-opsawg-operational-compute-metrics-08>.

Authors' Addresses

Kehan Yao
China Mobile
Email: yaokehan@chinamobile.com

L. M. Contreras
Telefonica
Email: luismiguel.contrerasmurillo@telefonica.com

Jordi Ros-Giralt
Qualcomm Europe, Inc.
Email: jros@qti.qualcomm.com