cats
Internet-Draft
Intended status: Informational
Expires: 8 January 2026

L. Contreras
Telefonica
M. Watts
Verizon
T. Jiang
China Mobile
7 July 2025

Compute-Aware Traffic Steering for Midhaul Networks
draft-lcmw-cats-midhaul-03

Abstract

Computing-Aware Traffic Steering (CATS) takes into account both computing and networking resource metrics for selecting the appropriate service instance to forwarding the service traffic. This document described the usage of Computing-Aware Traffic Steering (CATS) within Midhaul (MH) networks in the O-RAN architecture. It details how CATS can enhance traffic steering decisions between Distributed Units (DUs) and Centralized Units (CUs) by considering both compute resource metrics (e.g., CPU and memory utilization of CU instances) and network performance metrics (e.g., bandwidth, latency, reliability).

The document discusses the integration of CATS with O-RAN management frameworks, and the interplay with the Transport Network Manager (TNM) in O-RAN using standard interfaces defined by IETF (as for example the one for Network Slice Services for connectivity provisioning).

Table of Contents

1. Introduction

The radio functional split architecture proposed by O-RAN [ORAN-Arch] functionally separates the processing of the mobile radio signal originally performed in a single radio base station by placing functionality in three entities, namely the Radio Unit (RU), the Distributed Unit (DU) and the Centralized Unit (CU). Both DU and CU are typically deployed as service functions on virtualized compute nodes in the network.

The network segment between RU and DU is known as Fronthaul (FH), while the network segment between DU and CU is known as Midhaul (MH), or F1 interface according to 3GPP terminology. Both FH and MH have specific needs and characteristics in terms of latency and bandwidth, constrained by the nature of the data payload and the protocols intrinsic for the support of the radio functional split. More details can be found in [ORAN-Req]. The requirements on the FH are much stringent that the ones in MH.

In the current O-RAN framework, a DU selects a CU and then creates the association DU <> CU-UP (CU User Plane) in a dynamic way. Such association is established before a UE (or end device) comes to register with the mobile operator via the DU, and then the associated CU. From an architectural point of view, it is possible to consider scenarios where traffic flows from the DUs can be delivered to different CUs depending on the compute and network metrics observed during runtime. It is in these situations where CATS proposition can play a distinctive role at the time of ensuring proper delivery of the midhaul traffic and its processing. Thus, the DU <> CU-CP can be dynamic in a way that a DU might optimally select a CU based on compute and network metrics.

1.1. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT","SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

In addition, this document uses the terms defined in [I-D.ietf-cats-framework].

2. Midhaul Scenario

The connection of RU, DU and CU can be performed by means of an IP-based aggregation network. In O-RAN terminology [ORAN-Transport], the aggregation routers acting as PE-nodes are called Transport Network Elements (TNEs). The control and management of the TNEs is performed by a Transport Network Manager (TNM) [ORAN-TransportManagement]. Figure 1 illustrates a packet-switched based aggregation network in O-RAN.

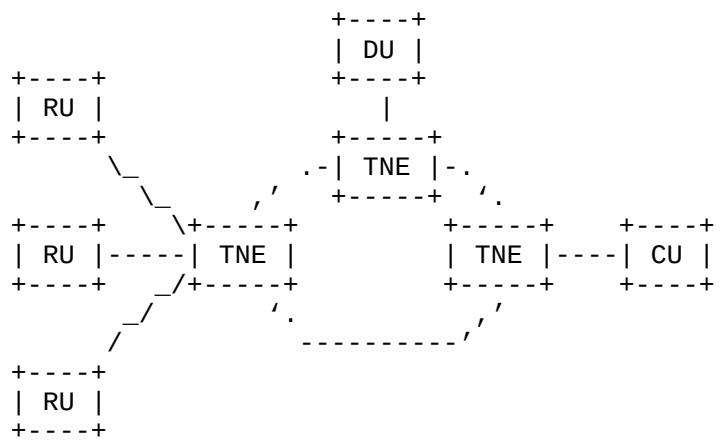

Figure 1: Midhaul Scenario

The FH segment connecting RUs and DUs is typically static in the sense that RUs are anchored with the same DU along the time. However, in the case of MH, the association between DUs and CUs could be more dynamic, subject to runtime situations such as DU and CU load, protection, workload migration (in the case of virtualized CU), energy efficiency, etc.

It is in these situations where the steering of the flows between DU and CU can take into consideration both service (including compute) and network metrics, as proposed by CATS. The focus in this document refers to the user plane of DU and CU connection (i.e., CU-UP).

The CUs can be deployed in different regions of the network, representing different service instances deployed in distinct service sites. For the illustration of the scenario, Figure 2 considers a number of CU instances in different Data Centers (DCs) and a DU running on a server, all of them interconnected by an aggregation network. Note that the DU could also run as an instance in a DC or even be a physical appliance.

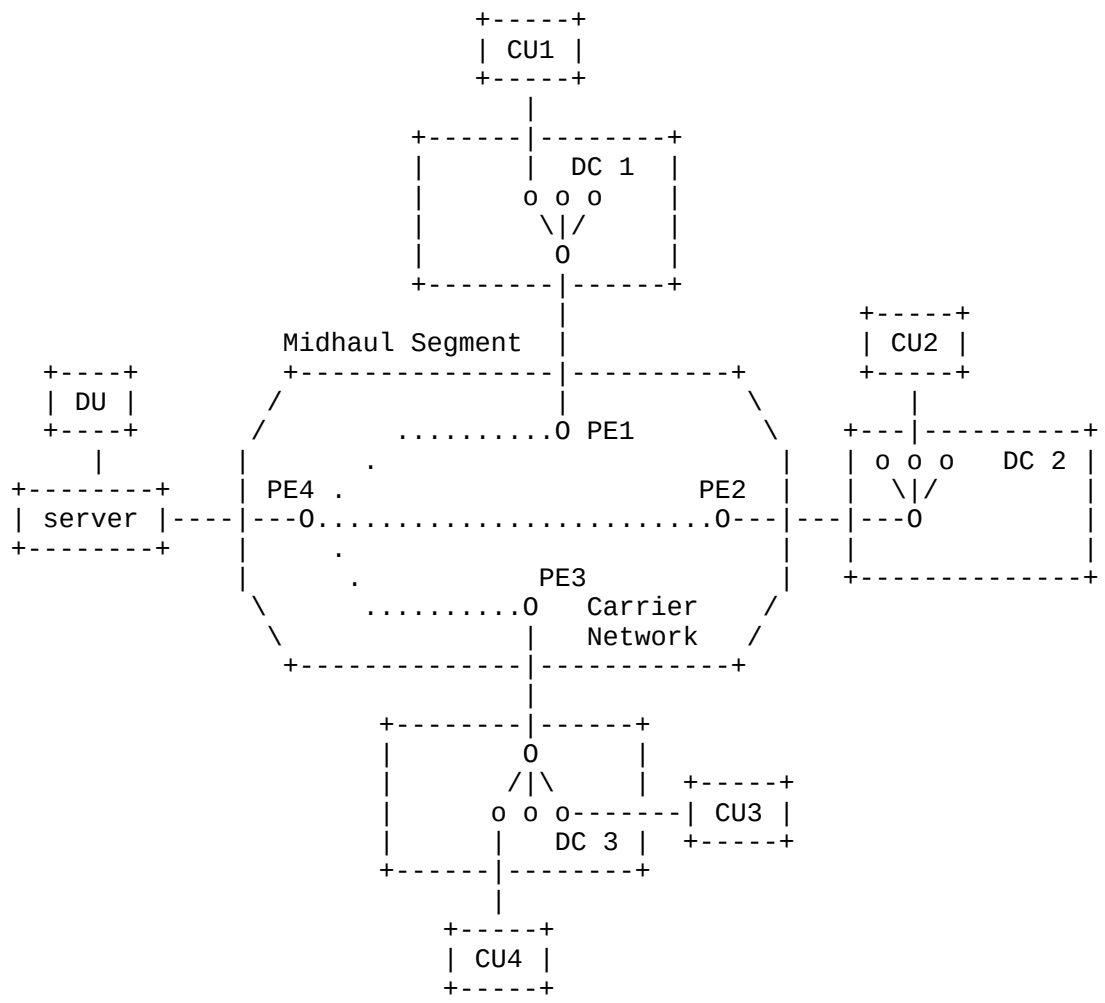

Figure 2: Midhaul Scenario

The aggregation network is IP-based, so the MH is realized by means of packet-switching technologies. This is consistent with the assumption in CATS that the underlay technology is IP/MPLS network. Figure 3 (according to the specification of the F1 / midhaul interface in [TS38.470] by 3GPP) illustrates the concern of CATS in the connection between DU and CU User Plane (CU-UP), considering as example an MPLS-based VPN connectivity.

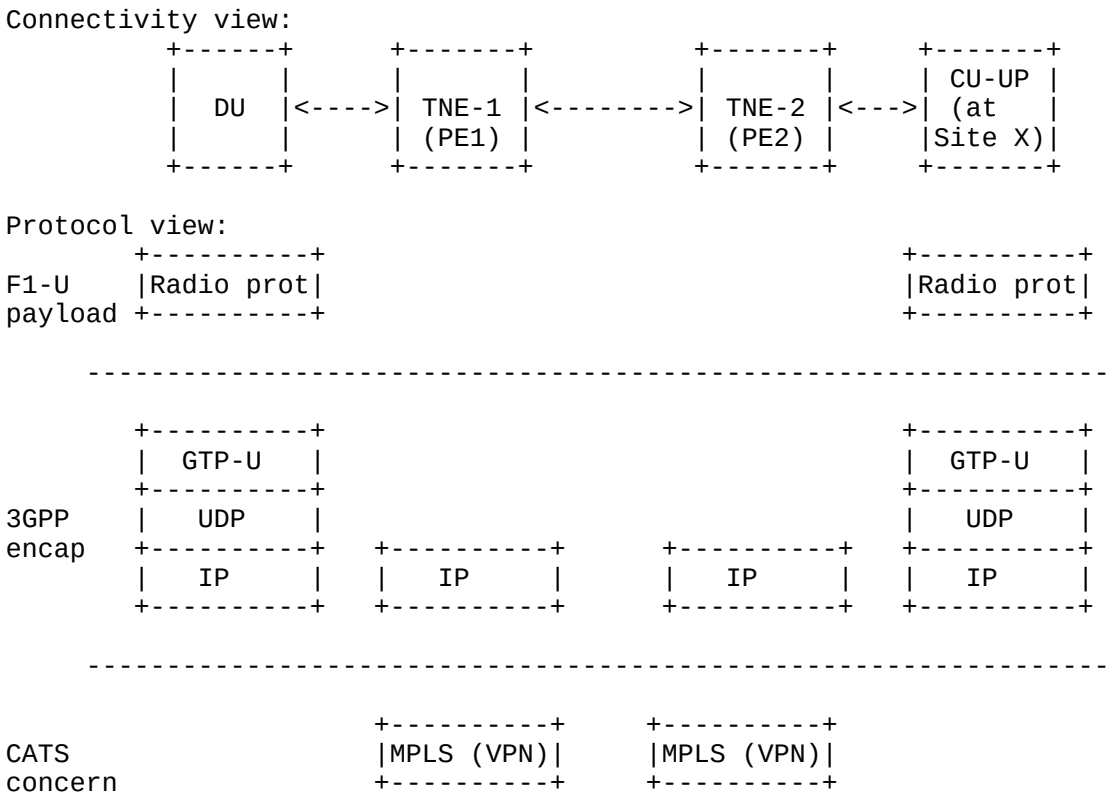

Figure 3: CATS concern

3. CATS framework applicability for Midhaul

The DU traffic cannot be separated in different flows. That is, the payload between DU and CU cannot be discriminated in individual flows since the payload represents a pre-processed analog radio signal, which will be entirely processed by the CU for obtaining the particular end-user flows. In this situation, the steering decision for the selection of a particular CU instance applies to the entire DU traffic. This simplifies the traffic classification since all the traffic from a DU is forwarded to the CU until any change is needed.

Note: Since all the midhaul traffic has the same service instance as destination (until any change applies), it is not necessary in this particular case the usage of CS-ID for accessing the service. The traffic classification is simple because all packets belong to the same service request.

The PE nodes (being TNEs in O-RAN terminology) in Figure 2 play the role of CATS-Forwarders. Each DC is expected to count with a CATS Service Metric Agent (C-SMA), while the network part is expected to count with a CATS Network Metric Agent (C-NMA). These agents will report different metrics and data to the CATS Path Selector (C-PS), which in this case can be assumed to be part of the TNM (i.e., considering that a centralized deployment model is followed, with the TNM playing the role of centralized control and management element).

Example of metrics related to compute could be the CPU average utilization or the memory usage of every CU-UP instance [ORAN-OCloud].

Figure 4 maps the CATS framework to the midhaul case.

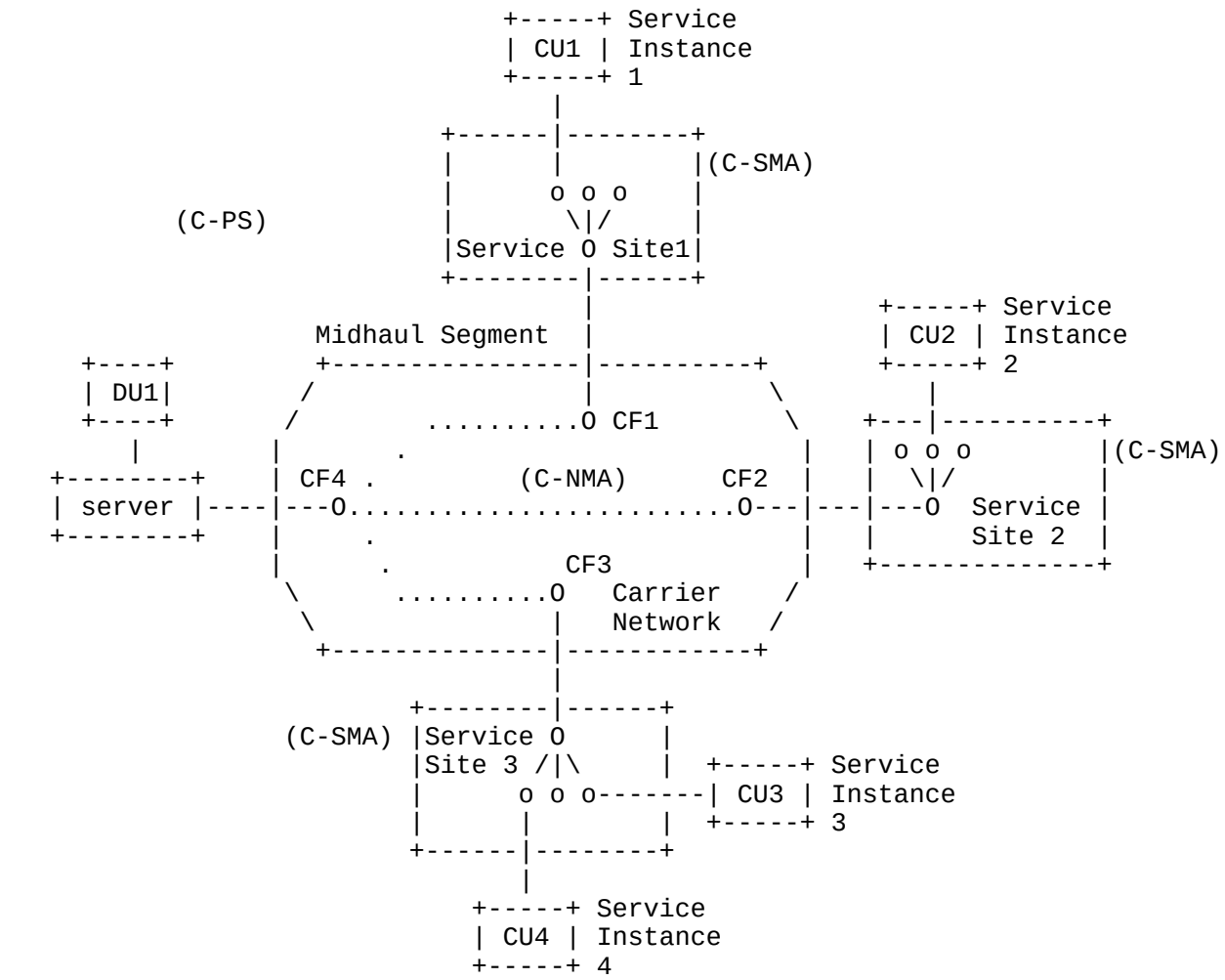

Figure 4: CATS applicability to Midhaul Scenario

3.1. Control plane interactions between O-RAN and IETF management entities

In the O-RAN architecture, the Service and Management Orchestrator (SMO) is responsible for RAN domain management and provides orchestration and management services.

The connectivity between O-RAN radio functional entities is assumed to be managed by a Transport Network Manager (TNM) in charge of the control and management of the network. The interplay between the O-RAN SMO and the TNM is currently under definition. The interplay

between TNM and SMO allows an integrated management of both the transport network and RAN elements, enhancing end-to-end service orchestration and operational efficiency.

The TNM function is assumed to be performed following IETF specifications. That role could be played, for instance, by the Network Slice Controller as defined in [RFC9543] for the provision of network slice services [ORAN-Transport]. In summary, TNM can incorporate standardized transport network data models and protocols to enrich O-RAN capabilities and interoperability. Figure 5 represents the relationship between O-RAN SMO and the TNM.

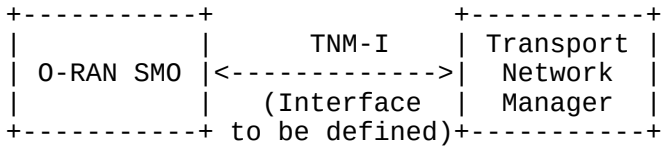

Figure 5: Interworking between SMO and TNM

The SMO defines the O1 interface for management functions such as fault, configuration, performance, and security management of O-RAN Network Functions (NFs), including O-DU and O-CU. For the specific case of CU-UP instance selection, every instance can be associated to various kinds of runtime information, i.e., including both (i) network metrics like bandwidth, delay, path-loss, reliability, etc., and (ii) compute or service metrics (of the CU-UP) like CPU load, memory, storage, service-load, etc. That information can assist on the selection of the most convenient CU-UP instance for a given DU.

The O-RAN architecture also defines the O-Cloud entity as the component concerned with infrastructure and workload-related metrics such as fault management, performance management (including possibly resource utilization, availability, throughput, latency), and configuration status of the cloud platform itself and the hosted O-RAN Network Functions. The SMO and the O-Cloud interact by means of the O2 interface.

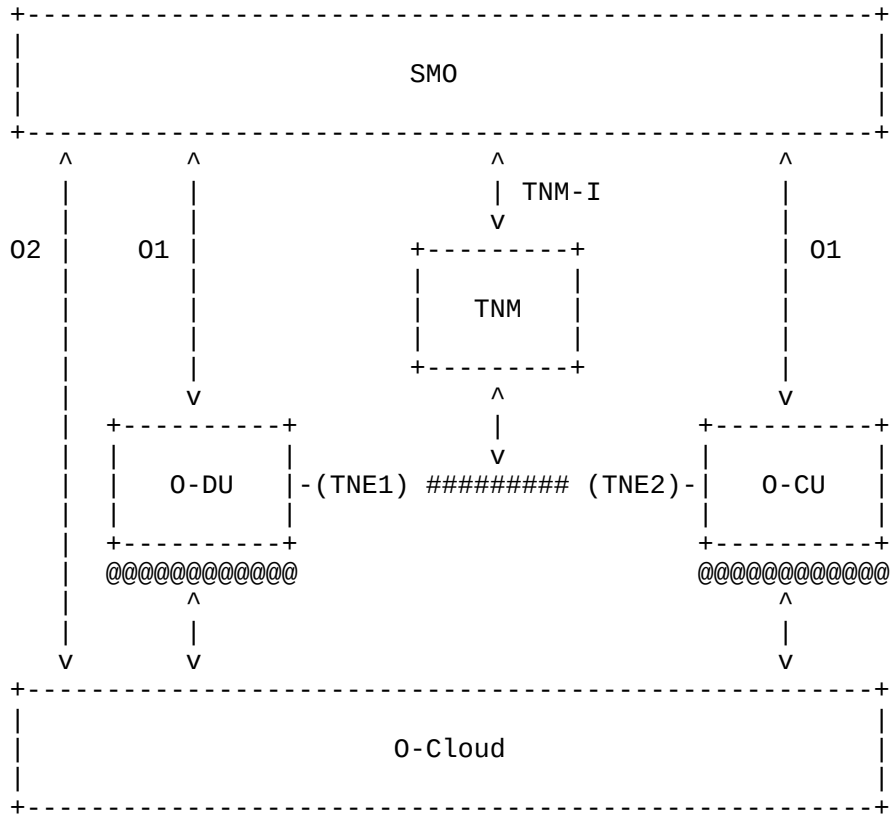

Figure 6: SMO, O-Cloud and TNM

After the selection of a given instance, the O-RAN SMO will proceed to the necessary configurations on the O-RAN functional entities and instruct the TNM for performing the traffic steering of the service flows. The TNM, in consequence, enforce the steering path for connecting DU and CU-UP.

When mapped to the CATS framework, this has the follwing implications:

* The compute-related metrics are provided by the O-Cloud component.

* The network-related metrics are provided by the TNM component.

* Two cases can be considered:

  - The processing of both kind of metrics is performed by the SMO, who finally determines the association of DU and CU-UP instances based on the metrics before. This means that the SMO gets network-related metrics from TNM (and probably some adidtional information, like topology), and that the CATS Path Selector functional component resides in the SMO.

  - The TNM gets from the SMO the compute-aware metrics performing the processing of both kind of metrics. Additional information can be required from SMO (for instance the location of the data center sites and/or the association policy). This means that the CATS Path Selector functional component resides in the TNM.

* After the association decision, the TNM for enforcing the steering path in the network (in the first case, instructed by the SMO, while in the second taking the enforcement decision by its own).

These approaches do not prevent other possible alternatives of mapping CATS entities in an O-RAN scenario.

3.2. Example of connectivity based on IETF Network Slice Service

The connectivity in the MH segment could be realized for instance by means of IETF Network Slice Services [RFC9543], as described in [I-D.ietf-teas-5g-network-slice-application] and [I-D.ietf-teas-5g-ns-ip-mpls], according to the Service Level Objectives of the Midhaul traffic. With that connectivity in place for each of the possible CUs as Service Instances, the C-PS could decided which slice to use for delivering the traffic to a specific CU. Note that the realization of the IETF Network Slice Service could be performed either by means of a common slice for connecting the DU with all the CUs, or a slice per DU to CU connection. Once the C-PS takes deicsion on which CU (or Service Instance) deliver all the DU traffic, a policy could be applied (e.g., usage of the IP address of the CU-UP instance as match criteria in [I-D.ietf-teas-ietf-network-slice-nbi-yang]) for mapping the DU traffic to the proper connectivty construct of the IETF Network Slice Service.

Thus, the IETF Network Slice Service could be defined as hub-and-spoke from each DU to any of the CU-UP instances, and realized, e.g., by means of a VPN. A potential definition of the slice service using [I-D.ietf-teas-ietf-network-slice-nbi-yang]) could be as follows in Figure 7:

```
            "connection-groups": {
              "connection-group": [
                {
                  "id": "matrix1",
                  "connectivity-type": "ietf-vpn-common:hub-spoke",
                  "connectivity-construct": [
                    {
                      "id": "1",
                      "p2mp-sender-sdp": "du1",
                      "p2mp-receiver-sdp": [
                        "cu-up1",
                        "cu-up2",
                        "cu-up3",
                        "cu-up4"
                      ],
                      },

                      "status": {}
                    }
                  ]
                }
              }
```

Figure 7: Definition of the CATS steering paths as IETF NEtwork Slice Service

Moreover, based on the metrics collected for both network and compute, the C-PS could take the decision of steering the traffic of the DU towards a particular CU-UP instance, properly configuring the math criteria, setting it to the destination IP address of the CU-UP instance of interest, as exemplified in Figure 8.

```
                "service-match-criteria": {
                  "match-criterion": [
                    {
                      "index": 1,
                      "match-type": "ietf-nss:destination-ip-prefix",
                      "value": ["2001:db8::1/64"],
                      "target-connection-group-id": "matrix1"
                    }
                  ]
                }
```

Figure 8: Enforcement of the path steering leveraging on match-criteria

4. Open points for discussion

This version is an initial attempt of applicability of CATS for Midhaul scenarios as defined in O-RAN. The followoing are identified open points for further discussion, which will be elaborated in next versions of the document.

* Actions / situations changing the service affinity in the case of midhaul (and potential interaction with O-RAN specific service orchestration capabilities).

5. Security Considerations

Same security considerations as in [I-D.ietf-cats-framework] apply also here.

6. Acknowledgements

TBC

7. References

7.1. Normative References

[RFC9543] Farrel, A., Ed., Drake, J., Ed., Rokui, R., Homma, S., Makhijani, K., Contreras, L., and J. Tantsura, "A Framework for Network Slices in Networks Built from IETF Technologies", RFC 9543, DOI 10.17487/RFC9543, March 2024, <https://www.rfc-editor.org/info/rfc9543>.

7.2. Informative References

[I-D.ietf-cats-framework] Li, C., Du, Z., Boucadair, M., Contreras, L. M., and J. Drake, "A Framework for Computing-Aware Traffic Steering (CATS)", Work in Progress, Internet-Draft, draft-ietf-cats-framework-10, 24 June 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-cats-framework-10>.

[I-D.ietf-teas-5g-network-slice-application] Geng, X., Contreras, L. M., Rokui, R., Dong, J., and I. Bykov, "IETF Network Slice Application in 3GPP 5G End-to-End Network Slice", Work in Progress, Internet-Draft, draft-ietf-teas-5g-network-slice-application-04, 3 March 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-teas-5g-network-slice-application-04>.

[I-D.ietf-teas-5g-ns-ip-mpls]
Szarkowicz, K. G., Roberts, R., Lucek, J., Boucadair, M., and L. M. Contreras, "A Realization of Network Slices for 5G Networks Using Current IP/MPLS Technologies", Work in Progress, Internet-Draft, draft-ietf-teas-5g-ns-ip-mpls-18, 3 April 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-teas-5g-ns-ip-mpls-18>.

[I-D.ietf-teas-ietf-network-slice-nbi-yang]
Wu, B., Dhody, D., Rokui, R., Saad, T., and J. Mullooly, "A YANG Data Model for the RFC 9543 Network Slice Service", Work in Progress, Internet-Draft, draft-ietf-teas-ietf-network-slice-nbi-yang-25, 9 May 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-teas-ietf-network-slice-nbi-yang-25>.

[ORAN-Arch]
"O-RAN Architecture Description, V11.00", February 2024.

[ORAN-OCloud]
"O-Cloud Information Model, V01.00", June 2024.

[ORAN-Req] "O-RAN Xhaul Transport Requirements, V01.00", February 2021.

[ORAN-Transport]
"O-RAN Xhaul Packet Switched Architectures and Solutions, V07.00", February 2024.

[ORAN-TransportManagement]
"O-RAN Management interfaces for Transport Network Elements, V07.00", October 2023.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[TS38.470] "F1 general aspects and principles, V16.2.0", 2020.

Authors' Addresses

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

Mark Watts
Verizon
Email: mark.t.watts@verizon.com

Tianji Jiang
China Mobile
Email: tianjijiang@yahoo.com