CATS WG   CJ. Bernardos
Internet-Draft   UC3M
Intended status: Standards Track   A. Mourad
Expires: 4 September 2025   InterDigital
                                                            3 March 2025

Service Mobility-Enabled Computing Aware Traffic Steering using IP address anchoring

draft-bernardos-cats-anchoring-service-mobility-03

Abstract

The IETF CATS WG addresses the problem of how the network infrastructure can steer traffic between clients of a service and sites offering the service, considering both network metrics (such as bandwidth and latency), and compute metrics (such as processing, storage capabilities, and capacity).

This document defines new extensions and procedures for a terminal connected to a network infrastructure, to benefit from transparent service migration adapting to specific connectivity and computing requirements, so traffic is always steered to an instance meeting both requirements. Both CATS-aware and -unaware terminals are considered. Exemplary signaling control messages and operation extending the well-known Proxy Mobile IPv6 protocol are also defined.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction and Problem Statement

1.1. Use case scenario

Let's consider a possible use case scenario, just for the sake of illustrating the scenario. Several nodes (UEs in this example) are acting as sensors in an Integrated Sensing and Communications (ISAC) case. The sensors generate/collect sensing data that needs to be processed timely and appropriately to generate an accurate sensing result. Part of this service is executed in the network infrastructure, posing some requirements on the connectivity (e.g., delay between the terminals and the node where the service is executed on the network infrastructure) and computing resources (e.g., capabilities to render the XR video within a certain latency budget). Within the network domain where the terminals are connected to there are multiple sites capable of hosting the service, each with potentially different connectivity and computing characteristics. Figure 1 shows an exemplary scenario. Considering the connectivity and computing latencies (just as an example of metrics), the best service site is #n-1 in the example used in the Figure.

Note that this is just an example, other services would also benefit from compute and connectivity traffic steering. For the sake of having a simpler service, we can also consider an AR/VR/XR service where a terminal connected to the network needs to instantiate a service in the network to aid in the AR/VR/XR service by providing computing capabilities with latency constraints.

Note on terminology. In this document we use the old terminology in which by ICR we mean Ingress CATS-Forwarder [I-D.ietf-cats-framework], and by ECR we mean Egress CATS-Forwarder.

```
                                ________________
                               (     ---------- )
                              (      |        |  )
                            (     ---------- |   )
      ________________     (     |        | |    )     _______________
     (     ---------- )    (   ---------- | |    )    (    ---------- )
    (      |        |  )   (   |service | |-     )   (     |        |  )
   (     ---------- |   )  (   |contact | |      )  (    ---------- |  )
   (     |        | |   )  (   |instance|--      )  (    |        | |  )
   (   ---------- | |   )   (  ----------       )   (  ---------- | |  )
   (   |service | |-    )    ( Serv. site #N-1 )    (  |service | |-   )
   (   |contact | |     )     -------+----------     (  |contact | |   )
   (   |instance|--    )   Computing  \              (  |instance|--   )
    (  ----------     )    delay:4ms   \              ( ----------     )
     ( Serv. site #1 )           --------+--           ( Serv. site #N )
      -------+--------       ----| ECR#N-1 |----        ---------+-----
              \  Computing --     -----------    --  Computing  /
               \ delay:10ms      Networking          delay:5ms /
             ---+-----           delay:7ms              ------+--
           ( | ECR#1 |            //                    | ECR#N | )
          (  ---------           //                     ---------  )
         ( Networking           //                       Networking )
        (  delay:5ms           //                         delay:15ms )
       (                      //                                      )
       (                     //                                       )
        (                   //                                       )
         (                 //                                       )
          (               //                                       )
           (       ---------                     ---------        )
            -------| ICR#1 |---------------------| ICR#2 |--------
                   ---------           __         ---------
                   (揃)   (揃)        / (  )           (揃)
                  (揃)   -------   -  (    )         (揃)
                 (揃)    | UE2 | /     (__) \      (揃)
                (揃)     -------    /         -   -------
               (揃)               /  (sensing  \  | UE3 |
             -------   ---------                 -------
             | UE1 | /
             -------
```

Figure 1: Exemplary scenario

1.2. Problem statement

The main problem that this document tries to address is the following. Networking systems do not have mechanisms yet to support service mobility optimized for connectivity and computing-aware traffic steering, which consider both connectivity and computing status to dynamically select where a service runs for a given terminal.

Based on the former, this document proposes solutions to enable the network to react and adapt to connectivity and computing conditions, triggering optimal service migration based on service-specific IP anchor mobility. In particular, this document addresses the following questions: (i) what mechanisms does the network need to implement to facilitate the migration of a service so its requirements in terms of computing and networking are maintained?; and, (ii) how to steer traffic to a new service instance location after moving the service, in a transparent manner to the terminal, by using IP anchor mobility?

2. Terminology

The following terms used in this document are defined by the IETF:

ECR: Egress CATS router. This refers to the Egress CATS-Forwarder as defined in [I-D.ietf-cats-framework].

ICR: Ingress CATS router. This refers to the Ingress CATS-Forwarder as defined in [I-D.ietf-cats-framework].

3. Enabling service continuity with IP anchor mobility for CATS

We describe next an example of operation and signaling for the network to perform service mobility. Three different solutions are described next, for variations (OPTIONS) of the procedures: terminal initiated, ECR-initiated and CATS-controller initiated. In addition to the functionality defined in [I-D.bernardos-cats-ip-address-anchoring], this documents defines a new functionality:

* Service mobility: it deals with the procedures required to (i) detect or predict a change of the current conditions, jointly considering computing and networking, requiring of a service mobility operation; (ii) selecting the best target service instance location, and (iii) triggering the service mobility by orchestrating the service anchor mobility and requesting service migration to a new site. For example, a terminal or ECR might use this functionality to perform active monitoring of a service with CATS agents running at the current ICR, ECR and or service site. It is also used to perform the actual service anchor mobility.

3.1. OPTION A: terminal-initiated

Next, we assume service mobility is triggered by a CATS-aware terminal. By having a CATS agent running on the terminal, it can perform different monitoring actions to predict or detect the need to migrate a service from one site to another. This CATS agent might, for example, interact with other CATS agents deployed on ICRs, ECRs and service sites.

In the following we describe a service anchor mobility procedure for CATS, initiated by a CATS-aware terminal. Following the terminal initiation, the network infrastructure is capable to select a target service instance meeting the connectivity and computing requirements of the service, with signaling procedures defined to perform a transparent anchor migration to a new site, facilitating the service migration in a transparent way for the terminal. Extensions and new behavior are highlighted. Note that variations are possible over this exemplary signaling diagram.

```
 +----+ +-----+ +-------------+ +---------------+ +-------------+ +----+
 |    | |     | |   site #1   | |   site #N-1   | |   site #N   | |CATS|
 |term| |ICR#1| |ECR#1 ag. SCI| |ECR#N-1 ag. SCI| |ECR#N ag. SCI| |ctrl|
 +----+ +-----+ +--+----+---+-+ +----+----+---+-+ +--+----+---+-+ +----+
    |       |       |    |   |        |    |   |       |    |   |      |
    |       |  O. Service instance running at site #n-1.    |   |      |
    |       |     Tunnel established between ICR#1 and ECR#N-1 |      |
    |       |       |    |   |        |    |   |       |    |   |      |
    |    1. An external or internal trigger is received regarding      |
    |        a change in the compute or networking conditions          |
    |       |       |    |   |        |    |   |       |    |   |      |
    |1a. CATS agent at terminal reports a perceived change on service|
    |揃揃揃揃揃>|       |    |   |        |    |   |       |    |   |      |
    |       |       |    |   |        |    |   |       |    |   |      |
    |       |2a.CATS query(service ID, terminal ID, ICR ID, CATS reqs.)
    |       |揃揃揃揃揃揃>|<揃揃>|    |        |    |   |       |    |   |      |
    |       |揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃
揃>|<揃揃>|    |       |
    |      3a.CATS response(service ID, terminal ID, ECR ID, CATS cond.)
    |       |<揃揃揃揃揃揃|    |   |        |    |   |       |    |   |      |
    |       |<揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃
揃揃|     |   |       |
    |       |       |    |   |        |    |   |       |    |   |      |
    |      4a. Service anchor/ECR@site#n is selected as best  |      |
    |       |       |    |   |        |    |   |       |    |   |      |
 5a.CATS request(service ID, terminal ID, ICR ID, CATS reqs., IP pref.)
    |揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃
揃揃揃揃揃>|     |   |       |
    |       |       |    |   |        |    |   |       |    |   |      |
    |6.CATS ACK(service ID, terminal ID, ECR ID, CATS cond., IP prefix)
    |<揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃揃
揃揃揃揃揃|     |   |       |
    |       |       |    |   |        |    |   |       |    |   |      |
    | 7.New tunnel for IP pref. established beetween ICR#1 and ECR#N |
    |       |       |    |   |        |    |   |       |    |   |      |
    |       | 8.Service migration from site#N-1 to site#N  |   |      |
    |       |       |    |   |        |    |   |       |    |   |      |
    |      9. Service specific traffic|    |   |       |    |   |      |
    |<---->|<---------------------------------------->|<------>|      |
    |       |       |    |   |        |    |   |       |    |   |      |
```

Figure 2: Exemplary signaling

Figure 2 shows the message sequence chart of the IP anchor mobility for CATS, initiated by a CATS-aware terminal (OPTION A), which is explained next:

0. A service/app has been already instantiated on a service site (for example, by using the mechanisms specified in [I-D.bernardos-cats-ip-address-anchoring]). In this example, the site where the service/app consumed by the terminal is currently instantiated is site #n-1. The terminal is connected to ICR #1

and there is a service tunnel between ICR#1 and ECR#N-1. This service/app requires some functionality to be run on the network infrastructure (e.g., an AR/VR/XR service). This service has specific requirements in terms of both connectivity and computing (CATS requirements).

1. An internal or external trigger is generated regarding a change in the computing of networking conditions, making the current selected service site not feasible for the running service (i.e., CATS requirements cannot be met). In this OPTION A, the terminal is CATS-aware, and it the node that detects the change in the conditions (how this is made is outside the scope of this document, but possible mechanisms include: monitoring at service/app level, in-situ monitoring by the terminal, etc.) and sends a trigger to the ICR.

2. The ICR sends a query to all (but currently used) ECRs of the domain, or a subset selected based on the location of the ICR. This query may include the following parameters:

    i. Service ID: an identifier of the service requested by the terminal. This allows to check if the service can be instantiated or it is already instantiated.

    ii. Terminal ID: an identifier of the terminal requesting the service. This is useful for example for affinity purposes. It might not include information that can be used to identify the user.

    iii. ICR ID: identifier of the requesting ICR.

    iv. CATS requirements: list of requirements, e.g., connectivity and computing requirements.

3. Each ECR, possibly after checking with the CATS agent of the site(s) it provides connectivity, responds, including the following information:

    i. The ICR sends a query to all ECRs of the domain, or a subset selected based on the location of the ICR. This query may include the following parameters:

        i. Service ID.

        ii. Terminal ID.

        iii. ECR ID: identifier of the ECR sending the response.

iv. CATS conditions: how the site meets each of the requirements included in the request.

v. (Optional): URI to get to the service instance.

A CATS agent at a site might be collocated with the ECR. Examples of a CATS agent at a site are network controllers or orchestrators at the site. Note that the way a CATS agent at an ECR may interact with the CATS agent of the site is out of the scope of this document. Examples include using monitoring and telemetry interfaces with an orchestrator managing the site.

ii. Based on the received responses, and considering both networking and computing metrics and policies, the ICR selects an ECR (#n).

iii. The ICR requests the proposed/selected ECR to establish a traffic steering session with it, sending a CATS request. This request includes the same information that was included in the CATS query (to facilitate stateless operation of the ECRs while being queried), plus the following additional parameters:

* ECR prefix: currently in use IP prefix IP to the terminal to reach the service instance.

* Lifetime: requested duration for the association between the ICR and the ECR.

iv. The selected ECR responds back with an acknowledgement, including the following information:

i. Service ID.

ii. Terminal ID.

iii. ECR ID: identifier of the ECR sending the response.

iv. CATS conditions: how the site meets each of the requirements included in the request.

v. IP prefix assigned for the terminal to use to reach the service instance. It should match the one included in the request.

vi. Lifetime: granted duration of the association between the ICR and the ECR.

v. An IP tunnel is established between the ICR and the new ECR. Optionally (not shown in the figure), the ICR might send a CATS request with zero lifetime to the old ECR to remove the old tunnel.

vi. The previous message triggers the service migration (from site #n to site #n-1). The specific mechanism is out of the scope of this document. Note that some preparation/ migration steps might be conducted in parallel (e.g., after messages #1 and #3) to accelerate the process, making this step just the final trigger for the service migration. At site #n, the prefix used by the terminal for accessing the service is configured to be used by migrated instance. This might requires routing updates to be perfomed in the site, potentially controlled by a CATS agent running in the site.

vii. Traffic of the service for this terminal is steered using the IP tunnel.

3.2. OPTION B: ECR-initiated

TBD.

3.3. OPTION C: CATS controller-initiated

TBD.

4. Proxy Mobile IPv6 signaling extensions to enable service mobility with IP address service-specific anchoring for CATS

The control plane extensions introduced in the previous section can be implemented over different protocols. This section specifies extensions to Proxy Mobile IPv6. Only new options additional to what is defined in [I-D.bernardos-cats-ip-address-anchoring] are included.

4.1. New ECR mobility option

The new ECR option has the following format:

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |      Type     |   Length      |   Reserved    | Addr. Length  |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    +                                                               +
    |                                                               |
    +                     New ECR IP address                        +
    |                                                               |
    +                                                               +
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Message fields:

* Option Type: TBA by IANA.

* Option Length: 8-bit unsigned integer. Length of the option, in octets, excluding the Option Type and Option Length fields.

* Addr. Length: 8-bit unsigned integer. Length of the New ECR IP address field, in octets.

* New ECR IP address: variable length field that includes IP address of the new ECR.

5. IANA Considerations

TBD.

6. Security Considerations

TBD.

7. Acknowledgments

The work of Carlos J. Bernardos in this document has been partially supported by the Horizon Europe MultiX (Grant 101192521), DISCO6G-CM (TEC-2024/COM-360) and UNICO I+D 6G-DATADRIVEN projects.

8. Informative References

[I-D.bernardos-cats-ip-address-anchoring]
Bernardos, C. J. and A. Mourad, "Computing Aware Traffic Steering using IP address anchoring", Work in Progress, Internet-Draft, draft-bernardos-cats-ip-address-anchoring-03, 26 February 2025, <https://datatracker.ietf.org/doc/html/draft-bernardos-cats-ip-address-anchoring-03>.

[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-05, 10 February 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-cats-framework-05>.

Authors’ Addresses

Carlos J. Bernardos
Universidad Carlos III de Madrid
Av. Universidad, 30
28911 Leganes, Madrid
Spain
Phone: +34 91624 6236
Email: cjbc@it.uc3m.es
URI: http://www.it.uc3m.es/cjbc/

Alain Mourad
InterDigital Europe
Email: Alain.Mourad@InterDigital.com
URI: http://www.InterDigital.com/