

CATS Problem Statement and Use cases

draft-ietf-cats-usecases-requirements-00

K. Yao, China Mobile

D. Trossen, Huawei

M. Boucadair, Orange

LM. Contreras, Telefonica

H. Shi, Y. Li, Huawei

S. Zhang, China Unicom

Draft status

Groundwork defined in CATS Charter

The CATS WG is chartered to work on the following items:

o Groundwork may be documented via a set of informational Internet-Drafts, not necessarily for publication as RFCs:

- Problem statement for the need to consider both network and computing resource status.
- Use cases for steering traffic from applications that have critical SLAs that would benefit from the integrated consideration of network and computing resource status.
- Requirements for commonly agreed computing metrics and their distribution across the overlay network, as well as the appropriate frequency and scope of distribution.

Table of Contents

1. Introduction	3
2. Definition of Terms	4
3. Problem Statement	5
3.1. Multi-deployment of Edge Sites and Service	5
3.2. Traffic Steering among Edges Sites and Service Instances	6
4. Use Cases	9
4.1. Computing-Aware AR or VR	10
4.2. Computing-Aware Intelligent Transportation	13
4.3. Computing-Aware Digital Twin	14
4.4. Computing-Aware SD-WAN	15
5. Requirements	17
5.1. Support dynamic and effective selection among multiple service instances	17
5.2. Support Agreement on Metric Representation	17
5.3. Support Moderate Metric Distributing	18
5.4. Support Flexible Use of Metrics	18
5.5. Support Session and Service Continuity	19
5.6. Preserve Communication Confidentiality	21
6. Security Considerations	21
7. IANA Considerations	21
8. Contributors	22
9. Acknowledgements	22
10. References	22
10.1. Normative References	22
10.2. Informative References	23
Authors' Addresses	23

Introduction

- User demands have driven the fast development of converged compute and network infrastructure
 - **low latency, high reliability**, as well as **stability**, etc...
- How to meet user requirements? Key problems:
 - **Service instance deployment** (need to think about locations, capability, etc...)
 - **Traffic scheduling** (dynamically steer traffic to the “best” service instance.)
- However, the **problem is the “closest” might NOT be the “best”**.

Key Definitions

➤ Revised definition on Service & Service instance:

- **Service:** An offering provided by a service provider, similar to the notion of a 'service function' in [RFC7665], which may or may not be of composite nature but appears in the problem space of CATS as a single service to which traffic needs to be steered.
- **Service instance:** A run-time environment (e.g., a server or a process on a server) that makes a service available. A particular service could be made available at multiple service instances at the same or different locations.

Problem Statement

➤ Multi-deployment of Edge Sites and Service:

➤ Factors need to be considered:

- Network and computing resource topology.
- Locations of users and service instances.
- Capacity of multiple edge nodes.
- Service category.

➤ Traffic Steering among Edges Sites and Service Instances:

➤ Problems when “closest” site is not the “best”:

- The closest site may not have enough resources, load may dynamically change.
- The closest site may not have related resource, heterogeneous hardware in different sites.
- The network path to the closest site might not provide the necessary network characteristics.

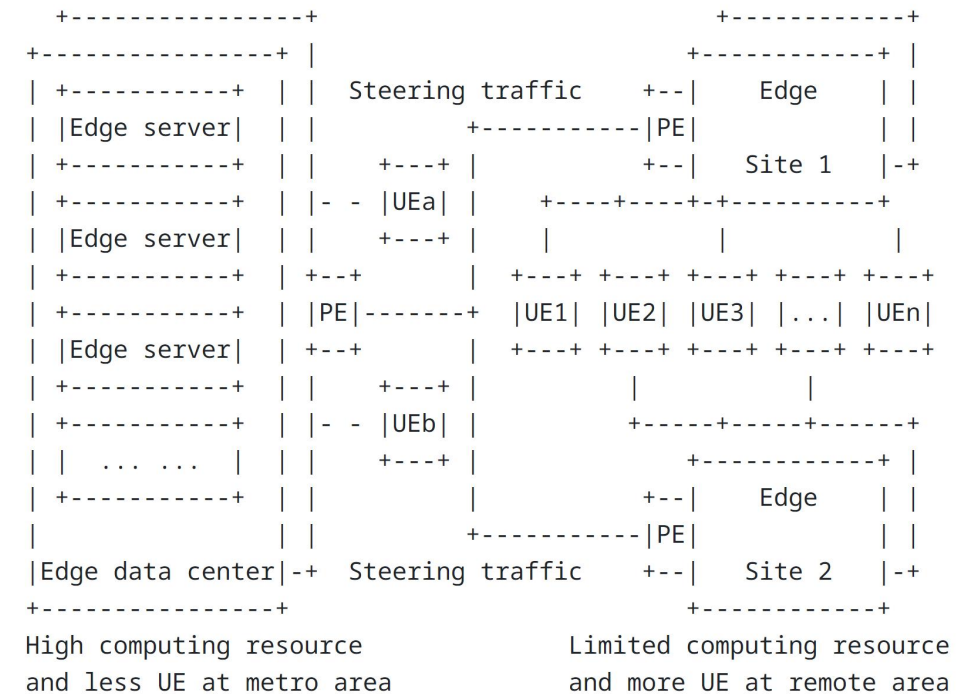
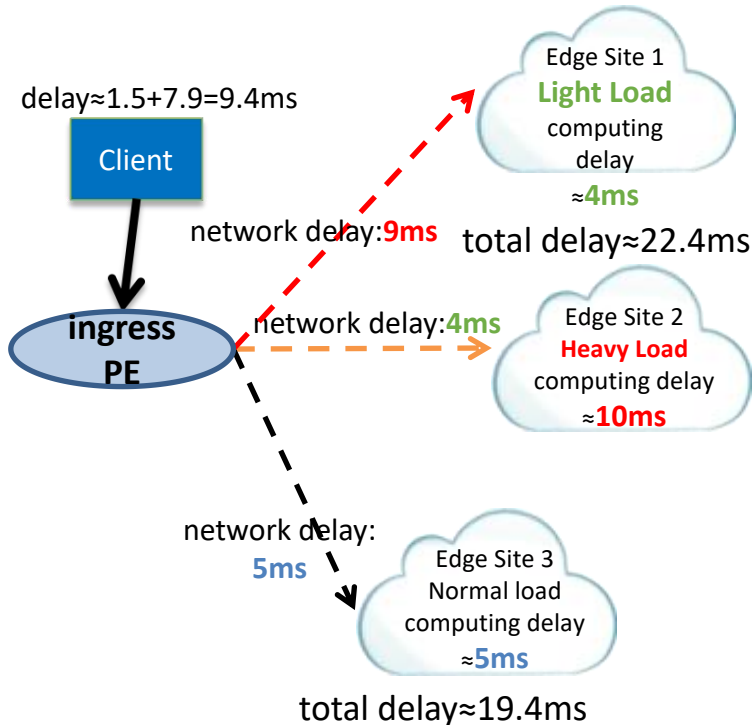


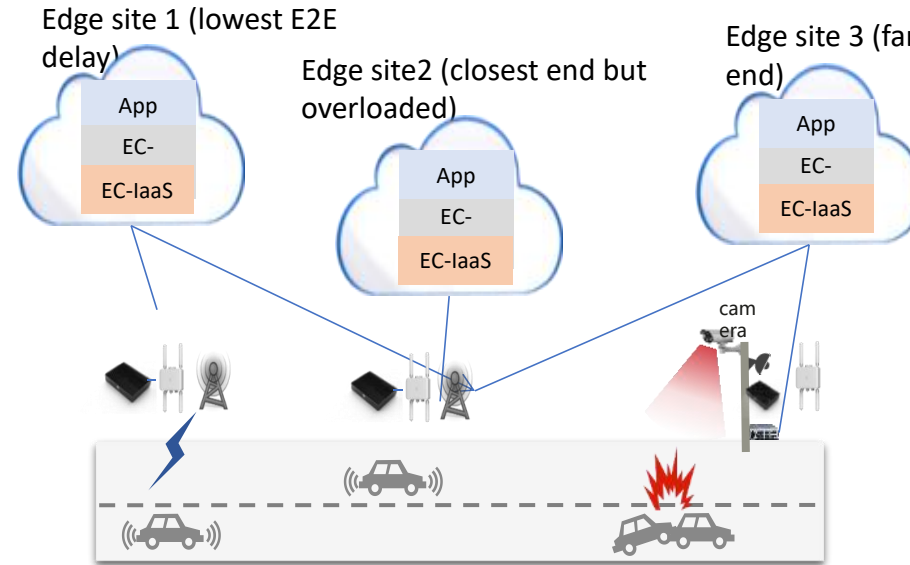
Figure 2: Steering Traffic among Edge Sites

Use cases:

Computing-Aware AR/VR



Computing-Aware V2X



Autonomous driving: Shorter latency, better safety.

For example. If the latency is reduced by 100 ms, the braking distance of a vehicle at 80 km/h can be reduced by **2.2 meter**.

Computing-Aware DT

- Programmable logic controller (PLC) may be virtualized and deployed at different edge sites.
- Several PLC instances may exist to enable load balancing and fail-over capabilities.
- High availability is needed.
- Storage instances deployed at multiple sites for digital twin should also be aware.

Require to dynamically steer traffic to the appropriate edge to meet the E2E delay requirements by considering **both network and computing resource status**

Use Cases: Computing-Aware SD-WAN

- Consider about cloudification deployment
- SD-WAN should be aware of the computing resources of vCPE

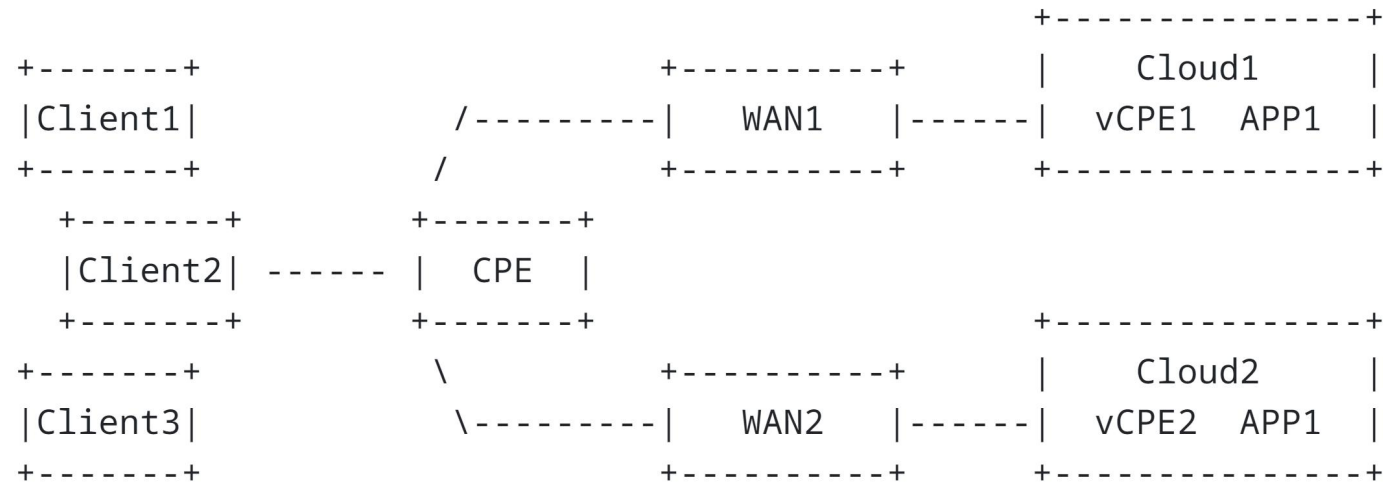


Figure 4: Illustration of Computing-aware SD-WAN for Enterprise Cloudification

Outstanding Comments:

- The description on problem statement and use cases is somewhat verbose, which should be more concise.
- There should be a clear definition on edge computing.
- Use cases appear to be very high-level, which should be more specific to derive requirements.
- There might be simpler ways to solve problems in SD-WAN use case.
- Policies should also be considered in SD-WAN use case.

Next Steps:

- Address these valuable comments in the next revised version.
- Consider the relationship with other newly submitted use cases drafts in the WG.
- Welcome discussions.

Thank you!