

Applicability of CATS Framework

T. Jiang, L. Contreras, M. Watts, CJ. Bernardos, Y. Shang, P. Liu

Nov. 01 ~ 07, 2025, IETF 124 @ Montreal,
Canada

Agenda

- CATS WG Documents:
 - Framework, Use Cases/Reqs., and Metrics
- CATS Framework & Applicability
- Scenario#1: Applicability of Framework to ACN (AI-agent Communication Network)
- Scenario#2: Applicability of Framework to 5G eEdge (Edge Enhancement)
- Scenario#3: Applicability of Framework to O-RAN Midhaul

CATS WG Drafts

Use cases & Requirements

- Provide a list of typical CATS scenarios (or use cases) and outline the requirements for the CATS to realize the scenarios.

Metrics

- Define a set of metrics from the computing domain used for CATS.

Framework

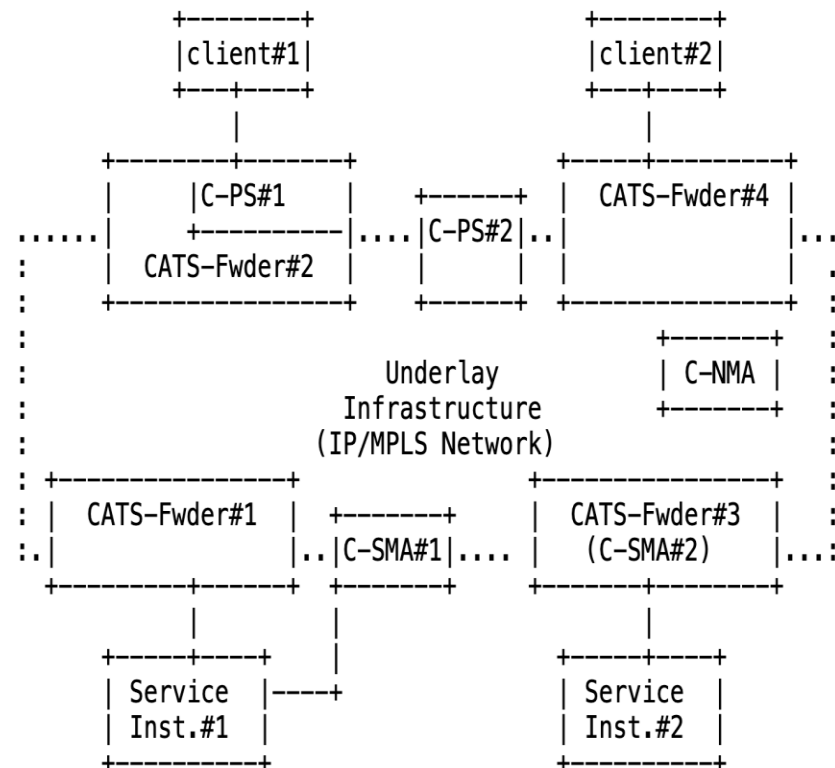
- Propose and define a general architecture for the distribution of network and compute metrics and for the transport of traffic from network edge to service instance
- Encompass various building blocks, realize CATS CP and DP for the objective of 'CATS optimization'.

CATS Framework: Standardize a general CATS Architecture

- ❖ Abstract & Identify the CATS components and their interactions; illustrate the workflows of both the CP and DP.
- ❖ Functional elements: CP (C-PS, C-SMA, C-NMA), DP (CATS-Fwder, C-TC, etc.)
 - (CP) C-SMA: gather info. about service sites (e.g. server resources), status of service instances, etc.
 - (CP) C-NMA: gather info. about the state of the underlay network.
 - (DP) CATS-Fwder (Ingress/Egress): steer the service traffic toward a 'suitable' service instance based on network- and compute- metrics

Deployment of CATS functional elements:

- Make no assumption of the model
- May only reflect the preferences and policies of the CATS service provider
- Bode well for drafting a document to illustrate the applicability of the CATS framework to various scenarios, e.g., ACN, 5G Edge, and the O-RAN Midhaul.



Scenario#1: Applicability of Framework to ACN (AI-agent Communication Network) (ACN Recap.)

AI agents:

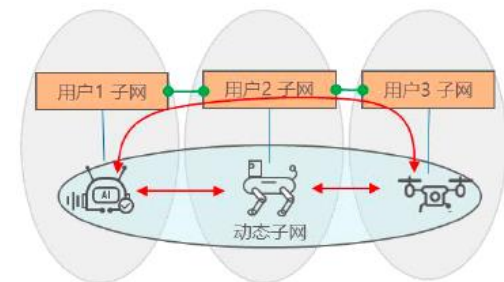
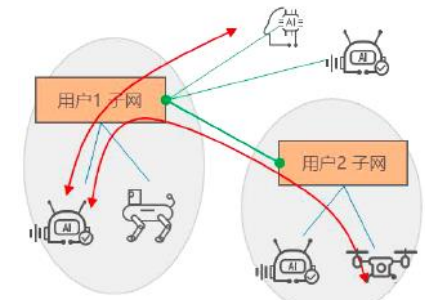
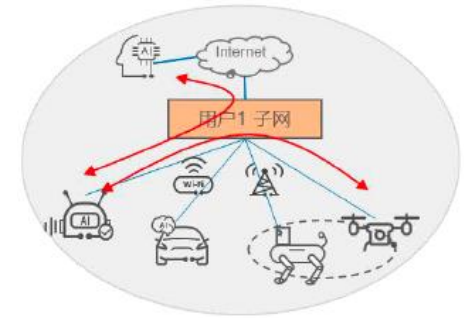
- SW-driven entities with embedded AI; instantiate either in the physical state as embedded HW devices (e.g., robots) or in the virtual state (e.g., SW-implemented Apps)
- Crucial in network domain (particularly in Mobile space)

AI-agent Communication Network (ACN):

- An interconnected network framework tailored to advance the communication among AI-agents. It satisfies the on-demand communication and could reach the global coverage.

ACN Realization: 3 network communication models

- Static intra-domain only ACN: the model for AI-agents residing within a *single administrative domain or network*.
- Static inter-domain ACN: the model for AI-agents that reside *across multiple administrative domains*.
- Dynamic multi-domain ACN: the model for AI-agents that may dynamically form a network group to handle a temporarily-generated task.

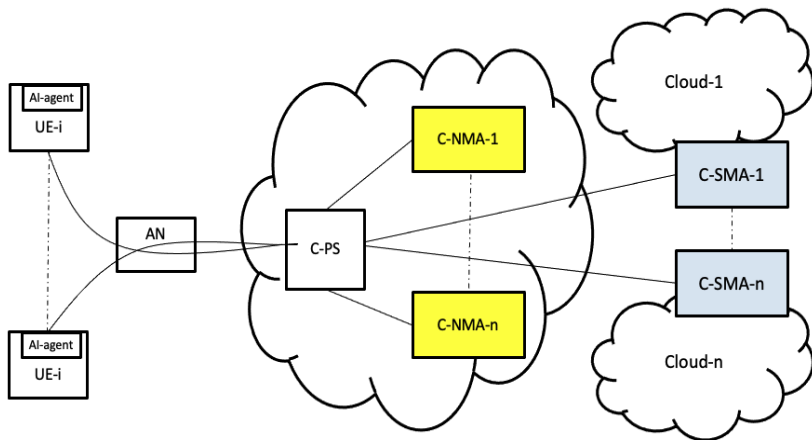


Scenario#1: Applicability of Framework to ACN

CATS-optimization for AI agent services:

Dramatic variations among AI demands, data processing tasks and capabilities of AI agents

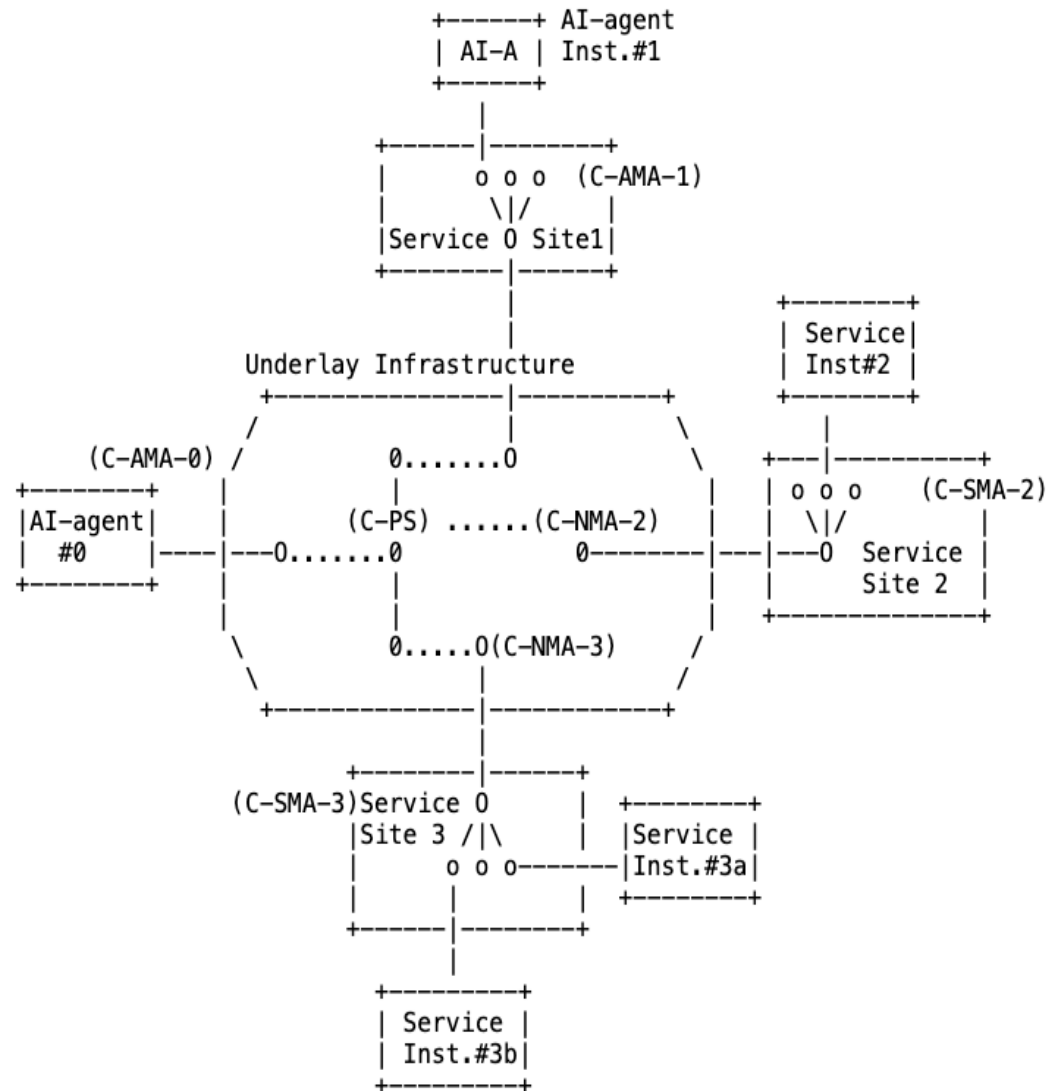
- Existing CATS functional elements (C-PS, C-NMA, C-SMA)
- Extension w/ ACN: C-AMA



CATS functional elements (w/ ACN extension)

- AI-agents (#0, #1)
- Service sites & service instances
- C-PS, C-NMA-x, C-SMA-y, C-AMA-z

C-AMA: CATS AI-agent Metric Agnet (May integrate with C-SMA)

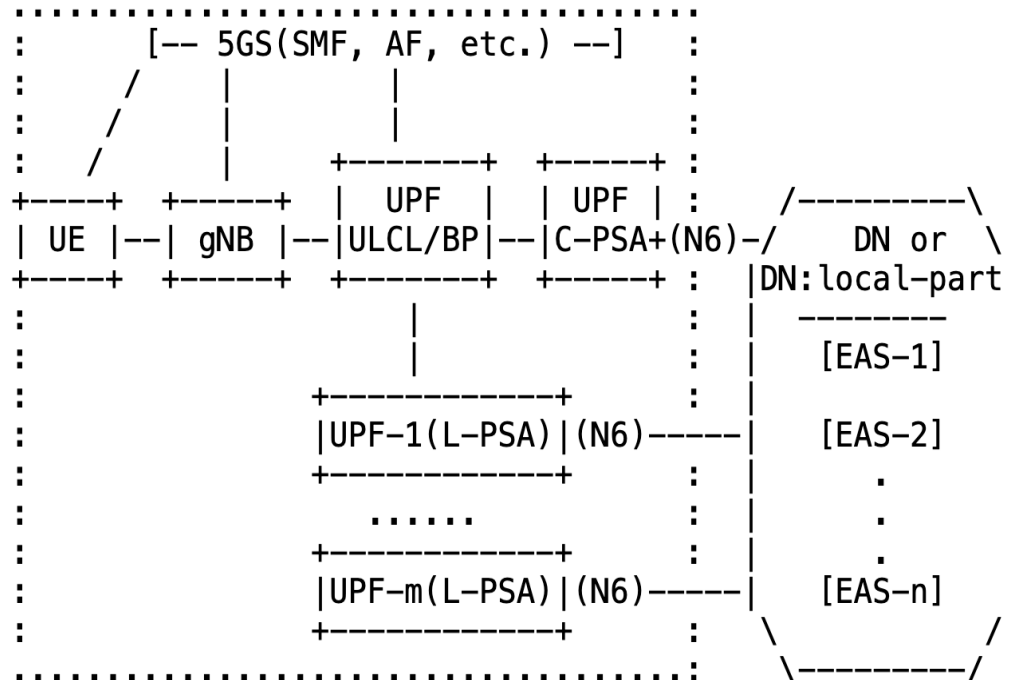


Scenario#2: Applicability of Framework to 5G eEdge

(Background) Computing-aware 5G eEdge:

- Services to be hosted close to an end device's access point of attachment
- Reduced end-to-end latency and load on the TN/transport network
- (CATS) Network Metric: N6 delay, over TN/transport network
- (CATS) Compute Metric: EAS-load (for future), with EAS'es located in local DN/DC.

C-PS <== SMF
 C-NMA, C-SMA <== SMF & AF
 CATS-Forwarder <== C/L-PSA UPF



C-PS <== SMF

- SMF selects the optimal UPF (from all candidates: UPF, UPF-1, ..., UPF-M) based on the network metric (N6-delay) as of now

C-NMA, C-SMA

- From AF-influenced procedure as specified in TS 23.502
- AF+SMF combination ==> C-NMA;
- C-SMA not-determined now, target for future.

CATS-Fwder <== C/L-PSA UPF

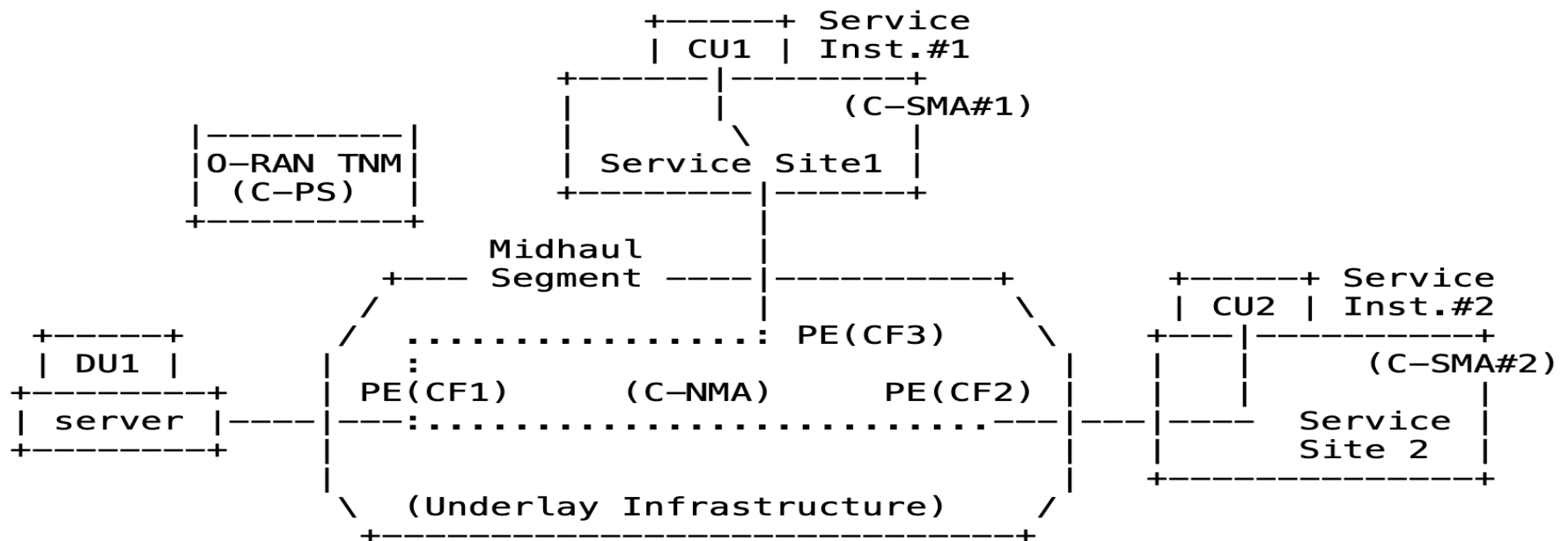
Scenario#3: Applicability of Framework to O-RAN Midhaul

(Background) CATS-optimization for O-RAN Midhaul:

- O-RAN splits radio functions into RU, DU, and CU; Midhaul connects DU to CU over an IP/MPLS aggregation network (the TN or transport network).
- Midhaul traffic steering can dynamically select CU instances based on both the CATS compute metrics (CPU, memory) and the network metrics (bandwidth, latency metrics).

Applicability of framework to O-RAN Midhaul:

- **C-NMA** and **C-SMA**: View on networking and compute metrics, e.g., CPU average utilization or the memory and energy usage of every CU-UP instance.
- O-RAN TNM (**C-PS**) : decide to which CU (or Service) instance to deliver the DU traffic
- O-RAN TNE (**CATS-Fwder**) as the PE node



Summary

- Discuss the applicability of the CATS framework
- Apply the CATS framework to some noteworthy scenarios:
 - ACN or AI-agent communication network, 5G eEdge, O-RAN midhaul.

Is the CATS WG interested in
pursuing *an applicability draft
applying the CATS framework?*

If yes, using this document as the
baseline?

Thank you!