# Exposure of Communication and Compute Information for Infrastructure-Aware Service Deployment and Selection

https://datatracker.ietf.org/doc/draft-rcr-opsawg-operational-compute-metrics/
https://datatracker.ietf.org/doc/draft-contreras-alto-service-edge/

Sabine Randriamasy (Nokia Bell Labs), Luis Contreras (Telefonica), Jordi Ros Giralt (Qualcomm Europe, Inc.)

# Content

Problem space: service lifecycle

What is this topic about?

Interest to IETF

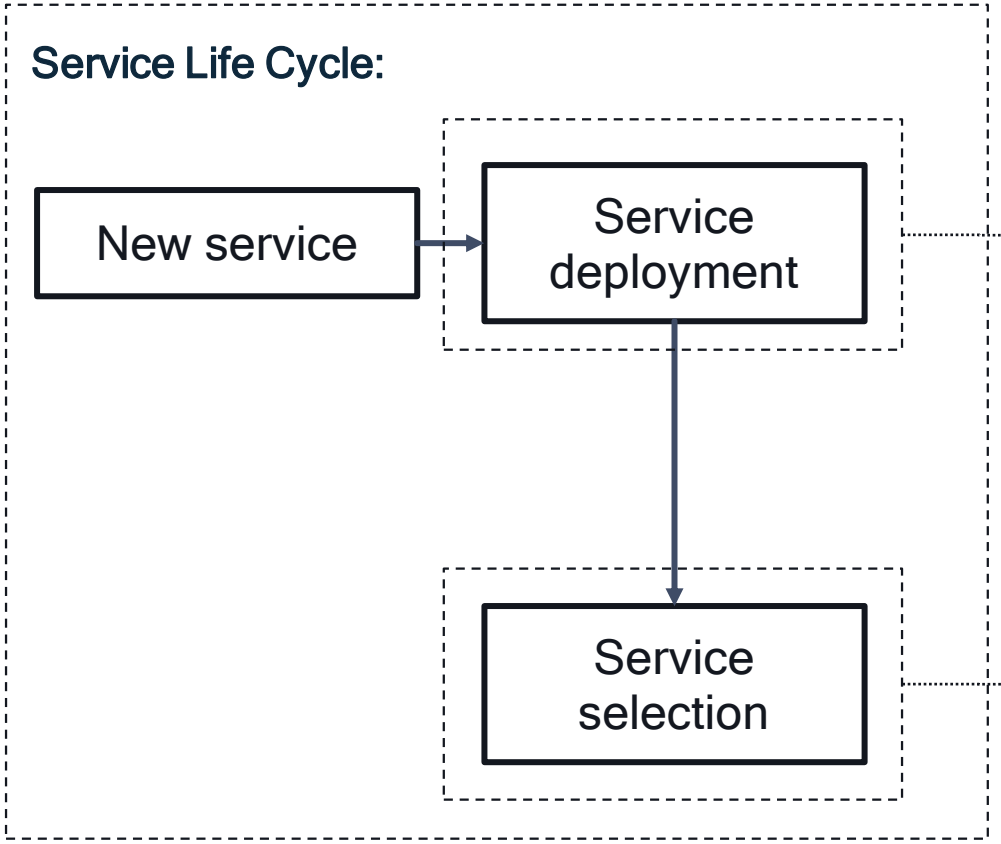
Use cases

Defining compute metrics at IETF (CATS and ALTO conversations)

Guiding principles

Desired outcome of this presentation

# Problem Space: Service Lifecycle and Information Exposure




| Action to take | Information needed | Who needs it | IETF WGs |
|---|---|---|---|
| • Service placement | • Compute<br>• communication | • Service provider | • ALTO |

| Action to take | Information needed | Who needs it | IETF WGs |
|---|---|---|---|
| • Service selection<br>• Path selection | • Compute<br>• communication | • Network provider | • CATS |
| | | • Application (client or proxy) | • ALTO |

# What is this topic about?

Seeking rough consensus on these four questions:

- Q1. Is it likely/viable that the network can expose communication and compute information to the service provider and application?
- Q2. Are there gaps in the entire service lifecycle (deployment/instantiation/selection) that are not currently being addressed and that are relevant?
- Q3. Would it make sense to define a common set of communication and compute metrics to address the various service lifecycle stages?
- Q4. If so, where should this effort be carried out within the IETF?

# Interest to the IETF

**Use cases**. The arrival of a new class of applications with stringent compute and communication requirements: distributed generative AI, XR/VR, vehicle networks, metaverse.

**Industry trend**.

- Linux CAMARA. “Reserve compute resources within the operator network”. “Influence the traffic routing from the user device toward the Edge instance of the Application”.

- GSMA Open Gateway. 21 operators to open up network APIs for developers.

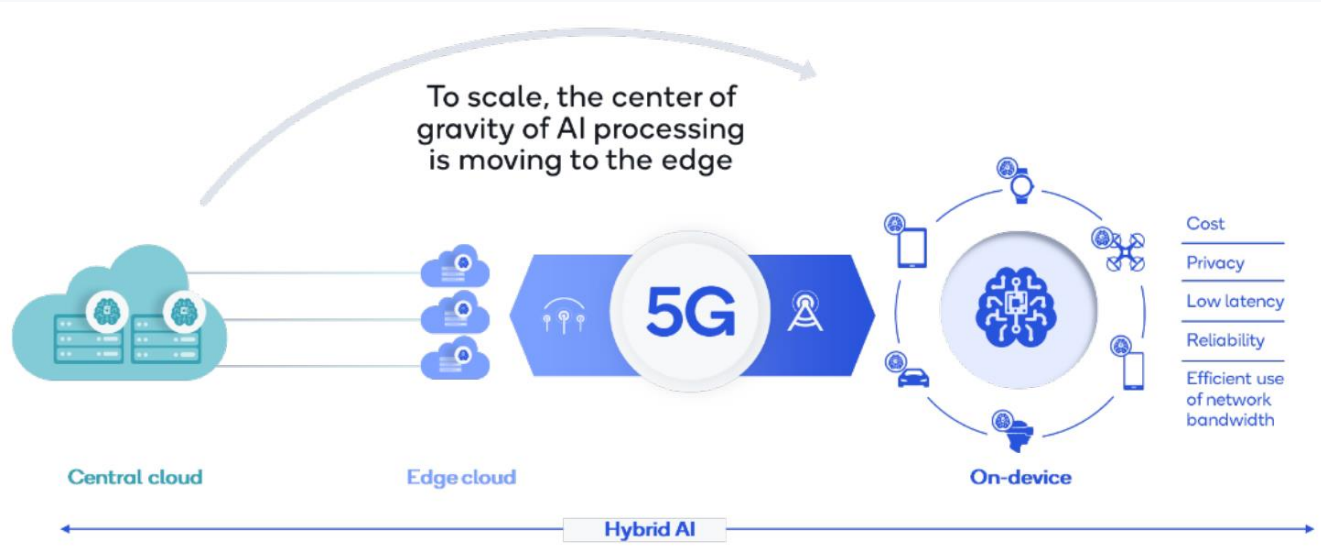
- 3GPP NEF. Enable exchange of information to/from an external application in a controlled and secure way.

**Posit.** There is a need for a structured/organized way to access this information from the network layer to avoid uncoordinated, ad hoc (thus inefficient) mechanisms.

# Examples of Use Cases

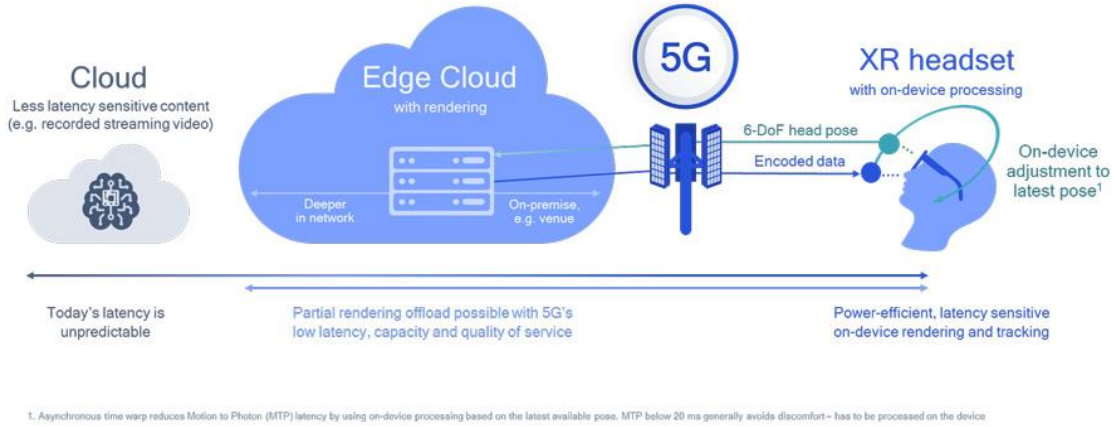
https://datatracker.ietf.org/doc/draft-contreras-alto-service-edge/

## Distributed AI computation




- Larger, mid-size, and smaller AI models are run in the cloud, the edge, and the device, respectively, enabling a trade-off between model accuracy and computational cost.
- To make proper service deployment/selection decisions at the application level, knowing compute information is key in today's edge computing applications. Without such information, resources and energy are wasted, and application performance severely degrades.

## Distributed XR computation




- On-device rendering is augmented by high-performance edge cloud graphics rendering over a high-capacity low-latency 5G connection.
- Select the best communication (e.g., 5G and Wi-Fi) and compute (device, edge, and cloud) combination to distribute processing between XR headset, edge, and cloud is crucial to avoid wasting energy and ensure the performance of the application.

# Defining Compute Metrics at the IETF

- Standardization of network information is quite mature but is in progress for compute information.
- There is a need to define a set of compute metrics to support various use cases being served in the IETF.
- Some ad hoc work exists in the IETF:
  - CATS (e.g., draft-du-cats-computing-modeling-description)
  - ALTO (e.g., draft-contreras-alto-service-edge)
  - OPSAWF (e.g., RFC 7666 MIB)
- Metrics are also defined in other bodies such as the Linux Foundation, DMTF, ETSI/NFVI:
  - Raw compute infrastructure metrics (e.g., processing, memory, storage)
  - Compute virtualization resources and service quality metrics (e.g., VNF resources in VMs)
  - Service metrics including compute-related information (e.g., service delay, availability)

# Defining Compute Metrics at the IETF: CATS/ALTO conversations IETF 117

- CATS charter https://datatracker.ietf.org/group/cats/about/
  - “There is a need for a general framework for the distribution of compute and network metrics and transport of traffic from network edge to service instance. It also is likely that some set of common metrics can be identified”.
  - “Exposure of network and compute conditions to applications is not in the scope of CATS”.
  - E.g., CATS drafts:
    - draft-du-cats-computing-modeling-description
    - draft-wang-opsawg-service-information-yang
- IETF 117 presentation by A Farrell to ALTO WG inviting to common work on defining metrics:
  - https://datatracker.ietf.org/doc/slides-117-alto-compute-aware-metrics-cats-working-with-alto/

# Guiding Principles

- P1. Leverage metrics across working groups to avoid reinventing the wheel. Examples:
  - RFC-to-be 9439 [I-D.ietf-alto-performance-metrics] leverages IPPM metrics from RFC 7679: https://datatracker.ietf.org/doc/draft-ietf-alto-performance-metrics/
  - Section 5.2 of [draft-du-cats-computing-modeling-description]: delay as a good metric (same units for compute and communication). ALTO defines network delay in its RFC-to-be 9439.
  - Section 6 of [draft-du-cats-computing-modeling-description]: “The network structure can be represented as graphs”. Similar to the ALTO map services (RFC 7285).
- P2. Ensure the combined efforts in the IETF don’t leave gaps in supporting the full lifecycle of service deployment and selection.
  - Example: CATS/ALTO potential cooperation/coordination on metrics to cover both service deployment and service/path selection:
    - CATS focus is on in-network service and path selection.
    - ALTO focus is on application-level service deployment and application-level service selection.

* Note: s/ALTO/X-WG/

# Desired Outcome

Seeking rough consensus on these four questions:

- Q1. Is it likely/viable that the network can expose communication and compute information to the service provider and application?
- Q2. Are there gaps in the entire service lifecycle (deployment/instantiation/selection) that are not currently being addressed and that are relevant?
- Q3. Would it make sense to define a common set of communication and compute metrics to address the various service lifecycle stages?
- Q4. If so, where should this effort be carried out within the IETF?