Joint Exposure of Network and Compute Information for Infrastructure-Aware Service Deployment

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

IETF 119 side meeting material: https://github.com/communication-compute-exposure/ietf-side-meetings/tree/main/ietf-119-side-meeting

Sabine Randriamasy (Nokia Bell Labs), Luis Contreras (Telefonica), Roland Schott (Deutsche Telekom), Jordi Ros-Giralt (Qualcomm Europe, Inc.)

Note Well

This is a reminder of IETF policies in effect on various topics such as patents or code of conduct. It is only meant to point you in the right direction. Exceptions may apply. The IETF's patent policy and the definition of an IETF "contribution" and "participation" are set forth in BCP 79; please read it carefully.

As a reminder:

- By participating in the IETF, you agree to follow IETF processes and policies.
- If you are aware that any IETF contribution is covered by patents or patent applications that are owned or controlled by you or your sponsor, you must disclose that fact, or not participate in the discussion.
- As a participant in or attendee to any IETF activity you acknowledge that written, audio, video, and photographic records of meetings may be made public.
- Personal information that you provide to IETF will be handled in accordance with the IETF Privacy Statement.
- As a participant or attendee, you agree to work respectfully with other participants; please contact the ombudsteam (https://www.ietf.org/contact/ombudsteam/) if you have questions or concerns about this.

Definitive information is in the documents listed below and other IETF BCPs. For advice, please talk to WG chairs or ADs:

- BCP 9 (Internet Standards Process)
- BCP 25 (Working Group processes)
- BCP 25 (Anti-Harassment Procedures)
- BCP 54 (Code of Conduct)
- BCP 78 (Copyright)
- BCP 79 (Patents, Participation)
- https://www.ietf.org/privacy-policy/(Privacy Policy)

IETF

Content

- Recap from Side Meeting IETF 118
- Use Cases
- Questions to be discussed in this side meeting

Note: we plan to request a mailing list, write to us if you are interested in joining it:

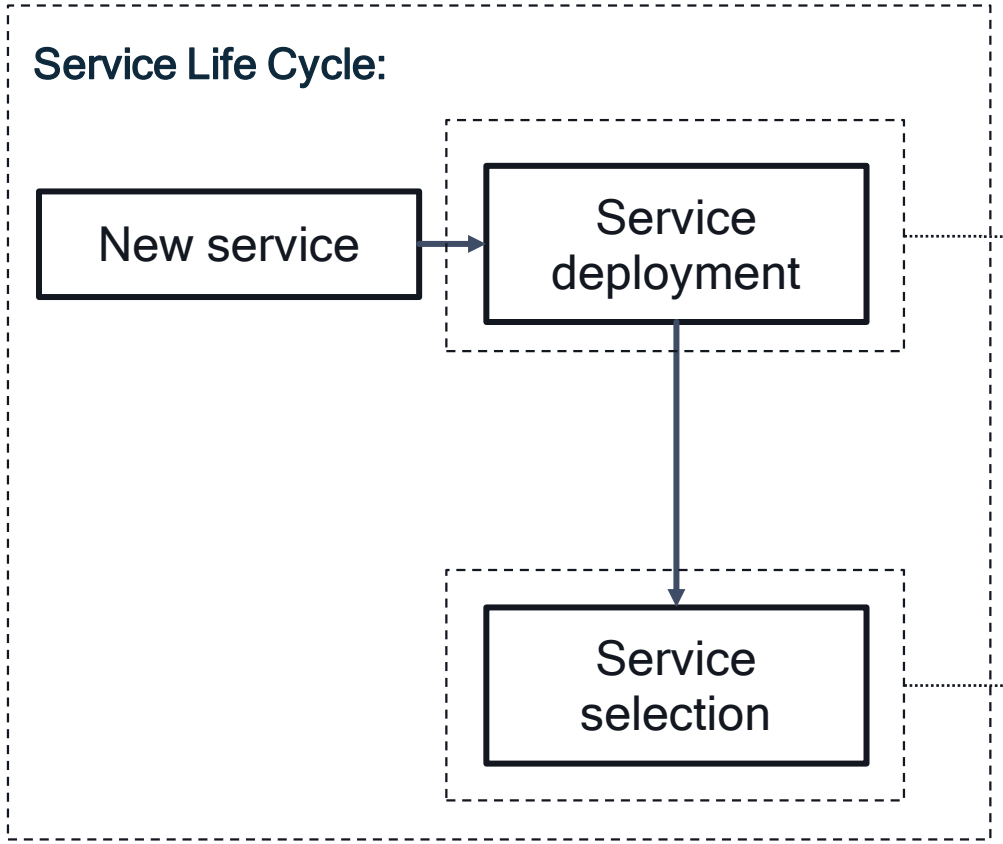
S. Randriamasy sabine.randriamasy@nokia-bell-labs.com

L. M. Contreras luismiguel.contrerasmurillo@telefonica.com

Roland Schott Roland.Schott@telekom.de

Jordi Ros-Giraltjros@qti.qualcomm.com

Problem Space: Service Lifecycle and Information Exposure



Action to take	Information needed	Who needs it
• Service placement	• Compute • communication	• Service provider

Action to take	Information needed	Who needs it
• Service selection • Path selection	• Compute • communication	• Network provider • Application (client or proxy)

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.

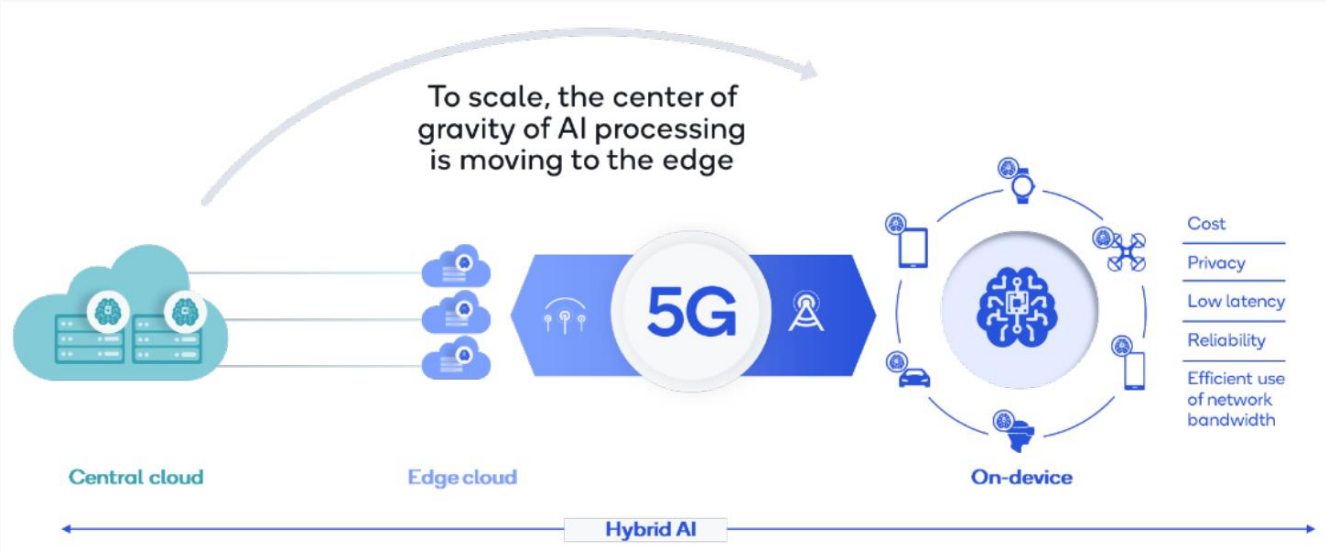
- Linux CAMARA. “Reserve compute resources within the operator network”. “Influence the traffic routing from the user device toward the Edge instance of the Application”.
- GSMA Open Gateway. 21 operators to open up network APIs for developers.
- 3GPP NEF. Enable exchange of information to/from an external application in a controlled and secure way.

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

Use Cases

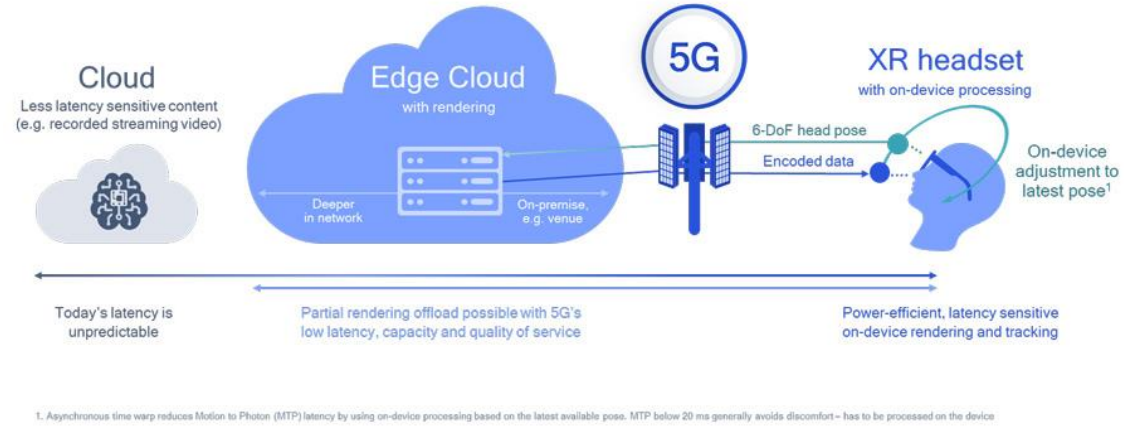
https://datatracker.ietf.org/doc/draft-contreras-alto-service-edge/

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.

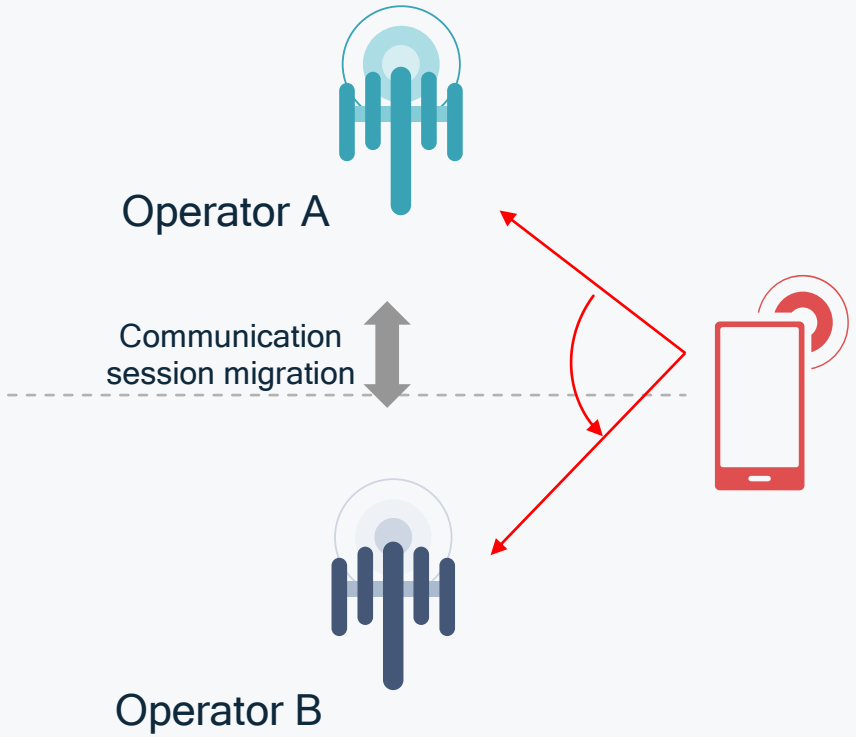
Distributed XR computation



- 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.

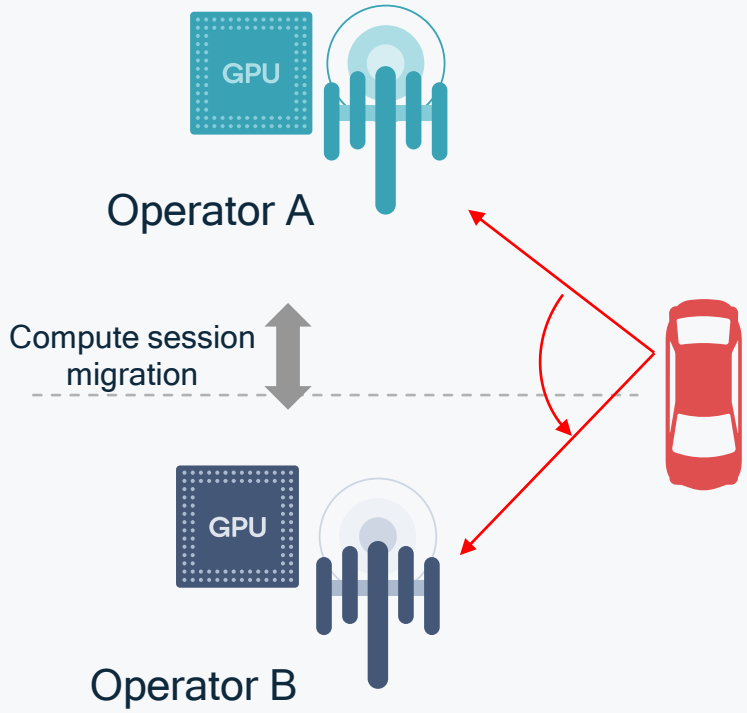
Use Cases

Voice: Seamless Communication Handover



- Today we are able to perform seamless communication handovers for applications such as voice communication.
- We already have open standards to cover this area.

Driverless Vehicles: Seamless Compute Handover



- To enable applications such as driverless vehicles, which require computational resources to perform AI inference (or even training), there will be a need to perform seamless compute handovers.
- Need to define open standards to cover the case of compute session handover.

Questions to be Discussed During this Side Meeting

- **Q1. Focus**. Defining the scope/bounderies of this effort:
  - E.g., on-path vs off-path
  - Leveraging existing RFCs for information exposure (e.g., ALTO) vs creating new work
- **Q2. Existing work.** Are you aware of this topic being addressed in any other WG?
- **Q3. Way forward**. Is there interest in moving this work forward? In which way?
  - Mailing list.
  - Within another WG?
  - BOF?

BACKUP SLIDES

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)

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/