



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/>

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

Content

What is this topic about? Problem space: Service lifecycle

Interest to the IETF

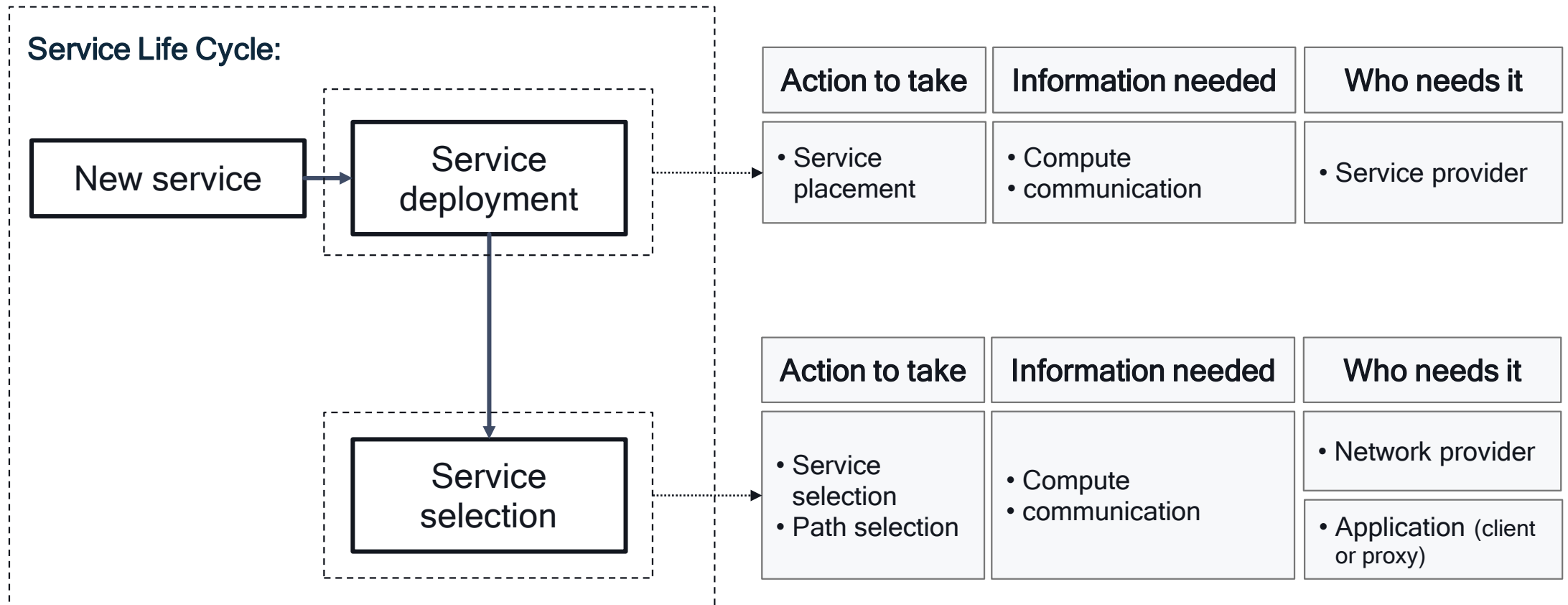
Use cases

Defining compute metrics at IETF (CATS and ALTO conversations)

Guiding principles

Desired outcome of this presentation

Problem Space: Service Lifecycle and Information Exposure



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.

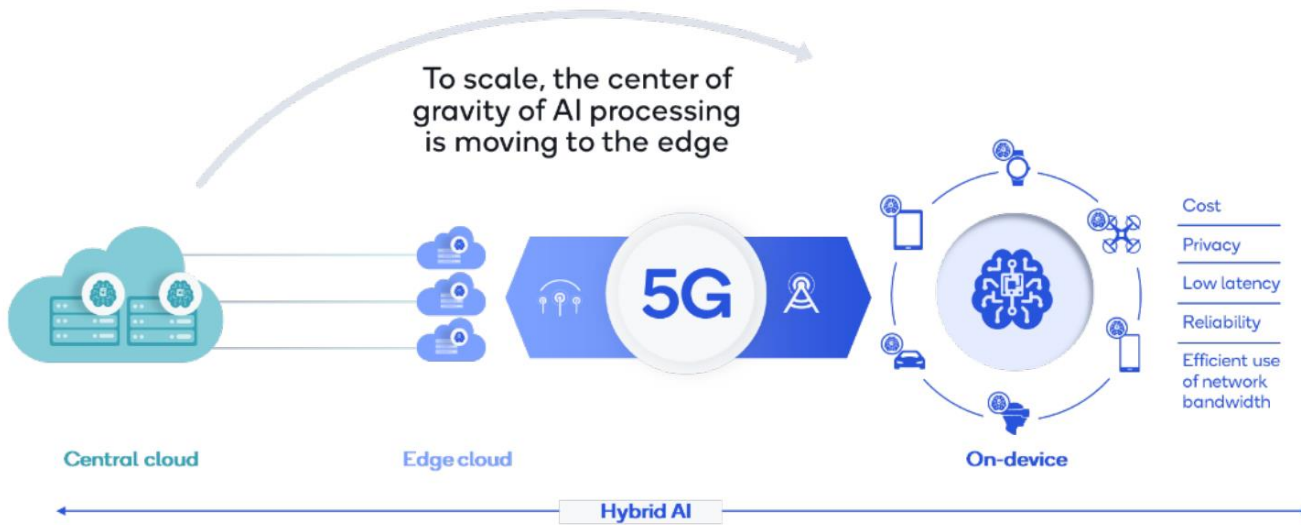
- Linux CAMARA. “Reserve compute resources within the operator network”. “Influence the traffic routing from the user device toward the Edge instance of the Application”.
- GSMA Open Gateway. 21 operators to open up network APIs for developers.
- 3GPP NEF. Enable exchange of information to/from an external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



Distributed XR computation



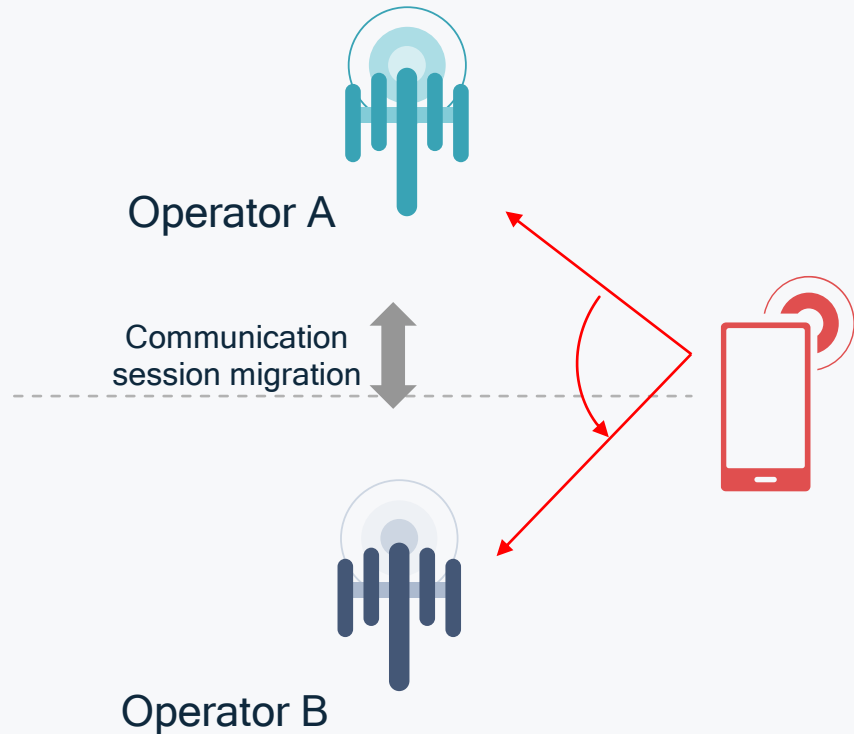
1. Asynchronous time warp reduces Motion to Photon (MTP) latency by using on-device processing based on the latest available pose. MTP below 20 ms generally avoids discomfort - has to be processed on the device

- 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.

- 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.

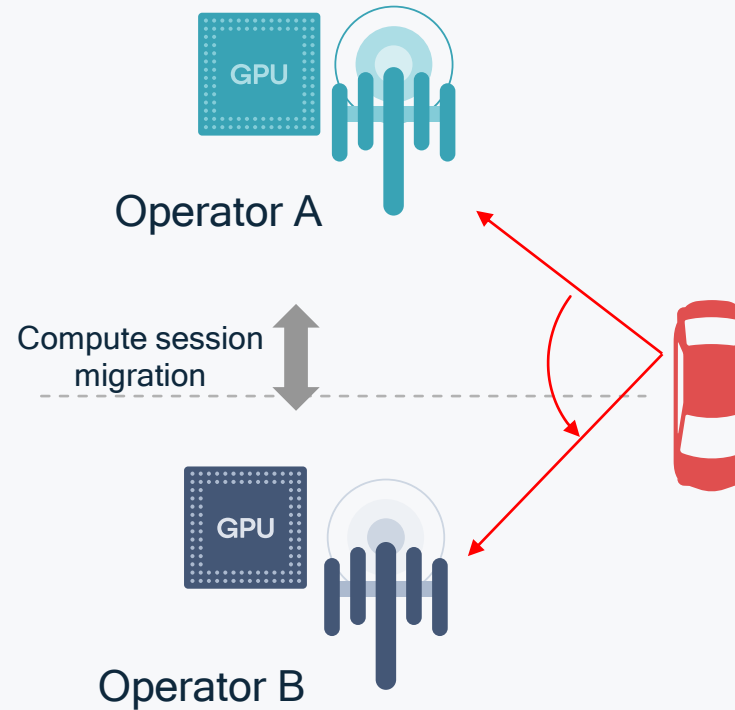
Use Cases

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.

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/

Side Meetings and Welcoming People to Join Discussions/Mailing List

- IETF 118 Side Meeting: <https://github.com/communication-compute-exposure/ietf-side-meetings/tree/main/ietf-118-side-meeting>
- IETF 119 Side Meeting: <https://github.com/communication-compute-exposure/ietf-side-meetings/tree/main/ietf-119-side-meeting>
- 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-Giralt jros@qti.qualcomm.com

BACKUP SLIDES

Example of possible Resource Definitions for cloud native Environments e.g., K8S

Resource Quota	Resource Type	Example
CPU	Compute	Processing capacity, for example defined range that can be used or is available (0.1 or 2 CPU). An additional parameter is the maximum amount of CPU requests or time availability of the resources. The unit format is the total.
Memory	Compute	Available Memory
Storage	Storage	Available Storage
Configmaps	Object	Configuration and Topology Maps
Secrets	Object	Total of all sensitive credentials (Secrets) that can be made available to Pods as environment variables or configuration files without the need to store these sensitive or changing data in version control systems or image registries. So, config maps and secrets allow you to separate code and configuration.
Pods	Object	Available Container & Worker Nodes
Jobs	Object	Total of concurrent Jobs
Services	Object	Total of concurrent Services
...		