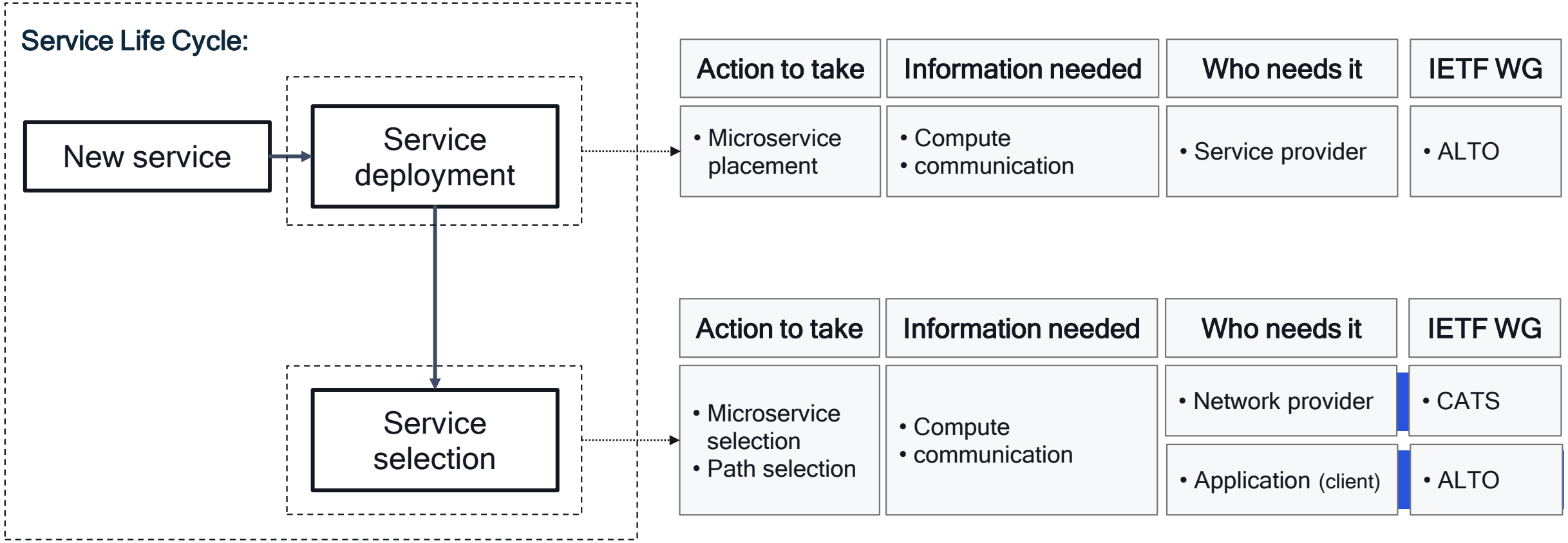
Exposure of Compute Information for Edge Service Placement and Selection

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

Luis Contreras (Telefonica), Sabine Randriamasy (Nokia Bell Labs), Dany Lachos (Benocs)
Jordi Ros Giralt (Qualcomm Europe, Inc.), Qi Xue (Qualcomm Technologies, Inc)

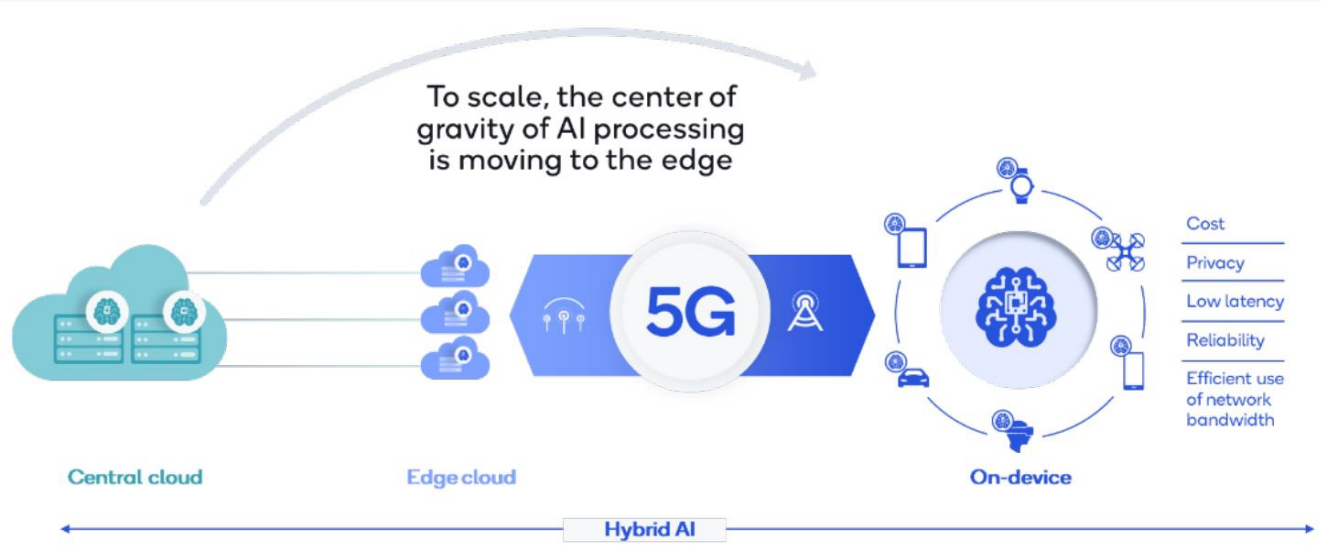
Service Deployment and Selection: Problem Space



IETF ALTO Edge Service Draft: Examples of Key Use Cases

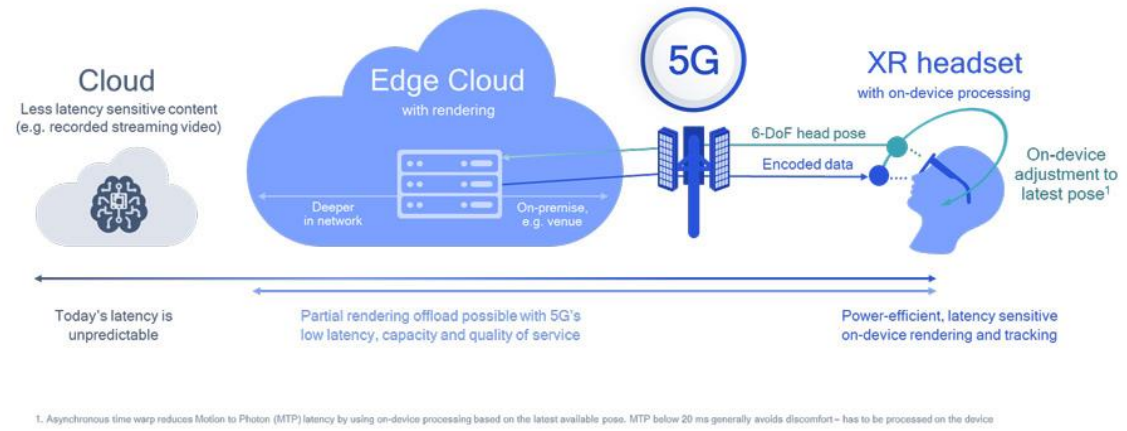
https://datatracker.ietf.org/doc/draft-contreras-alto-service-edge/

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.

WG Work Item Proposal and Participants

- Members of the ALTO WG propose to work on writing an informational draft with the following objectives:
  - GAP analysis focusing on scoping the necessary metrics and representations needed to cover the full life cycle of service deployment and selection.
  - Analysis of other related activities in the IETF to avoid reinventing the wheel.
  - Feasibility analysis to extend ALTO with compute information:
    - Southbound interface feasibility
    - Northbound interface feasibility
  - Analysis of interactions between ALTO and other IETF WGs (e.g., CATS). Draft can be cross-reviewed by member of these other WGs.
- Participants/champions:
  - Luis Contreras (Telefonica), Jordi Ros Giralt (Qualcomm Europe, Inc.), Sabine Randriamasy (Nokia Bell Labs), Dany Lachos (Benocs), Qi Xue (Qualcomm Technologies, Inc), Richard Yang (Yale University), and other members of the ALTO WG.

BACK UP SLIDES

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. Aim for simplicity, while ensuring the combined efforts in the IETF don’t leave gaps in supporting the full life cycle of service deployment and selection.
  - CATS/ALTO cooperation/coordination on metrics to cover both service deployment and service/path selection:
    - CATS focus appears to be on in-network service and path selection.
    - ALTO focus is on application-level service deployment and application-level service/path selection.
- P3. “Permutable” interaction between ALTO and CATS.
  - ALTO as a source of network (and/or compute) information to CATS.
  - CATS as a source of compute (and/or network) information to ALTO.