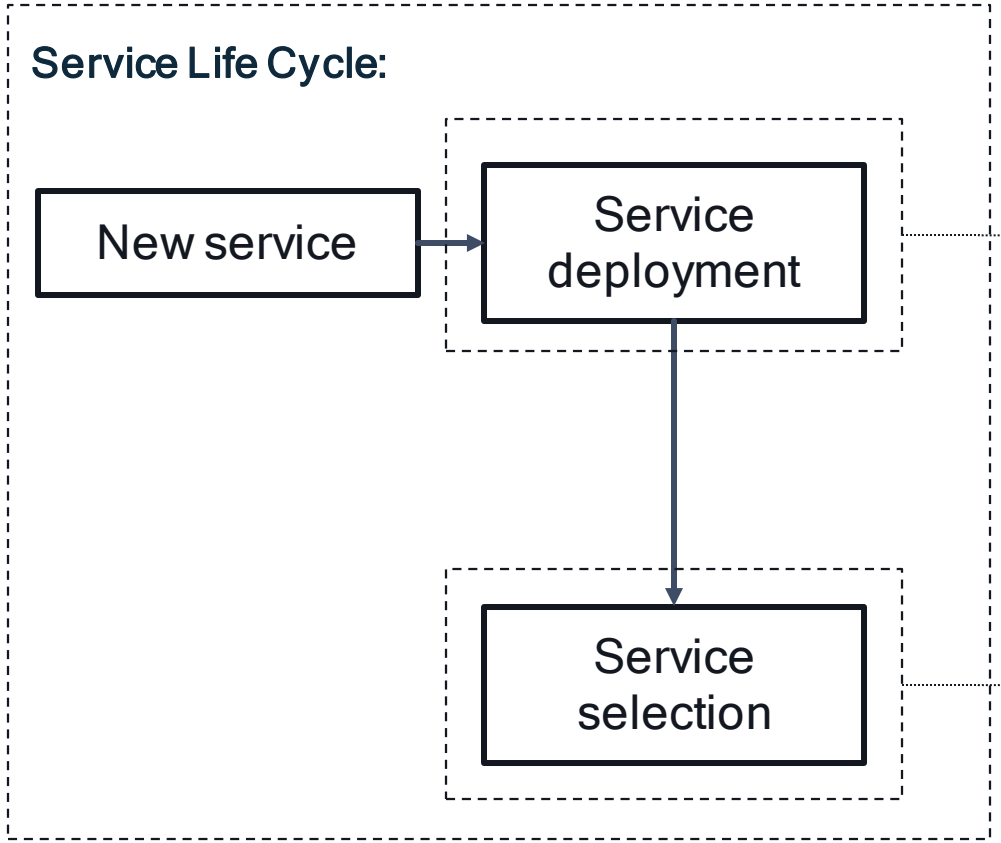
Exposure of Communication and Compute Information for Infrastructure-Aware Service Deployment and Selection

https://datatracker.ietf.org/doc/draft-rcr-opsawg-operational-compute-metrics/

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

Problem Space: Service Lifecycle and Information Exposure



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

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

Defining compute metrics at IETF

- Joint exposure of network and compute metrics requires a common unsterstanding of the exposed information
- Standardization of network information quite mature but is in progress for compute information
- First step: define a common set of compute metrics to support the various use cases being served in the IETF
- Related work exists in IETF and other bodies such as ETSI, to provide
  - 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)

OPSAWG as a common ground for compute metrics

- Consumers of compute information can be manyfold and located at different levels
  - Applications, users, controllers, routers that need information at different granularities
- Thus, the set and scope of applicable metrics is manifold
  - Capabilities, actual/estimated/predicted state
  - Aggregated processing delay, C/GPU performance, etc.
- Compute metrics are being defined in several bodies and work is in progress
  - IETF, ETSI, Linux Foundation, Cloud providers, etc.
- This calls for a common framework to specify standardized metrics
  - To support trustable compute capability assessment and benchmarking
- We think OPSAWG could be the appropriate venue
  - Leveraging contributions and methodology proposed in IETF WGs and other standard and opensource bodies
  - Gathering use cases and identify gaps
- Side meeting to discuss metrics and their exposure on Wed at 15:30, Karlin 4
  - 15:30-17:00 | Karlin 4 | Edge Computing | https://github.com/compute-exposure/ietf-118-side-meeting
  - Exposure of Network and Compute information to Support Edge Computing Applications

Backup slides

Content

Problem space: service lifecycle

What is this topic about?

Interest to IETF

Use cases

Previous work

Guiding principles

Expected outcomes

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.

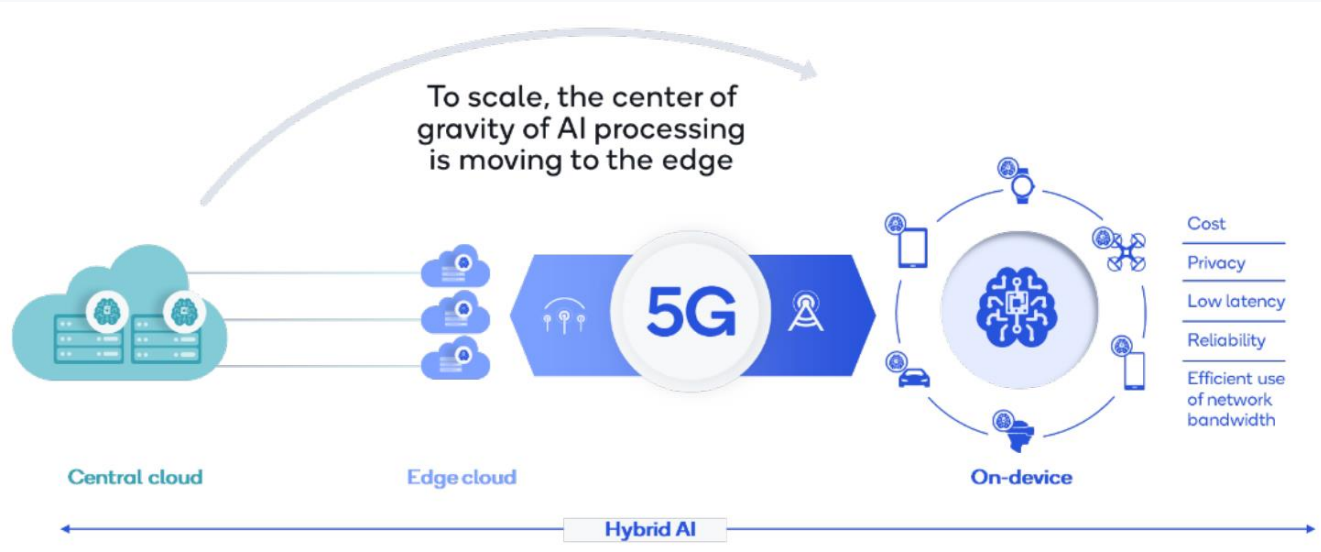
- Linux CAMARA. "Reserve compute resources within the operator network". "Influence the traffic routing from the user device toward the Edge instance of the Application"
- GSMA Open Gateway. 21 operators to open up network APIs for developers
- 3GPP NEF. Enable exchange of information to/from an external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

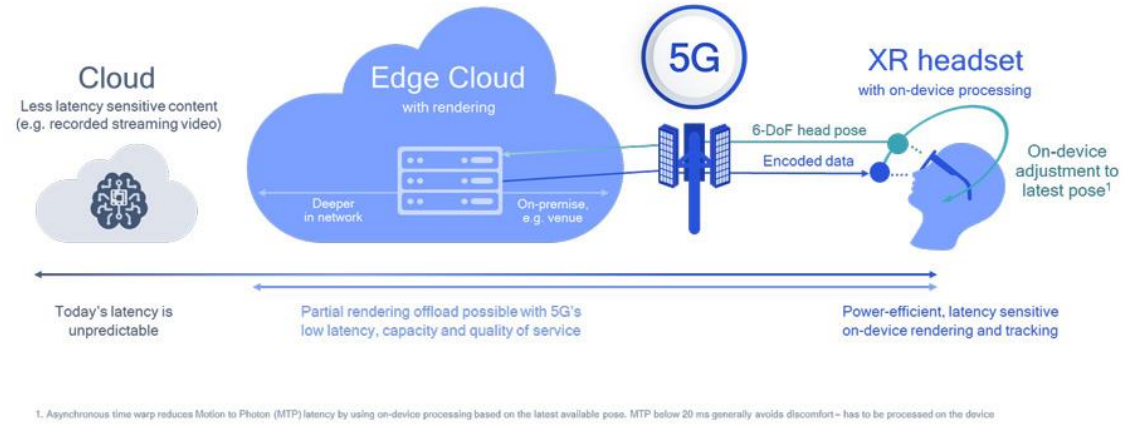
https://datatracker.ietf.org/doc/draft-contreras-alto-service-edge/

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.

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 selection.

Potential to Work within OPSAWG

- Developing a YANG model for exposure of communication and compute information.
- Examples of related proposed YANG models
  - A YANG Data Model for Service Information: https://datatracker.ietf.org/doc/draft-wang-opsawg-service-information-yang/
  - A YANG Data Model for Intermediate System to intermediate System (IS-IS) / Open Shortest Path First (OSPF) Topology: https://datatracker.ietf.org/doc/draft-ogondio-opsawg-isis-topology/ and https://datatracker.ietf.org/doc/draft-ogondio-opsawg-ospf-topology/
  - YANG Data Models for the Application-Layer Traffic Optimization (ALTO) Protocol: https://datatracker.ietf.org/doc/draft-ietf-alto-oam-yang/

Expected Outcomes

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?