Information Exposure for Service Deployment at the Edge

<draft-rcr-opsawg-operational-compute-metrics>

Luis Contreras (*Telefonica*), Jordi Ros Giralt (*Qualcomm Europe, Inc.*), Roland Schott (*Deutsche Telekom*), Julien Maisonneuve (*Nokia Bell Labs*)

IETF 122, Bangkok, March 2024

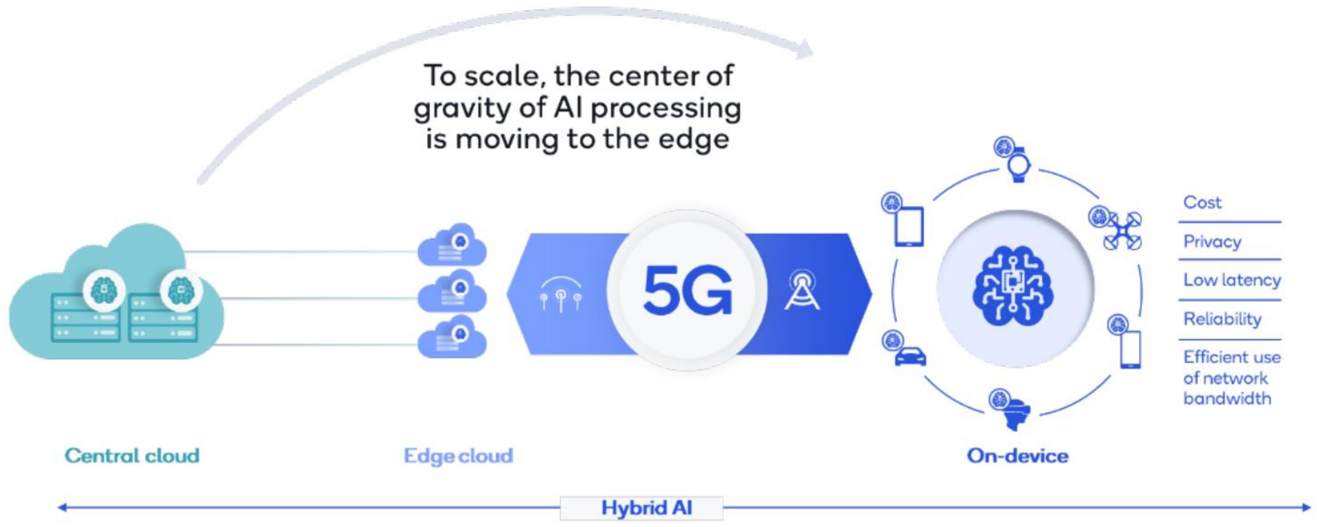
Motivation

- While the standardization of protocol interfaces to expose network information is quite mature, it lags behind for compute information.
- There is a need to define a compute model and set of compute metrics to support various use cases being served in the IETF.
- Some work exists in the IETF:
    - CATS (draft-ysl-cats-metric-definition)
    - ALTO (e.g., draft-contreras-alto-service-edge)
    - OPSAWF (e.g., RFC 7666 MIB, draft-rcr-opsawg-operational-compute-metrics)
- Metrics have also been defined in other bodies such as the Linux Foundation, DMTF, ETSI NFV, etc:
    - 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)

Examples of Use Cases

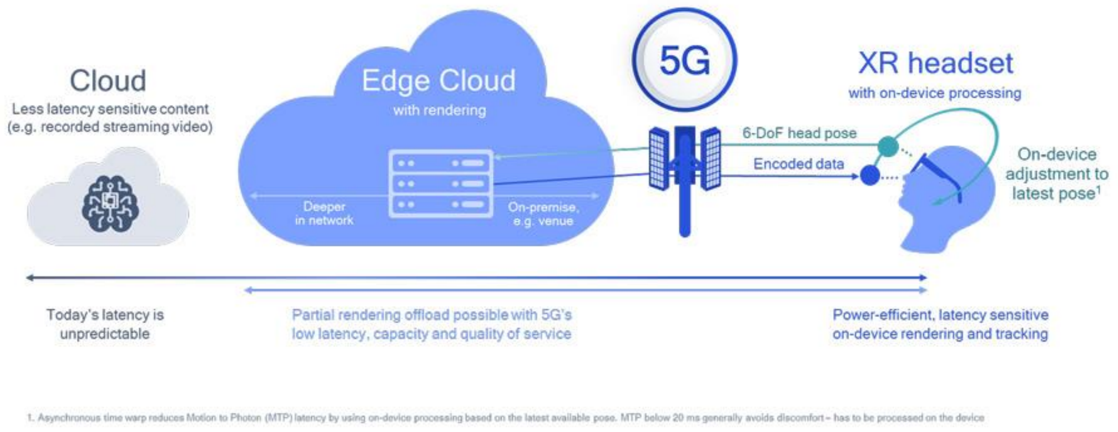
https://datatracker.ietf.org/doc/draft-contreras-alto-service-edge/

Distributed AI computation



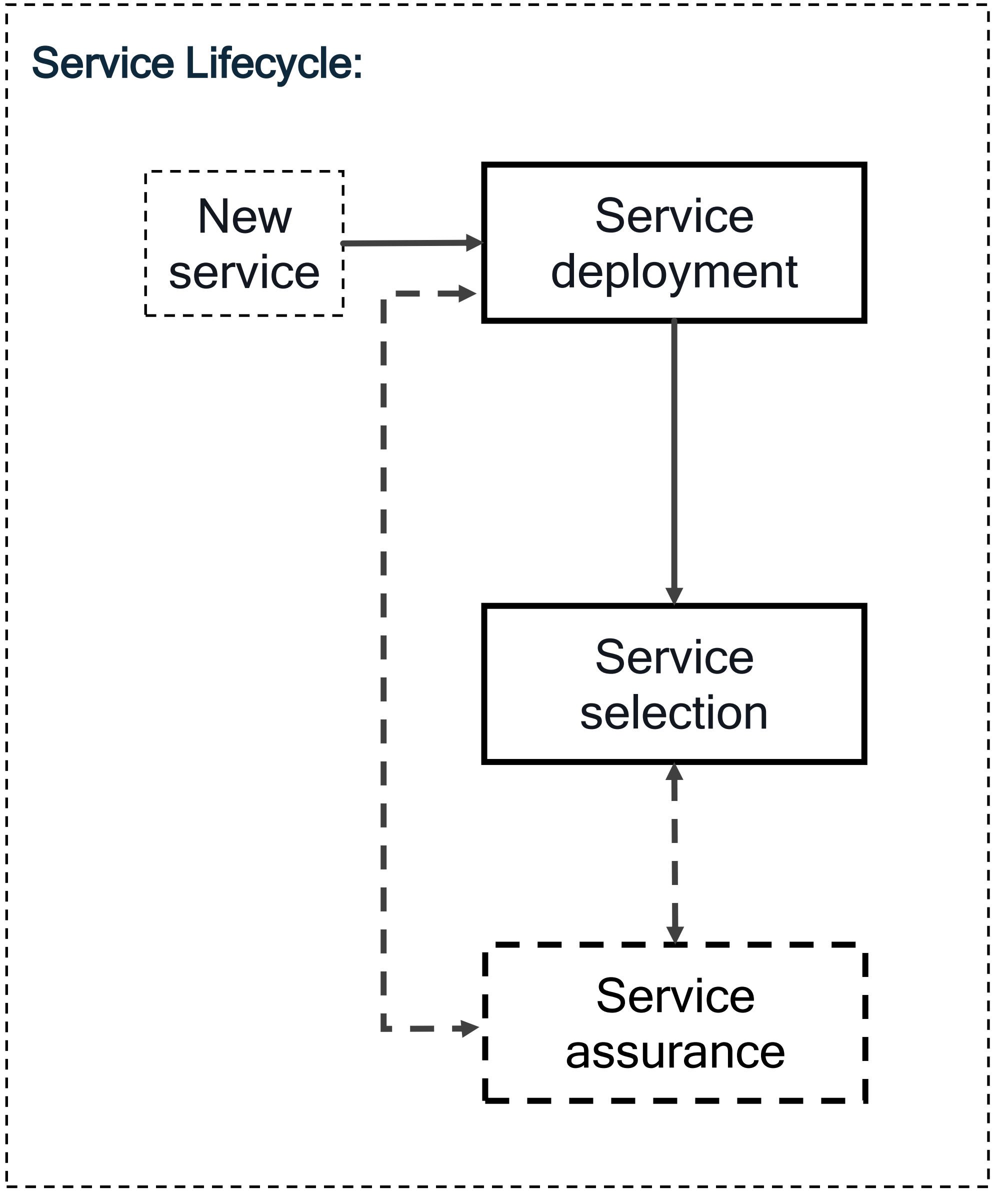
- Larger, mid-size, and smaller AI models are run in the cloud, the edge, and the device, respectively, enabling a trade-off between model accuracy and computational cost.
- To make proper service deployment/selection decisions at the application level, knowing compute information is key in today's edge 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.

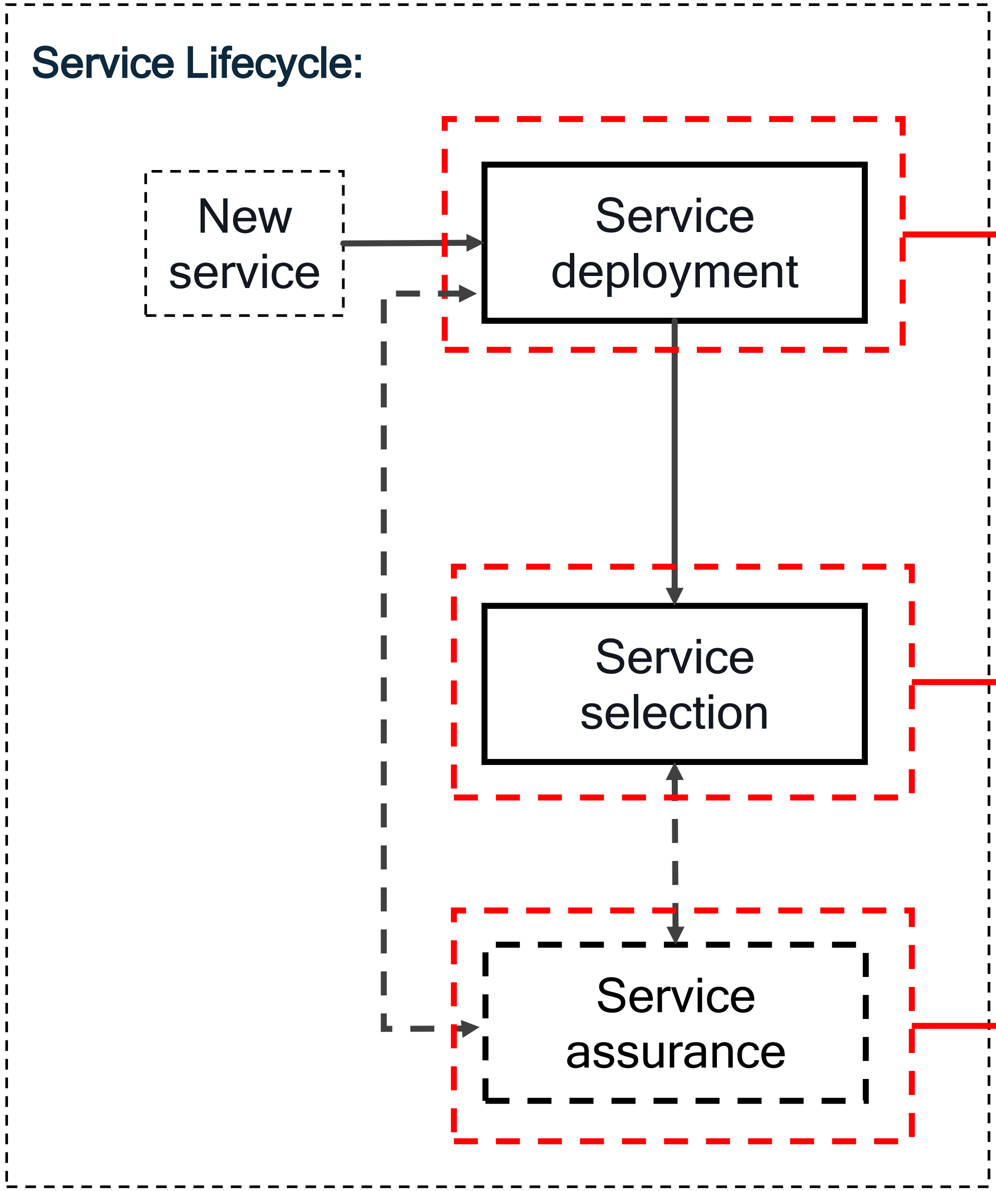
General Problem Space: Service Lifecycle and Information Exposure



Action to take	Information needed	Who needs it
(1) Service placement	Compute and communication	Service provider
(2.a) Service selection: compute node selection	Compute and communication	Network provider, service provider or application
(2.b) Service selection: path selection	Communication	Network provider or application
(3) Service assurance	Compute and communication	Network provider, service provider or application

Table 1: Problem space, needs, and stakeholders.

General Problem Space: Service Lifecycle and Information Exposure



Needed but not being covered in the IETF yet

Action to take	Information needed	Who needs it
(1) Service placement	Compute and communication	Service provider
(2.a) Service selection: compute node selection	Compute and communication	Network provider, service provider or application
(2.b) Service selection: path selection	Communication	Network provider or application
(3) Service assurance	Compute and communication	Network provider, service provider or application

Table 1: Problem space, needs, and stakeholders.

IETF CATS problem space

Problem Space: Metric Definition and Exposure Mechanism

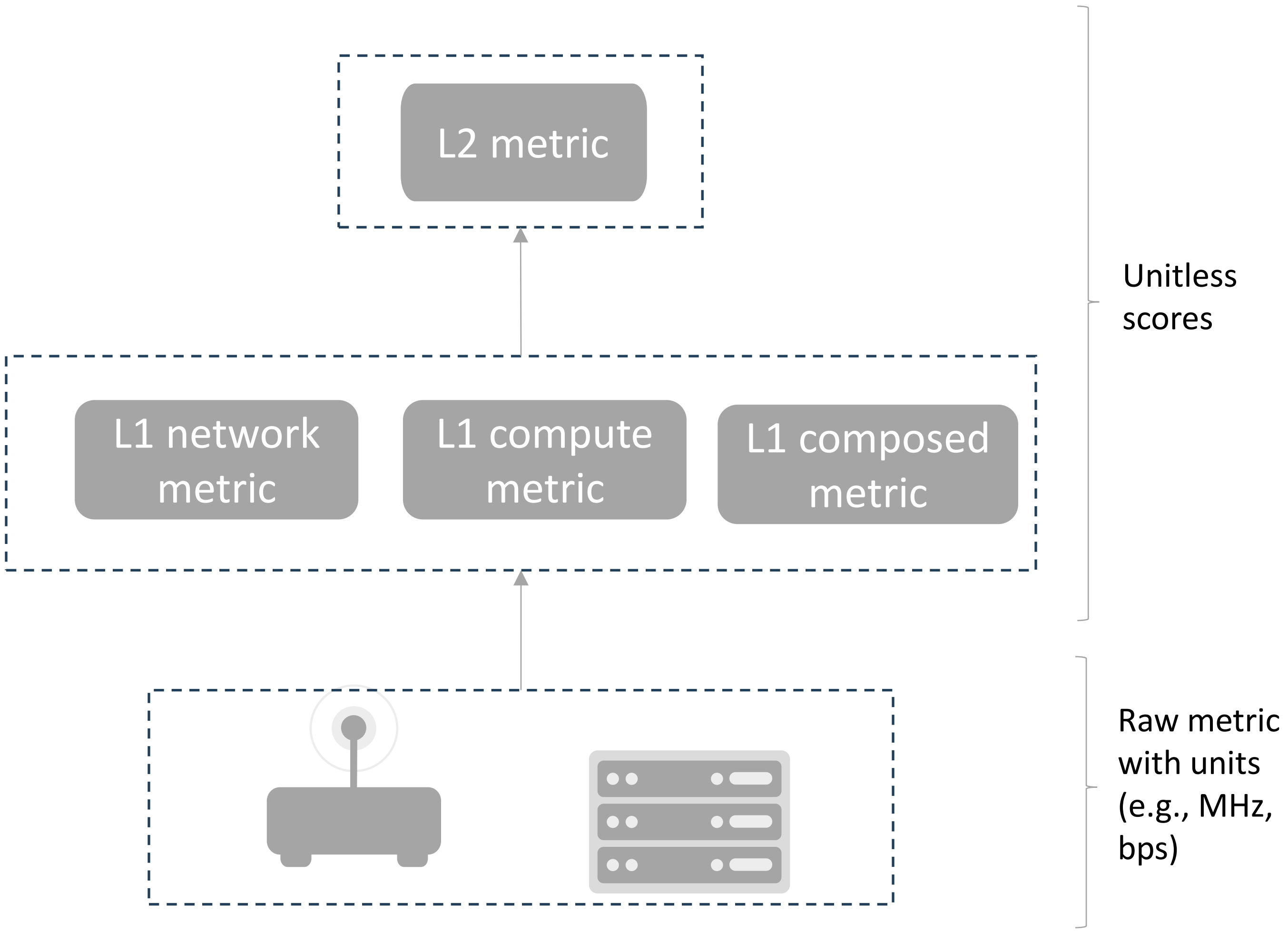
Two main problems to work on:

(1) Definition of the compute model and the compute metrics

(2) Definition of the protocol interface to expose the metrics to the consumer

Definition of the Metrics: Summary of Approach

CATS Metric Model: A 3-level framework to meet the trade-off interoperability vs scalability vs usefulness.



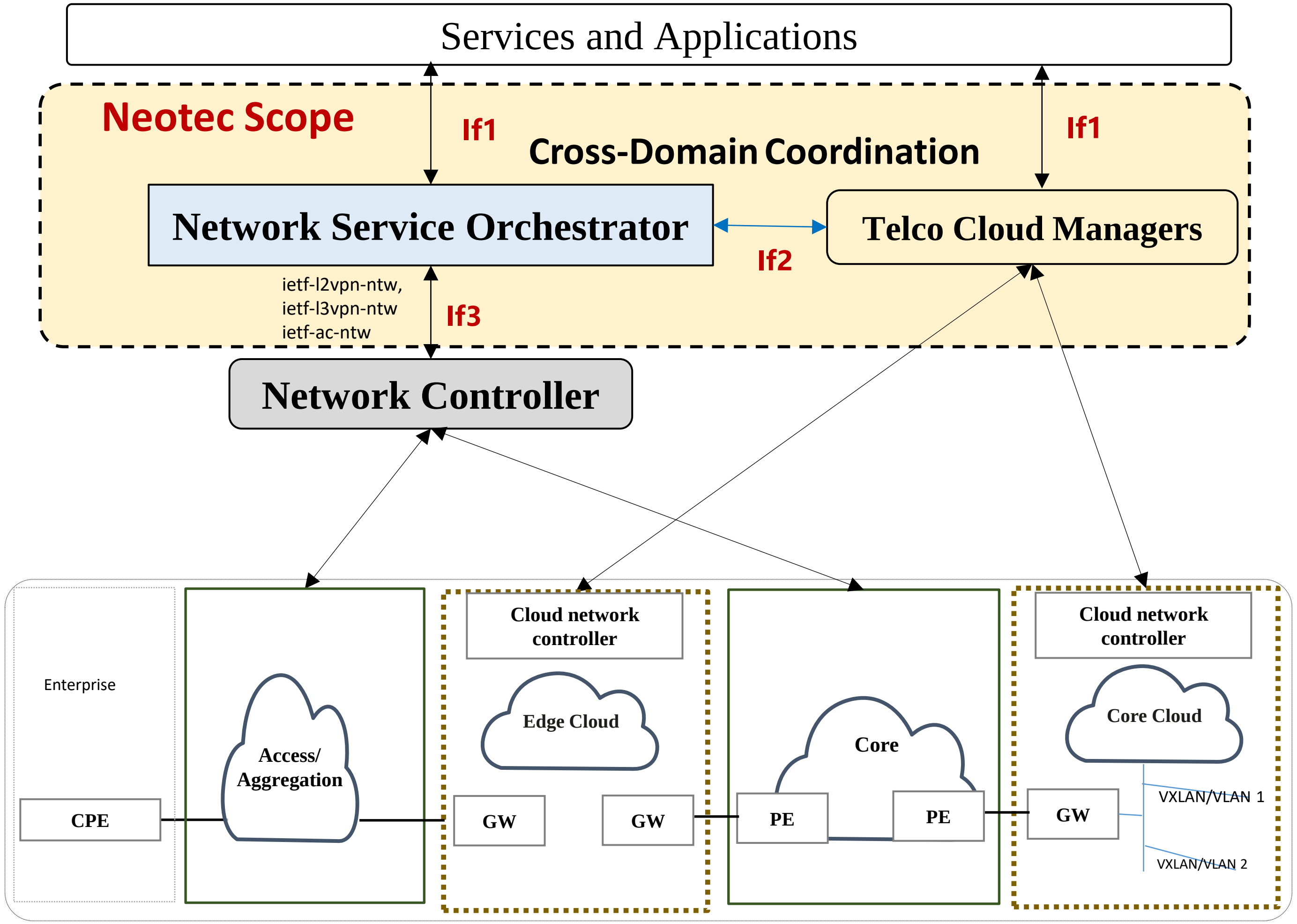
Analogy: Works similarly to the University Grade Point Average system. Every university abstracts out a single score for each student that is specific to its country (e.g., country A score goes from 1 to 10, country B score goes from 1 to 5). When a student travels to another country, the score can be translated to that country's metric. This allows for each country to independently implement their own metrics without global coordination, while achieving global interoperability.

Definition of the Interface to Expose the Metrics

- I-D.ldbc-cats-framework presents three CATS models: distributed, centralized and hybrid. Their corresponding distribution mechanism are:
  - Distributed: Directly distributed to the network devices.
  - Centralized: Collected by a centralized control plane.
  - Hybrid: Some directly, some centralized.
- Optimal choice depends on dynamicity: higher-frequency metric updates tend to favor a centralized collection approach, and vice versa.
- For decentralized approach, draft-ll-idr-cats-bgp-extension and draft-ietf-idr-5g-edge-service-metadata propose using BGP.
- For centralized approach, a potential candidate solution is to leverage ALTO (e.g., RFC7285, RFC9240)

Positioning of this Work within the Neotec Architecture

- Interface 1(If 1) : 1)Intent-driven service deployment and scaling policy with service and SLO requirements can be directly mapped to cloud-network alliance policies. E.g. low-latency 100ms service, the system automatically selects edge nodes whose latency is less than 100 ms and reserves dedicated network bandwidth for the node. 2) Cloud aware network topology and metrics information (Luis)
- Interface 2(If 2) : Cloud exposes the resource and operation metrics to the orchestrator, for network aware service placement and scaling policies
- Interface 3(If 3) : Network exposes the resource and operation metrics to the orchestrator for cloud resource aware network connectivity's and service QoS policy



Neotec Goal:

① Central Cloud: Elastic Scaling on Demand (Millisecond-Level Scaling)

② Edge Cloud: Deterministic Low Latency (<10ms End-to-End)

③ Global Efficiency: Cross-Domain Resource Utilization Improved, and agile Domain Coordination

Positioning of this Work within the Neotec Architecture

- Interface 1(If 1) : 1)Intent-driven service deployment and scaling policy with service and SLO requirements can be directly mapped to cloud-network alliance policies. E.g. low-latency 100ms service, the system automatically selects edge nodes whose latency is less than 100 ms and reserves dedicated network bandwidth for the node. 2) Cloud aware network topology and metrics information (Luis)
- Interface 2(If 2) : Cloud exposes the resource and operation metrics to the orchestrator, for network aware service placement and scaling policies
- Interface 3(If 3) : Network exposes the resource and operation metrics to the orchestrator for cloud resource aware network connectivity's and service QoS policy