# 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

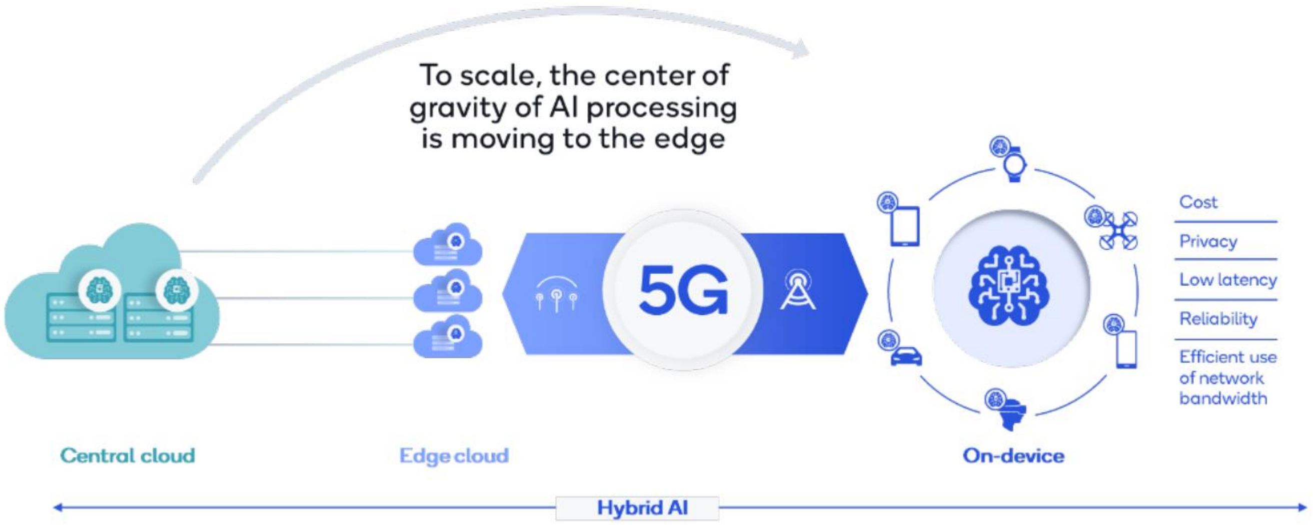
# Motivation

- While the standardization of protocol interfaces to expose network information is quite mature, it lags behind for compute information.
- This draft was presented in OPSAWG during IETF 121. Quote by the chair from the meeting minutes: “Interesting topic, this needs to be solved somewhere. (…) Said that it would be done in CATS, which is about steering. If the answer is yes, then take the work to CATS, but if no, then maybe there is something to be brought here.”
- 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)
  - OPSAWG (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.

# Examples of Use Cases

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

Distributed AI computation

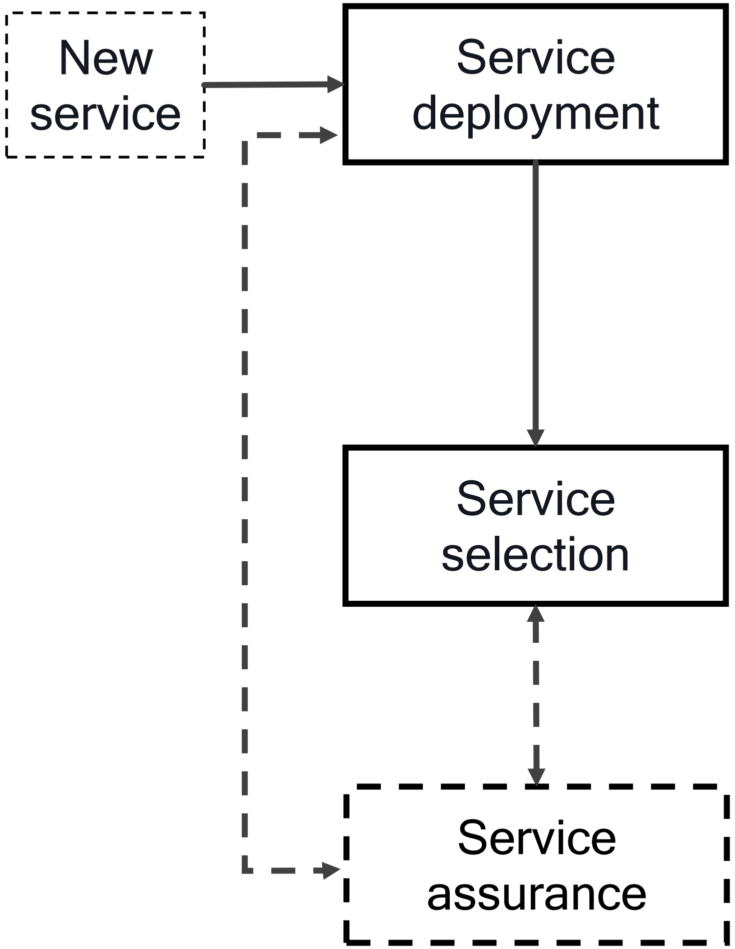

To scale, the center of gravity of AI processing is moving to the edge
5G
Cost
Privacy
Low latency
Reliability
Efficient use of network bandwidth
Central cloud
Edge cloud
On-device
Hybrid AI

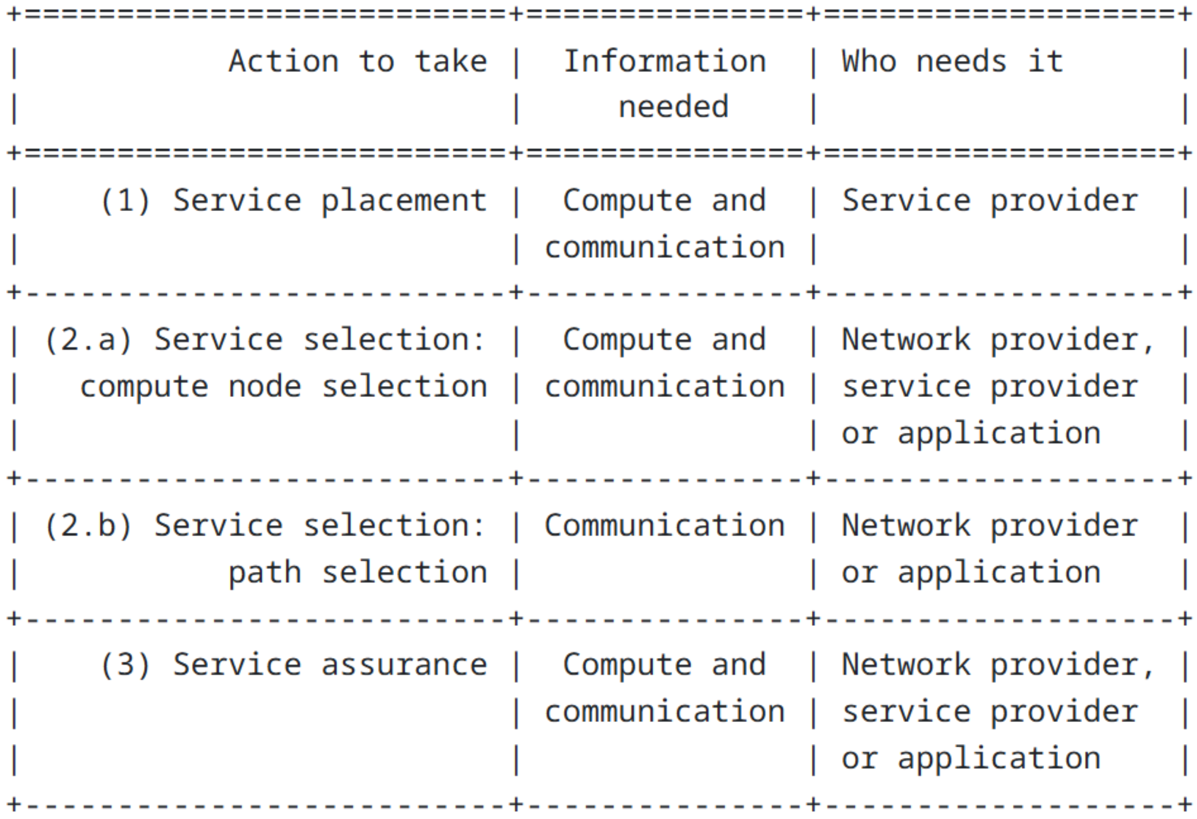
- Larger, mid-size, and smaller AI models are run in the cloud, the edge, and the device, respectively, enabling a trade-off between model accuracy and computational cost.
- Exposing IETF domain metrics such as network resources (latency, bandwidth, topology, etc.) combined with compute metrics from the cloud domain is key to enable proper service deployment decisions.

- CATS resolves the problem of load balancing requests to the computing nodes. It standardizes both the framework for traffic steering and the abstraction of the metrics needed for the CATS use case.
- CATS assumes the compute service replicas are already deployed.
- No WG is looking into the support needed from the IETF domain to assist in the actual deployment of the compute service replicas.
- E.g., to feasibly deploy an LLM in the edge cloud, one needs to gather communication and compute information such as latency, bandwidth, compute resources, memory resources.

# General Problem Space: Service Lifecycle and Information Exposure

Service Lifecycle:

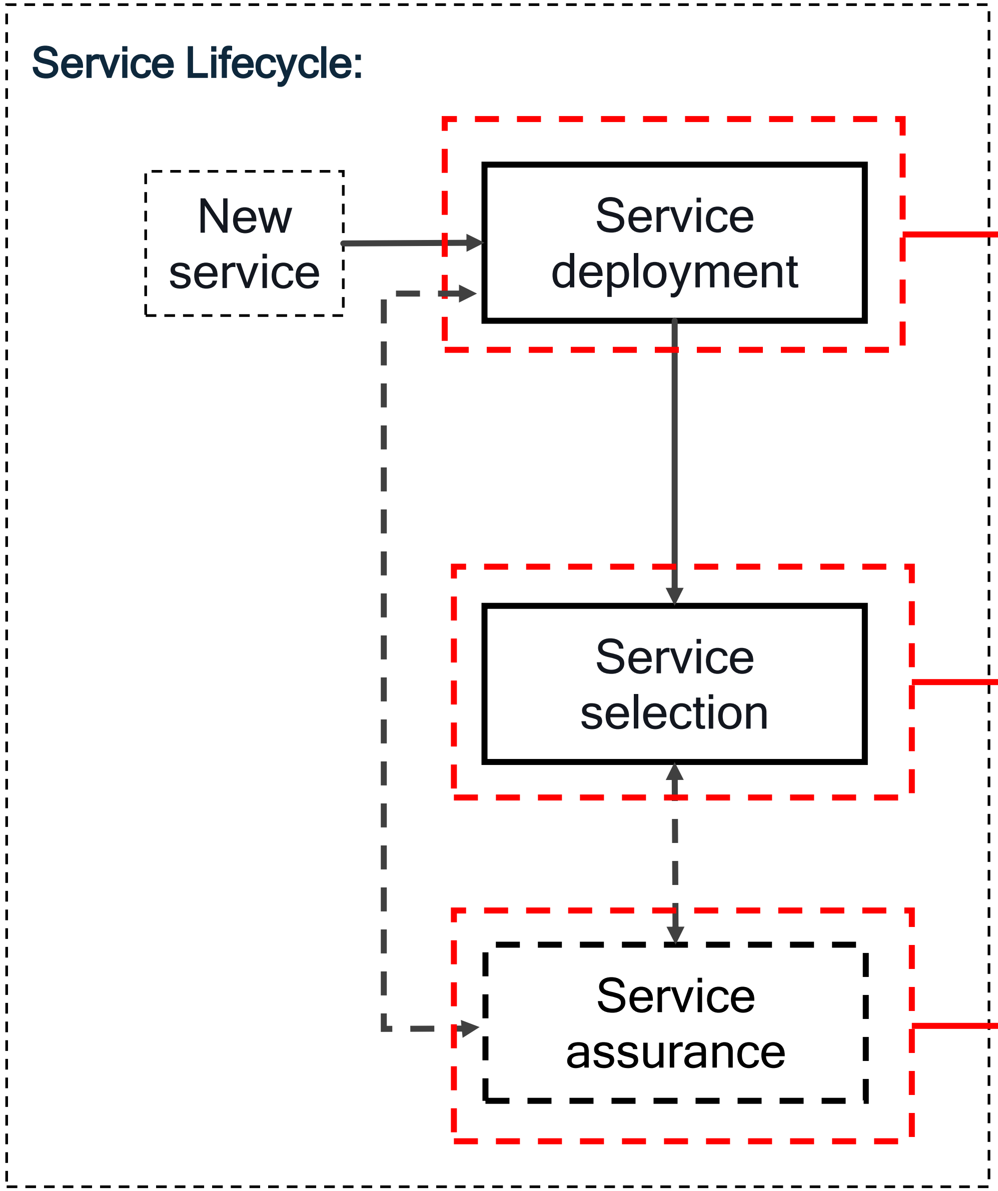




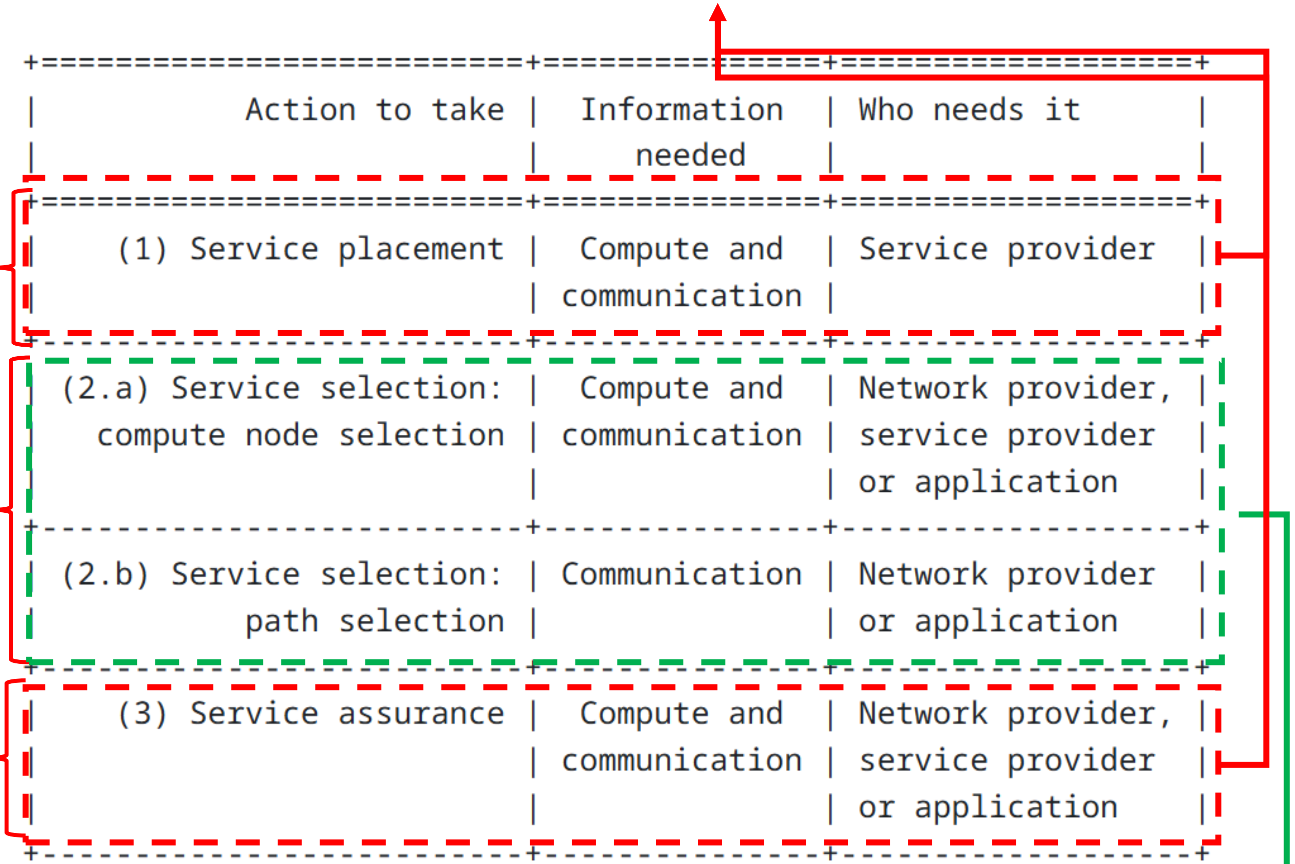
| Action to take | Information needed | Who needs it |
|---|---|---|
| (1) Service placement | Compute and communication | Service provider |
| (2.a) Service selection: compute node selection | Compute and communication | Network provider, service provider or application |
| (2.b) Service selection: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



IETF information domain needed but not being covered by 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.

Covered by IETF CATS

## Problem Space: Metric Definition and Exposure Mechanism

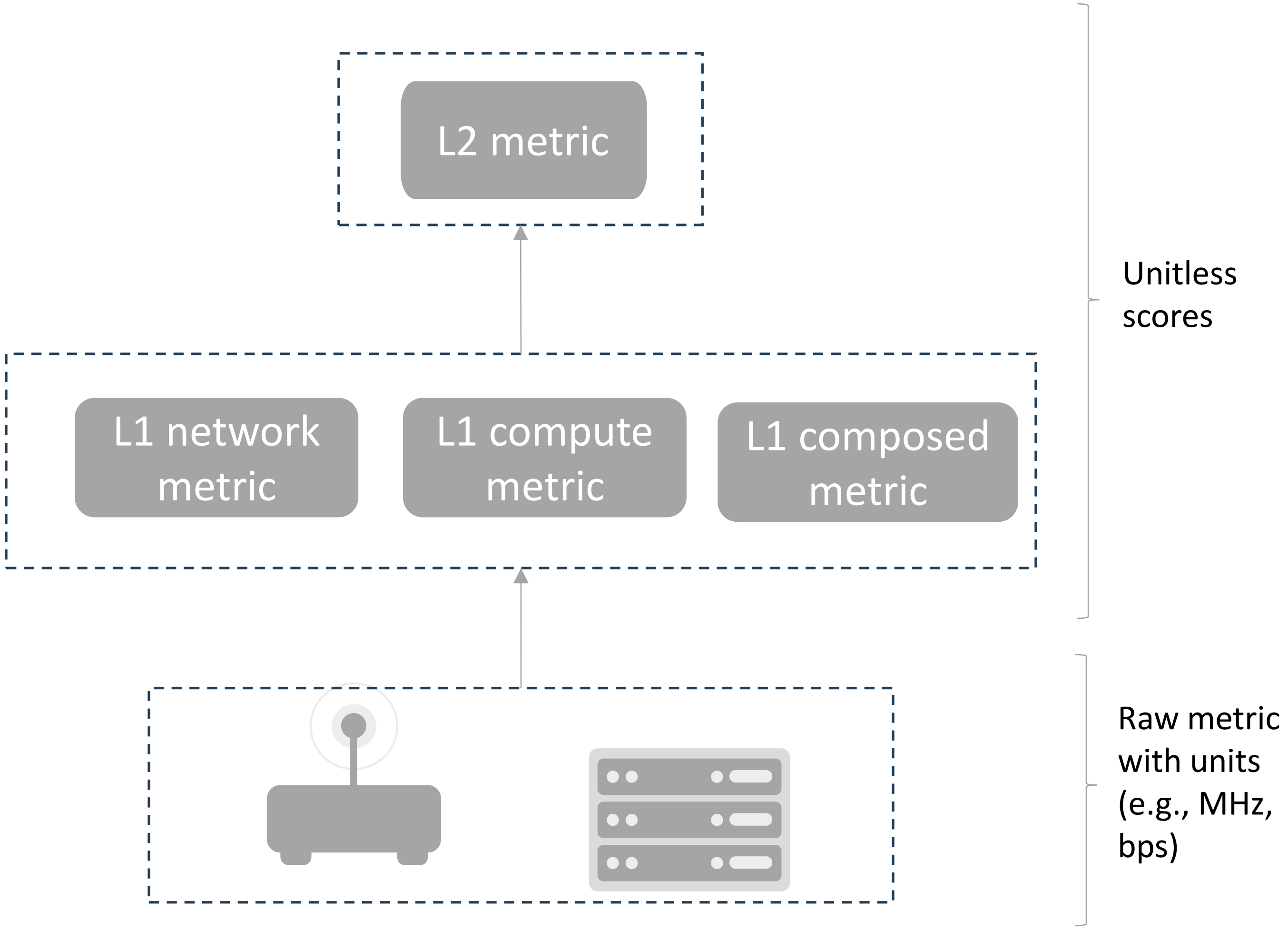
Two main problems to work on:

(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.



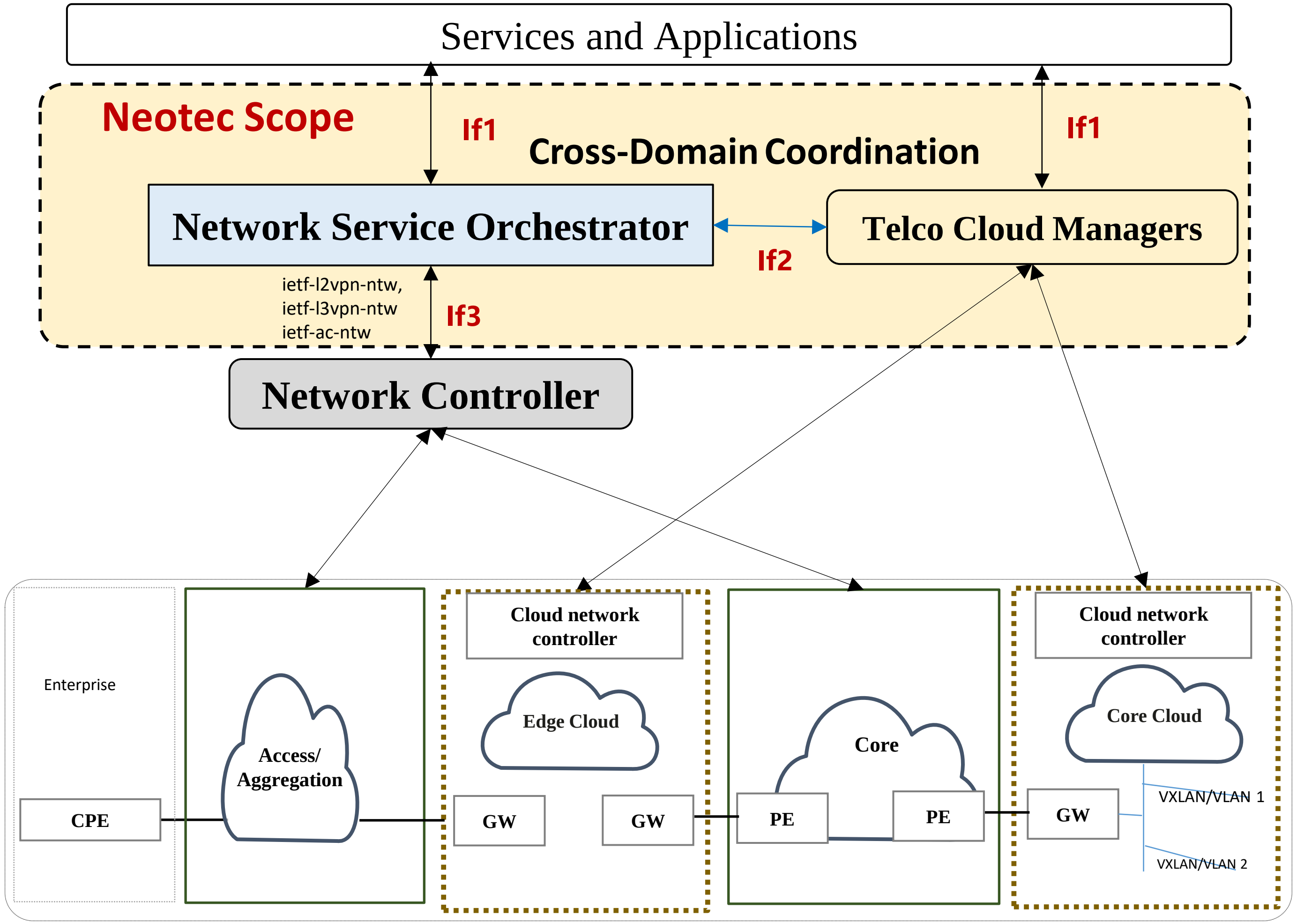
**Analogy**: Works similarly to the University Grade Point Average system. Every university abstracts out a single score for each student that is specific to its country (e.g., country A score goes from 1 to 10, country B score goes from 1 to 5). When a student travels to another country, the score can 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:

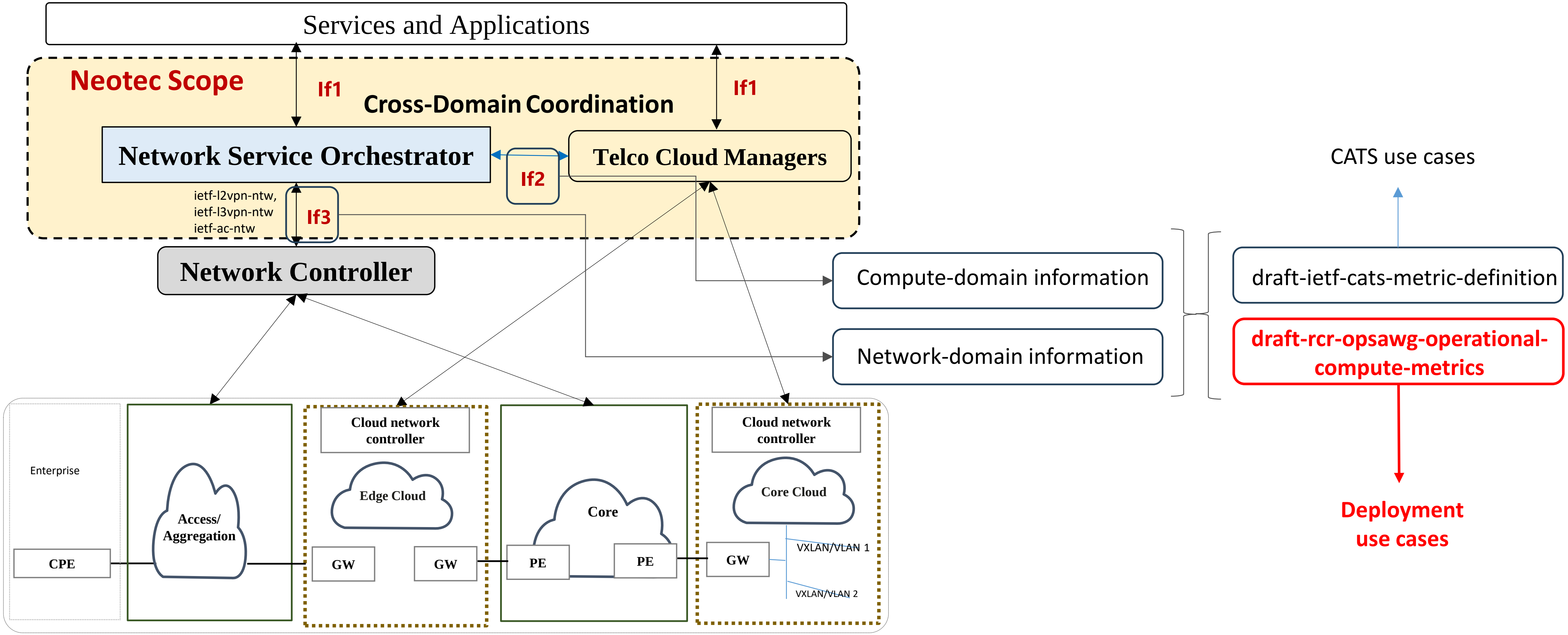
① Central Cloud: Elastic Scaling on Demand (Millisecond-Level Scaling)

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


Services and Applications
Neotec Scope
If1
Cross-Domain Coordination
If1
Network Service Orchestrator
If2
Telco Cloud Managers
ietf-l2vpn-ntw,
ietf-l3vpn-ntw
ietf-ac-ntw
If3
Network Controller
Compute-domain information
Network-domain information
CATS use cases
draft-ietf-cats-metric-definition
draft-rcr-opsawg-operational-compute-metrics
Deployment use cases
Enterprise
CPE
Access/ Aggregation
Cloud network controller
Edge Cloud
GW
GW
Core
PE
PE
Cloud network controller
Core Cloud
GW
VXLAN/VLAN 1
VXLAN/VLAN 2