Neotec Working Group
Internet-Draft
Intended status: Standards Track
Expires: 11 December 2025

X. Li
C. Li
China Telecom

9 June 2025

Unified Network and Cloud Orchestration Framework
draft-li-unco-framework-02

Abstract

This draft introduces the Unified Network and Cloud Orchestration Framework (UNCO), which is designed to enable real-time and joint orchestration of network and computing resources in 5G and future-generation networks. UNCO framework addresses inefficiencies in current resource scheduling mechanisms, resolves objective conflicts across domains, and provides unified policy and security management. It is applicable in emerging scenarios such as ultra-reliable low-latency communications (URLLC), mobile edge computing (MEC), and network slicing, where service quality and operational efficiency are paramount.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 11 December 2025.

Copyright Notice

Table of Contents

1. Introduction

As next-generation telecom networks evolve to support latency-sensitive, compute-intensive, and highly dynamic applications across metro networks, backbone networks, mobile networks, and beyond, traditional siloed orchestration mechanisms are no longer sufficient. The integration of network and computing resources is essential to enable real-time, adaptive service provisioning across diverse deployment environments. Current industry efforts such as ETSI NFV [NFV033], 3GPP MEC, and IETF service chaining [RFC8969] have made progress in specific domains, but a holistic orchestration framework that bridges network and computing domains with unified security and policy governance remains lacking.

In addition, Telecom Clouds introduce new operational complexities that differ significantly from public cloud deployments. Unlike public clouds, which rely on third-party network providers, Telecom Clouds operate under a single administrative domain where both network and cloud infrastructure are tightly coupled and managed by the same operator. This integration opens up opportunities for real-time coordination between cloud service scaling events and network policy adjustments. However, most existing network management systems can not ajust with dynamic cloud states, which can lead to inefficient load balancing, suboptimal routing, and SLA violations for critical services like AI/ML pipelines, video streaming, and 5G slice traffic.

To address these limitations, the UNCO framework introduces a telemetry-driven mechanism whereby cloud-side resource and service status can be abstracted and delivered to network controllers in near real-time. This mechanism enables the dynamic adjustment of network policies such as UCMP and load balancing, based on ongoing changes in cloud resource availability or service deployment state. Unlike existing IETF efforts (e.g., TEAS [draft-ietf-teas-ietf-network-slice-framework], OPSAWG [draft-ietf-opsawg-service-assurance-architecture], CATS [draft-ietf-cats-framework]), which offer valuable foundations for traffic engineering and service-aware routing, UNCO builds upon and extends them by incorporating real-time cloud-derived metrics directly into the orchestration logic. This approach ensures SLA-compliant, fine-grained orchestration of both network and compute infrastructure in multi-cloud and Telecom Cloud environments.

The Unified Network and Cloud Orchestration framework (UNCO) addresses these gaps by enabling:

* Unified orchestration of computing and network resources.

* Dynamic, SLA-driven scheduling of heterogeneous resources.

* Cross-domain policy alignment and enforcement.

* Real-time observability and security management across domains.

UNCO introduces a layered architectural model with well-defined functional modules and interfaces to facilitate standardization and interoperability among diverse vendor ecosystems.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119] [RFC8174].

3. Terminology

The following terms are used in this draft:

* UNCO: Unified Network and Cloud Orchestration Framework.

* NSOS: Network Service Orchestration and Scheduling System.

* MEC: Multi-access Edge Computing, a framework that extends cloud capabilities to the edge of the network.

* URLLC: Ultra-Reliable Low-Latency Communications, a category of 5G use cases requiring high reliability and very low latency.

* SLA: Service-Level Agreement, a formalized agreement on expected service performance metrics.

* UCMP: Unequal-Cost Multi-Path routing, a technique that uses paths with different costs simultaneously.

* TEAS: Traffic Engineering Architecture and Signaling, an IETF working group focused on traffic engineering mechanisms.

* CATS: Computing-Aware Traffic Steering, an emerging framework for steering traffic based on computing availability.

* NFV: Network Functions Virtualization, an architecture for virtualizing network functions previously implemented in hardware. [NFV033]

* YANG: Yet Another Next Generation, a data modeling language used to model configuration and state data manipulated by the NETCONF protocol.[RFC8969]

* RBAC: Role-Based Access Control, a policy-neutral access control mechanism defined around roles and privileges.

* IAM: Identity and Access Management, the security discipline that enables the right individuals to access the right resources at the right times.

* SDN: Software Defined Networking, an approach to networking that uses software-based controllers to direct traffic on the network.

* API: Application Programming Interface, a set of definitions and protocols for building and integrating application software.

* QoS: Quality of Service, the description or measurement of the overall performance of a service.

* AR/VR/XR: Augmented Reality / Virtual Reality / Extended Reality, technologies for immersive digital experiences.

4. Problem Overview

4.1 Real-Time and Dynamic Resource Scheduling

Modern applications, such as immersive reality, smart manufacturing, and vehicular communication systems, demand rapid provisioning and adjustment of both compute and network resources. Traditional orchestrators often pre-allocate resources statically or based on historical models, which are ill-suited to handle:

* Burst surges of user demands (e.g., traffic spikes in live streaming).

* Elastic scaling requirements (e.g., AI inference workload offloading).

* Edge-cloud resource handoff and failover scenarios.

These limitations lead to under-utilization of expensive infrastructure and inconsistent quality of experience (QoE).

4.2 Contradictions Among Different Objectives

Multiple stakeholders often have conflicting optimization goals. For instance:

* Maximizing compute utilization may increase network path redundancy.

* Reducing latency by routing over low-latency paths may overload specific compute clusters.

* Minimizing operational costs may sacrifice redundancy and resilience.

A successful orchestration strategy must balance these trade-offs dynamically, based on service priorities and system state.

4.3 Lack of Joint Effectiveness Evaluation

Scheduling strategies are often evaluated independently in the context of either network performance (e.g., throughput, delay) or computing performance (e.g., CPU usage, task completion time). However, next-gen services require holistic metrics that combine:

* End-to-end latency from user device to compute execution node.

* Task success rate under constrained bandwidth and CPU cycles.

* Adaptive resource reallocation under failure or congestion.

Such unified metrics are crucial for validating orchestration policies.

4.4 Security and Strategy Fragmentation

Network policy (e.g., firewalls, ACLs, segmentation) and cloud security policy (e.g., IAM, security groups) are traditionally managed in isolation. This results in:

* Inconsistent access controls between compute and data planes.

* Increased cross-domain attack surface.

* Complexity in policy auditing, validation, and enforcement.

UNCO proposes a unified security model to enforce coherent policies across cloud and network domains.

5. Overview of the UNCO framework

This section provides an overview of the UNCO framework and an introduction to its key components. The high-level framework overview of UNCO is shown in Figure 1.

UNCO is composed of three primary modules:

1. NSOS (Network Service Orchestration and Scheduling System): The central decision-making and coordination entity responsible for managing service deployment, orchestrating cross-domain resources, and enforcing global policies.

2. Cloud Manager: A cloud-native resource controller that abstracts heterogeneous computing resources (VMs, containers, GPUs, NPUs, etc.) across edge and central cloud domains. It acts as the compute-plane orchestrator, reporting availability and enforcing workload deployment.

3. Network Controller: A domain-specific SDN or legacy-compatible controller that governs routing, QoS, and telemetry. It operates on the data plane and acts as a programmable policy agent for traffic forwarding, service chaining, and SLA-aware path selection.

These components are deployed in a logically centralized but physically distributed manner to support scalability and fault tolerance. They interact via well-defined interfaces and protocols to deliver seamless joint orchestration.

UNCO is designed to operate across hybrid infrastructures:

* Public Cloud: Multi-cloud environments (e.g., AWS, Azure, Alibaba Cloud) .

* Private Cloud/Enterprise DC: Bare-metal and virtualized compute clusters .

* Edge Computing: Regional micro-DCs or device-near nodes .

* Transport and Access Networks: L2/L3 infrastructure supporting MPLS, SRv6, or P4-based forwarding.

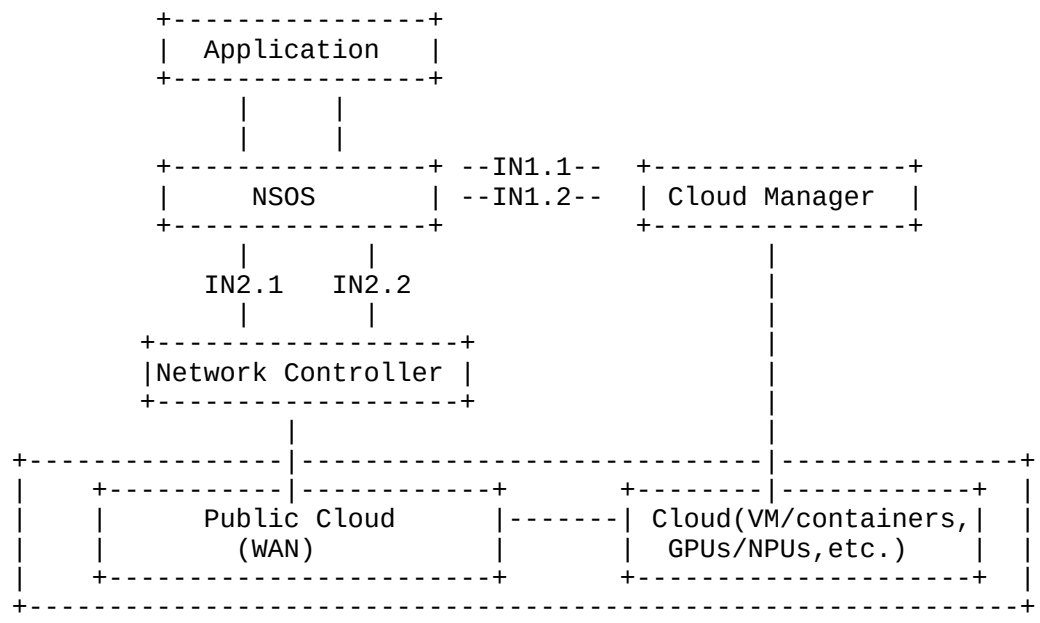

Figure 1 The overall framework of UNCO

Each module can scale independently, supporting multi-tenancy, high availability, and flexible deployment topologies. NSOS typically includes a policy engine, resource graph model, service catalog, and intent resolution logic. It may integrate with external OSS/BSS systems for commercial service integration.

5.1. NSOS

The NSOS (Network Service Orchestration and Scheduling System) serves as the brain of the UNCO framework. It is designed to perform centralized decision-making while maintaining awareness of service requirements, real-time resource availability, and policy enforcement across domains. NSOS is capable of translating high-level application intents into concrete actions such as workload placement, bandwidth allocation, and route optimization.

It plays a vital role in translating service-level requirements into programmable tasks, ensuring optimal resource usage while maintaining SLA commitments. The NSOS also maintains a overall view of global topology and performance state of both computing and networking infrastructure, enabling end-to-end orchestration decisions. Moreover, it ensures feedback-driven loop closure, adapting orchestration actions based on monitored outcomes. Through coordination with both the Cloud Manager and the Network Controller, the NSOS can adjust deployments in response to failures, demand surges, or SLA violations.

The NSOS is a logically centralized orchestrator with the following extended capabilities:

* Service Parsing & Decomposition: Translates high-level service intents into fine-grained resource requirements.

* Topology Awareness: Maintains a live map of compute, storage, and network nodes with performance telemetry.

* Feedback Loops & SLA Assurance: Continuously collects performance metrics to adapt placements and routing in real-time.

* Security Federation: Validates policy consistency across cloud-native RBAC and network access lists.

5.2. Cloud Manager

The Cloud Manager is the dedicated module responsible for managing the full lifecycle of cloud-side computing resources, including virtual machines, containers, GPUs, FPGAs, and NPUs, deployed across centralized, regional, and edge datacenters. It plays a passive but essential role in the UNCO architecture by exposing resource states and executing scheduling directives issued by the NSOS.

It provides the following capabilities:

* Resource Management: Monitors and manages compute, storage, and accelerator resources (e.g., CPU, GPU, FPGA).

* Telemetry Exposure to Network Provides fine-grained metrics such as CPU/GPU usage, memory availability, disk IOPS, and thermal/load levels. Metrics can be sampled per second, per event, or on demand, and are tagged with contextual identifiers (e.g., service instance ID, tenant ID, SLA level).

* Execution of NS-OSS Scheduling Instructions Accepts compute deployment instructions from NSOS, including resource types (e.g., CPU, GPU, FPGA), workload type (training, infehy rence, storage, HPC), number of instances, placement constraints (region/zone/ affinity), image and network configuration, and reservation mode.

The Cloud Manager operates at the same architectural level as the Network Controller, but with a compute-focused scope. It does not make orchestration decisions but serves as an intelligent agent for resource reporting and enforcement. All interactions with the network plane occur indirectly via the NSOS, ensuring separation of concerns and a clean interface model.

5.3. Network Controller

The Network Controller in UNCO serves as a programmable interface between orchestration logic and the physical or virtual network infrastructure. It is responsible for interpreting policies and traffic engineering directives from NSOS and translating them into actionable configurations on network devices or SDN agents.

As the network-facing component, the controller collects real-time metrics from the underlying transport and access networks, including traffic utilization, link health, congestion indicators, and routing anomalies. These insights feed back into NSOS to enable adaptive reconfiguration in response to network dynamics. The controller also supports integration with emerging technologies such as P4 programmable data planes and segment routing protocols, allowing fine-grained per-flow steering based on SLA metadata or service tags.

The Network Controller performs programmable data-plane management and service-aware traffic engineering:

* Telemetry-Driven Path Optimization: Continuously monitors link quality (bandwidth, jitter, RTT, congestion).

* Dynamic QoS Enforcement: Applies differentiated service policies (e.g., priority queues, rate limits, ECN) based on slice and service IDs.

* Programmable Fabric Support: Interfaces with SDN controllers, P4 switches, or segment routing agents for granular traffic steering.

* Inter-Domain Routing Federation: Coordinates with external network controllers (e.g., IP/MPLS, BGP peers) for path stitching across domains.

The Network Controller, like the Cloud Manager, is coordinated by the NSOS. While the Cloud Manager provides visibility into compute supply, the Network Controller ensures that the transport infrastructure aligns with compute demand. Together, they enable closed-loop orchestration in real-time, multi-domain environments.

6. Standard Interfaces and Functional Requirements

6.1. Standard Interfaces

The UNCO framework defines standard interfaces between its components to support unified orchestration and closed-loop control across cloud and network domains. The interfaces are categorized as follows:

1) Cloud Manager - NSOS Interface

This interface enables the Cloud Manager to provide real-time cloud resource status to NSOS.

* IN1.1 Resource Metrics Report (Cloud Manager → NSOS)
  - Parameters:
    - Resource Identification: VM ID/Container Group/Storage Volume ID.
    - Indicator type: CPU utilization/memory usage/disk IOPS/GPU load.
    - Sampling period: seconds/minutes/event triggered.
    - Related service tags: Service/Tenant/SLA level.
  - Purpose: Enables NSOS to assess cloud-side resource availability and support informed scheduling decisions.
* IN1.2 Service Status Report (Cloud Manager → NSOS)
  - Parameters:
    - Computing power requirements: computing power types (CPU/GPU/FPGA), Resource quantity (number of CPU cores/memory/GPU model and quantity), Scenarios (training/inference/storage/high-performance computing) .

- Network status: topology, bandwidth, latency and other information Deployment configuration: availability data center, image identification (operating system/preset image ID), network configuration (VPC ID/subnet ID/security group rule summary).

- Resource pre-occupation: resource pool type (public cloud/ private cloud/hybrid cloud), pre-occupation mode (on-demand/ reserved instance), storage configuration (type/capacity/ IOPS).

- Purpose: Supports service lifecycle management, monitoring, and fault recovery.

3) IN2: NSOS - Network Controller Interface

This interface allows the NSOS to dynamically program the network according to real-time cloud and service state and requirements.

* IN2.1 Issuing of Network Control Policy (NSOS → Network Controller)

  - Parameters:

    - Link identifier: source/destination node ID, logical link name.

    - Cloud Serivce instance ID, a globally unique identifier assigned to each cloud-based service instance (such as a virtual machine, container, or function) deployed within the Telecom Cloud. This ID is used for tracking, management, and associating network policies to specific service instances.

    - Target bandwidth required (Mbps/Gbps) .

    - Effective method: immediate effect/smooth transition (rate gradient time window).

  - Purpose: Issuing of Network Control Policy.

* IN2.2 Report of Network Status (Network Controller→ NSOS)

  - Parameters:

    - Link ID: Logical link globally unique identifier.

  - Real-time bandwidth utilization: current traffic percentage (%) .
  - Delay and packet loss: Avg/Max delay (ms) and packet loss rate (%) in the most recent sampling period.
  - Timestamp: Data collection time.

- Purpose: Provides telemetry for closed-loop network control and orchestration optimization.

6.2. Functional Requirements

To ensure UNCO support a wide range of networked applications across edge, cloud, and transport environments, it defines a set of functional requirements that guide its architectural design and interface behaviors. These requirements emphasize responsiveness, reliability, and compatibility across multi-vendor, multi-domain infrastructures. The following functions are essential to enable joint orchestration of computing and networking resources while preserving service quality, optimizing resource utilization, and maintaining policy consistency.

Here are some functional requirements:

* FR1: SLA-compliant orchestration for computing, network, and storage resources.
* FR2: Elastic demand-driven scheduling based on real-time data and service intent.
* FR3: Inter-domain policy normalization and conflict mitigation across compute and network planes.
* FR4: Observability and feedback mechanisms for orchestration decisions.
* FR5: Unified access control, audit trails, and policy enforcement across domain.

7. Conclusion

Cloud computing has become a foundational component in the infrastructure of modern telecom operators. With the increasing deployment of cloud-based AI services and edge-native applications, it is essential to support integrated orchestration of cloud and network resources as well as end-to-end security management. UNCO addresses these requirements by providing mechanisms to incorporate cloud-related information into network control and policy decision-making, enabling dynamic, SLA-driven service management.

However, the lack of standardized interfaces and models for exchanging cloud telemetry across the network domain remains a key obstacle. Cross-domain collaboration is often hindered by proprietary APIs, inconsistent abstractions, and limited interoperability. These limitations result in delayed network adjustments and fragmented service delivery.

UNCO addresses these challenges by proposing a unified framework and standardized interfaces that bring real-time cloud awareness into network orchestration. Its ability to coordinate compute and network resources holistically enables more resilient, efficient, and SLA-compliant service delivery across public clouds, private datacenters, and edge platforms.

As UNCO continues to evolve, its ability to bridge these gaps through telemetry integration, policy abstraction, and multi-domain orchestration will be critical. Potential application scenarios include:

* Elastic AI/ML service hosting at the edge and core, requiring workload-aware bandwidth and path adjustments.

* Immersive applications (AR/VR/XR, cloud gaming, real-time collaboration) that rely on strict latency and jitter guarantees.

* Dynamic multi-cloud interconnection for enterprise-grade network slicing and hybrid connectivity, etc.

These emerging services demand orchestration frameworks like UNCO that go beyond siloed resource management and offer unified, programmable, and standards-aligned operational control.

UNCO presents a comprehensive framework for integrating computing and networking orchestration in modern networks. By addressing dynamic scheduling, multi-objective trade-offs, cross-domain policy harmonization, and end-to-end security, UNCO provides a strong foundation for enabling future-ready services.

8. IANA Considerations

TBD

9. Acknowledgement

TBD

10. Normative References

[draft-ietf-cats-framework]
"Computing-Aware Traffic Steering Framework".

[draft-ietf-opsawg-service-assurance-architecture]
"draft-ietf-opsawg-service-assurance-architecture Service Assurance Architecture".

[draft-ietf-teas-ietf-network-slice-framework]
"draft-ietf-teas-ietf-network-slice-framework IETF Network Slice Framework", September 2019.

[NFV033] "ETSI GS NFV-IFA 033-2020", September 2010.

[RFC2119] "RFC2119".

[RFC8174] "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words".

[RFC8969] "A Framework for Automating Service and Network Management with YANG".

Authors' Addresses

Xueting Li
China Telecom
Beiqijia Town, Changping District
Beijing
Beijing, 102209
China
Email: lixt2@foxmail.com

Cong Li
China Telecom
Beiqijia Town, Changping District
Beijing
Beijing, 102209
China

Email: licong@chinatelecom.cn