NeoTec
Internet-Draft
Intended status: Informational
Expires: 7 January 2026

Jiayuan. Hu, Ed.
F. Zhang, Ed.
Y. Zhu, Ed.
C. Xie
China Telecom
6 July 2025

Use cases in network operations in telco cloud
draft-hu-neotec-usecases-notc-00

Abstract

This document presents two network operations in telco cloud orchestration use case for AI-based video recognition in smart city management and dynamic high-bandwidth transport. Key innovations include dynamic resource scheduling across heterogeneous computing (GPU/NPU) and network domains, centralized training with distributed inference, and low-latency data transmission compliant with data sovereignty requirements. Additionally, the use case demonstrates elastic bandwidth provisioning and failover mechanisms to ensure reliability. The framework highlights the need for standardized interfaces between cloud and network controllers to optimize performance, resource utilization, and QoS in telecom cloud environments.

Table of Contents

1. Introduction

This document presents two network operations in telco cloud scheduling use case including AI-based video recognition in smart city management and dynamic high-bandwidth transport. The AI-based video use case addresses critical urban governance challenges including illegal street vending, unauthorized parking, garbage disposal, and waste classification through intelligent video analysis.

dynamic high-bandwidth transport is an innovative network solution to address the challenges of large-scale, cross-regional data migration for high-performance computing (HPC), AI training, scientific research, and enterprise applications. It provides on-demand, elastic, and secure high-bandwidth connectivity tailored for temporary or periodic bulk data transfers, significantly reducing costs and improving efficiency compared to traditional methods like physical hard disk shipping or fixed-bandwidth dedicated lines. It enables instant setup and teardown of connections, allowing users to request bandwidth (1G-100G) only when needed (e.g., for scheduled

nighttime transfers). Supports multi-dimensional billing (bandwidth, duration, distance, traffic volume, or usage frequency). Moreover, dynamic high-bandwidth transport can dynamically adjust bandwidth (e.g., from 30M to 10G in seconds) to match data transfer demands and implements network slicing (FlexE/IPv6+), SRv6 tunneling, and encryption to ensure data isolation and integrity.

One of the example of dynamic high-bandwidth transport is LHAASO cosmic ray observatory, it transfers 11PB data from Sichuan to Beijing for processing per year. Reduced a 1.6TB transfer over 2,000 km to 40 minutes (vs. days via hard disks). In AI/ML training area, dynamic high-bandwidth transport supports large-scale dataset migration to GPU clusters for distributed training. Moreover, PB-scale video rendering can be uploaded to cloud-based post-production studios in hours (e.g., 2TB/day via 10Gbps) in media production scenario.

2. Conventions Used in This Document

2.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2.2. Abbreviations

LHAASO: Large High Altitude Air Shower Observatory

UCMP: Unequal-Cost Multi-Path routing

3. Problem Statement

Telecom Clouds integrate compute, storage, and networking resources to deliver low-latency, high-bandwidth services such as 5G, AI/ML workloads, and real-time media processing. Unlike public clouds that depend on third-party networks, Telecom Clouds are operated under a single administrative domain, enabling tight coupling between cloud infrastructure and network operations. However, existing network management systems lack real-time visibility into dynamic cloud resource states, resulting in suboptimal performance, inefficient resource utilization, and SLA violations. Key challenges include:

1. Network controllers remain unaware of cloud-side scaling events (e.g., VM/container orchestration, GPU resource allocation), preventing dynamic adjustments to load balancing, UCMP routing, or QoS policies.

2. While cloud platforms (e.g., AWS CloudWatch, Azure Monitor) expose resource metrics, no standardized APIs or data models exist for network controllers to ingest and act on this telemetry in real time.

3. AI/ML pipelines, 5G network slicing, and inter-cloud traffic exhibit highly variable patterns. Without real-time coordination between cloud resource availability and network state, traffic engineering becomes reactive, leading to congestion, unbalanced resource usage, and degraded QoE.

4. standardized interface for informing routing decisions like UCMP weight adjustments, flow steering, or bandwidth allocation.[draft-li-unco-framework]

5. Traditional network orchestrators often pre-allocate resources statically or based on historical models, but modern applications demand rapid provisioning and adjustment of both compute and network resources. Real-Time and dynamic resource scheduling ability is needed.[draft-li-unco-framework]

To solving the problem is critical to achieving true cloud-network convergence, where dynamic cloud workloads and network resources are orchestrated as a unified system.

4. Use Cases

4.1. Example 1: AI-based Video Recognition for City Management

This use cases leverages cloud-network-computing integration to enable intelligent urban governance through real-time video analytics. Key Applications include

1. Illegal Street Vending Detection: Identifies static objects (e.g., tables, chairs) left in restricted zones for prolonged periods, indicating unauthorized vending activities.

2. Unauthorized Parking Monitoring: Detects vehicles parked in no-parking areas by analyzing predefined zones in video feeds.

3. Litter and Waste Management: Flags scattered waste (bottles, paper, bags) on streets and overflowing/uncovered trash bins.

4. Public Space Compliance: Monitors violations like disorderly wiring and shopfront obstructions.

For the cameras used in urban management, the main network structure is shown in the following figure:

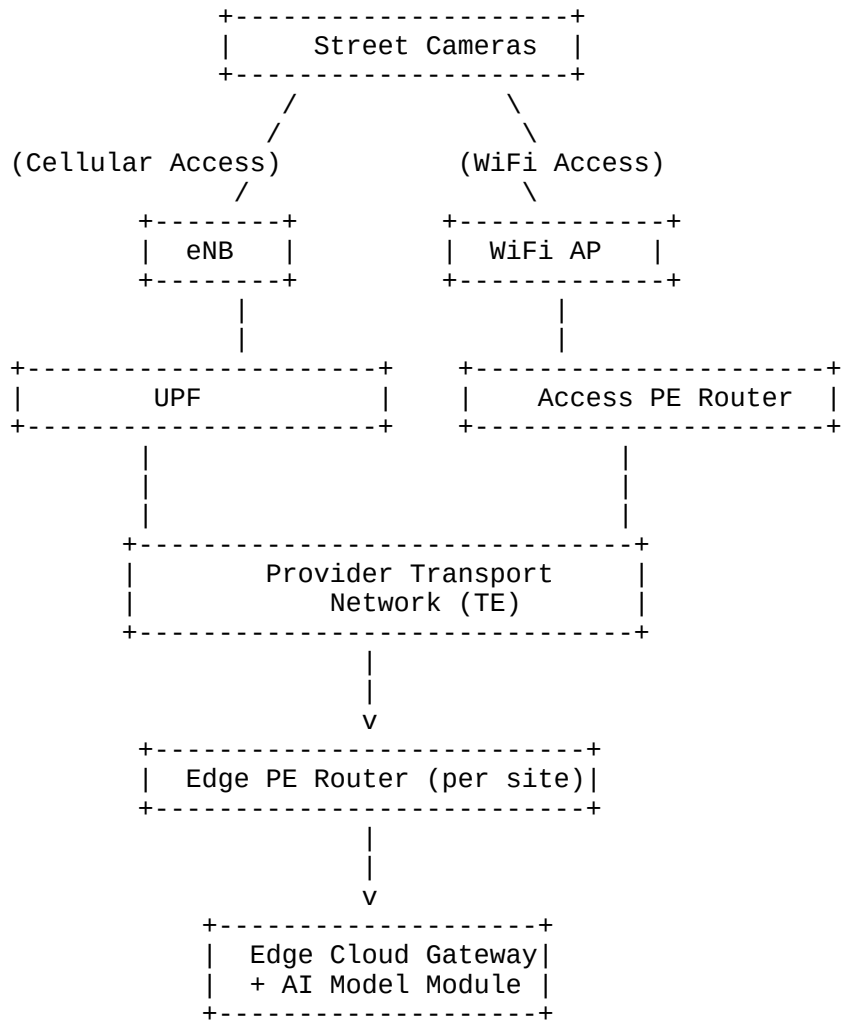

Figure 1: The framework of AI-based video recognition for city management

In the cloud-network convergence architecture, AI cameras transmit data via cellular networks (e.g., through eNBs/gNBs and User Plane Functions (UPFs)) or WiFi Access Points.For cellular access, data is

forwarded via GTP-U tunnels from eNBs to UPFs, which are often co-located with Edge Cloud sites. Data traverses the provider's transport network between the access point (PE router) and the Edge Cloud PE router. The Edge PE router connects to the Edge Cloud Gateway or compute node hosting the AI workload (e.g., real-time inference modules). A Cloud Manager evaluates end-to-end paths (bandwidth, latency, topology) between cameras and Edge Cloud sites to select optimal deployment locations for AI models.Network controllers dynamically adjust UCMP (Unequal Cost Multipath) load-balancing algorithms to meet performance constraints (e.g., XX Gbps bandwidth, YY ms delay) for inter-site data exchange. [draft-dunbar-neotec-ac-te-applicability]

This architecture ensures low-latency, high-throughput data transmission for real-time AI processing while enabling dynamic resource allocation based on network-aware metrics. The solution leverages edge computing infrastructure to deploy AI inference models closer to data sources, enabling real-time processing of high-resolution video streams with millisecond-level response times. Key technical components include:

1.Centralized training at the group data center with distributed edge inference

2.Dynamic resource orchestration across heterogeneous computing facilities (GPU/NPU-enabled edge nodes)

3.Cloud-aware network optimization ensuring low-latency data transmission

4.Data sovereignty compliance through localized processing

4.2. Example 2: dynamic high-bandwidth transport

The dynamic high-bandwidth transport is an innovative network solution designed to address the challenges of large-scale, high-efficiency data migration for scenarios such as scientific computing, AI training, and cross-regional data transfers. Typical Use Cases like East-to-West Data Storage: Low-cost cold/backup data transfers to western data centers. Scientific Computing: Supports projects like the LHAASO cosmic ray observatory (11PB/year data) with high-speed links to supercomputing centers. AI/Media Production: Accelerates raw footage (e.g., 2TB/day) or AI model training data transfers.

The network architecture of thedynamic high-bandwidth transport line service comprises three layers: the service enabling layer, the service core layer, and the business carrying layer. It can provide

services for various data transmission businesses such as gene sequencing, scientific computing, cloud-to-cloud storage, film and television production, artificial intelligence, and more. The service enabling layer, through user-oriented unified APIs, SDKs, or service platforms, invokes network capabilities and allocates network resources on demand based on the business requirements transmitted by various applications, generating a combination of network capabilities and business capabilities. The service provides users with network capabilities such as elastic bandwidth, security isolation, flexible networking, deterministic resource assurance, and flexible billing based on usage, according to the business requests transmitted by the service enabling layer. The business carrying layer, as the physical carrier providing data transportation, builds on-demand, deterministic, secure, and reliable network channels between communicating parties, including network functional entities such as access terminals, super business gateways, and routers. The specific network architecture is shown in Figure 2.

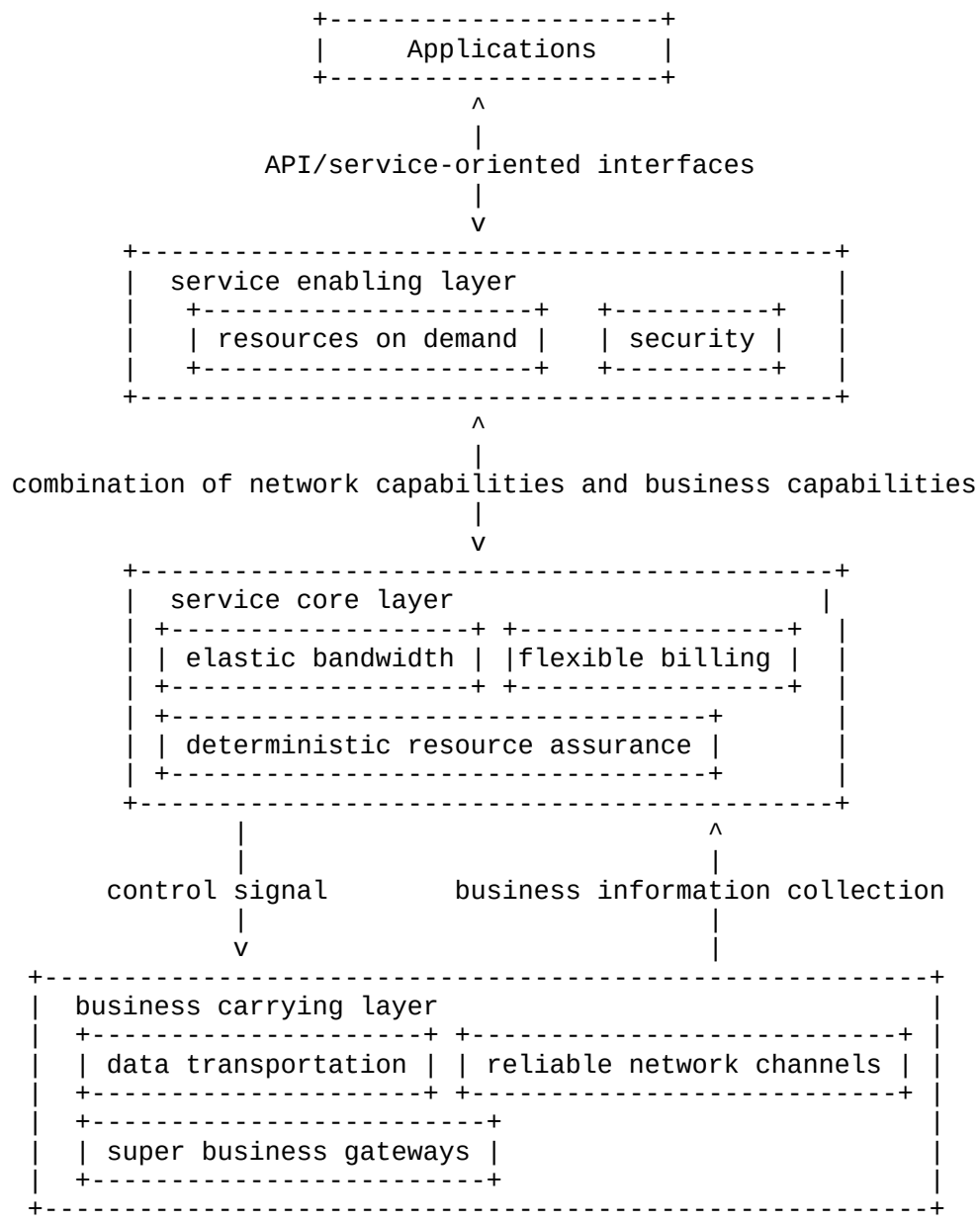

Figure 2: The framework of dynamic high-bandwidth transport

Overall, dynamic high-bandwidth transport can significantly reducing costs and improving efficiency compared to traditional methods like physical hard disk shipping or fixed-bandwidth dedicated lines. Here’s an overview of its key aspects:

1. Task-Based On-Demand Service: Supports instant setup and teardown of connections (1G-100G bandwidth) for temporary or scheduled data transfers (e.g., nighttime off-peak usage).

2. Elastic Bandwidth: Allows dynamic adjustment of bandwidth (e.g., from 100M to 10G) to meet burst demands while maintaining cost efficiency.

3.High-Bandwidth and Low-Latency: Optimizes protocols (e.g., TCP/UDP), leverages wide-area RDMA for lossless transmission, and uses load balancing (e.g., SRv6 UCMP) to maximize throughput.

4. Security and Reliability: Ensures end-to-end isolation via FlexE slicing and VPNs, with built-in encryption (IPSec) and route authentication (RPKI).

5. Cross-Domain Coordination: Enables multi-domain/operator collaboration through centralized or distributed control planes for seamless resource scheduling.

5. Requirements

To enable seamless cloud-network integration across edge, core, and transport environments, cloud-network integration framework establishes a set of functional requirements that drive its architecture and interface design. These requirements prioritize:

1. To achieve dynamic resource scheduling, the system MUST support real-time elastic scaling of computing resources (e.g., GPU containers) and network bandwidth based on AI workload fluctuations.

2. Upon detecting GPU node failures or BGP route oscillations, the system SHOULD automatically migrate services to back up nodes and activate OTN protection rings within 60 seconds.

These requirements emphasize responsiveness, reliability, and compatibility in multi-domain environments ensures cloud-native applications (e.g., AI/ML, XR) achieve deterministic performance while maintaining operational efficiency in cloud-network fused environments.

6. IANA Considerations

TBC

7. Security Considerations

TBC

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[draft-li-unco-framework] "Unified Network and Cloud Orchestration Framework".

[draft-dunbar-neotec-ac-te-applicability] "Applying Attachmet Circuit and Traffic Engineering YANG Data Model to Edge AI Use Case".

Contributors

Thanks to all of the contributors.

Authors' Addresses

Jiayuan Hu (editor)
China Telecom
109, West Zhongshan Road, Tianhe District
Guangzhou
Guangdong, 510000
China
Email: hujy5@chinatelecom.cn

Fan Zhang (editor)
China Telecom
109, West Zhongshan Road, Tianhe District
Guangzhou
Guangdong, 510000
China
Email: zhangf52@chinatelecom.cn

Yongqing Zhu (editor)
China Telecom
109, West Zhongshan Road, Tianhe District
Guangzhou
Guangdong, 510000
China
Email: zhuyq8@chinatelecom.cn

Chongfeng Xie
China Telecom
Beiqijia Town, Changping District
Beijing
102209
China
Email: xiechf@chinatelecom.cn