Draft new Recommendation ITU-T Y.3169 (ex.Y.REOUPF)

Resource Efficiency Optimization for managing User Plane Function in IMT-2020 networks and beyond

Summary

This Recommendation specifies the requirements, architecture, and procedures for resource efficiency optimization for managing user plane function (UPF) in IMT-2020 networks and beyond. The specified architecture enables dynamic resource orchestration of containerized UPFs in distributed edge computing environments through monitoring, predictive analytics, and policy-driven control to enhance resource utilization while ensuring service stability and performance.

Keywords

Resource efficiency, management and orchestration, user plane function (UPF)

Table of Contents

1. Scope 3

2. References 3

3. Definitions 3

3.1 Terms defined elsewhere 3

3.2 Terms defined in this Recommendation 4

4. Abbreviations and acronyms 4

5. Conventions 5

6. Overview 5

7. Requirements for resource efficiency optimization for managing UPF 6

7.1 General requirements 6

7.2 Functional requirements 6

8. Architecture of resource efficiency optimization for managing UPF 7

8.1 Architecture 7

8.2 Functional entities 8

8.3 Reference points 9

9. Procedures for resource efficiency optimization for managing UPF 10

9.1 Procedure for resource efficiency data collection and analysis 10

9.2 Procedure for resource efficiency policy generation and enforcement 10

9.3 Procedure for traffic steering and load balancing 11

9.4 Procedure for resource elastic scaling 12

10. Security considerations 12

Bibliography 13

Draft new Recommendation ITU-T Y.3169 (ex.REOUPF)

Resource Efficiency Optimization for managing User Plane Function in IMT-2020 networks and beyond

1. Scope

This Recommendation specifies resource efficiency optimization for managing user plane function (UPF) in IMT-2020 networks and beyond. This Recommendation addresses the following aspects of resource efficiency optimization in IMT-2020 networks and beyond:

- Requirements for resource efficiency optimization for managing UPF in IMT-2020 networks and beyond;
- Architecture of resource efficiency optimization for managing UPF in IMT-2020 networks and beyond;
- Procedures for resource efficiency optimization for managing UPF in IMT-2020 networks and beyond;
- Security considerations for resource efficiency optimization for managing UPF in IMT-2020 networks and beyond.

2. References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published.

The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T Y.3100] Recommendation ITU-T Y.3100 (2017), *Terms and definitions for IMT-2020 network*.

[ITU-T Y.3104] Recommendation ITU-T Y.3104 (2018), *Architecture of the IMT-2020 network*.

[ITU-T Y.3111] Recommendation ITU-T Y.3111 (2017), *IMT-2020 network management and orchestration framework*.

[ITU-T Y.3150] Recommendation ITU-T Y.3150 (2020), *High-level technical characteristics of network softwarization for IMT-2020*.

[ITU-T Y.3158] Recommendation ITU-T Y.3158 (2022), *Local shunting for multi-access edge computing in IMT-2020 networks.*

[ITU-T Y.3535] Recommendation ITU-T Y.3535 (2022), *Cloud computing-Functional requirements for a container*.

3. Definitions

3.1 Terms defined elsewhere

This document uses the following terms defined elsewhere:

3.1.1 container [ITU-T Y.3535]: A set of software to provide isolation, resource control and portability for virtualization processing of an application.

NOTE – A container runs on the kernel in a bare-metal machine or virtual machine.

NOTE – "Application" implies business logic including a required library or binary to run in a container.

3.1.2 IMT-2020 [ITU-T Y.3100]: Systems, system components, and related technologies that provide far more enhanced capabilities than those described in [b-ITU-R M.1645].

3.1.3 management [ITU-T Y.3100]: In the context of IMT-2020, the processes aiming at fulfilment, assurance, and billing of services, network functions, and resources in both physical and virtual infrastructure including compute, storage, and network resources.

3.1.4 network function [ITU-T Y.3100]: In the context of IMT-2020, a processing function in a network.

NOTE 1 – Network functions include but are not limited to network node functionalities, e.g., session management, mobility management and transport functions, whose functional behaviour and interfaces are defined.

NOTE 2 – Network functions can be implemented on a dedicated hardware or as virtualized software functions.

NOTE 3 – Network functions are not regarded as resources, but rather any network functions can be instantiated using the resources.

3.1.5 network softwarization [ITU-T Y.3100]: An overall approach for designing, implementing, deploying, managing and maintaining network equipment and/or network components by software programming.

NOTE – Network softwarization exploits the nature of software such as flexibility and rapidity all along the lifecycle of network equipment and/or components, for the sake of creating conditions that enable the re-design of network and services architectures, the optimization of costs and processes, self-management and bring added values in network infrastructures.

3.1.6 orchestration [ITU-T Y.3100]: In the context of IMT-2020, the processes aiming at the automated arrangement, coordination, instantiation and use of network functions and resources for both physical and virtual infrastructure by optimization criteria.

3.2 Terms defined in this Recommendation

None.

4. Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

CPU	Central Processing Unit
DMA	Data Monitoring Agent
DMF	Data Monitoring Function
M&O	Management and Orchestration
NF	Network Function
NFV	Network Function Virtualization
PMF	Policy Management Function
QoS	Quality of Service
REC	Resource Efficiency Controller
RP	Reference Point
RPF	Resource Prediction Function
SCF	Service Classification Function

SDN	Software Defined Network
TSF	Traffic Steering Function
UMF	Unified Metrics Function
UPF	User Plane Function
VM	Virtual Machine
VNF	Virtualized Network Function

5. Conventions

In this Recommendation:

The keywords "**is required to**" indicate a requirement which must be strictly followed and from which no deviation is permitted if conformance to this document is to be claimed.

6. Overview

Emerging vertical industrial applications (e.g., vehicular communication application, IoT application, remote surgery application) impose differentiated demands for network resources on IMT-2020 networks and beyond, particularly on the UPF deployed at the network edge, which handles user plane data routing between end users and external data networks. Network softwarization [ITU-T Y.3150] abstracts heterogeneous infrastructure as virtualized network, computing and storage resources through software-defined networking (SDN) and network function virtualization (NFV), shifting network functions (NFs) such as UPF from dedicated hardware to virtualized network functions (VNFs) decoupled from general purpose servers. Edge computing [ITU-T Y.3158] enables distributed UPF deployment with local traffic shunting and low-latency routing, while containerization technologies [ITU-T Y.3535] implement UPF via lightweight container instances, enhancing resource efficiency through spatial optimization and granular resource allocation. However, the dynamic traffic patterns and user mobility necessitate flexible, elastic and highly efficient resource orchestration for UPF while ensuring service stability. Managing diverse resources across geographically distributed edge nodes with constrained capabilities increases orchestration complexity and difficulty in IMT-2020 networks and beyond, hindering optimal resource utilization and service stability.

Therefore, introducing resource efficiency optimization for managing the UPF can be a promising solution to improve resource utilization while ensuring service stability in IMT-2020 networks and beyond. Figure 6-1 shows overview of the resource efficiency optimization for UPF management. The resource efficiency controller (REC) generates optimization policies through continuous monitoring and predictive analytics. Based on these policies, the IMT-2020 network management and orchestration (M&O) system then performs dynamic UPF resource orchestration to ensure that service requirements can be satisfied during workload variations, while enhancing overall resource efficiency.

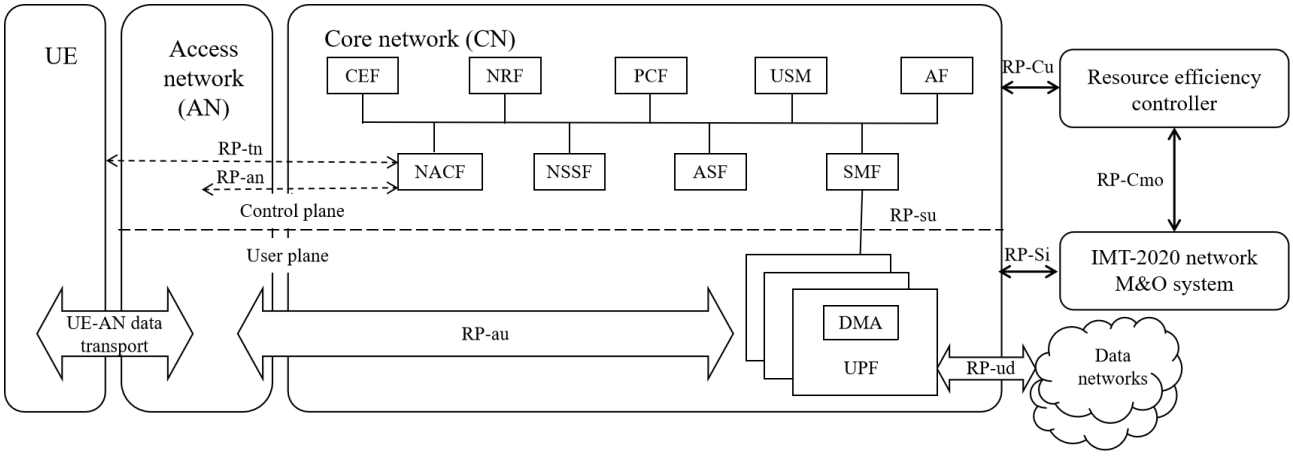

Figure 6-1 Overview of resource efficiency optimization for managing UPF

7. Requirements for resource efficiency optimization for managing UPF

7.1 General requirements

[REQ-GR1]: The resource efficiency optimization architecture is required to be compatible with existing IMT-2020 network management and orchestration system.

[REQ-GR2]: The resource efficiency optimization architecture is required to support multi-domain interoperability for UPF instances deployed in distributed edge environments.

[REQ-GR3]: The UPFs are required to support container-based deployment in edge computing environments along with other edge computing services in IMT-2020 networks and beyond.

[REQ-GR4]: The resource efficiency optimization orchestration is required to ensure end-to-end service stability and performance during resource reconfiguration and scaling operations.

7.2 Functional requirements

7.2.1 Requirements for resource efficiency controller (REC)

[REQ-REC1]: The REC is required to support functions of data monitoring, service classification, united metrics, resource prediction, policy management and traffic steering.

[REQ-REC2]: The REC is required to support integration with the IMT-2020 network management and orchestration system.

[REQ-REC3]: The REC is required to support machine learning-based algorithms for predictive analytics and proactive resource optimization.

7.2.2 Requirements for data monitoring function

The data monitoring function (DMF) is required to support the collection, processing, and management of multi-dimensional monitoring data from UPF instances and their underlying infrastructure through a coordinated operation between a centralized function within the REC and distributed data monitoring agents (DMAs). The DMAs are lightweight software agents deployed at UPF instances, responsible for local data collection, initial pre-processing, and transmission of the processed data to the central data monitoring function.

[REQ-DM1]: The DMA is required to collect multi-dimensional monitoring data from the local network infrastructure, including resource metrics (e.g., central processing unit (CPU), memory, and network bandwidth capacity and utilization), UPF performance metrics (e.g., packet loss, session statistics, forwarding latency), and service characteristics (e.g., service priority, quality of service (QoS), user location, traffic patterns such as burst, periodic, continuous).

NOTE – The local network infrastructure includes the containers, clusters, and hosts supporting the UPF instance where the DMA is deployed.

[REQ-DM2]: The DMF is required to consolidate, correlate, and manage the monitored data to provide a unified basis for resource optimization.

NOTE – This includes capabilities for coordinating data aggregation from distributed DMAs, performing data correlation, and managing stored data for analysis and reporting.

[REQ-DM3]: The DMF is required to support capabilities for data pre-processing, including but not limited to deduplication, imputation, and error correction.

[REQ-DM4]: The DMF is required to ensure the integrity, confidentiality, and availability of collected data throughout its lifecycle, from collection at the DMA to storage and processing at the REC.

[REQ-DM5]: The DMA is required to support real-time detection and reporting of anomalies based on configured thresholds or locally statistical analyzed metrics.

[REQ-DM6]: The DMA is required to be lightweight and designed to minimize resource consumption on the UPF instances and host infrastructure.

7.2.3 Requirements for service classification function

[REQ-SC1]: The service classification function (SCF) is required to categorize services into different service types (e.g., high-priority online, low-priority online, and offline services) based on monitoring data including service priority, QoS requirements, and traffic patterns.

[REQ-SC2]: The SCF is required to support the data visualization capabilities such as cluster portraits, resource portraits, and service portraits.

[REQ-SC3]: The SCF is required to dynamically update service categories based on real-time traffic patterns and application demands.

7.2.4 Requirements for unified metrics function

[REQ-UM1]: The unified metrics function (UMF) is required to establish a set of indicators for collected monitoring data.

[REQ-UM2]: The UMF is required to support customizable and recommended indicators for specific scenarios such as UPF scaling and edge computing resource allocation.

[REQ-UM3]: The UMF is required to provide normalized metrics to facilitate cross-domain and cross-layer resource optimization.

7.2.5 Requirements for resource prediction function

[REQ-RP1]: The resource prediction function (RPF) is required to forecast utilization trends and levels based on historical and real-time data.

[REQ-RP2]: The RPF is required to predict the future peak/valley values of resource utilization and their time of occurrence based on the fluctuation of traffic workloads over a monitoring period.

[REQ-RP3]: The RPF is required to support both active (proactive) and passive (reactive) prediction modes.

[REQ-RP4]: The RPF is required to employ machine learning models and algorithms for efficient forecasting.

7.2.6 Requirements for policy management function

[REQ-PM1]: The policy management function (PMF) is required to automatically generate the resource efficiency optimization policies.

[REQ-PM2]: The PMF is required to support the flexible resources recommendation policy based on predictive insights.

[REQ-PM3]: The PMF is required to generate policies for real-time resource release, waste identification, exception and conflict detection.

[REQ-PM4]: The PMF is required to generate CPU preemption policies for high priority services and active eviction for low priority services.

7.2.7 Requirements for traffic steering function

[REQ-TS1]: The traffic steering function (TSF) is required to dynamically route traffic to optimal UPF instances based on service requirements, user location, and network conditions.

[REQ-TS2]: The TSF is required to distribute traffic loads across UPF instances to prevent congestion and ensure efficient resource utilization.

[REQ-TS3]: The TSF is required to support seamless handover of user sessions between UPF instances during mobility events or resource reallocation.

8. Architecture of resource efficiency optimization for managing UPF

8.1 Architecture

The architecture for UPF resource efficiency optimization comprises three functional entities that interact through defined reference points, as illustrated in Figure 8-1.

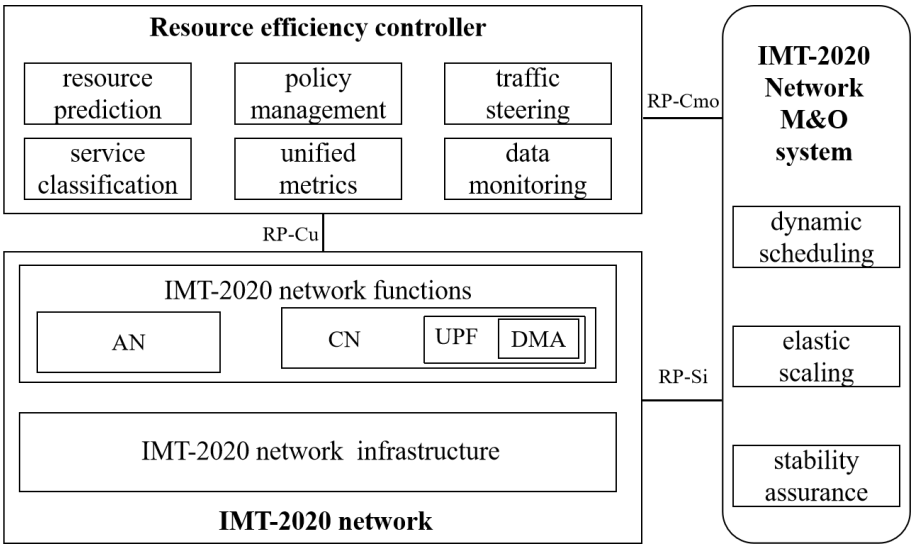

Figure 8-1 Architecture of resource efficiency optimization for managing UPF in IMT-2020 networks and beyond

(1) The **IMT-2020 network M&O system** is responsible for the dynamic orchestration of network resources in accordance with [ITU-T Y.3111]. It executes elastic scheduling and scaling operations (e.g., instantiating, terminating, or scaling UPF container/virtual machine (VM) instances and replicas) based on optimization policies received from the REC. The IMT-2020 network M&O system dynamically prioritizes resource assignment based on the policies with service-aware classification to ensure end-to-end service stability throughout all resource reconfiguration processes.

(2) The **REC** integrates several functions (as detailed in clause 8.2.2) to collect and analyze monitoring data, predict resource trends, generate resource optimization policies, and dynamically steer traffic. It operates based on machine learning-driven analytics to proactively and reactively optimize UPF resource utilization.

(3) The **IMT-2020 network** encompasses the physical and virtual infrastructure, including the access network and core network functions as defined in [ITU-T Y.3104]. DMAs are deployed within its edge-deployed UPF instances. These agents continuously collect and perform initial pre-processing of multi-dimensional performance and resource data, which serves as the foundational input for the optimization processes performed by the REC.

This architecture enables cooperative operation: The DMAs in the IMT-2020 network provide data to the REC; the REC analyzes the data and generates optimization policies; and the IMT-2020 network M&O system executes these policies, with confirmation and updated status fed back to the REC for continuous refinement.

8.2 Functional entities

8.2.1 IMT-2020 network management and orchestration system

The IMT-2020 network M&O system is enhanced to provide policy-driven orchestration of UPF through its integration with REC. It executes workload-aware and network-topology-aware scheduling and scaling of UPF resources (including container/VM instances and replicas) through predictive analytics. Concurrently, it enforces service stability constraints during resource reconfiguration processes, ensuring uninterrupted operations while optimizing physical resource utilization across distributed edge environments.

8.2.2 Resource efficiency controller

The resource efficiency controller (REC) optimizes UPF resource utilization through data analytics and policy-driven control. The REC integrates six functions:

(1) The DMF coordinates with distributed DMAs deployed in UPF instances to collect, aggregate, and pre-process multi-dimensional monitoring data. The DMF consolidates data from multiple DMAs to provide a unified view of the resource status, UPF performance, and service characteristics across the infrastructure.

(2) The SCF categorizes service traffic flows traversing UPFs into high priority online, low priority online, and offline types based on standardized QoS parameters and service priority, and generates cluster, resource, service profiles for visualization and analysis.

NOTE 1 – Offline services are non-latency-sensitive processing tasks (e.g., big data analytics).

NOTE 2 – Cluster profile provides a topological visualization of interconnected hosts/VMs/containers executing UPF instances, depicting operational statuses (active/standby/failed), load distribution, and network latency matrices between nodes.

NOTE 3 – Resource profile represents algorithmic analysis of IMT-2020 network infrastructure resource utilization trends, generating time-series dashboards for CPU/memory/storage usage patterns across hosts, containers, and VMs based on data from DMF.

NOTE 4 – Service profile provides a correlation between classified service types (high/low-priority online, offline) and resource consumption by mapping key quality indicators (e.g., packet loss, latency) to underlying resource utilization metrics for service-resource dependency analysis.

(3) The UMF establishes a set of evaluation indicators from raw monitoring data, enabling customizable indicator definitions for scenario-specific optimization (e.g., UPF instance scaling thresholds).

(4) The RPF uses machine learning models to forecast resource utilization trends and demand patterns proactively and reactively supporting predictive resource optimization.

(5) The PMF generates resource optimization policies based on predictive insights and standardized metrics (e.g., instance/replica count recommendations, capacity planning parameters, and real-time resource release/detection rules), and interfaces with the IMT-2020 network M&O system for policy enforcement.

(6) The TSF dynamically routes traffic to optimal UPF instances by evaluating the real-time service requirements, user location, and network performance indicators, while distributing traffic loads across UPF instances to prevent congestion and ensure efficient resource utilization.

8.2.3 DMA of UPF

DMAs are lightweight functional entities deployed within containerized UPF instances at the network edge. They are responsible for continuous collection and local pre-processing of monitoring data, including resource metrics, UPF performance metrics, and services characteristics. The condensed data is transmitted to the DMF within the REC. DMAs also support real-time anomaly detection and alerting based on locally configured thresholds.

8.3 Reference points

8.3.1 Reference point RP-Cmo

RP-Cmo establishes the resource efficiency optimization policy coordination interface between IMT-2020 network M&O system and REC. It conveys machine-readable resource optimization policies, including predicted scaling parameters, instance deployment blueprints, and stability assurance mechanisms, from the policy management in REC to the IMT-2020 network M&O system, while simultaneously enabling IMT-2020 network M&O system to feed policy execution status and infrastructure constraints back to the REC for closed-loop optimization.

8.3.2 Reference point RP-Si

RP-Si is enhanced to support real-time scheduling instructions transmission in accordance with [ITU-T Y.3111]. RP-Si conveys the resource orchestration commands between the IMT-2020 network M&O system and the IMT-2020 network infrastructure. Through this reference point, the IMT-2020

network M&O system issues real-time scheduling instructions, such as workload redistribution commands, topology-aware placement adjustments, and container/VM instantiation/scaling orders for UPF, to the IMT-2020 network infrastructure, which subsequently return resource utilization confirmations and report orchestration exceptions to IMT-2020 network M&O system for state reconciliation.

8.3.3 Reference point RP-Cu

RP-Cu provides the monitoring and control interface between the REC and the UPF in IMT-2020 network. It facilitates bidirectional communication, transporting three types of information. Uplink transmission delivers pre-processed monitoring data from DMAs to the DMF. Downlink transmission distributes traffic steering rules generated by TSF to forward traffic flows to optimal UPF instances. And infrastructure capability advertisements broadcasting edge node computing capacities and topological constraints to inform REC decision-making processes.

9. Procedures for resource efficiency optimization for managing UPF

9.1 Procedure for resource efficiency data collection and analysis

This clause describes the procedures of resource efficiency optimization data collection and analysis for managing UPF in IMT-2020 networks and beyond, as shown in Figure 9-1.

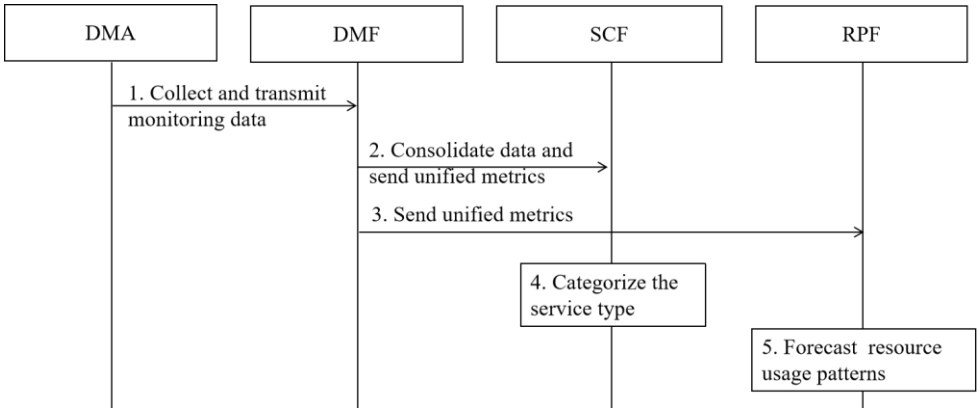

Figure 9-1 Procedure for resource efficiency optimization data collection and analysis

Step 1: The DMAs continuously collect multi-dimensional monitoring data of distributed UPF instances and perform initial pre-processing (e.g., deduplication, error correction, anomaly detection based on local metrics). The pre-processed monitoring data is transmitted to the DMF within the REC.

Step 2: DMF consolidates the monitoring data from multiple DMAs and sends the unified metrics to SCF.

Step 3: DMF sends the unified metrics to RPF.

Step 4: SCF categorizes the services carried by UPF into three types: the high priority online services, low priority online services, and offline services, based on QoS parameters, priority levels, and traffic pattern of the received metrics.

Step 5: RPF analyzes historical and real-time data to generate proactive/reactive forecasts regarding resource usage trends peak/average/minimum utilization time points, and spatial-temporal resource demand patterns for UPF instances.

9.2 Procedure for resource efficiency policy generation and enforcement

This clause describes the procedure for generating and enforcing resource efficiency optimization policies, as shown in Figure 9-2.

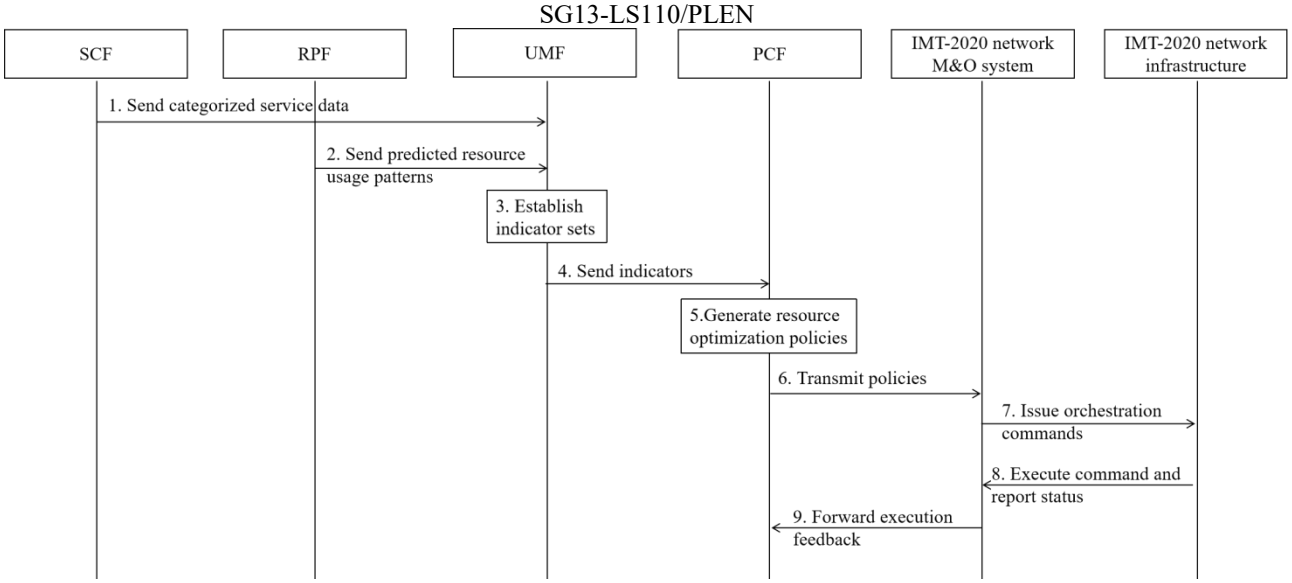

Figure 9-2 Procedure for resource efficiency policy generation and enforcement

Step 1: SCF sends the categorized service data to UMF.

Step 2: RPF sends the predicted resource usage patterns to UMF.

Step 3: Based on these inputs, UMF establishes a set of evaluation indicators.

Step 4: UMF sends evaluation indicators to PMF.

Step 5: PMF generates resource optimization policies (e.g., scaling recommendations, resource release instructions, exception handling rules) based on these indicators.

Step 6: PMF transmits the machine-readable optimization policies to the IMT-2020 network M&O system.

Step 7: The IMT-2020 network M&O system interprets the policies and issues corresponding orchestration commands (e.g., instantiate, scale, migrate, or terminate UPF instances) to the IMT-2020 network infrastructure.

Step 8: The IMT-2020 network infrastructure executes the commands and reports back the execution status and updated resource metrics to the IMT-2020 network M&O system.

Step 9: The IMT-2020 network M&O system forwards the execution feedback to the PMF within the REC for closed-loop policy refinement and validation.

9.3 Procedure for traffic steering and load balancing

This clause describes the procedure for dynamic traffic steering and load balancing across UPF instances, as shown in Figure 9-3.

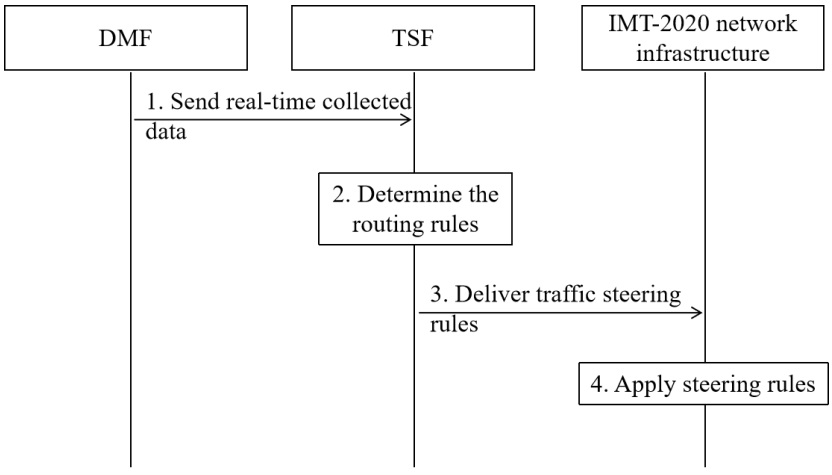

Figure 9-3 The procedure for traffic steering and load balancing

Step 1: DMF within the REC sends real-time network status, user location data, and service requirements to the TSF.

Step 2: TSF determines the optimal UPF instance for routing traffic based on current load, latency, resource availability, and service priority.

Step 3: TSF delivers the traffic steering rules to the relevant UPF instances within the IMT-2020 network infrastructure.

Step 4: The target UPF instances in the IMT-2020 network infrastructure apply the steering rules to forward traffic accordingly.

9.4 Procedure for resource elastic scaling

This clause describes the procedure for elastic scaling of UPF resources based on predicted utilization trends of 3rd party application, as shown in Figure 9-4.

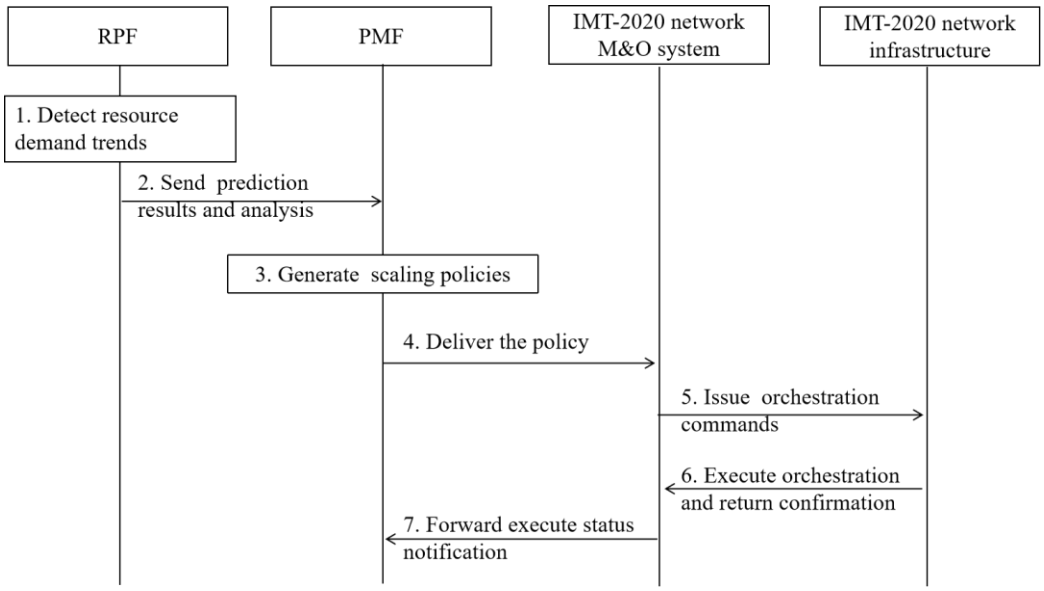

Figure 9-4 Procedure for resource elastic scaling

Step 1: RPF detects a trend indicating increasing resource demand or potential possibility of overload in a specific UPF instance or cluster.

Step 2: RPF sends the prediction results and relevant analysis to PMF.

Step 3: PMF generates a scaling policy, including indication of horizontal scaling (e.g., adding or removing container replicas) or vertical scaling (e.g., adjusting CPU/memory allocations) of UPF resources.

Step 4: PMF sends the policy to the IMT-2020 network M&O system.

Step 5: The IMT-2020 network M&O system initiates scaling operations by issuing the appropriate commands to the IMT-2020 network infrastructure.

Step 6: The IMT-2020 network infrastructure executes the commands and returns an execution confirmation (indicating success or failure) to the IMT-2020 network M&O system.

Step 7: Upon receiving confirmation, the IMT-2020 network M&O system forwards this execution status notification to PMF within the REC.

10. Security considerations

This Recommendation is recognized as an enhancement of IP-based mobile networks to achieve the resource efficiency optimization. Thus, it is assumed that security considerations in general are based on the security considerations identified by [b-ITU-T Y.2701] and [ITU-T Y.3158].

Bibliography

[b-ITU-T Y.2701] Recommendation ITU-T Y.2701 (2007), *Security requirements for NGN release 1*.

[b-ITU-R M.1645] Recommendation ITU-R M.1645 (2003), *Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000*.