ETSI GR NFV-EVE 027 V6.1.1 (2026-04)

GROUP REPORT

Network Functions Virtualisation (NFV) Release 6; Evolution and Ecosystem; Report on Model-as-a-Service (MaaS) in NFV

Disclaimer

The present document has been produced and approved by the Network Functions Virtualisation (NFV) ETSI Industry Specification Group (ISG) and represents the views of those members who participated in this ISG.
It does not necessarily represent the views of the entire ETSI membership.

Reference
DGR/NFV-EVE027

Keywords
AI, MaaS, management, NFV

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI IPR online database.

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™**, **LTE™** and **5G™** logo are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Foreword

This Group Report (GR) has been produced by ETSI Industry Specification Group (ISG) Network Functions Virtualisation (NFV).

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

1 Scope

The present document investigates Model-as-a-Service (MaaS) for AI-based applications in the context of telco cloud management. It describes and analyses a set of relevant use cases, with a focus on the definition and role of MaaS within the telco cloud management.

The present document also describes key issues, potential solutions, and where applicable, it also provides recommendations for enhancements to the NFV architecture and its functionality. The proposed solutions are aiming to provide further support for how MaaS can be introduced to the NFV framework and how MaaS can assist the management of telco cloud services.

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents may be useful in implementing an ETSI deliverable or add to the reader's understanding, but are not required for conformance to the present document.

[i.1] ETSI GR NFV 003: "Network Functions Virtualisation (NFV); Terminology for Main Concepts in NFV".

[i.2] ETSI GS NFV-IFA 047: "Network Functions Virtualisation (NFV) Release 5; Management and Orchestration; Management data analytics Service Interface and Information Model Specification".

[i.3] ETSI GS NFV-IFA 049: "Network Functions Virtualisation (NFV) Release 5; Architectural Framework; VNF generic OAM functions and other PaaS Services specification".

[i.4] ETSI GS NFV-IFA 050: "Network Functions Virtualisation (NFV) Release 5; Management and Orchestration; Intent Management Service Interface and Information Model Specification".

[i.5] ETSI GR NFV-IFA 054: "Network Functions Virtualisation (NFV) Release 6; Architecture; Report on architectural support for NFV evolution".

[i.6] ETSI GR ENI 045 (V4.1.1): "Experiential Networked Intelligence (ENI); Research on Application Scenarios of Network Large Language Models for Operation, Administration, Maintenance, and Performance".

[i.7] ETSI GS ENI 050 (V4.5.1): "Experiential Networked Intelligence (ENI); Lifecycle management for large model".

[i.8] 3GPP TR 28.858: "Study on Artificial Intelligence / Machine Learning (AI/ML) management Phase 2 (Release 19)".

[i.9] ETSI TS 128 105: "5G; Management and orchestration; Artificial Intelligence/ Machine Learning (AI/ML) management (3GPP TS 28.105 Release 19)".

[i.10] TMForum TMF780: "MaaS API Profile".

[i.11] ETSI GR NFV-IFA 046: "Network Functions Virtualisation (NFV) Release 5; Architectural Framework; Report on NFV support for virtualisation of RAN".

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the terms given in ETSI GR NFV 003 [i.1] and the following apply:

NOTE: A term defined in the present document takes precedence over the definition of the same term, if any, in ETSI GR NFV 003 [i.1].

large model application: type of application that invokes large models according to specific requirements to accomplish tasks within a particular application scenario

large-scale model: type of artificial neural network model with complex structure and large number of parameters, enabling it to learn complex patterns and perform tasks with high accuracy

large-scale pre-trained model: type of a large-scale model that is already trained on massive datasets before being adapted to specific tasks

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the abbreviations given in ETSI GR NFV 003 [i.1] and the following apply:

AI	Artificial Intelligence
LLD	Low-Level Design
MaaS	Model-as-a-Service
ML	Machine Learning
RAG	Retrieval-Augmented Generation

4 Introduction and overview

4.1 Background information

4.1.1 Introduction to large model technology

Currently, the integration of Artificial Intelligence (AI) technology to assist in telco cloud management has entered the maturity phase. In the context of NFV AI technology can be utilized to enhance management within various domains, including support for the management data analytics service defined in ETSI GS NFV-IFA 047 [i.2], the generic OAM functions defined in ETSI GS NFV-IFA 049 [i.3], the intent management service defined in ETSI GS NFV-IFA 050 [i.4], among others. With the rapid advancements in AI technology, particularly the ongoing breakthroughs in reinforcement learning, large models, and content generation, the adoption of large models as the foundation for downstream tasks (such as anomaly detection, intelligent maintenance, and intent recognition) has emerged as a new paradigm in AI-based application domains.

The industry often follows a progression from foundational large models to industry-specific large models. Initially, large models are trained to be applicable to multiple domains and tasks, equipping them with robust generalization capabilities that enable extrapolation from one instance to another. Subsequently, industry-related data is used for pre-training or fine-tuning to enhance performance and accuracy within that specific domain.

Based on application scenarios and functionalities, large models can be categorized into several types: large language models (for processing natural language text), vision large models (for processing images and videos), structured large models (for processing structured data), and multimodal large models (for processing multi-modal data such as text, images, audio, etc.). These models have the potential to empower various application scenarios within telco cloud management.

Compared to traditional small models, large models are better equipped to handle the uncertainties and complexities of the network domain, while small models still possess noteworthy advantages in tasks related to resource utilization, interpretability, and stability. Therefore, it is necessary to rationally orchestrate the workflows within the network domain, leveraging the collaboration between large and small models to achieve complementary advantages and enhance the overall system's performance and efficiency.

In the realm of telco cloud management, leveraging the generalization capabilities of general-purpose large models and incorporating exclusive domain-specific data from the telco cloud management field, capabilities such as intent understanding, human-computer interaction, task decomposition, text generation, analysis, and reasoning can be embedded into different application scenarios. Such scenarios include for example operation and maintenance knowledge querying, data statistics analysis, network fault sensing, event delimitation and localization, and fault handling closed loops.

Based on this foundation, large models can be used to gain a deep understanding of the complex dynamics of cloudified network environments, providing robust support for real-time optimization of network resources, fault prediction, and continuous improvement of service quality. This ultimately enhances the intelligence and automation capabilities of telco cloud management processes significantly.

4.1.2 Challenges in large model technology

As large models continue to advance rapidly and application needs grow, the technical and economic costs have risen sharply, bringing along complex challenges.

Firstly, deploying large models requires massive computational resources and data processing capabilities. With the rapid expansion of model parameters, both training and subsequent model inference, which require significant computational support, are leading to increasing costs.

Secondly, large models are technically complex regarding training, optimization, inference, and deployment. They have stricter requirements by means of dataset structure and quality. Additionally, new techniques such as prompt engineering (the process of structuring instructions that can be interpreted and understood by a large model) have further raised the technical bar. This is due to the fact that prompt engineering requires a deep understanding of large model capabilities and limitations, as well as the ability to craft instructions that effectively combine linguistic and technical knowledge.

Furthermore, redundancy in model development is a notable issue, resulting not only in the waste of valuable resources such as computational power, data storage, and development time, but also leading to a reduction in overall efficiency. This occurs when multiple models are created that perform similar tasks, or when models are developed from scratch without leveraging existing solutions or frameworks.

Lastly, the challenges of adapting large models to different scenarios and improving development efficiency urge for attention. Due to their complexity and diversity, efficiently deploying large models in specific business contexts remains a technical challenge to overcome.

NOTE: Distinctions between large models and small models in NFV for existing activities like in ETSI GS NFV-IFA 047 [i.2], ETSI GS NFV-IFA 049 [i.3], and ETSI GS NFV-IFA 050 [i.4] will be explored in future releases.

4.1.3 MaaS enables seamless AI deployment and utilization

MaaS (Model as a Service) encapsulates AI models and their associated capabilities into reusable services, enabling users to swiftly build, deploy, monitor, and invoke models without the need to develop and maintain underlying foundational capabilities. It offers a comprehensive suite of platform tools that streamline model training, tuning, and deployment, empowering users to efficiently customize and bring models into operation. Additionally, MaaS in principle integrates an extensive model library and dataset, thereby eliminating redundant development work.

Furthermore, MaaS features robust application development capabilities, providing platforms or tools tailored to specific scenarios. These enable users to rapidly construct AI applications. The MaaS services can be used to support a range of model enhancement techniques, such as Retrieval-Augmented Generation (RAG), collaboration between large and small models, delivering high-quality services in the form of intelligent agents tailored to different scenarios.

Intelligent agents in NFV can be considered autonomous software entities embedded within the NFV framework that leverage AI/ML and data-driven reasoning to optimize, manage, and control virtualised network resources and services.

MaaS's capabilities encompass three key aspects: pure model inference services (providing direct access to trained models for inference without involving model training or optimization), application-invoked model composite agent services, and integrated management of application/model research, development, and operations. These comprehensive services not only lower technical barriers and encourage model sharing but also enhance application adaptability. Therefore, MaaS facilitates the widespread adoption and efficient utilization of AI models, driving transformation and growth across various industries.

4.1.4 Relevant work in other SDOs

Many SDOs have already carried out research work related to large models and MaaS. In ETSI GR ENI 045 [i.6], research was conducted on how to leverage large model technologies to assist in communication network operations and management. In ETSI GS ENI 050 [i.7], research was conducted on how to encapsulate large models as a service and manage the lifecycle of such services. In 3GPP TR 28.858 [i.8], research on requirements for generative AI (involving large models) is also included, while ETSI TS 128 105 [i.9] defines the ML model lifecycle, ML model lifecycle management capabilities and information model definitions for AI/ML management. In TMForum TMF780 [i.10], APIs related to MaaS are defined.

The open-source organizations CAMARA under the Linux® Foundation and ONAP® (Open Network Automation Platform) under Linux® Foundation Networking have carried out development work on functions and requirements related to MaaS.

NOTE: Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries.

The present document builds upon the existing work of the aforementioned standards and open-source organizations to investigate how MaaS can be introduced to the NFV framework and how MaaS can assist the management of Telco clouds.

5 Use cases

5.1 Overview

Packaging models could assist in various Telco cloud scenarios. For example the combination of large models, small models, various tools, knowledge bases, and other components can be used to form diverse large model applications supporting tasks like in telco cloud maintenance management and operation, among others.

In the context of maintenance management, large model services facilitate automated querying of performance metrics and alarms within Telco clouds. Furthermore, by leveraging Telco cloud fault alarms and related performance data, large models help pinpoint the root causes of faults.

Within operational scenarios, large model technology can assist in the deployment of Telco services. For instance, during the assisted generation of LLD (Low-Level Design) documentation for Telco clouds, a large model specific to Telco clouds provides guidance on configuring and selecting options for LLD documentation creation, thus enabling efficient automation of the documentation process.

5.2 Use case #1: Root cause identification of Telco cloud failures

5.2.1 Introduction

As the scale of Telco cloud equipment increases, the difficulty in rapidly locating and addressing faults in Telco clouds is continuously rising, making it challenging for traditional manual operation and maintenance modes to meet the high-reliability requirements. In response to this situation, large models can be leveraged to enable intelligent fault location identification and automated processing for Telco clouds, thereby reducing fault handling time.

5.2.2 Actors and roles

Table 5.2.2-1 describes the use case actors and roles.

Table 5.2.2-1: Use case #1 actors and roles

#	Actor and role	Description
1	Operator	The user who submits the Telco cloud fault diagnosis request.
2	Large model application	An application leveraging large models for fault diagnosis fault localization and resolution and root cause analysis in Telco cloud management.
3	Large model	A type of artificial neural network with a complex structure and large-scale parameters, facilitating intelligent processing and analysis to support fault diagnosis in Telco cloud management.
4	Operations and maintenance system	A system within NFV-MANO, responsible for raising alarm data queries. It enables the retrieval of alarm data related to fault events in Telco cloud management.
5	Fault diagnosis knowledge base	A knowledge base storing fault operation and maintenance information (e.g. based on manuals) from operators and vendors, including alarm titles, alarm lists, fault detection methods, and handling recommendations. Used to support fault diagnosis and root cause analysis within NFV-MANO.
6	Faulty equipment	The virtual and physical resources managed by the NFVI that are defective.

NOTE: The knowledge base is owned and managed by the network operator who controls the access.

5.2.3 Trigger

Table 5.2.3-1 describes the use case trigger.

Table 5.2.3-1: Use case #1 trigger

Trigger	Description
1	Operator requests to diagnose a Telco cloud fault.

5.2.4 Pre-conditions

Table 5.2.4-1 describes the use case pre-conditions.

Table 5.2.4-1: Use case #1 pre-conditions

#	Pre-condition	Description
1	Both the large model application and the large model are available.	No additional description.
2	Operations and maintenance system is available	No additional description.
3	A pre-built fault diagnosis knowledge base, saving operator and multiple manufacturers' fault operation and maintenance manuals, mainly including various fault alarm lists and corresponding fault detection plans and processing suggestions.	No additional description.
4	Failure in infrastructure causes a fault in the NFVI.	No additional description.

5.2.5 Post-conditions

Table 5.2.5-1 describes the use case post-conditions.

Table 5.2.5-1: Use case #1 post-conditions

#	Post-condition	Description
1	Root cause is identified and fault handling suggestions are provided.	The large model application has identified the root cause of the Telco cloud failure and provided the operator with the relevant fault handling procedures followed or notified the operator if the root cause could not be identified.

5.2.6 Flow description

Table 5.2.6-1 describes the use case flow.

Table 5.2.6-1: Use case #1 flow description

#	Actor/Role	Action/Description
Begins when	Operator -> Large model application	The operator submits a fault diagnosis request in natural language, such as "The alarm count has increased in resource pool A, diagnose and identify the root cause." The request is sent to the large model application.
1	Large model application -> Large model	The large model application processes the request using the large model, extracting key parameters required for alarm data queries and fault analysis.
2	Large model application -> Operations and maintenance system	Using the extracted parameters, the large model application queries the operations and maintenance system, an NFV-MANO component, to retrieve alarm data (e.g. alarms triggered within a specific time range before the fault occurred). This interaction can rely on standardized NFV-MANO alarm object types (see ETSI GS NFV-IFA 045 [i.6]).
3	Large model application -> Fault diagnosis knowledge base	Based on the alarm name obtained in step 2 (e.g. 'memory overload'), the large model application searches the fault diagnosis knowledge base to retrieve the most likely fault types and the corresponding operating procedures for fault location.
4	Large model application -> Large mode	The large model application sends the collected information to the large model, which includes the fault type (retrieved in step 3) and the faulty device information (obtained in step 2). The large model analyses this data and selects the most likely fault type, then obtains the corresponding operating procedures for fault localization. The large model integrates this information with the fault location method.
5	Large model application -> Faulty equipment	The large model application accesses the faulty NFVI equipment managed by NFV-MANO in accordance with the operating procedures for fault localization obtained in step 4. It interacts with the faulty equipment to retrieve detailed diagnostic information.
6	Large model application -> Large model	The large model application sends the collected data, including the fault type and equipment information, to the large model for analysis. The large model determines whether the fault location can be determined. If the fault is localized, the process proceeds to step 7. If the root cause cannot be identified, the large model application reanalyse alternative fault location methods using alarm information. If all possible methods have been tried, the large model application output that fault location cannot be determined.
7	Large model application -> Operator	Once the fault root cause is located, the large model application reports the root cause and affected equipment to the operator. Additionally, the large model application retrieves the most relevant fault handling procedures from the fault diagnosis knowledge base and provides actionable recommendations. If all diagnostic attempts fail and the fault root cause cannot be identified, the large model application notifies the operator of the failure and may suggest alternative diagnostic paths.

5.3 Use case #2: Query of Telco cloud operational metrics

5.3.1 Introduction

Within Telco cloud management, performance and fault management data are typically stored in databases. Querying these databases to extract and monitor key operational metrics often involves complex rule definitions. By leveraging large model technology, query rules can be described in a semantic manner, enabling flexible and context-aware queries for extracting and summarizing operational monitoring indicators. This approach allows operators and administrators to retrieve relevant monitoring data and generate summary reports without directly interacting with database query interfaces or writing query scripts. Consequently, this reduces the technical complexity of Telco cloud operations and maintenance, while improving efficiency, automation, and intelligence in monitoring and troubleshooting workflows.

5.3.2 Actors and roles

Table 5.3.2-1 describes the use case actors and roles.

Table 5.3.2-1: Use case #2 actors and roles

#	Actor and role	Description
1	Operator	The user who submits the query for Telco cloud operational metrics.
2	Large model application	Leveraging large models for processing operator queries related to operational metrics.
3	Large model	A type of artificial neural network with a complex structure and large-scale parameters, enabling the processing of query data and generation of insights for operational metrics.
4	Metrics names knowledge base	A knowledge base storing a list of standard and alternative metric names, along with their relationships, used for efficient querying and identification (e.g. user input like "hard disk" or "storage volume" is mapped to the system's database terminology such as "hard disk".
5	Operational knowledge base	A knowledge base within the Telco cloud management domain that stores key operational data, such as performance indicators and resource utilization, supporting fault detection and performance management.

5.3.3 Trigger

Table 5.3.3-1 describes the use case trigger.

Table 5.3.3-1: Use case #2 trigger

Trigger	Description
1	Operator requests to query the Telco cloud operational metrics.

5.3.4 Pre-conditions

Table 5.3.4-1 describes the use case pre-conditions.

Table 5.3.4-1: Use case #2 pre-conditions

#	Pre-condition	Description
1	Both the large model application and the large model are available.	No additional description.
2	Operational knowledge base is available.	No additional description.
3	Pre-built metrics names knowledge base, storing synonym lists for various metrics names, primarily converting metrics names described in users' natural language into system-recognizable metrics names.	No additional description.

5.3.5 Post-conditions

Table 5.3.5-1 describes the use case post-conditions.

Table 5.3.5-1: Use case #2 post-conditions

#	Post-condition	Description
1	Providing operational metric query results	Providing the corresponding operational metric data and summary reports according to the operators request.

5.3.6 Flow description

Table 5.3.6-1 describes the use case flow.

Table 5.3.6-1: Use case #2 flow description

#	Actor/Role	Action/Description
Begins when	Operator -> Large model application	The operator submits a query request in natural language, specifying the type of metrics, resource pools, filtering rules, and time range. The request is sent to the large model application.
1	Large model application -> Large model	The Large model application sends to the large model the metrics type, time range, and filtering rules based on keywords in the operator's natural language request, translating it into the relevant parameters for database query commands.
2	Large model application -> Metrics names knowledge base	The large model application uses the metric type parsed in step 1 to query the metrics names knowledge base, retrieving the corresponding standardized metric names.
3	Large model application -> Large model	The large model embeds the parsed metric parameters from steps 1 and 2 into the specific query parameters of a query command to generate a database query command for operational metrics.
4	Large model application -> Operational knowledge base	The large model application uses the generated query command to retrieve the requested operational data (e.g. CPU, memory usage) from the operational knowledge base.
5	Large model application -> Large model	The large model application passes the query results along with additional prompt words (e.g. report generation requirements) to the large model for processing the data and generating a report.
6	Large model application -> Operator	The large model application returns the metrics report to the operator, which includes a natural language or graphical summary of the operational metrics, along with any insights gained from the process followed.

5.4 Use case #3: Smart deployment plan generation for Telco cloud

5.4.1 Introduction

The deployment plan details the deployment specifications, storage information, network configurations, and other relevant components. In the NFV context, this includes the configuration of VNFs, their interconnections, and the provisioning of NFVI resources, such as compute, storage, and networking. The plan also outlines the configuration of both VNFs and NFVI resources, to meet high-availability and performance requirements. In addition to initial deployment, the plan also covers configuration changes needed throughout the lifecycle, ensuring that the system adapts to evolving network requirements.

However, the strict high-availability requirements of the deployment plan also present a significant challenge and a high technical barrier for operators to fill out. In response to this situation, large model technology, which supports interaction with users in natural language, extracts configuration parameters required for cloud deployment through multiple rounds of Question & Answer. This technology ultimately assists in generating the deployment plan based on a predefined deployment plan template. The template provides detailed configuration parameters for each component, including VNFs, NFVI resources, and their interconnections, ensuring that high-availability, fault tolerance, and scalability requirements are met. As a result, the technical threshold for formulating deployment plans is lowered, allowing experts to easily generate deployment plans without needing to understand complex technical details. This improvement simplifies the deployment process and increases the efficiency and adaptability of Telco cloud services.

5.4.2 Actors and roles

Table 5.4.2-1 describes the use case actors and roles.

Table 5.4.2-1: Use case #3 actors and roles

#	Actor and role	Description
1	Operator	The user who submits the telco cloud configuration change or new service application configuration deployment request.
2	Large model application	Leveraging large models for processing operator queries provides assistance in generating deployment plans.
3	Large model	A type of artificial neural network with a complex structure and large-scale parameters, enabling intelligent parsing of operator queries and assisting in extracting deployment parameters for generating deployment plans.
4	Metrics names knowledge base	A knowledge base storing a list of standard and alternative metric names, along with their relationships, used for efficient querying and identification (e.g. user input like "hard disk" or "storage volume" is mapped to the system's database terminology such as "hard disk").
5	Deployment plan knowledge base	A knowledge base storing deployment plan template information. It is used in the NFV-MANO framework to standardize and automate the generation of deployment plans.

5.4.3 Trigger

Table 5.4.3-1 describes the use case trigger.

Table 5.4.3-1: Use case #3 trigger

Trigger	Description
1	Operator requests to change telco cloud configuration or deploy new service applications.

5.4.4 Pre-conditions

Table 5.4.4-1 describes the use case pre-conditions.

Table 5.4.4-1: Use case #3 pre-conditions

#	Pre-condition	Description
1	Both the large model application and the large model are available.	No additional description.
2	Pre-built metrics names knowledge base, storing synonym lists for various metrics names, primarily converting metrics names described in users' natural language into system-recognizable metrics names.	No additional description.
3	Pre-built deployment plan knowledge base, saves the information of deployment plan templates.	No additional description.

5.4.5 Post-conditions

Table 5.4.5-1 describes the use case post-conditions.

Table 5.4.5-1: Use case #3 post-conditions

#	Post-condition	Description
1	Generate deployment plan	Based on the operator's requirements and deployment plan templates, successfully generate deployment plan.

5.4.6 Flow description

Table 5.4.6-1 describes the use case flow.

Table 5.4.6-1: Use case #3 flow description

#	Actor/Role	Action/Description
Begins when	Operator -> Large model application	The operator describes the Telco cloud configuration change or new service application configuration deployment requirements in natural language and sends the requirements to the large model application.
1	Large model application -> Large model	The Large model application sends to the large model the parameter information such as metric type, resource planning and virtualisation layer configuration based on prompt words (describing the requirements for deployment plan parameters) from the operator's natural language request.
2	Large model application -> Metrics names knowledge database	The large model application uses the metric type parsed in step 1 to query the metrics names knowledge database, retrieving the corresponding standardized metric names.
3	Large model application -> Deployment plan knowledge base	Based on the parameter information from step 1 and metrics names obtained in step 2, retrieve the reference deployment plan template with the highest similarity to the configuration parameters from deployment plan knowledge base.
4	Large model	The large model compares the parameter information obtained in steps 1 and 2 with the deployment plan template obtained in step 3 to confirm if there are any required parameters that need to be supplemented.
5	Large model -> Large model application -> Operator	If there are deployment plan template required parameters that have not been resolved, request the operator to supplement the parameters through multiple rounds of dialogue, and parse the operator-supplied content. Return to step 4 for re-comparison after supplementary parameters are added. Repeat the above steps until all required parameters are successfully resolved.
6	Large model application -> Operator	After all required parameters are resolved, the large model application automatically generates a deployment plan that adheres to NFV-MANO standards.

5.5 Use case #4: Optimizing intent negotiation with large models

5.5.1 Introduction

In ETSI GS NFV-IFA 050 [i.4] use Case #5 describes how the Intent Owner negotiates with the IM (Intent Management) on achievable intent object parameters values. These parameters can involve performance requirements, geographical location, isolation requirements, or security considerations. To optimize this process, large models can be integrated to assist in analysing historical data, predicting feasible parameter values, and suggesting optimal configurations.

5.5.2 Actors and roles

Table 5.5.2-1 describes the use case actors and roles.

Table 5.5.2-1: Use case #1 actors and roles

#	Actor and role	Description
1	Intent Owner	Determines an intent object identifying the requirements, constraints and characteristics it needs for NS functionality and captures them in the parameters related to NFV intent object, and initiate a negotiation request for the above parameters.
2	IM	Interprets the intent object and maps it to corresponding NS operation(s) (e.g. instantiate a new NS or update an existing NS, etc.), and confirm the feasibility of the corresponding NS operation through NFV-MANO. If it is unfeasible, provide suggestions for modifying the intent parameters based on the feasible NS operation confirmed by NFV-MANO. The IM uses the large model to predict the most likely feasible values and recommend adjustments to achieve a successful negotiation outcome.
3	Large model	A type of artificial neural network with a complex structure and large-scale parameters, facilitating intelligent processing and analysis to support the intent negotiation process and feasibility assessment.
4	NFV-MANO	Confirms the feasibility of the corresponding NS-related operation(s) (e.g. NS LCM, NS PM, NS FM, etc.).

5.5.3 Trigger

Table 5.5.3-1 describes the use case trigger.

Table 5.5.3-1: Use case #1 trigger

Trigger	Description
1	The Intent Owner submits a request to negotiate intent parameters.

5.5.4 Pre-conditions

Table 5.5.4-1 describes the use case pre-conditions.

Table 5.5.4-1: Use case #1 pre-conditions

#	Pre-condition	Description
1	Both the IM and the large model are available.	No additional description.

5.5.5 Post-conditions

Table 5.5.5-1 describes the use case post-conditions.

Table 5.5.5-1: Use case #1 post-conditions

#	Post-condition	Description
1	The result of negotiation is completed and shared between the Intent Management and the Intent Owner.	No additional description.

5.5.6 Flow description

Table 5.5.6-1 describes the use case flow.

Table 5.5.6-1: Use case #1 flow description

#	Actor/Role	Action/Description
Begins when	Intent Owner	The Intent Owner determines an intent object which contains the expectations for desired NS(s). For example, performance requirements (e.g. the minimal incoming/outgoing data rate of a certain SAP), geographical location, isolation requirements (e.g. whether or not it is allowed to share any resources with other NS(s), special security requirements (e.g. use of secure enclaves).
1	Intent Owner -> IM	The Intent Owner sends the desired NFV intent object to be negotiated towards the IM.
2	IM	After receiving the intent object from the Intent Owner, that is to be evaluated, IM interprets the requirements and maps them to corresponding NFV-MANO operation(s).
3	Large model → IM	Large model is used to evaluate the best values that can be achieved for the negotiated requirement or expectation and communicates the result to IM. If for any reason, it fails to provide the result intent object, relevant reasons of negotiation failure are communicated to IM.
4	IM <-> Large Model	IM shares the feasible parameters with the large model. The large model suggests further optimizations or adjustments to the parameters based on historical data and the feasibility results.
5	IM-> Intent Owner	IM would share either the intent object with the confirmed feasible parameters or the reason for the negotiation failure with the Intent Owner.
6	Intent Owner <-> IM	The Intent Owner after receiving the proposals on the feasible values that can be achieved for the negotiation request, if the proposal does not meet its requirements then it would start with a new variant of intent and ask for renewed proposal, by returning to step 1.
7	IM-> Intent Owner	If the Intent Owner and IM determine that the intent negotiation result is completed, or the Intent Owner abandons the negotiation, the intent negotiation process ends.

6 Key issue analysis

6.1 Key issue on providing large models and large model applications for Telco cloud management purposes

Key issue #1.1: Adapting large models for Telco cloud management purposes

The use cases outlined in the present document describes various Telco cloud management operations that leverage the capabilities of the large model, different assistance is provided for various tasks. Therefore, it is necessary to first examine what the essential processes are to ensure the alignment of large models with Telco cloud management requirements. Questions that can be formulated related to this key issue are:

- In Telco cloud management, existing large models can be used directly, or is it more effective to use a customized large model?
- How can large models be optimized to better suit the specific requirements of Telco cloud management?

Key issue #1.2: Issues in constructing and managing large model applications

In use case #1, a large model application is used to enable intelligent fault location and automated processing for telco clouds. In use case #2, a large model application is used to retrieve relevant monitoring data and generate summary reports. In use case #3, a large model application is used to assist in generating the deployment plan based on a predefined deployment plan template. Therefore, it is necessary to construct different large model applications and organize their associated components to support various telco cloud scenarios. Questions that can be formulated related to this key issue are:

- What are the typical steps for constructing large model applications in different scenarios?
- How are various components registered, referenced, and coordinated in the context of large model applications?

6.2 Introducing and exposing large models as a service

Key issue #2.1: Introducing large model into the telco cloud architecture

In Use case #1, Use case #2, and Use case #3, based on the capabilities of the large model, assistance is provided for various aspects of telco cloud operations, including supporting OAM functions within the telco cloud platform to improve efficiency and automation. In Use case #4, large models assist in optimizing intent processing within telco cloud service orchestration. Therefore, it is necessary to study how large models can be introduced into the existing telco cloud architecture. Questions that can be formulated related to this key issue are:

- How can large models be introduced into the telco cloud architecture to enhance the intelligence and automation of telco cloud management?
- How can large models provide effective support for the telco cloud platform?
- How can large models provide effective support for telco cloud service orchestration?

Key issue #2.2: Exposing large models as services for large scale applications

In Use case #1, Use case #2, Use case #3, and Use case #4, large models provide different types of assistance for Telco cloud management. Therefore, it is necessary to explore how large models can be exposed as services to provide various capabilities. This involves understanding how these capabilities, along with other tools and components, can be combined to form comprehensive large model applications that deliver enhanced functionality. Questions that can be formulated related to this key issue are:

- What are the typical steps for constructing large model applications in different scenarios?
- How are various large models and different components registered, referenced, and coordinated in the context of large model applications?
- How can different large models be exposed as services within the telco cloud architecture?
- What additional capabilities should the architecture provide to integrate various components related to large model applications and enable the construction of large model applications?
- What are the typical interaction processes that large models assist with in Telco cloud management?

Key issue #2.3: Management of large scale models

- There is a need to describe packaging, onboarding, LCM and OAM of larger models and large model applications

7 Framework and potential solutions

7.1 Introduction

Clause 7 documents potential solutions addressing the key issues discussed in clause 6 of the present document. Each solution is organized as follows:

- introduction describing the background and the conceptual information underlying the solution;
- description of the solution;
- reference to the key issues tackled by the solution; and
- identification of the gaps in the ETSI NFV architectural framework and/or referenced ETSI NFV specifications, if applicable.

7.2 Potential solutions

7.2.1 Solution #1: Integrating large models for telco cloud management

7.2.1.1 Introduction

Key issue #1.1 deals with integrating large models for telco cloud management, Key issue #1.2 deals with constructing and managing large model applications. Key issue #2.1 and Key issue #2.2 focus on introducing large models into the telco cloud architecture and exposing large models as services for the telco cloud architecture. The following clauses provide solutions related to supporting the above functions through MaaS Service.

7.2.1.2 Solution description

In the NFV architecture, a MaaS Service is provided to support the integration of large models and the construction of large model applications, while exposing the capabilities of large models and large model applications to other network cloud management-related functions.

Figure 7.2.1.2-1 shows how the MaaS Service integrates components like large models and provides corresponding assistance to other functions related to Telco cloud management.

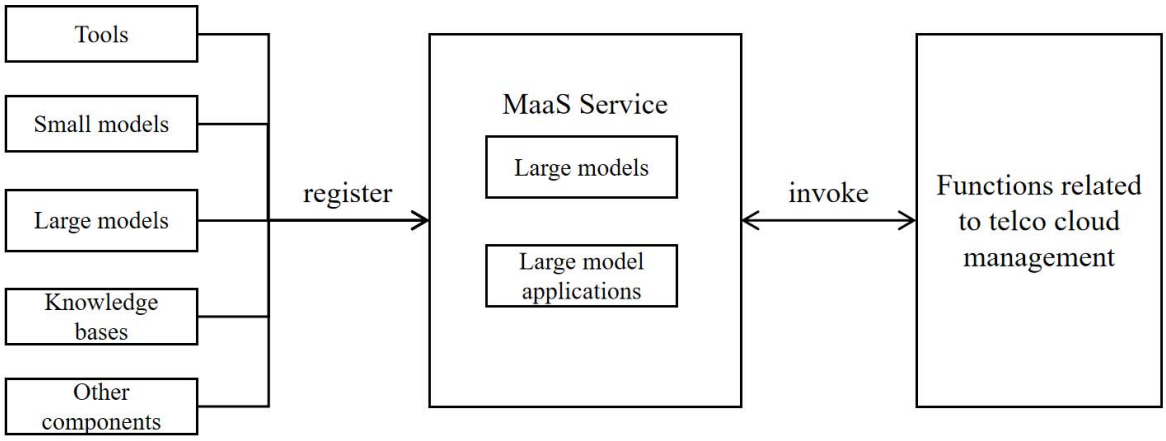

Figure 7.2.1.2-1: Integrating large models for Telco cloud management

The MaaS Service can provide the following capabilities:

- Large models, small models, tools, knowledge bases, and other components needed to construct a large model application, either inside or outside NFV-MANO, can be registered to the MaaS Service. The registration information includes basic details of the components, such as the type of large models, the purpose of tools, and the calling methods.
- Various functions within Telco cloud management can discover the currently supported large models and component information (e.g. metadata and description about the functionalities of the component) through the MaaS Service and, based on their specific requirements, request the MaaS Service to construct large model applications from various components like large models, small models, tools, and knowledge bases, etc. The various functions within Telco cloud management use large models or large model applications to assist in providing specific capabilities.
- Providing lifecycle management for large models and large model applications.

Based on the simplified view of the target architectural framework proposed in clause 7.4.3 of ETSI GS NFV-IFA 054 [i.5], the Telco Cloud Platform produces application and management-oriented platform services of the Telco Cloud to Telco Cloud Service Orchestration, Telco Cloud Applications, or other OSS/BSS. Figure 7.2.1.2-2 shows how the MaaS Service is integrated into the telco cloud architecture framework.

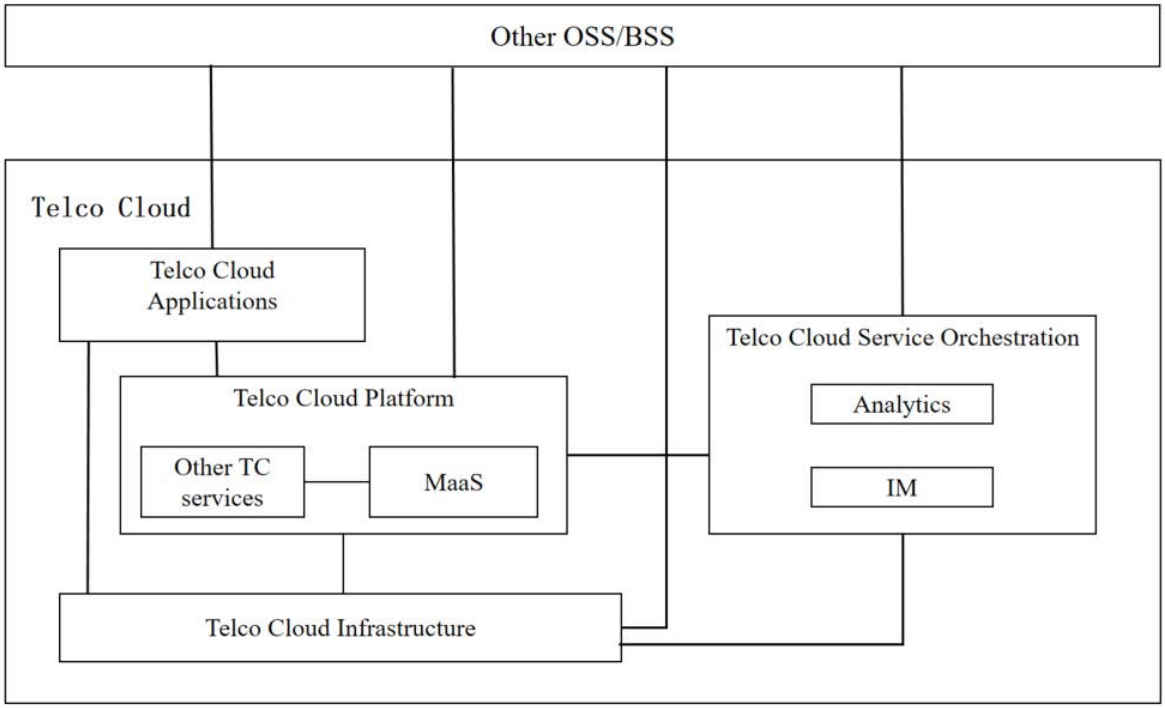

Figure 7.2.1.2-2: Integrating MaaS service into the telco cloud architecture framework

The MaaS Service can be provided as a service by the Telco Cloud Platform to offer large model and large model application capabilities for telco cloud management, including the following:

- Providing large model and large model application capabilities for other Telco cloud services in Telco Cloud Platform.
- Providing large model and large model application capabilities for Telco Cloud Service Orchestration (e.g. Analytics, Intent Management).
- Providing large model and large model application capabilities for Telco Cloud Applications.
- Providing large model and large model application capabilities for Telco Cloud Infrastructure.

7.2.1.3 Key issues address

The present solution aims at addressing aspects of the following key issues described in clause 6:

- Key issue #1.1,
- Key issue #2.1, and
- Key issue #2.2.

7.2.1.4 Gap analysis

The referenced ETSI NFV specifications in the present solution do not specify:

Gap #1.1: The NFV system needs to define the interfaces for onboarding and registering components in the MaaS Service (e.g. large models), needed to construct a large model application.

Gap #1.2: The MaaS Service needs to define the process and interfaces for consumers to invoke the MaaS service.

7.2.2 Solution #2: large model application deployment with separation of models and application logic components

7.2.2.1 Introduction

Large model application developers can leverage the resources of the operator's network to deploy large model applications, these applications can logically be considered to consist of two parts:

- Application Logic Component: The application logic components can utilize the capabilities of large models to provide services to users, such as leveraging large language models to offer intelligent customer service, translation, code generation.
- Large Model: large Model provides analytical and predictive intelligence capabilities, which can be decoupled from the application logic components.

large models are generally quite large, requiring significant bandwidth and time for transmission. Additionally, large models have compatibility dependencies on the runtime environment. This solution aims to provide a method for deploying the application logic components and the large model separately.

7.2.2.2 Solution description

Common application images (whether OS containers or VMs) usually contain complete functionalities. However, due to the large size of large models, including the entire large model in the large model application image would significantly increase the storage space required by the image and transmission bandwidth in case it is transferred. Therefore, the large model application image can be designed to only include the application logic components, while the application descriptor provides the metadata and download path of the large model.

NOTE: ETSI GR NFV-IFA 046 [i.11] discusses how to provide ML modelling information in NFV. Solution SOL-C1-2 is about integrating ML models into VNF Packages as non-MANO artifacts. Solution SOL-C1-3 is specific to MDA. Solution SOL-C1-1 assumes that ML models can be referenced within the VNFD, but it does not detail the specific information elements and procedures, which are provided in this solution.

During the on-boarding phase of the large model application, a model management function in the NFV system can determine the model to be used by the large model application based on metadata, and locate the large model from an internal large model repository . If the required model is not available in the large model repository, the model management function can download the large model using a provided download path and perform a security check.

During the instantiation phase of the large model application, an application management function in the NFV system allocates resources to instantiate the application logic components of the application based on the application descriptor, and waits for the model to be ready. The application provider could specify the model mount path for the large model in the application descriptor. The application management function accesses the internal large model repository to fetch the model, stores it at the model mount path, and then notifies the application logic component that the model is ready.

Figure 7.2.2.2-1 describes the process of large model application deployment.

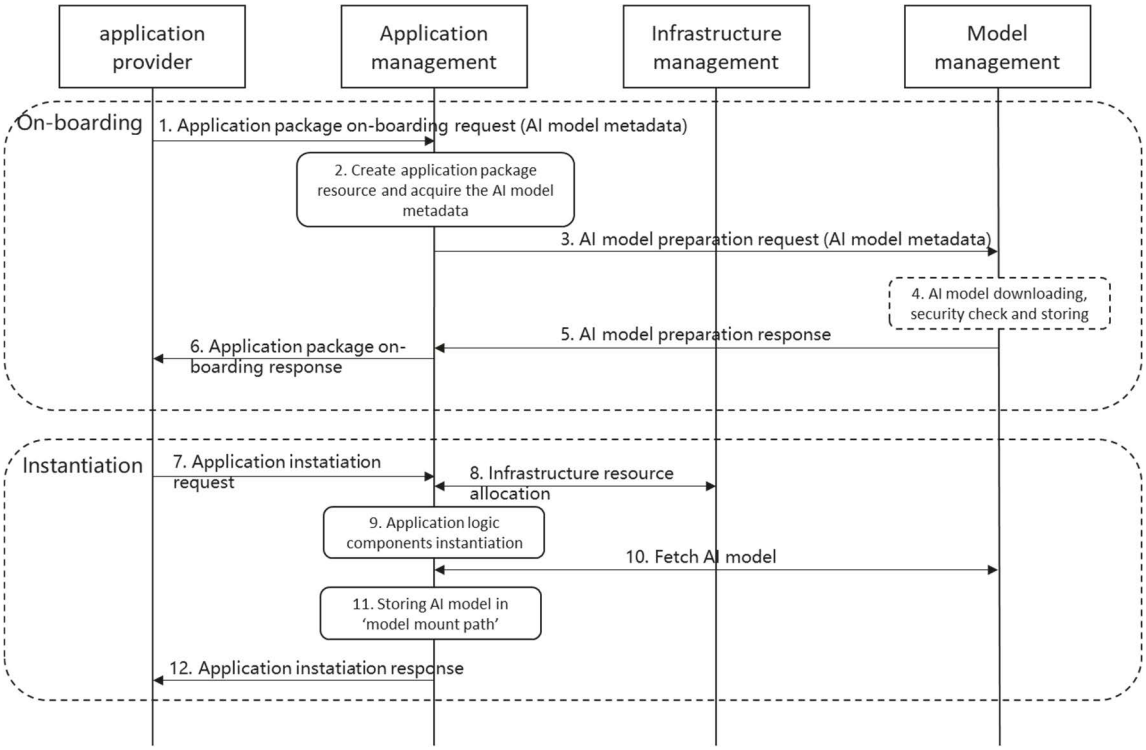

Figure 7.2.2.2-1: AI Application deployment with separation of models and application logic components

The application provider could provide the following content and information in the application descriptor:

- **Application logic component image or download URL:** Contains the code for the application function.
- **large model metadata:** large model name, large model version, large model download path, etc.
- **Model mount path:** The path through which the application logic component accesses the large model.
- **large model runtime framework:** A software toolkit that loads trained large models and executes inference efficiently.
- **Infrastructure resource description:** Infrastructure resources required for the large model application.
- **Network configuration information:** Information such as the exposed port and the protocol used for large model service.

7.2.2.3 Key issues address

The present solution aims at addressing aspects of the following key issues described in clause 6:

- Key issue #1.2.

7.2.2.4 Gap analysis

The referenced ETSI NFV specifications in the present solution do not specify:

Gap #2.1: Application descriptor needs to be enhanced to support large model applications with separated deployment of large models and application logic components.

Gap #2.2: The NFV system needs to support the management of large models and allow downloading large models from external sources.

Gap #2.3: The NFV system needs to support the lifecycle management of large model applications.

8 Recommendations

8.1 Overview

The recommendations are structured and elaborated as follows:

- aspects related to the architecture and framework (refer to clause 8.2);
- interfaces and associated information/data model (refer to clause 8.3);
- descriptors (refer to clause 8.4).

8.2 Recommendations related to the NFV architectural framework

The present clause documents recommendations intended to enhance the NFV architectural framework by identifying potential new functions or functional blocks, and interactions among these functional blocks and functions.

Tables 8.2-1 provides recommendations related to the NFV architectural framework.

Table 8.2-1: Recommendations related to the NFV architectural framework

Identifier	Recommendation description	Comments and/or traceability
maas.arch.001	It is recommended to specify a requirement for the NFV architectural framework to include a function that provides the MaaS service.	Refer to gaps #1.1 and #1.2
maas.arch.002	It is recommended to specify a requirement for the NFV architectural framework to include a function that provides large model and large model application LCM and OAM operations.	Refer to gaps #2.2 and #2.3

8.3 Recommendations related to interfaces and information model

The present clause provides recommendations focusing on interfaces and associated information.

Tables 8.3-1 provides the recommendations related to interfaces and associated information for the MaaS service.

Table 8.3-1: Recommendations related to interfaces and information model

Identifier	Recommendation description	Comments and/or traceability
maas.if.001	It is recommended to specify a requirement for the interfaces exposed by NFV-MANO to be able to support the registration of components needed to construct a large model application.	Refer to gap #1.1
maas.if.002	It is recommended to specify a requirement for the interfaces exposed by the MaaS Service to be able to support telco cloud management services and telco cloud applications in invoking large models and large model applications through these interfaces.	Refer to gap #1.2

8.4 Recommendations related to NFV descriptors

The present clause provides recommendations focusing on NFV descriptors.

Tables 8.4-1 provides the recommendations related to NFV descriptors.

Table 8.4-1: Recommendations related to NFV descriptors

Identifier	Recommendation description	Comments and/or traceability
maas.desc.001	It is recommended to specify a requirement for the VNFD to support describing the metadata and download path of the large model.	Refer to gap #2.1

9 Conclusion

The present document investigates MaaS for AI-based applications in the context of telco cloud management and telco cloud applications. Use cases and key issues associated to certain use cases are described and analysed in this regard. Potential solutions for addressing the identified key issues are proposed, and finally recommendations for potential enhancements to the NFV architectural framework are summarized.

History

Version	Date	Status
V6.1.1	April 2026	Publication