ITUPublications
Recommendations

International Telecommunication Union
Standardization Sector

Recommendation

ITU-T Q.5008 (12/2023)

SERIES Q: Switching and signalling, and associated measurements and tests

Signalling requirements and protocols for IMT-2020 – Signalling requirements and architecture of IMT-2020

Signalling requirements and architecture to support artificial intelligence-based vertical services in future networks including IMT-2020 and beyond

ITU-T Q-SERIES RECOMMENDATIONS

Switching and signalling, and associated measurements and tests

SIGNALLING IN THE INTERNATIONAL MANUAL SERVICE	Q.1-Q.3
INTERNATIONAL AUTOMATIC AND SEMI-AUTOMATIC WORKING	Q.4-Q.59
FUNCTIONS AND INFORMATION FLOWS FOR SERVICES IN THE ISDN	Q.60-Q.99
CLAUSES APPLICABLE TO ITU-T STANDARD SYSTEMS	Q.100-Q.119
SPECIFICATIONS OF SIGNALLING SYSTEMS NO. 4, 5, 6, R1 AND R2	Q.120-Q.499
DIGITAL EXCHANGES	Q.500-Q.599
INTERWORKING OF SIGNALLING SYSTEMS	Q.600-Q.699
SPECIFICATIONS OF SIGNALLING SYSTEM NO. 7	Q.700-Q.799
Q3 INTERFACE	Q.800-Q.849
DIGITAL SUBSCRIBER SIGNALLING SYSTEM NO. 1	Q.850-Q.999
PUBLIC LAND MOBILE NETWORK	Q.1000-Q.1099
INTERWORKING WITH SATELLITE MOBILE SYSTEMS	Q.1100-Q.1199
INTELLIGENT NETWORK	Q.1200-Q.1699
SIGNALLING REQUIREMENTS AND PROTOCOLS FOR IMT-2000	Q.1700-Q.1799
SPECIFICATIONS OF SIGNALLING RELATED TO BEARER INDEPENDENT CALL CONTROL (BICC)	Q.1900-Q.1999
BROADBAND ISDN	Q.2000-Q.2999
SIGNALLING REQUIREMENTS AND PROTOCOLS FOR THE NGN	Q.3000-Q.3709
SIGNALLING REQUIREMENTS AND PROTOCOLS FOR SDN	Q.3710-Q.3899
TESTING SPECIFICATIONS	Q.3900-Q.4099
PROTOCOLS AND SIGNALLING FOR PEER-TO-PEER COMMUNICATIONS	Q.4100-Q.4139
PROTOCOLS AND SIGNALLING FOR COMPUTING POWER NETWORKS	Q.4140-Q.4159
PROTOCOLS AND SIGNALLING FOR QUANTUM KEY DISTRIBUTION NETWORKS	Q.4160-Q.4179
SIGNALLING REQUIREMENTS AND PROTOCOLS FOR IMT-2020	Q.5000-Q.5049
Signalling requirements and architecture of IMT-2020	**Q.5000-Q.5019**
Protocols for IMT-2020	Q.5020-Q.5049
COMBATING COUNTERFEITING AND STOLEN ICT DEVICES	Q.5050-Q.5069

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T Q.5008

Signalling requirements and architecture to support artificial intelligence-based vertical services in future networks including IMT-2020 and beyond

Summary

Recommendation ITU-T Q.5008 provides the signalling requirements and architecture to support artificial intelligence (AI) based vertical services in future networks including IMT-2020 and beyond. These requirements include the signalling information over each of the reference points and service procedures for a high-level AI platform.

History *

Edition	Recommendation	Approval	Study Group	Unique ID
1.0	ITU-T Q.5008	2023-12-14	11	11.1002/1000/15730

Keywords

AI, AI platform, MEC, vertical services.

* To access the Recommendation, type the URL https://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents/software copyrights, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T Q.5008

Signalling requirements and architecture to support artificial intelligence-based vertical services in future networks including IMT-2020 and beyond

1 Scope

This Recommendation provides signalling requirements and architecture to support artificial intelligence (AI) based vertical services in future networks including IMT-2020 and beyond. It addresses the following subjects:

– Signalling architecture and requirements for supporting AI-based vertical services;

– High-level AI platform service procedures;

– Signalling information for each of the reference points.

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 M.3080] Recommendation ITU-T M.3080 (2021), *Framework of artificial intelligence enhanced telecom operation and management (AITOM).*

[ITU-T Q.5003] Recommendation ITU-T Q.5003 (2022), *Signalling requirements and architecture for federated multiaccess edge computing.*

[ITU-T Y.3172] Recommendation ITU-T Y.3172 (2019), *Architectural framework for machine learning in future networks including IMT-2020.*

[ITU-T Y.3174] Recommendation ITU-T Y.3174 (2020), *Framework for data handling to enable machine learning in future networks including IMT-2020.*

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 machine learning (ML) [ITU-T Y.3172]: Processes that enable computational systems to understand data and gain knowledge from it without necessarily being explicitly programmed.

NOTE 1 – Definition adapted from [b-ETSI GR ENI 004].

NOTE 2 – Supervised machine learning and unsupervised machine learning are two examples of machine learning types.

3.1.2 multi-access edge computing (MEC) [b-ETSI GS MEC 001]: System which provides an IT service environment and cloud-computing capabilities at the edge of an access network which contains one or more type of access technology, and is in close proximity to its users.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 artificial intelligence (AI) application: Application that can be instantiated on a user equipment (UE) within the AI system and can potentially provide or consume AI services.

3.2.2 artificial intelligence (AI) service: Service provided via the AI platform.

3.2.3 artificial intelligence (AI) platform: A full stack of technologies that enables AI service providers to support automated AI modelling and services for the AI-based applications.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations:

AI Artificial Intelligence

API Application Programming Interface

DBMS Database Management System

FE Functional Entity

MEC Multi-access Edge Computing

ML Machine Learning

NPU Neural Processing Unit

OAuth Open Authorization

PACS Picture Archiving and Communication Systems

UE User Equipment

URI Uniform Resource Identifier

URL Uniform Resource Locator

5 Conventions

None

6 Introduction

An increasing number of companies are adapting artificial intelligence (AI) technologies to their business model. Appendix I introduces a number of AI service use cases in vertical domains. These AI services utilize AI models to analyse datasets, find patterns and make predictions without the need for human inference. As described in Appendix II, the development and implementation of AI services involves data collection and processing, AI modelling, training, and inferencing. Most AI modelling and training is carried out using machine learning (ML) [ITU-T Y.3172], [ITU-T Y.3174] and [b-ISO/IEC 22989].

The development and implementation of AI services requires huge amounts of high-performance AI dedicated cloud resources and expertise. However, for individual AI-based service providers and even for small and medium-sized businesses (enterprises), operating in such a required environment by themselves often becomes onerous. IMT-2020 network operators run large scale cloud computing facilities in their networks including edge cloud or multi-access edge computing (MEC).

AI platform which supports automated AI modelling and services, built in the network operator's cloud facilities, can provide the required resources and services for AI-based service providers to develop and deploy necessary AI models and applications. However, as AI services are very diverse

and components for an AI service are distributed over the network, it is necessary to have a standardized signalling architecture that efficiently interlinks them to fulfil the purpose of each service.

In addition to the signalling architecture to provide required resources and services for AI-based service providers, this Recommendation defines a layered architectural model with reference points, and specifies the signalling flows and message information. For each layer, functional entities (FEs) and requirements are also defined. The signalling flows and message information are specified in accordance with the high-level AI platform service procedures to support AI-based services.

7 Signalling requirements and architecture

The architectural model for AI services in future networks including IMT-2020 and beyond is required to support the following functionalities to:

– analyse and process the data for the user equipment (UE) if the UE has the AI infrastructure function;

– analyse the data in the edge access network layer for the MEC [b-ETSI GS MEC 001] and [ITU-T Q.5003] environment or the core network layer for non-MEC environment and then provide the specific AI service by building the AI model;

– provide the specific AI service, in the edge access network layer or in the core network layer, by building the AI model with analysis using the data from the UE.

The raw data and processed data from the UE and from each of the predefined network layers, such as the core network layer and the edge access network layer, may be used independently or together for AI services.

7.1 Architectural model to support AI-based services

Figure 7-1 depicts an architectural model to support AI-based vertical services in future networks including IMT-2020 and beyond. As shown in the figure, the architectural model consists of four functional layers, the service layer, the core network layer, the edge access network layer and the UE.

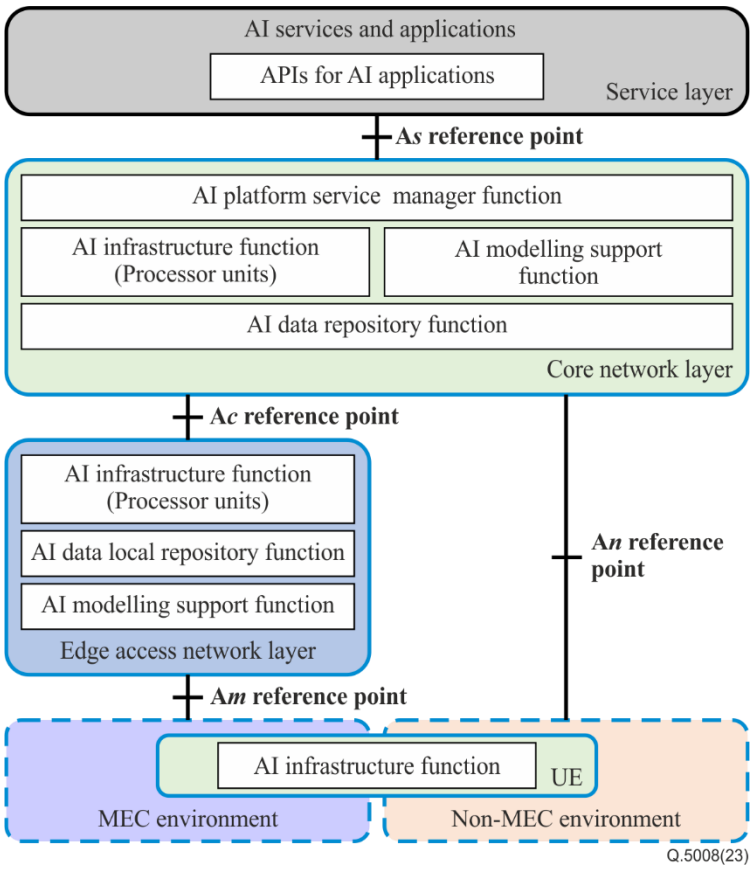

Figure 7-1 – Architectural model to support AI-based services

7.1.1 Functional entities and requirements

7.1.1.1 Service layer

The service layer is required to be responsible for providing AI platform services to various service providers by using application programming interfaces (APIs) for AI applications. In this way various service providers, such as security, media, medical and manufacturing service providers, etc., can utilize the AI services provided by the AI platform.

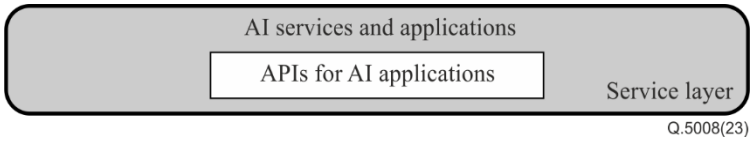

Figure 7-2 – Functional entities for the service layer

7.1.1.2 Core network layer

The core network layer is required to provide functionalities which are necessary to process AI platform services such as AI data processing, modelling, and data repository services which may require the most complex AI processing costs and high performance levels. The functional entities (FEs) in the core network layer, shown in Figure 7-3, are required to perform functionalities as follows:

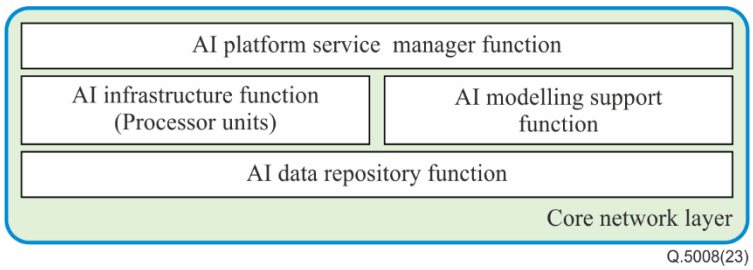

Figure 7-3 – Functional entities for the core network layer

– AI platform service manager functional entity (FE) is a centralized focal point which interacts with the service layer through the As reference point. The FE generally manages and maintains up to date information on all the APIs, AI models, and labelled raw/processed public data, as well as information on the AI platform resources such as availability, performance and utilization, etc. It may also manage the information on the subscriptions to the APIs and resources. The FE posts or deletes the information and APIs for AI models and other AI platform services on behalf of individual AI platform FEs in the core network layer and edge access network layer. The FE also handles resource allocation requests for the AI platform infrastructure and data repository;

– AI infrastructure FE processes raw data or various training data, using AI processor units;

– AI modelling support FE generates AI model to be suitable for the AI-based application services that the application service providers want to provide, based on the analysed data;

– AI data repository FE stores raw data, training data, testing data, or meta data that are generated during the AI process pipeline (e.g., data collection – analysing – modelling – service).

7.1.1.3 Edge access network layer

The edge access network layer is required to provide functionalities which are necessary for the AI service processing tasks such as AI data processing, modelling, and data repository services which may require mid-level complexity and performance, meaning that it could perform the AI pipelines similar to that of the core network but with less resources.

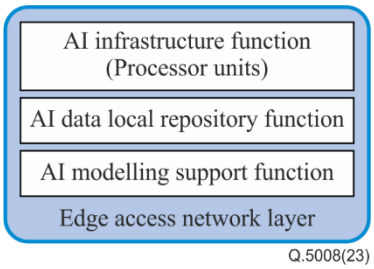

Figure 7-4 – Functional entities for the edge access network layer

In order to build the AI pipelines, the functional entities in the edge access network layer shown in Figure 7-4, are required to perform functionalities as follows:

– AI infrastructure FE processes raw data or various training data, using AI processor units;

– AI modelling support FE generates the AI model suitable for the AI services that the service providers want to provide, based on analysed data;

– AI data repository FE stores raw data, training data, testing data, or meta data that are generated during the AI process pipeline (e.g., data collection – analysing – modelling – service).

However, due to the relatively lower computational power and capacity compared to those of the core network layer, AI services generated from the edge access network layer should be much simpler than those from the core network layer.

7.1.1.4 User equipment (UE)

UE, such as smartphones and edge AI cameras, provide the most simple and low-level AI services by themselves, such as simple inferencing. Thereby, it consists only of AI infrastructure FEs with embedded AI functions such as object detection. However, by interworking with AI platform FEs in the edge access network layer or in the core network layer, UE can also provide more complex AI services to users.

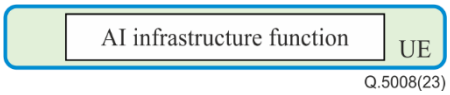

Figure 7-5 – Functional entities for the UE

7.2 High-level AI platform service procedures

7.2.1 Procedures to post/delete APIs

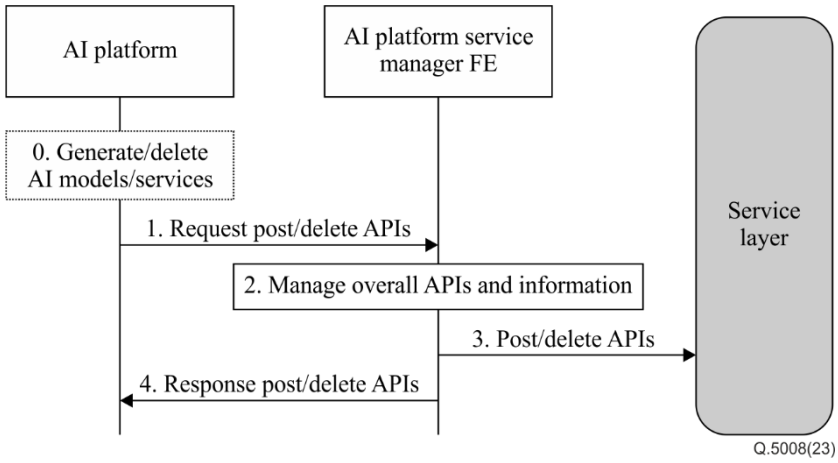

Figure 7-6 – AI platform service procedures to post/delete APIs

Figure 7-6 shows AI platform service procedures to post/delete APIs.

0. AI platforms in the edge access network layer (i.e., MEC environment) or in the core network layer (i.e., non-MEC environment) generate or delete AI models, AI platform services, and raw/processed labelled public data for the training practices.

1. The AI platform requests post/delete APIs with related information such as API name, API ID, API version, URI/URL (e.g., URL for web APIs), required parameters, and API descriptions, etc. Updating an API may be accomplished by deleting an API and posting a new one with the same API ID but with a new version.

– APIs for AI models may include, for example, APIs to download AI models and to use the AI models for the AI applications to analyse or to inference from input data, etc.

– APIs for AI platform services may also include, for example, APIs to request AI platform resources and to use the allocated resources.

– For the raw/processed labelled public data, users may download them to their facilities, or place them in the AI data repository FE and practice training using AI Infrastructure and AI modelling support FE tools.

2. Centralized AI platform service manager FE in the core network layer manages overall APIs and related information.

3. The AI platform service manager FE posts/deletes APIs towards the service layer according to the requests from the AI platforms.

4. The AI platform service manager FE responses to the requested AI platform with the result.

7.2.2 Procedures to request AI platform resources

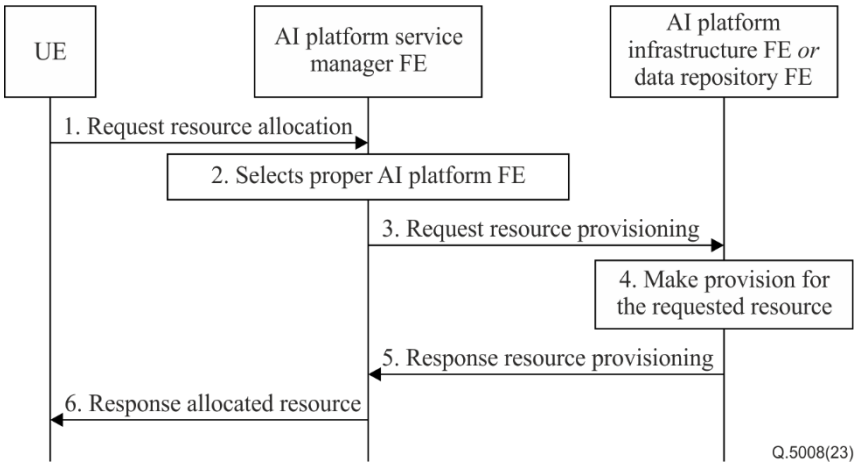

Figure 7-7 – Procedures to request AI platform resources

Figure 7-7 shows procedures to request AI platform resources.

1. Using posted APIs, the UE requests AI platform resources allocation such as AI processor units and data repository storage capacity, etc. The request message may include resources type, amount of capacity and technical specifications such as performance, throughput, latency and resiliency, etc. The API calls directly to the AI platform service manager FE as the API target endpoint.

2. Upon receiving the resources allocation request API call, the AI platform service manager FE selects a proper AI platform FE which fulfils the request in terms of availability, utilization and performance, etc.

3. Then the AI platform service manager FE sends a resources provisioning request to the selected AI platform FE. The request message contains required resources information specified by the resource allocation request from the UE.

4. According to the resource provisioning request, the selected FE secures required resources.

5. Then the selected FE responds to the AI platform service manager FE with the information on the provisioned resources such as resource ID, APIs which contain the uniform resource identifiers (URIs) to access the resource, and allocated resource capacity, etc.

6. The AI platform service manager FE responds to the requested UE with the allocated resource information received from the selected FE.

7.2.3 Procedures to release AI platform resources

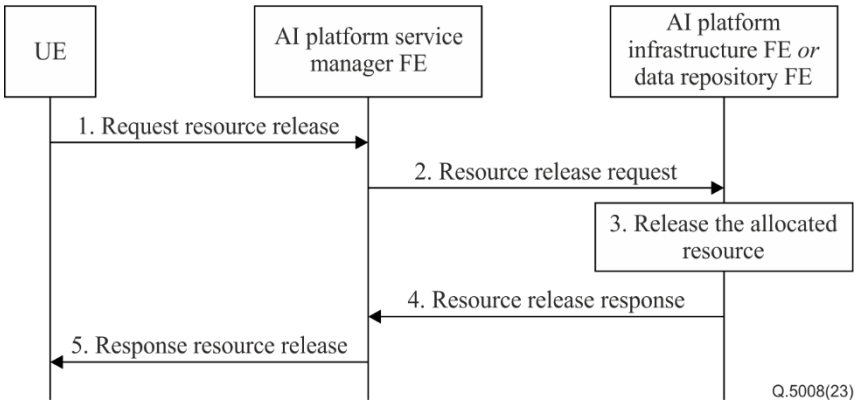

Figure 7-8 – Procedures to release AI platform resources

Figure 7-8 shows procedures to release AI platform resources.

1. Using posted APIs, UE requests allocated AI platform resources such as AI processor units and data repository storage capacity, etc. The request message may include the AI platform ID which provisioned the resource, resource type, allocated resource ID and UE application ID, etc. The API call directs to the AI platform service manager FE as the API target endpoint.

2. Upon receiving the resource release request API call, the AI platform service manager FE sends a resource release request to the target AI platform FE.

3. The target AI platform FE, e.g., AI infrastructure or data repository according to the resource type specified in the release request, releases the allocated resource as requested.

4. Then the target AI platform FE responds to the AI platform service manager FE with the resource release result.

5. The AI platform service manager FE responds to the requested UE with resource release result received from the FE which released the allocated resource as requested.

7.2.4 Procedures to respond API calls

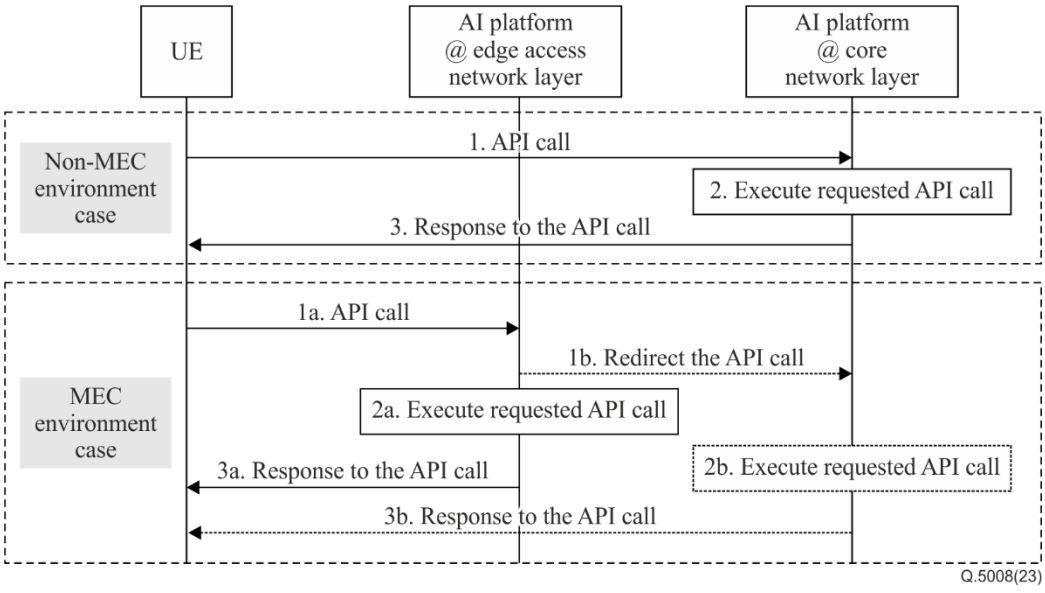

Figure 7-9 – AI platform service procedures to respond API calls

Figure 7-9 shows AI platform service procedures to respond API calls.

For UEs attached to the network access point which does not support the MEC environment:

1. UE triggers an API call with the required parameters. For example, when an AI application on the UE triggers the API call to analyse the input data using a trained AI model, the API call may include the AI application ID and input data (or URI/URL to it), etc. As the API endpoint URI/URL contained in the call can be served at the core network layer AI platform, the API call is directed to it.
2. The AI platform executes the requested API call.
3. The AI platform sends a response to the requested UE. For the above example in step 1, the response message may include analysed result, output data if any, and the requested AI application ID, etc.

For UEs attached to the network access point which supports the MEC environment:

1a. UE triggers an API call with required parameters. If the API endpoint URI/URL contained in the call can be served at the local AI platform on the attached edge access network layer MEC, the API call is directed to the AI platform. The example API call parameters given in step 1 of the above non-MEC environment case, also applies to this case.

1b. If the local MEC AI platform cannot serve the API call, the AI platform redirects the API call to the AI platform at the core network layer AI platform. This is the case when the API call execution requires higher capability than that of the local MEC AI platform in terms of complexity, AI processing performance and AI data repository capacity, etc.

2a. The local MEC AI platform executes the requested API call.

2b. The AI platform at the core network layer executes the requested API call.

3a. The local MEC AI platform sends a response to the requested UE. The information components of the response message given in step 3 of the above non-MEC environment case, also applies to this case.

3b. The AI platform at the core network layer sends a response to the requested UE. The information components of the response message given in step 3 of the above non-MEC environment case, also applies to this case.

8 Signalling information for each of the reference points

For interworking between the service layer, core network layer, edge access network layer, and UE, the following reference points for signalling and message exchange between layers should be defined:

– A*s* reference point between the service layer and core network layer;

– A*c* reference point between core network layer and edge access network layer;

– A*m* reference point between edge access network layer and UE in the MEC environment;

– A*n* reference point between the core network layer and UE in the non-MEC environment.

8.1 Signalling information for reference point A*s* (Service layer – Core network layer)

In order to provide AI services and applications to users, the A*s* reference point is required to exchange request/response messages between the AI services and the core network layer:

– To authorize the use of AI platform services and resources, the A*s* reference point is required to allow request/response message exchange for the authentication code and the access tokens;

– As reference point is required to exchange AI infrastructure resource information regarding AI processor units, AI model and API information that performs services, and information on raw data and processed data;

– In addition, to maintain the best service quality, functions such as updating APIs and API model information, or deleting unnecessary APIs and AI models should be provided.

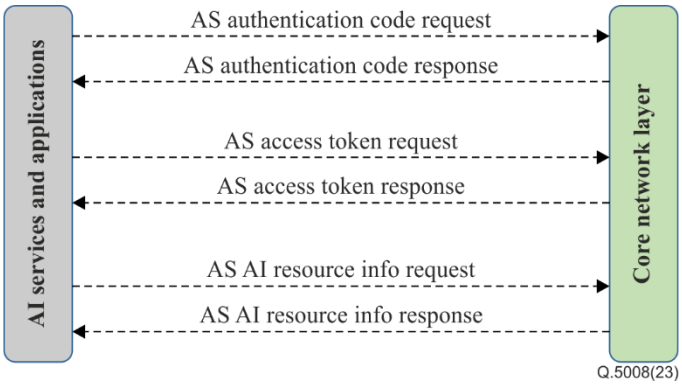

Figure 8-1 – Signalling flows for reference point A*s*

8.1.1 AS authentication code request and response messages

The AS authentication code message is defined as AS-AUTHENTICATION-CODE message. This message is sent by the AI application to the core network layer for requesting the AS authentication (*CRUD[1]) code.

8.1.1.1 AS authentication code request message

The AS authentication code request information flow is sent by AI applications to the core network layer to obtain the authentication code. It contains the following information components:

Message format:

```
< AS-AUTHENTICATION-CODE-REQUEST-Message> ::= < Message Header >
{ Application-ID }
{ Service-Session-ID }
{ AS-ID }
{ AS-Name }
   { AS-Description}
   { AS-Argument-Info }
      {Service-Name, Access-Right (CRUD) }
```

8.1.1.2 AS authentication code response message

The AS authentication code response information flow is sent by the core network layer to AI applications to provide the authentication code. It contains the following information components:

Message format:

```
< AS-AUTHENTICATION-CODE-RESPONSE-Message> ::= < Message Header >
{ Application-ID }
{ Service-Session-ID }
{ AS-ID }
```

[1] C: Create / R: Read / U: Update / D: Delete

```
{ AS-Name }
    { AS-Description}
    { AS-Argument-Info }
        {Service-Name, Access-Right (CRUD) }
    { AS-Result }
        {AS-Authentication-Code }
```

8.1.2 AS access token request and response messages

The AS access token message is defined as AS-ACCESS-TOKEN message. This message is sent by the AI application to the core network layer for requesting the AS access token.

8.1.2.1 AS access token request message

The AS access token request information flow is sent by AI applications to the core network layer to obtain the AS access token. It contains the following information components:

Message format:

```
< AS-ACCESS-TOKEN-REQUEST-Message> ::= < Message Header >
{ Application-ID }
{ Service-Session-ID }
{ AS-ID }
{ AS-Name }
    { AS-Description}
    { AS-Argument-Info }
        {AS-Authentication-Code }
```

8.1.2.2 AS access token response message

The AS access token response information flow is sent by the core network layer to AI applications to provide the AS access token. It contains the following information components:

Message format:

```
< AS-ACCESS-TOKEN-RESPONSE-Message> ::= < Message Header >
{ Application-ID }
{ Service-Session-ID }
{ AS-ID }
{ AS-Name }
    { AS-Description}
    { AS-Argument-Info }
        {AS-Authentication-Code }
    { AS-Result }
    {AS-Access-Token}
```

8.1.3 AS AI resource info request and response messages

The AS AI resource info message is defined as AS-AI-RESOURCE-INFO message. The AS-AI-RESOURCE-INFO message is sent by an AI application to the core network layer to request the AI resource information such as AI processor units, AI model and API information that performs services, and information on raw data and processed data provided by the core network layer. The core network layer then responds to the application through the response message with the requested AI resource information.

8.1.3.1 AS AI resource info request message

The AS AI resource info request message is sent by AI applications to the core network layer to obtain the AI resource information. It contains the following information components:

Message format:

```
<AS-AI-RESOURCE-INFO-REQUEST-Message> ::= < Message Header >
{ Application-ID }
{ Service-Session-ID }
{ AS-ID }
{ AS-Name }
   { AS-Description}
   { AS-Argument-Info }
      {Processor-Unit?AI-model?API-Info },
   {AS-Access-Token }
```

8.1.3.2 AS AI resource info response message

The AS AI resource info response message is sent by the core network layer to AI applications to provide the requested AI platform resource information. It contains the following information components:

Message format:

```
<AS-AI-RESOURCE-INFO-RESPONSE-Message> ::= < Message Header >
{ Application-ID }
{ Service-Session-ID }
{ AS-ID }
{ AS-Name }
   { AS-Description}
   { AS-Argument-Info }
      { Processor-Unit?AI-model?API-Info },
   {AS-Access-Token }
   { AS-Result }
      {AS-ID#1, AS-Name, Processor-Unit, AI-model-Type, API-Info (CRUD)},
      {AS-ID#N, AS-Name, Processor-Unit, AI-model-Type, API-Info (CRUD)}
```

8.2 Signalling information for reference point A*c* (Core network layer – Edge access network layer)

The A*c* reference point should allow the core network layer to cooperate with different sub-network AI platform FEs in the edge access network layer MECs, and to check the availability of each sub-network AI platform FEs and resources.

– A*c* reference point is required to exchange AI infrastructure resource information regarding processor units, etc., AI model information, and information on raw data and processed data.

– Also, up to date information on the supported APIs from each edge access network layer should be exchanged through the A*c* reference point, thereby the AI platform service manager FE in the core network layer can manage and maintain the latest status of all the AI platform service availabilities, providing optimal service configuration in the core network layer. Figure 8 2 shows signalling flows for reference point A*c*.

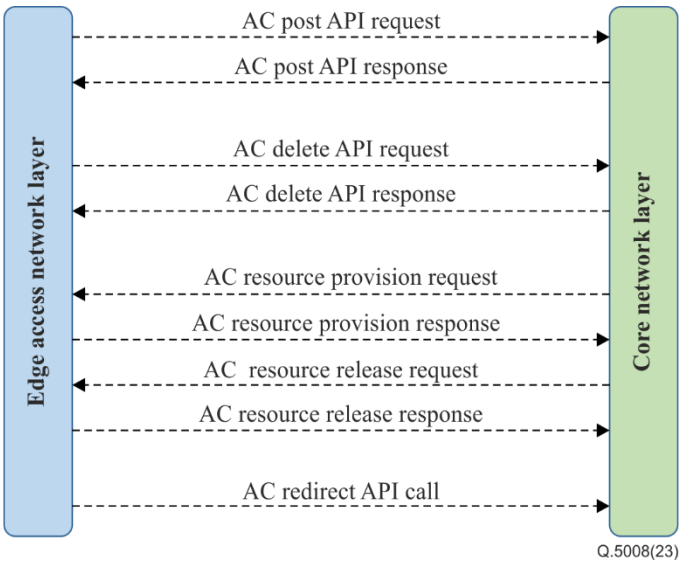

Figure 8-2 – Signalling flows for reference point A*c*

8.2.1 AC post API request and response messages

As specified in clause 7.2.1 and shown in Figure 7-6, AI platforms in the edge access network layer may exchange the request/response messages to post APIs with the core network layer.

8.2.1.1 AC post API request message

AI platforms at the edge access network layer need to post APIs which allows access to their AI infrastructure, data repository, AI models, AI modelling support function and labelled raw/processed public data. Posting and deleting the APIs on the service layer is the AI platform service manager's role at the core network layer as defined in clause 7.1.1.2. To post the APIs, each corresponding FE in the edge network layer AI platform sends a post API request message to the AI platform service manager through reference point A*c*.

The message can be used to update the posted APIs. To update the posted API, the FE which posted the API deletes the API, then posts a new one with updated API information including new API version. The message may include following information:

- AI platform ID to identify the requesting AI platform
- Information on the API ID, API version, API name, etc.
- API type which classifies the targeted AI Platform resource and service types
  - API to access/use AI Infrastructure; or
  - API to access/use local data repository; or
  - API to access/use labelled raw/processed public data; or
  - API to access/use AI models; or
  - API to access/use AI modelling support functions
- API descriptor which specifies the information required to access/use AI platform services including:
  - Resource URI/URL
  - Protocol/HTTP method
  - Pre-conditions
  - OAuth
  - Data format
  - Parameters (e.g., Required/Optional, Query string parameters, Response parameter, etc.)

Message format:

```
<AC-POST-API-REQUEST-Message>::=<Message Header>
    {AI-Platform-ID}
    {API-ID}
    {API-Version}
    {API-Name}
    {API-Type}
    {API-Descriptor}
```

8.2.1.2 AC post API response message

The AI platform service manager at the core network layer posts the requested API on the service layer through the A*s* reference point, then responds to the post API request message. In this way, the function generally manages and maintains up to date information on all the APIs for AI infrastructure resources, AI models, AI modelling support functions and labelled raw/processed public data from AI platforms. The post API response message may include the following information:

- AI platform ID to identify the requested AI platform
- Information on API ID, API version, API name, etc.
- API type which classifies the targeted AI Platform resource and service types, such as
  - API to access/use AI infrastructure; or
  - API to access/use local data repository; or
  - API to access/use labelled raw/processed public data; or
  - API to access/use AI models; or
  - API to access/use AI modelling support functions
- Post API result

Message format:

```
<AC-POST-API-RESPONSE-Message>::=<Message Header>
    {AI-Platform-ID}
    {API-ID}
    {API-Version}
    {API-Name}
    {API-Type}
    {Post-API-Result}
```

8.2.2 AC delete API request and response messages

As specified in clause 7.2.1 and shown in Figure 7-6, AI platforms in the edge access network layer may exchange the request/response messages to delete APIs with the core network layer.

8.2.2.1 AC delete API request message

For the posted APIs, each corresponding FE in the edge network layer AI platform can delete APIs which are not available anymore or that need to be updated. For this, the FE sends a delete API request message to the AI platform service manager at the core network layer through reference point A*c*. The delete API request message may include following information:

- AI platform ID to identify the requesting AI platform
- Information on API ID, API version, API name, etc.
- API type which classifies the targeted AI platform resource and service types, such as

  - API to access/use AI infrastructure; or
  - API to access/use local data repository; or
  - API to access/use labelled raw/processed public data; or
  - API to access/use AI models; or
  - API to access/use AI modelling support functions
- Reason to delete the API

Message format:

```
<AC-DELETE-API-REQUEST-Message>::=<Message Header>
    {AI-Platform-ID}
    {API-ID}
    {API-Version}
    {API-Name}
    {API-Type}
    {Reason}
```

8.2.2.2 AC delete API response message

The AI platform service manager at the core network layer deletes the API on the service layer through A*s* reference point as requested, then responds to the delete API request message. In this way, the function generally manages and maintains up to date information on all the APIs for AI infrastructure resources, AI models, AI modelling support functions and labelled raw/processed public data from AI platforms. The delete API response message may include following information:

- AI platform ID to identify the requested AI platform
- Information on API ID, API version, API name, etc.
- API type which classifies the targeted AI Platform resource and service types, such as
  - API to access/use AI infrastructure; or
  - API to access/use local data repository; or
  - API to access/use labelled raw/processed public data; or
  - API to access/use AI models; or
  - API to access/use AI modelling support functions
- Delete API result

Message format:

```
<AC-DELETE-API-RESPONSE-Message>::=<Message Header>
    {AI-Platform-ID}
    {API-ID}
    {API-Version}
    {API-Name}
    {API-Type}
    {Delete-API-Result}
```

8.2.3 AC resource provision request and response messages

As specified in clause 7.2.2 and shown in Figure 7-7, AI platforms in the edge access network layer may exchange the resource provision request/response messages with the core network layer.

8.2.3.1 AC resource provision request message

For the resource allocation request from an UE, the AI platform service manager sends a resource provision request message to a selected AI platform. The message may include the following information:

- AI platform ID to identify the target AI platform
- Information on the UE which requests the resource allocation, such as:
  - UE identifier
  - UE context information
  - UE authentication code
  - API ID, API version, API name, API type which are used to request the resource allocation
  - Access token
- Resource allocation request message information, such as:
  - UE application ID which requests the resource allocation
  - Requested resource type
  - Amount of required capacity and technical specifications such as performance, throughput, latency and resiliency, etc.
  - CRUD information

Message format:

```
<AC-RESOURCE-PROVISION-REQUEST-Message>::=<Message Header>
    {AI-Platform-ID}
    {UE-Info}
    {Resource-Allocation-Request-Info}
```

8.2.3.2 AC resource provision response message

For the resource provision request from the AI platform service manager, an AI platform FE corresponding to the requested resource type, such as AI infrastructure, AI data repository, etc. in the edge access network layer, sends a resource provision response message to the AI platform service manager with the information on the provisioned resources. The message may include following information:

- AI platform ID to identify the AI platform which provisioned the requested resource
- Information on the UE which requests the resource allocation, such as:
  - UE identifier
  - UE context information
  - UE authentication code
  - API ID, API version, API name, API type which are used to request the resource allocation
  - Access token
- Provisioned resource information, such as:
  - UE application ID which requests the resource allocation
  - Resource type
  - Resource ID
  - APIs each of which contains the URI/URL to access and utilize the resource

  - Resource descriptor which contains the amount of provisioned capacity and technical specifications such as performance, throughput, latency and resiliency, etc.
  - CRUD information
  - Constraint information
- Resource provision result

Message format:

```
<AC-RESOURCE-PROVISION-RESPONSE-Message>::=<Message Header>
    {AI-Platform-ID}
    {UE-Info}
    {Provisioned-Resource-Info}
    {Resource-Provision-Result}
```

8.2.4 AC resource release request and response messages

As specified in clause 7.2.3 and shown in Figure 7-8, AI platforms in the edge access network layer may exchange the resource release request/response messages with the core network layer.

8.2.4.1 AC resource release request message

For the resource release request from UE, the AI platform service manager sends a resource release request message to the AI platform which allocated the corresponding resource. The message may include following information:

- AI platform ID to identify the target AI platform
- Information on the UE which requests the resource release, such as:
  - UE identifier
  - UE context information
  - UE authentication code
  - API ID, API version, API name, API type which are used to request the resource release
  - Access token
- Resource release request message information, such as:
  - UE application ID which requests the resource release
  - Allocated resource type to release
  - Allocated resource ID to release
  - Reason to release

Message format:

```
<AC-RESOURCE-RELEASE-REQUEST-Message>::=<Message Header>
    {AI-Platform-ID}
    {UE-Info}
    {Resource-Release-Request-Info}
```

8.2.4.2 AC resource release response message

For the resource release request from the AI platform service manager, an AI platform FE corresponding to the requested resource release, such as AI infrastructure, AI data repository, etc. in the edge access network layer, sends a resource release response message to the AI platform service manager with the result of the requested resource release. The message may include following information:

- AI platform ID which released the resource

- Information on the UE which requests the resource release, such as:
  - UE identifier
  - UE context information
  - UE authentication code
  - API ID, API version, API name, API type which are used to request the resource release
  - Access token
- Released resource information, such as:
  - UE application ID which requests the resource release
  - Released resource type
  - Released resource ID
- Resource release result

Message format:

```
<AC-RESOURCE-RELEASE-RESPONSE-Message>::=<Message Header>
    {AI-Platform-ID}
    {UE-Info}
    {Released-Resource-Info}
    {Resource-Release-Result}
```

8.2.5 AC redirect API call message

If an AI platform in the edge access network layer cannot serve a received API call, the AI platform redirects it to the AI platform at the core network layer using a redirect API call message, as specified in clause 7.2.4 and shown in Figure 7-9. The message may include information such as:

- AI platform ID which redirects the API call
- Information on the UE which invokes the API call, such as:
  - UE identifier
  - UE context information
  - UE authentication code
  - Called API ID, API version, API name, API type
  - Access token
- Encapsulated API call

Message format:

```
<AC-RESOURCE-PROVISION-REQUEST-Message>::=<Message Header>
    {AI-Platform-ID}
    {UE-Info}
    {Encapsulated-API-Call}
```

8.3 Signalling information for reference point A*m* (Edge access network layer – UE)

The A*m* reference point between UE and the edge access network layer allows UE to invoke diverse API calls of which URI/URL direct to the API resource at the edge access network Layer AI platform. By invoking the API calls, AI applications running on the UE utilize the support of the necessary AI platform service. Figure 8-3 shows signalling flows for reference point A*m*.

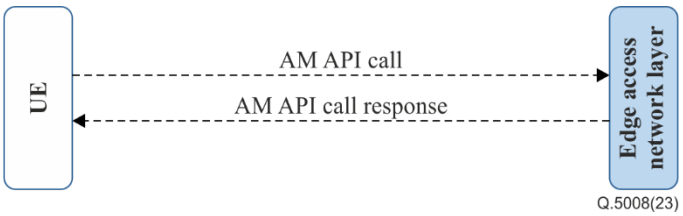

Figure 8-3 – Signalling flows for reference point A*m*

8.3.1 AM API call and response messages

As specified in clause 7.2.4 and shown in Figure 7-9 for the MEC environment case, AI applications running on the UE invoke API calls to utilize the support of necessary AI platform services in the edge access network layer. In order to request/response the services, UE and AI platform in the edge access network layer exchange the API call invoke/response messages.

8.3.1.1 AM API call message

UE invokes API calls to analyse and process data, and to access and use resources including trained AI models and other AI services which are provided by AI platforms in the edge access layer. Each API call is made with the required parameters as specified in the posted API description. Although the required parameters may differ according to API, the API call message may include the following basic information:

- UE application ID which invokes the API call
- Access token
- URI/URL to the API resource
- Input data information (e.g., URI/URL to upload payload, etc.)
- Required parameters for the API call

Message format:

```
<AM-API-CALL-Message>::=<Message Header>
    {Application-ID}
    {Access-Token}
    {API-URI/URL}
    {Input-Data-Info}
    {Request-Parameters}
```

8.3.1.2 AM API call response message

AI platforms respond to the invoked API calls with API call response messages. Response parameters for each API are made up in accordance with the response parameters as specified in the posted API description. Although the response parameters may differ depending on the API, the API call response message may include basic information such as:

- UE application ID which invoked the API call
- Access token
- Uploaded input data ID
- Job ID of the API execution
- Output data information (e.g., URI/URL to output/meta data, etc.)
- Response parameters for the API call
- Error code and description if failed

Message format:

```
<AM-API-CALL-RESPONSE-Message>::=<Message Header>
    {Application-ID}
    {Access-Token}
    {Input-Data-ID}
    {Job-ID}
    {Output-Data-Info}
    {Response-Parameters}
    {Error-Code}
```

8.4 Signalling information for reference point A*n* (Core network layer – UE)

The A*n* reference point between UE and the core network layer allows UE to request AI platform resources and to release the allocated resources. It also allows UE to invoke diverse API calls such as URI/URL direct to the API resource at core network layer AI platform. Scaling up or down the allocated resource is achieved by invoking the APIs related to the allocated resource usage.

Figure 8-4 shows signalling flows for reference point An.

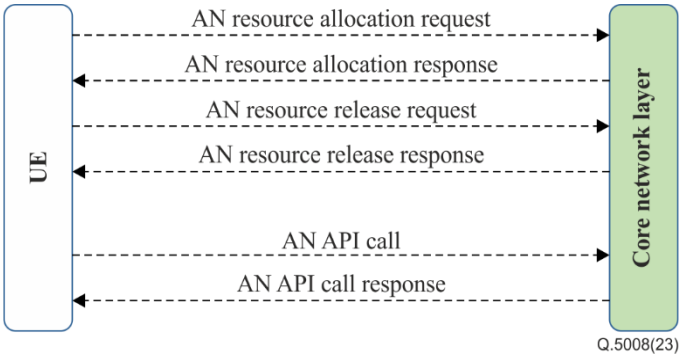

Figure 8-4 – Signalling flows for reference point A*n*

8.4.1 AN resource allocation request and response messages

As specified in clause 7.2.2 and shown in Figure 7-7, AI applications running on the UE may request the allocation of AI platform resources in the core network layer or in the edge access network layer. In order to request/response the resource allocation, the UE and AI platform service manager FE in the core network layer exchange the resource allocation request/response messages.

8.4.1.1 AN resource allocation request message

To request the required AI platform resources such as AI infrastructure (e.g., AI processors), AI data repository storage, AI models, etc., UEs send resource allocation request messages to the AI platform service manager. The message may include information such as:

- UE application ID which requests the resource allocation
- Authentication code
- Resource type for the allocation request
- Required resource descriptor which contains the amount of required capacity and technical specifications such as performance, throughput, latency and resiliency, etc.
- CRUD information

Message format:

```
<AN-RESOURCE-ALLOCATION-REQUEST-Message>::=<Message Header>
    {Application-ID}
```

{Authentication-Code}

{Resource-Type}

{Resource-Descriptor}

{CRUD-Info}

8.4.1.2 AN resource allocation response message

The AI Platform Service Manager at Core Network Layer responds to the UE's resource allocation request with Resource Allocation Response message. The message may include following information such as:

- ID of the AI platform which provisions the requested resource
- UE application ID which requests the resource allocation
- Allocated resource type
- Allocated resource ID
- APIs each of which contains the URI/URL to access and utilize the allocated resource
- Allocated resource descriptor which contains the amount of required capacity and technical specifications such as performance, throughput, latency and resiliency, etc.
- CRUD information
- Constraint information for the access and the use of the allocated resource
- Resource provision result

Message format:

<AN-RESOURCE-ALLOCATION-RESPONSE-Message>::=<Message Header>

{AI-Platform-ID}

{Application-ID}

{Resource-Type}

{Resource-ID}

{APIs}

{Resource-Descriptor}

{CRUD-Info}

{Constraint}

{Resource-Provision-Result}

8.4.2 AN resource release request and response messages

As specified in clause 7.2.3 and shown in Figure 7-8, AI applications running on the UE may request the release of AI platform resources in the core network layer or in the edge access network layer. In order to request/response the resource release, the UE and AI platform service manager FE in the core network layer exchange the resource release request/response messages.

8.4.2.1 AN resource release request message

To release the allocated resource, UEs send a resource release request message to the AI platform service manager at the core network layer. The message may include information such as:

- ID of the AI platform which provisions the allocated resource
- UE application ID which requests the resource release
- Authentication code
- Allocated resource type to release
- Allocated resource ID to release

- Reason to release

Message format:

```
<AN-RESOURCE-RELEASE-REQUEST-Message>::=<Message Header>
    {AI-Platform-ID}
    {Application-ID}
    {Authentication-Code}
    {Resource-Type}
    {Resource-ID}
    {Reason}
```

8.4.2.2 AN resource release response message

The AI platform service manager at the core network layer responds to the UE's allocated resource release request with a resource release response message. The message may include information such as:

- ID of the AI platform which released the allocated resource as requested
- UE application ID which requested the resource release
- Released resource type
- Released resource ID
- Resource release result

Message format:

```
<AN-RESOURCE-RELEASE-RESPONSE-Message>::=<Message Header>
    {AI-Platform-ID}
    {Application-ID}
    {Resource-Type}
    {Resource-ID}
    {Result}
```

8.4.3 AN API call and response messages

As specified in clause 7.2.4 and shown in Figure 7-9 for the non-MEC environment case, AI applications running on the UE invoke API calls to utilize the support of necessary AI platform services in the core network layer. In order to request/response the services, UE and AI platform in the core network layer exchange the API call invoke/response messages.

8.4.3.1 AN API call message

The UE invokes API calls to analyse and process data, to access and use resources including trained AI models and other AI services. which are provided by AI platforms at the core network layer. Each API call is made with the required parameters as specified in the posted API description. Although the required parameters may differ depending on the API, the API call message may include the following basic information:

- UE application ID which invokes the API call
- Access token
- URI/URL to the API resource
- Input data information (e.g., URI/URL to upload payload, etc.)
- Required parameters for the API call

Message format:

```
<AN-API-CALL-Message>::=<Message Header>
    {Application-ID}
    {Access-Token}
    {API-URI/URL}
    {Input-Data-Info}
    {Request-Parameters}
```

8.4.3.2 AN API call response message

AI platforms respond to the invoked API calls with an API call response message. Response parameters for each API are made up in accordance with the response parameters as specified in the posted API description. Although the response parameters may differ depending on the API, the API call response message may include basic information such as:

- UE application ID which invoked the API call
- Access token
- Uploaded input data ID
- Job ID of the API execution
- Output data information (e.g., URI/URL to output/meta data, etc.)
- Response parameters for the API call
- Error code and description if failed

Message format:

```
<AN-AI-CALL-Response-Message>::=<Message Header>
    {Application-ID}
    {Access-Token}
    {Input-Data-ID}
    {Job-ID}
    {Output-Data-Info}
    {Response-Parameters}
    {Error-Code}
```

9 Security considerations

This Recommendation provides signalling requirements and architecture to support AI-based vertical services in future networks including IMT-2020 and beyond. These networks are subject to security and privacy measures, and sensitive information should be protected as a high priority in order to avoid leaking and unauthorized access. Security and privacy concerns should be aligned with the requirements specified in [b-ITU-T Y.3101] and [b-ITU-T Y.2701], together with the security framework from network function virtualization [b-ITU-T X.1046] and [b-ETSI GS NFV-SEC 022].

AI modelling and data handling for the AI platform services, in addition, require to be carried out with considerations for the security requirements such as those specified in [ITU-T M.3080], [ITU-T Y.3172], [ITU-T Y.3174] and [b-ITU-T X.1601].

Appendix I

Use cases for AI service in vertical domains

(This appendix does not form an integral part of this Recommendation.)

AI can be used in different types of services across diverse industry verticals where optimized AI service provision is required.

This appendix describes use cases of AI in vertical industry domains to illustrate the concept and service flow in domains, such as the public security, media, manufacturing and medical industry domains. Also, this appendix provides the flow of how raw vision data, image or video, is used by an AI to help improve traditional video surveillance-based monitoring systems.

I.1 Use cases for AI service in vertical domains: Public security

I.1.1 Use case: AI service based public security service to improve safety

As shown in Figure I.1, AI function is applied to edge devices equipped with cameras, such as drone, surveillance cameras or a closed-circuit television (CCTV) system, or light detection and ranging (LiDAR) captures and analyses video data, or streams out the captured video data to cloud servers for analysis. Analysed data generated, either from devices or cloud servers, helps agents in the control centre of the security service providers to narrow down their focus on important events or issues from hundreds of monitors. It is hoped that the service based on this system will eventually contribute to building a safer society.

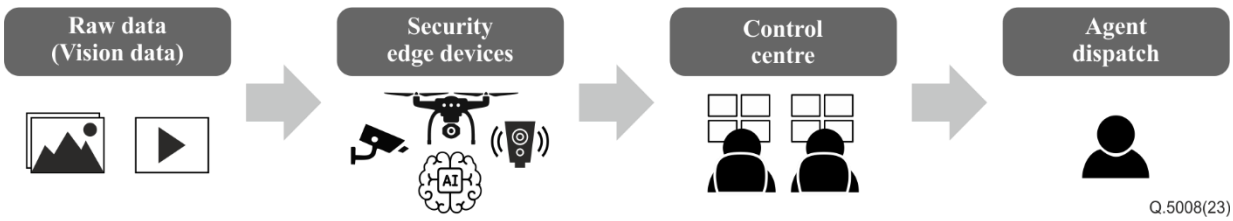

Figure I.1 – Concept of AI service in the security domain

The service will be largely provided to the areas such as public safety and security, protection of national infrastructures and facilities, or prevention of serious industrial disasters.

I.1.2 Service scenario: AI in security domain

Figure I.2 illustrates a service scenario based on AI in a security domain. The scenario constitutes two parts. First a 'learning and modelling' part followed by a 'service provision' part.

a) Learning and modelling

An AI automation platform labels the very first sets of data generated from edge devices such as CCTV, drone, or LiDAR and trains to create an initial AI model. Through validation, testing and additional training, the AI platform reinforces the AI model with higher accuracy.

b) Service provision

In the service provision step, the generated AI model is deployed to diverse edge devices in the fields where the security solutions are required, such as roadways, construction sites, or national facilities with high importance. Edge devices capture images or video in each site and stream the data to the security control centre. Then, the agents in the control centre analyse the image and video data, signal check for false alarms, and most importantly can focus on important events or issues that have occurred. The set of meta data generated during the validation process can be used to improve the accuracy of the AI model.

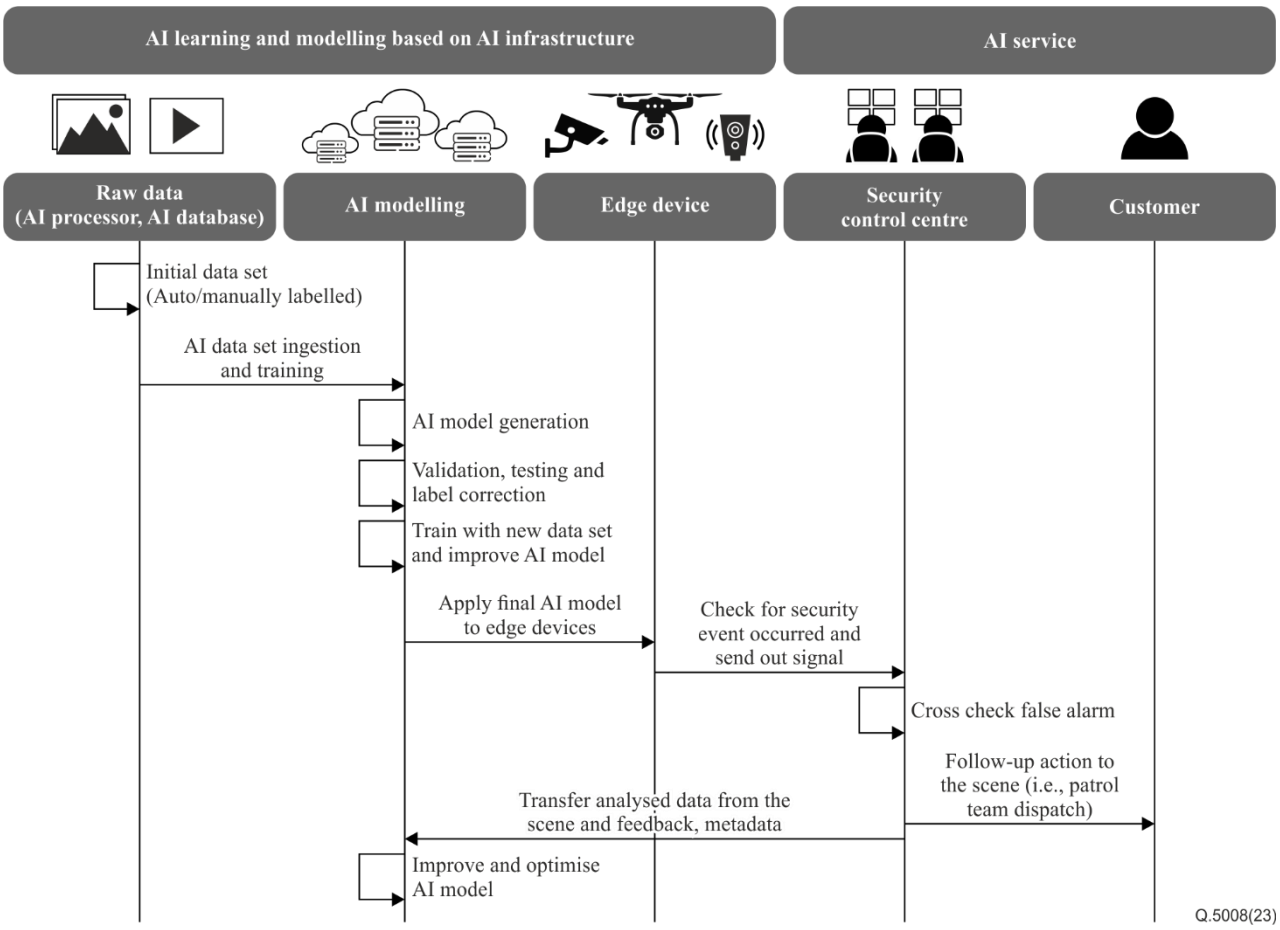

Figure I.2 – AI service scenario in the security domain

I.2 Use cases for AI service in vertical domains: Media

This appendix describes the flow how the source data, image or video, is used by an AI to help improve streaming services, such as internet protocol television (IPTV) or over-the-top (OTT) streaming services.

I.2.1 Use case: AI-based streaming service to upscale video data

As shown in Figure I.3, AI is used in video upscaling and streaming services. As display technologies are advancing rapidly, old videos that were taken in low resolution cannot catch-up with display requirements, stimulating the industry's needs in video upscaling solutions. Appendix II explains how AI can help significantly improve the video upscaling process and how media service providers can enhance the quality of streamed videos.

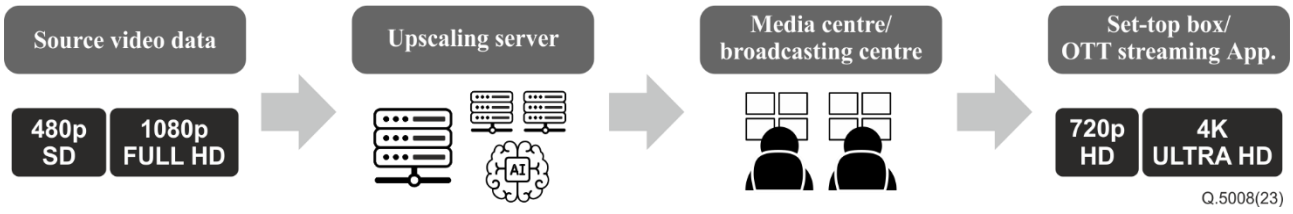

Figure I.3 – Concept of AI full stack-based service in the media domain

I.2.2 Service scenario: AI service in media domain

Figure I.4 illustrates a service scenario based on AI in the media domain. The scenario constitutes two parts, 'learning and modelling' and 'service provision'.

a) Learning and modelling

High-quality video data sources are ingested to an AI automation platform and help the platform generate the initial AI model. Through the repetitive training-validation-testing process, the AI platform reinforces the AI model with higher accuracy, and generates the final AI model that can be used in commercial-grade video upscaling services.

b) Service provision

In the service provision step, the generated AI model is deployed to media upscaling servers in the data centre of media service providers, such as broadcasting stations, IPTV service providers or even OTT service providers. The deployed AI model upscales video, or images into high-quality media data. In the case of video, the model can upscale SD quality video to HD, and FHD quality video to 4K UHD grade. After the video upscaling is finished, the media centre or broadcasting centre transmits the upscaled video data to customer-side set-top boxes or streaming applications to provide an enhanced media experience. The same upscaling scheme can be also used to franchise movie theatre networks, by sending out video data to each theatre's micro-data centres in the site.

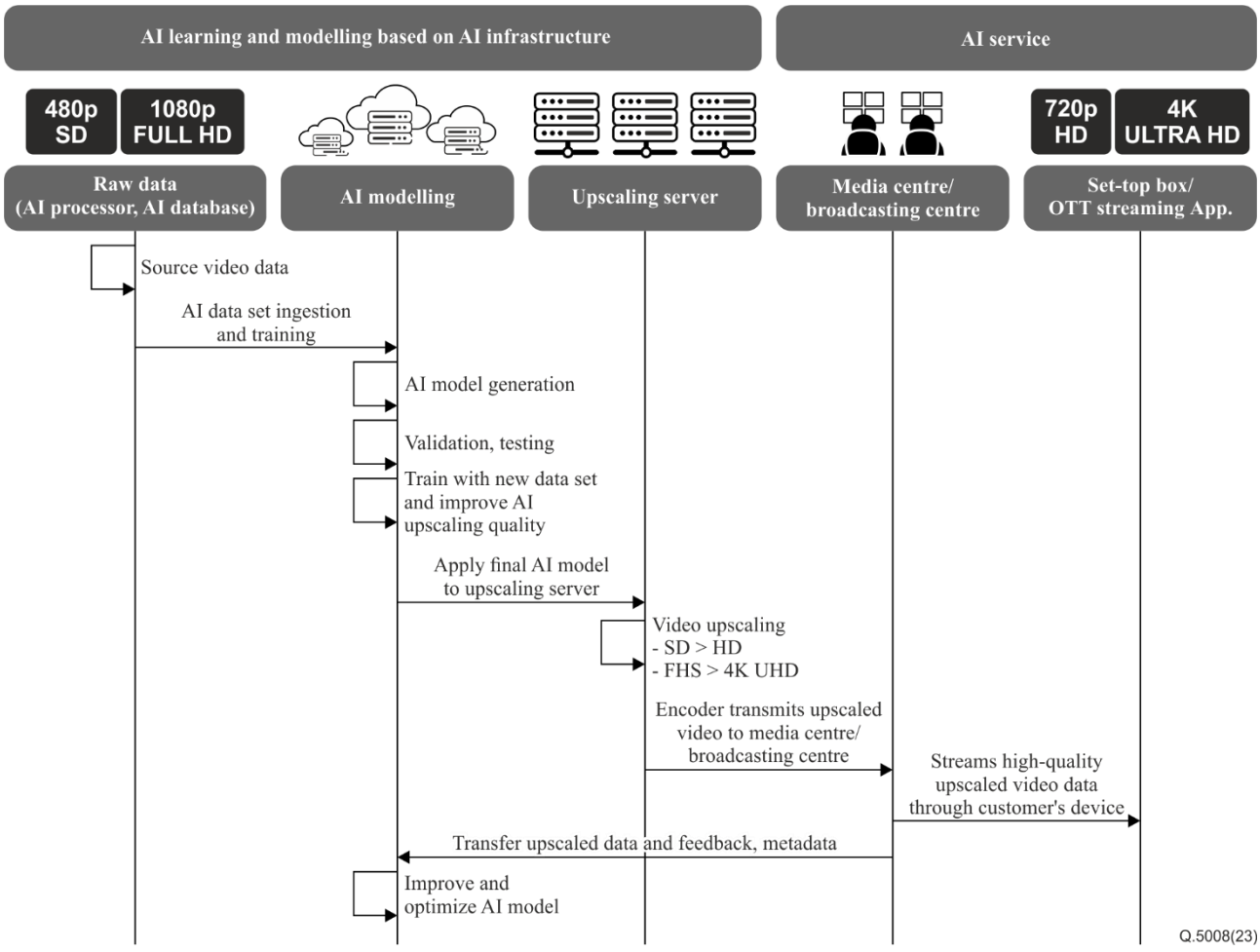

Figure I.4 – AI service scenario in the media domain

I.3 Use cases for AI service in vertical domains: Industry (manufacturing/logistics)

This appendix describes the flow of how source data is used by an AI full stack to help improve efficiency in industries such as manufacturing and logistics.

I.3.1 Use case: AI service to add intelligence to factory/logistics machines

As shown in Figure I.5, an AI full stack is used to upgrade equipment such as robot-based depalletization, or 5G-AI machine vision solutions that are used in smart factories and smart logistics centres. Appendix I.3 explains how AI can add intelligence to industry-grade robots or equipment, and help employees to enhance workplace efficiency and productivity, which will eventually lead to revenue increase.

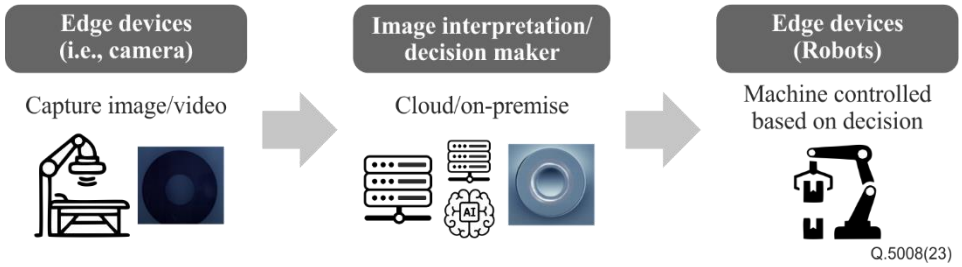

Figure I.5 – Concept of AI service in industry (manufacturing/logistics) domain

I.3.2 Service scenario: AI service in industry (manufacturing and logistics) domain

Figure I.6 illustrates a service scenario based on AI in an industry (manufacturing and logistics) domain. This scenario constitutes two parts, a 'learning and modelling' part, where the edge devices take images or videos of manufactured items for analysis, and a 'service provision' part, where the edge devices perform their tasks following the decisions made.

a) Learning and modelling

High-quality video data sources that illustrate 'defects' of manufactured goods are ingested to an AI automation platform and help the platform generate the initial AI model. Through the repetitive training-validation-testing process, the AI platform reinforces the AI model with higher accuracy to detect defects. The final AI model generated is deployed to smart factory solutions or smart logistics solutions.

b) Service provision

In the service provision step, the generated AI model is deployed to decision/interpretation servers, which are a part of smart factory solutions. The AI model can also be loaded to robot vision services, such as (de)palletization, to help edge devices such as brainless robots or machine vision solutions to perform designated jobs by receiving control signals from the decision or interpretation server.

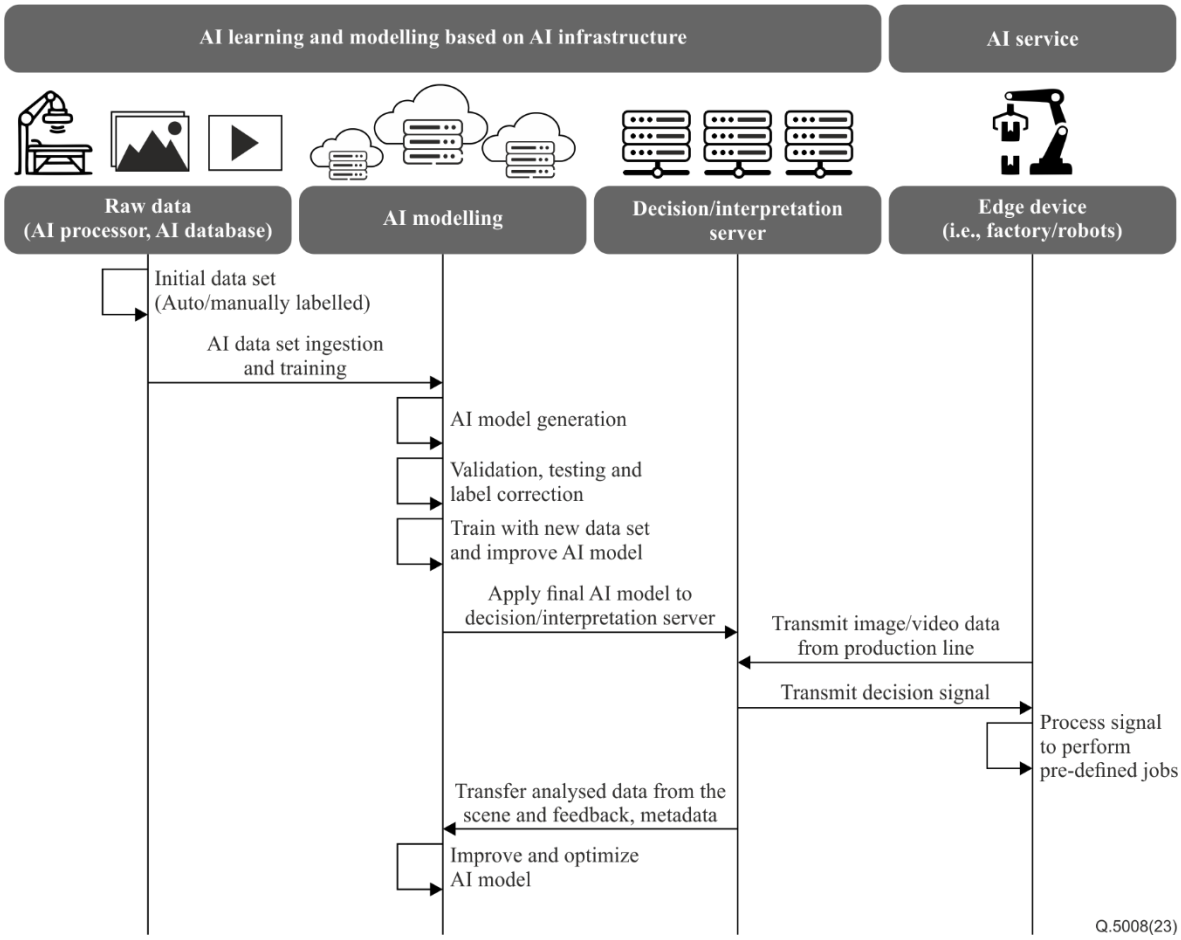

Figure I.6 – AI full stack-based service scenario in the industry domain

I.4 Use cases for AI service in vertical domains: Medical

This appendix describes the flow of how the source data is used by an AI full stack to help doctors to get assistance from medical AI solutions in disease diagnosis [b-ITU-T Y-Sup.55].

I.4.1 Use case: AI service to add intelligence to diagnosis assistance solutions

As shown in Figure I.7, AI is used to help doctors by adding intelligence to medical devices that assist decision-making in diagnosing diseases such as brain disease or pneumonia where doctors have to look at and check the images or videos taken from the devices manually. By deploying AI engines to the diagnosis assistance solution, the overall accuracy of a doctor's diagnosis can be assisted through a diagnosis assistance solution and so accuracy when making a diagnosis is enhanced.

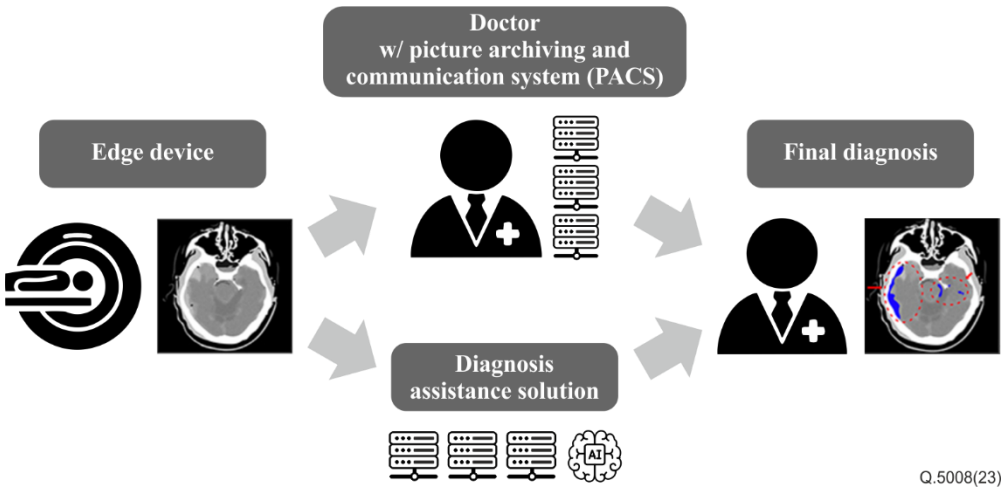

Figure I.7 – Concept of AI service in medical domain

I.4.2 Service scenario: AI service in medical industry domain (Medical)

Figure I.8 illustrates a service scenario based on an AI full stack in the medical industry domain. The scenario constitutes two parts, the 'learning and modelling' part where the AI model learns medical vision AI data from X-rays and MRIs to build the final AI model for disease analysis. In the 'service provision', part the AI model deployed in the diagnosis assistance solutions receives data from edge devices to perform the diagnosis assistance functions, and transmits analysed results to the picture archiving and communication systems (PACS), or to doctors directly, so that the doctor who received the analysed data can sum up diagnosis resulting from visual inspection, and the result from the diagnosis assistance solution. This will eventually help improve the accuracy rate of disease diagnosis.

A) Learning and modelling

High-quality video and image data sources already taken by medical devices, such as X-rays and MRIs, which illustrate certain diseases inside the human body, such as brain disease or pneumonia. Utilizing the ingested raw data, the AI automation platform generates an initial AI model and through the repetitive training-validation-testing process, the AI platform reinforces an AI model with higher accuracy to detect disease. The final AI model generated is deployed to the diagnosis assistance solutions.

b) Service provision

In the service provision step, the generated AI model is deployed to the diagnosis assistance solution servers, to detect disease which the model has learned through the learning process. As medical edge devices take new images or videos from the patient, it transmits data through PACS, and the diagnosis assistance solution uses received data to interpret and make analysis results on each ingested data. Simultaneously, the doctor also conducts a visual inspection using PACS, manually to make his or her own decision. Finally, as the doctor sums up the analysed result created by the AI model and visual inspection, and the doctor comes up with final diagnosis result. Through this process, medical institutions, such as hospitals, can reach higher accuracy in disease diagnosis.

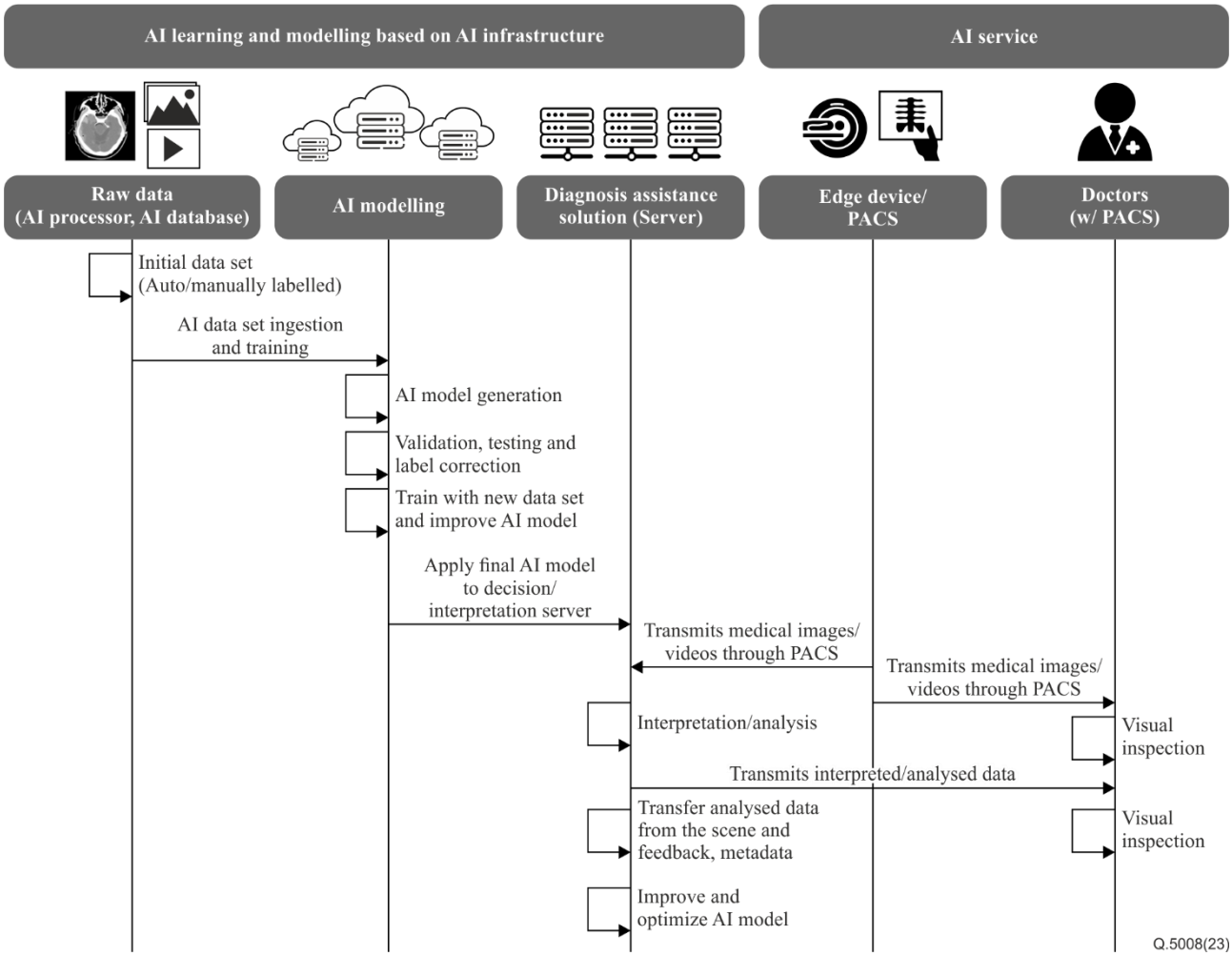

Figure I.8 – AI service scenario in the medical domain

Appendix II

AI base service implementation model

(This appendix does not form an integral part of this Recommendation.)

Generating AI models for the AI-based services requires a huge amount of high-performance AI dedicated cloud resources and expertise.

Running such a required environment is often a heavy burden for individual AI-based service providers and even for small and medium sized businesses (enterprises).

IMT-2020 network operators run large-scale cloud computing facilities in their networks including edge cloud. Automated AI modelling and services support platform (i.e., AI platform), in their cloud facilities and can provide required resources and services for the AI-based service providers to develop and deploy the necessary AI models and applications. [b-ITU-T Y.3100]

Figure II.1 shows an AI-based service implementation model.

II.1 AI platform

Technical components of the AI platform, which are required to support automated AI modelling and services, are:

- AI data repository and database management system (DBMS)

  AI data repository and DBMS stores and searches raw data, trained data, and meta data for AI learning and modelling.
- Automated AI modelling tools

  Automated AI modelling tools perform the general AI modelling suitable for the AI services based on the labelled data.
- Data warehouse

  Data warehouse builds the pipeline of data collection, annotation and labelling, AI learning and modelling, and AI model deployment process.
- AI processors and accelerators

  Dedicated AI processors, i.e., neural processing units (NPUs) and accelerators provide optimized high-performance AI algorithm operations for deep learning and inference, etc.
- APIs for AI service

  Open API for the interworking with 3rd party systems and devices.

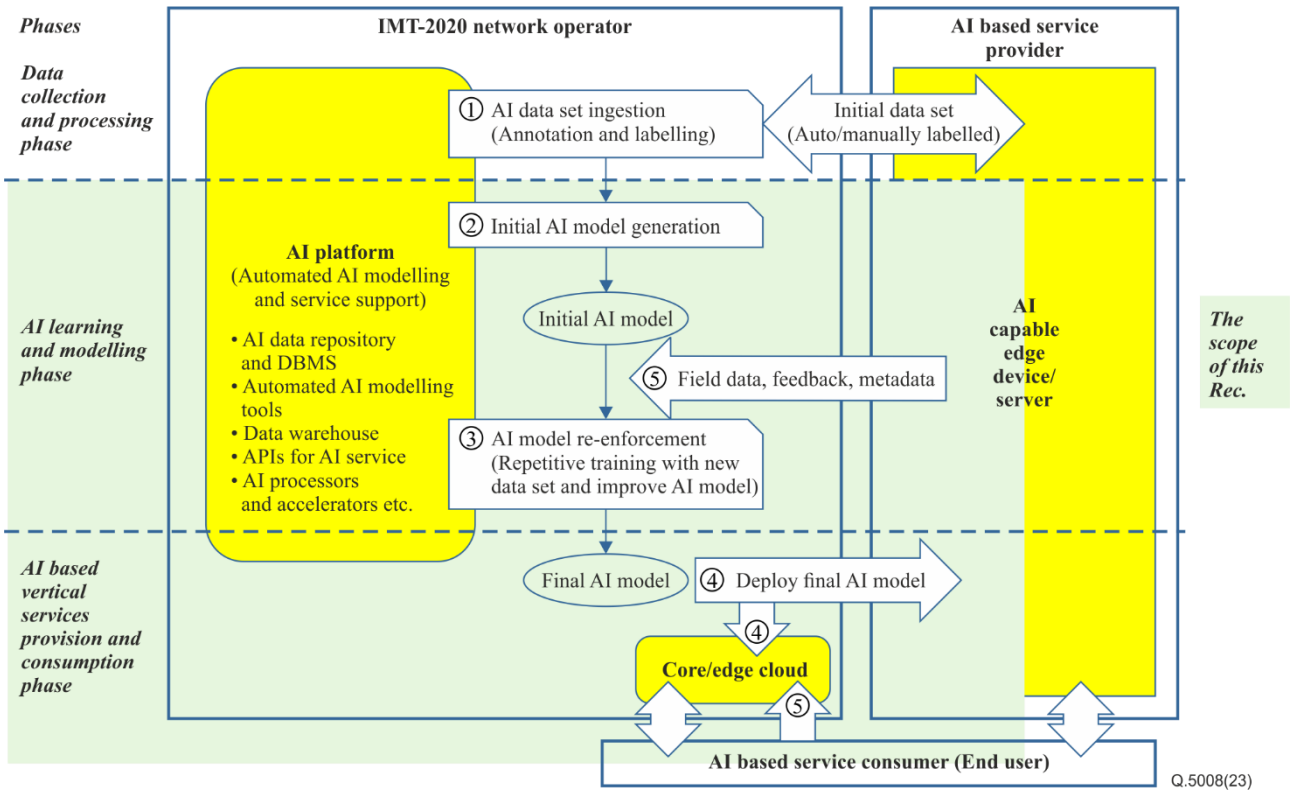

Figure II.1 – AI-based service implementation model

II.2 Roles in the AI-based service implementation model

From a logical point of view, AI-based service implementation involves following three distinct roles:

1) AI-based service provider

The AI-based service provider is an entity that provides services to the AI-based service consumer e.g., end user customers or devices in the field.

It hosts end user consumers and provides AI-based services (i.e., AI applications) developed based on the AI models. AI models to be used for the AI applications are generated with the support of the IMT-2020 network operator's AI platform and can be deployed either in the IMT-2020 network operator's edge/core network cloud or in the AI-based service provider's own facility.

If an enterprise or an institution builds the AI-based services for its own use, it plays both the AI-based service provider role and the consumer role at the same time.

To provide AI-based services to the end user consumers, the service provider may run AI capable edge devices/servers on which the AI model is loaded. For example, the drones, surveillance cameras or LiDAR in the 'public security services' use case can be considered as the AI-based service provider's edge devices. In the use case scenarios of Appendix I, the 'upscaling server in video upscaling and streaming service' use case, the 'decision/interpretation server in manufacturing and logistics' use case and the 'diagnosis assistance solution server in intelligent diagnosis assistance medical service' use case can be considered as examples of an AI-based service provider's servers.

For the AI model reinforcement, AI-based service providers transfer analysed data from the field, feedback and metadata to the AI platform of the IMT-2020 network operator. In the use case scenarios of the Appendix I, this is performed by the 'decision/interpretation server' of the security control centre in the 'public security service' use case, and the media centre/broadcasting centre in the 'video upscaling and streaming service' use case. In contrast with other use case scenarios, the 'diagnosis assistance solution server' performs the AI model reinforcement in the AI-based service provider's facility instead of the AI platform of the IMT-2020 network operator.

2) IMT-2020 network operator

The IMT-2020 network operator provides automated AI modelling support for the AI-based service provider role entities. Its AI platform generates trained AI models to be deployed. AI-based service providers load them on their edge devices and servers through APIs for AI services.

The IMT-2020 network operator can also host and serve AI-based service end user consumers directly at its edge/core cloud. In such cases, it plays both the IMT-2020 network operator role in the sense of providing the AI platform and also plays the AI-based service provider role in the sense of providing the service to the end user consumer.

3) AI-based service consumer (end user)

The end user is an entity that consumes the service hosted by AI-based service providers. It can be an end user equipment or a customer such as, in the use case scenarios of Appendix I, a set-top box/OTT streaming App in the 'video upscaling and streaming service' use case, factory/robots in the 'manufacturing and logistics' use case, PACS and doctors with PACS in the 'intelligent diagnosis assistance medical service' use case.

When the factory/robots are run by an AI-based service provider, they are considered as the end user equipment from a logical point of view even though they are edge devices of the service provider.

II.3 AI-based service implementation phases

AI-based services are implemented in the following three phases:

1) Data collection and processing phase

AI-based service providers provide the initial AI data set for learning and modelling. The initial AI data set, which is captured and analysed either from their AI functions running on the edge devices or cloud servers, may be labelled manually. Then the IMT-2020 network operator's AI platform ingests the initial AI data set by performing the annotation and labelling process.

This data collection and ingestion procedure between AI-based service providers and IMT-2020 network operator's AI platform is out of scope of this Recommendation.

2) AI learning and modelling phase

With the ingested initial AI data, the AI platform generates the initial AI model. Through the repetitive training with validation and testing processes for the new data set, the AI platform reinforces the AI model with improved accuracy, and generates the final AI model.

3) AI-based services provision and consumption phase

The generated AI model is deployed to diverse AI capable edge devices in the field and servers in the AI-based service provider's facility such as security control centre, media/broadcasting centre, or on-premises cloud server, etc. as described in Appendix. According to circumstances, they can be deployed in the core/edge network clouds too.

End user customers or devices consume the service provided by the AI applications. AI-based service provider's edge devices or servers on which AI model or application is loaded, capture and analyse the field data during the provider's own validation process and generate a set of meta data and feedback. They transfer these data to the AI platform to be used for the AI model reinforcement.

Bibliography

[b-ITU-T X.1046] Recommendation ITU-T X.1046 (2020), *Framework of software-defined security in software-defined networks/network functions virtualization networks.*

[b-ITU-T X.1601] Recommendation ITU-T X.1601 (2015), *Security framework for cloud computing*

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

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

[b-ITU-T Y.3101] Recommendation ITU-T Y.3101 (2018), *Requirements of the IMT-2020 network.*

[b-ITU-T Y-Sup.55] ITU-T Y.3170-series Recommendations – Supplement 55 (2019), *Machine learning in future networks including IMT-2020: use cases.*

[b-ETSI GR ENI 004] ETSI GR ENI 004 V1.1.1 (2018), *Experiential Networked Intelligence (ENI); Terminology for Main Concepts in ENI*.

[b-ETSI GS MEC 001] ETSI GS MEC 001 (2019), *Multi-access edge computing (MEC) terminology.*

[b-ETSI GS NFV-SEC 022] ETSI GS NFV-SEC 022 (2020), *Network functions virtualisation (NFV) release 2; Security; Access token specification for API access*.

[b-ISO/IEC 22989] ISO/IEC 22989:2022, *Information technology – Artificial intelligence – Artificial intelligence concepts and terminology.*

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	**Switching and signalling, and associated measurements and tests**
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Published in Switzerland
Geneva, 2024