ITUPublications
Recommendations

International Telecommunication Union
Standardization Sector

Recommendation

ITU-T F.748.38 (06/2024)

SERIES F: Non-telephone telecommunication services

Multimedia services

Technical specification for artificial intelligence cloud platform: General architecture

ITU-T F-SERIES RECOMMENDATIONS

Non-telephone telecommunication services

TELEGRAPH SERVICE	F.1-F.109
Operating methods for the international public telegram service	F.1-F.19
The gentex network	F.20-F.29
Message switching	F.30-F.39
The international telemessage service	F.40-F.58
The international telex service	F.59-F.89
Statistics and publications on international telegraph services	F.90-F.99
Scheduled and leased communication services	F.100-F.104
Phototelegraph service	F.105-F.109
MOBILE SERVICE	F.110-F.159
Mobile services and multidestination satellite services	F.110-F.159
TELEMATIC SERVICES	F.160-F.399
Public facsimile service	F.160-F.199
Teletex service	F.200-F.299
Videotex service	F.300-F.349
General provisions for telematic services	F.350-F.399
MESSAGE HANDLING SERVICES	F.400-F.499
DIRECTORY SERVICES	F.500-F.549
DOCUMENT COMMUNICATION	F.550-F.599
Document communication	F.550-F.579
Programming communication interfaces	F.580-F.599
DATA TRANSMISSION SERVICES	F.600-F.699
MULTIMEDIA SERVICES	**F.700-F.799**
ISDN SERVICES	F.800-F.849
UNIVERSAL PERSONAL TELECOMMUNICATION	F.850-F.899
ACCESSIBILITY AND HUMAN FACTORS	F.900-F.999

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

Recommendation ITU-T F.748.38

Technical specification for artificial intelligence cloud platform: General architecture

Summary

Recommendation ITU-T F.748.38 provides technical specifications and capability requirements for artificial intelligence cloud platform. It specifies the capabilities of artificial intelligence cloud platforms from service providers in the following six aspects: resource management, model development, model deployment, high availability, performance and platform security.

History*

Edition	Recommendation	Approval	Study Group	Unique ID
1.0	ITU-T F.748.38	2024-06-13	16	11.1002/1000/15917

Keywords

Artificial intelligence, cloud computing, distributed machine learning, high availability, machine learning, platform security, resource scheduling.

* 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 F.748.38

Technical specification for artificial intelligence cloud platform: General architecture

1 Scope

This Recommendation focuses on technical specifications and capability requirements of artificial intelligence cloud platform. It is suitable for guiding cloud vendors to build machine learning service specifications and providing references for enterprise users to select artificial intelligence cloud products. The Recommendation includes six aspects: resource management, model development, model deployment, high availability, performance and platform security.

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.

None.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

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

3.1.2 machine learning model [b-ITU-T Y.3172]: Model created by applying machine learning techniques to data to learn from.

3.2 Terms defined in this Recommendation

None.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ASIC Application Specific Integrated Circuit

CPU Central Processing Unit

FPGA Field Programmable Gate Array

GPU Graphics Processing Unit

IaaS Infrastructure as a Service

ONNX Open Neural Network Exchange

PMML Predictive Model Markup Language

5 Conventions

The following conventions are used in this Recommendation:

The keywords "is recommended" indicate a requirement that is recommended, but which is not absolutely required to claim conformance with this Recommendation.

6 General architecture of artificial intelligence cloud platform

Figure 6-1 shows that the general architecture of artificial intelligence cloud platform consists of platform security, performance, model development, model deployment, resource management and high availability.

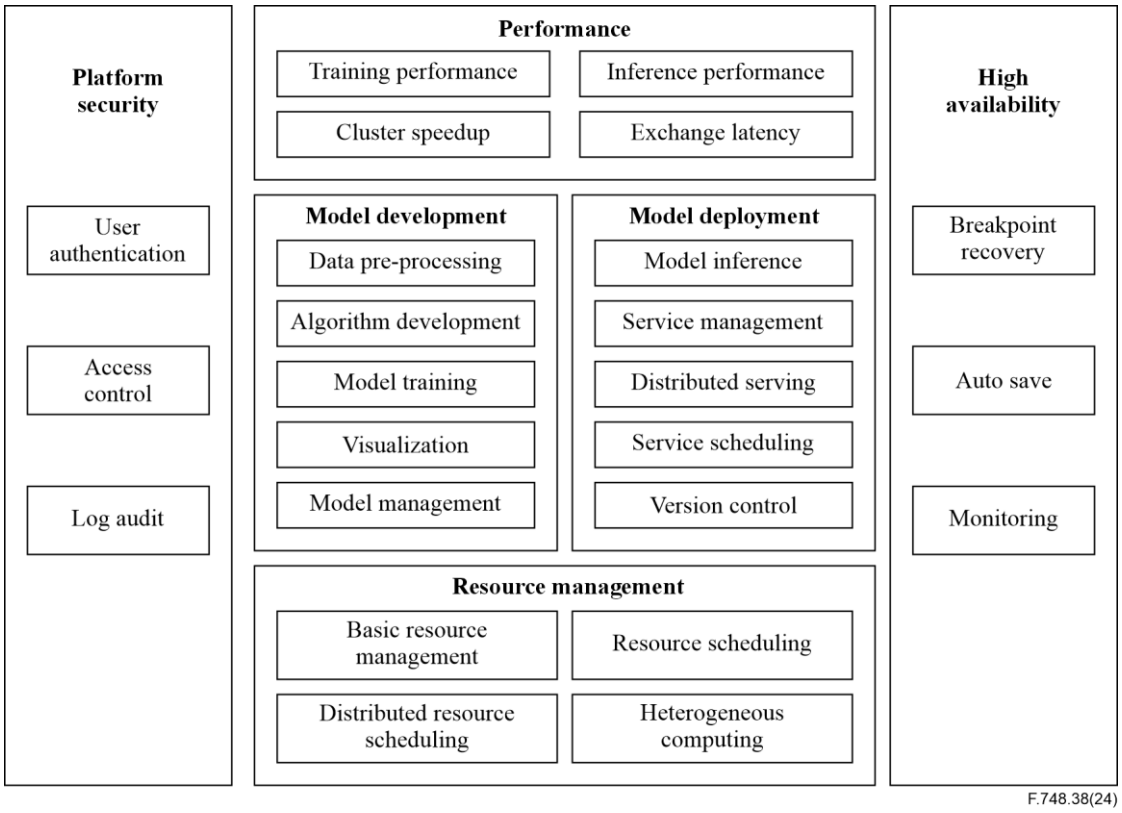

Figure 6-1 – General architecture of artificial intelligence cloud platform

7 Resource management

7.1 Basic resource management

The platform needs to effectively manage the resources that rely on the infrastructure layer (IaaS) to support machine learning model training, inference and other tasks. The platform is recommended to support the following minimum set of features:

a) Support monitoring of the platform IaaS resources, including monitoring of virtual machines, graphics processing units (GPUs), memory, storage and network.

b) Support monitoring of customers' container resources, including monitoring of GPUs, memory, storage and network.

7.2 Resource scheduling

The platform needs to support user customized configuration during resource creation and modification. Resources can be increased/decreased according to specific tasks, and idle resources can be automatically recovered. The platform is recommended to support the following minimum set of features:

a) Allow users to increase/decrease the number of GPUs or the allocation of elastic compute service.

b) Automatic recovery of idle resource after training or inference session ends.

c) Scheduling and management of distributed resources, e.g., multi-cluster.

d) Optimization of distributed computing, e.g., network, transmission, storage, etc.

7.3 Heterogeneous computing

The platform needs to process massive data and perform large-scale calculations for training and inferencing machine learning tasks. It needs to support parallel computing with heterogeneous resources in order to accelerate tasks. The platform is recommended to support the following minimum set of features:

a) GPU-accelerated computing and GPU resource scheduling.

b) Field programmable gate array (FPGA)-accelerated computing and FPGA resource scheduling.

c) Application specific integrated circuit (ASIC)-accelerated computing and ASIC resource scheduling.

8 Model development

8.1 Data pre-processing

The platform is recommended to have certain processing ability for massive data, including support for source data format collection and the whole data processing flow. The platform is recommended to support the following minimum set of efficient data-access strategies:

a) Processing of structured or unstructured data, e.g., picture, video, audio, etc.

b) Mass upload of large-scale data; support breakpoint recovery for data transmission.

c) Complete data pre-processing flow.

8.2 Algorithm development

For flexible model editing, the platform provides management tool of algorithm development, helping reduce the complexity of model development process. The platform is recommended to support the following minimum set of features:

a) Development methods in various forms, e.g., drag-and-drop tool, notebook, etc.

b) Real-time monitoring of performance metrics during model development.

c) Creation, deletion, start and stop of algorithm services; support service version management.

d) Query of service history and current status.

e) Design, publication and utilization of user-customized algorithms, bringing greater flexibility to users.

8.3 Model training

The platform is recommended to handle model training tasks in an effective and reliable manner.

a) Support training task management, e.g., create, delete, update, retrieve.

b) Support real-time monitoring of performance metrics during model training process.

c) Utilize reliability strategies, e.g., checkpoint, breakpoint recovery, task retry.

8.4 Visualization

The platform is recommended to provide visualization charts of models, data and parameters, which is convenient for performance comparison on the user side.

a) Support multiple displays of model performance indicators, such as tables and charts.

b) Support visualization of both structured and unstructured data.

8.5 Model management

The platform is recommended to support the following minimum set of model management operations:

a) Model creation, storage and deletion.

b) Display of model performance metrics, e.g., accuracy.

c) The import and export of model files in common standard formats, e.g., predictive model markup language (PMML), TFserving, open neural network exchange (ONNX), etc.

9 Model deployment

9.1 Model inference

Model deployment is the process of integrating the AI models into production to make decisions on real-world data. The platform is recommended to support the following minimum set of model inference operations:

a) Accelerate and standardize the process of model inference.

b) Build scalable and high-performance services of model inference.

9.2 Service management

The platform provides service management components to help users quickly and conveniently complete operations such as the release, deployment of AI models. The platform is recommended to support service management operations.

Including but not limited to:

a) Provides functions of standard API encapsulation.

b) Supports unified interface to create tasks, start training, model publishing and other functions.

9.3 Distributed serving

The platform's distributed scheduling system distributes workflows to different nodes in multiple clusters, providing elastically scalable computing capabilities for model inference. The platform is recommended to support the following minimum set of distributed serving features:

a) parallel processing of multiple tasks;

b) load balancing of distributed tasks.

9.4 Service scheduling

The platform is recommended to support the following minimum set of service scheduling features:

a) multi-instance concurrent scheduling of models, e.g., batch parameters, manual scheduling, rerun, etc.;

b) load balancing for service scheduling and elastic scalability of corresponding resources;

c) distributed service scheduling strategy.

9.5 Version control

The platform is recommended to support the following minimum set of version control operations features:

a) service version management, including A/B testing, model rollback, grey released, etc.;

b) service version management of algorithm services, service history records, query of current status of services, etc.

10 High availability

10.1 Breakpoint recovery

To improve the stability of the training process, the platform is recommended to support saving of the current state if a task is interrupted or stopped accidentally to enable task continuation after state recovery. Support of the following minimum set of features is recommended:

a) breakpoint resume;

b) automatic retries of tasks.

10.2 Auto saving

The platform needs to support automatic saving during model building, training and inference, in order to prevent any significant loss caused by mis-operation. Support of the following minimum set of features is recommended:

a) Configurable auto-saving policy.

10.3 Monitoring

The platform conducts multi-dimensional statistics of basic metrics, such as usage of central processing unit (CPU), memory, network, disk, etc. By accurately reflecting real-time traffic and health status, it provides references for evaluation of resources consumption and service load during a training or inference task. Operations personnel can observe real-time status of the platform, thus improving the efficiency of online operation and maintenance. Support of the following minimum set of features is recommended:

a) single task monitoring, including cluster resources consumption and task instance status;

b) single task life cycle management;

c) user customized alarm configuration, including methods of alarm notification, e.g., e-mail, short message, etc.;

d) provide visualization charts of platform operation status and resource utilization status.

11 Performance

11.1 Training performance

Mainly refers to performance indicators of the platform processing model training tasks, providing reference for improvement and optimization of training process. Tracking of the following minimum set of metrics is recommended:

a) Platform throughput: the maximum number of tasks that the platform supports for simultaneous training.

b) Training speed: the time that the platform takes to complete same training tasks compared with others, the less the better.

c) Data dimension processing: support data processing for high vector dimension.

11.2 Inference performance

Mainly refers to performance indicators of the platform processing model inference tasks, providing reference for improvement and optimization of inference process. Tracking of the following minimum set of metrics is recommended:

a) Throughput: the number of queries processed per second.

b) Delay: the process time of each inference request.

c) Inference speed: the time that the platform takes to complete same inference tasks compared with others, the less the better.

11.3 Cluster speed-up

The platform provides a distributed scheduling strategy to make full use of the advantages of multi-machine and multi-chips in clusters, improving its cluster speed-up ratio. Tracking of the following minimum set of metrics is recommended:

a) The ratio between performance growth and resource growth, i.e., the amount of processing acceleration upon increasing computing resources for the task.

11.4 Exchange latency

The delay of parameter exchange or data exchange in distributed training should be tracked and the following minimum set of metrics is recommended:

a) processing time of parameter exchange;

b) processing time of data exchange.

12 Platform security

12.1 User authentication

The platform is recommended to utilize access-control methods to ensure the security of the platform. Support of the following minimum set of features is recommended:

a) identification and authentication of system users;

b) definition of password complexity rules and password change frequency;

c) identification of administrators with two or more verification techniques combined.

12.2 Access control

The platform is recommended to achieve resource isolation for different tenants by controlling permissions of different users and roles. Support of the following minimum set of features is recommended:

a) security policy-based user access control over data and resource; support application isolation, data isolation, resource isolation and operation isolation among tenants;

b) role-based user permission authorization;

c) multi-tenant management;

d) sensitivity label for important data and resources, and support security policy based strict user operation control over data and resources with sensitivity labels.

12.3 Log audit

The platform is recommended to be able to record a series of user operations, and be able to query and display the operation log for debugging and root causing purposes. Support of the following minimum set of features is recommended:

a) record and audit of platform tasks;

b) data analysis and corresponding audit report based on recorded data;

c) log management system that supports a view of training log and resource usage log, log visualization and log comparison.

Bibliography

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

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