

International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

Y.3538

(09/2022)

SERIES Y: GLOBAL INFORMATION
INFRASTRUCTURE, INTERNET PROTOCOL ASPECTS,
NEXT-GENERATION NETWORKS, INTERNET OF
THINGS AND SMART CITIES

Cloud Computing

**Cloud computing – Global management
framework of distributed cloud**

Recommendation ITU-T Y.3538

ITU-T



ITU-T Y-SERIES RECOMMENDATIONS

GLOBAL INFORMATION INFRASTRUCTURE, INTERNET PROTOCOL ASPECTS, NEXT-GENERATION NETWORKS, INTERNET OF THINGS AND SMART CITIES

GLOBAL INFORMATION INFRASTRUCTURE	
General	Y.100–Y.199
Services, applications and middleware	Y.200–Y.299
Network aspects	Y.300–Y.399
Interfaces and protocols	Y.400–Y.499
Numbering, addressing and naming	Y.500–Y.599
Operation, administration and maintenance	Y.600–Y.699
Security	Y.700–Y.799
Performances	Y.800–Y.899
INTERNET PROTOCOL ASPECTS	
General	Y.1000–Y.1099
Services and applications	Y.1100–Y.1199
Architecture, access, network capabilities and resource management	Y.1200–Y.1299
Transport	Y.1300–Y.1399
Interworking	Y.1400–Y.1499
Quality of service and network performance	Y.1500–Y.1599
Signalling	Y.1600–Y.1699
Operation, administration and maintenance	Y.1700–Y.1799
Charging	Y.1800–Y.1899
IPTV over NGN	Y.1900–Y.1999
NEXT GENERATION NETWORKS	
Frameworks and functional architecture models	Y.2000–Y.2099
Quality of Service and performance	Y.2100–Y.2199
Service aspects: Service capabilities and service architecture	Y.2200–Y.2249
Service aspects: Interoperability of services and networks in NGN	Y.2250–Y.2299
Enhancements to NGN	Y.2300–Y.2399
Network management	Y.2400–Y.2499
Computing power networks	Y.2500–Y.2599
Packet-based Networks	Y.2600–Y.2699
Security	Y.2700–Y.2799
Generalized mobility	Y.2800–Y.2899
Carrier grade open environment	Y.2900–Y.2999
FUTURE NETWORKS	Y.3000–Y.3499
CLOUD COMPUTING	Y.3500–Y.3599
BIG DATA	Y.3600–Y.3799
QUANTUM KEY DISTRIBUTION NETWORKS	Y.3800–Y.3999
INTERNET OF THINGS AND SMART CITIES AND COMMUNITIES	
General	Y.4000–Y.4049
Definitions and terminologies	Y.4050–Y.4099
Requirements and use cases	Y.4100–Y.4249
Infrastructure, connectivity and networks	Y.4250–Y.4399
Frameworks, architectures and protocols	Y.4400–Y.4549
Services, applications, computation and data processing	Y.4550–Y.4699
Management, control and performance	Y.4700–Y.4799
Identification and security	Y.4800–Y.4899
Evaluation and assessment	Y.4900–Y.4999

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

Recommendation ITU-T Y.3538

Cloud computing – Global management framework of distributed cloud

Summary

Recommendation ITU-T Y.3538 introduces the framework and functional requirements for the global management of distributed cloud. The global management framework includes resource management, data management, platform service management, application service management, operation and maintenance management and risk management.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T Y.3538	2022-09-29	13	11.1002/1000/15061

Keywords

Cloud computing, distributed cloud, global management framework.

* To access the Recommendation, type the URL <http://handle.itu.int/> in the address field of your web browser, followed by the Recommendation's unique ID. For example, <http://handle.itu.int/11.1002/1000/115061>.

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 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/>.

© ITU 2022

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

	Page
1 Scope	1
2 References.....	1
3 Definitions	1
3.1 Terms defined elsewhere	1
3.2 Terms defined in this Recommendation.....	2
4 Abbreviations and acronyms	2
5 Conventions	3
6 Overview of global management of distributed cloud	3
6.1 Global management framework of distributed clouds	3
6.2 Operation model using global scheduling for distributed cloud	4
7 Functional requirements of global management for distributed cloud.....	5
7.1 Resource management.....	5
7.2 Data management	6
7.3 Platform service management	7
7.4 Application service management	7
7.5 Operation and maintenance management.....	8
7.6 Risk management	8
8 Security considerations.....	9
Appendix I – Use case of global management for distributed cloud	10
I.1 Use case of resource management.....	10
I.2 Use case of data management.....	10
I.3 Use case of platform service management	11
I.4 Use case of application service management	11
I.5 Use case of operation and maintenance management	12
I.6 Use case of risk management	13
I.7 Use case of auto-scaling for distributed cloud	13
I.8 Use case of resource allocation in global management for distributed cloud	14
I.9 Use case for an AI application in global management of distributed cloud...	15
I.10 Use case for migration in a distributed cloud.....	17

Recommendation ITU-T Y.3538

Cloud computing – Global management framework of distributed cloud

1 Scope

This Recommendation provides the framework and functional requirements for the global management of distributed cloud. The scope of this Recommendation includes:

- Overview of the global management of distributed cloud:
 - Global management framework of distributed cloud;
 - Operation model using global scheduling for distributed cloud;
- Functional requirements of the global management of distributed cloud.

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

[ITU-T Y.3500] Recommendation ITU-T Y.3500 (2014) | ISO/IEC 17788:2014, *Information technology – Cloud computing – Overview and vocabulary*.

[ITU-T Y.3508] Recommendation ITU-T Y.3508 (2019), *Cloud computing – Overview and high-level requirements of distributed cloud*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 cloud computing [ITU-T Y.3500]: Paradigm for enabling network access to a scalable and elastic pool of shareable physical or virtual resources with self-service provisioning and administration on-demand.

NOTE – Examples of resources include servers, operating systems, networks, software, applications and storage equipment.

3.1.2 cloud service customer [ITU-T Y.3500]: Party which is in a business relationship for the purpose of using cloud services.

NOTE – A business relationship does not necessarily imply financial agreements.

3.1.3 cloud service provider [ITU-T Y.3500]: Party which makes cloud services available.

3.1.4 core cloud [ITU-T Y.3508]: A part of distributed cloud, which is located in the core network and provides cloud computing capabilities.

NOTE – Cloud computing capabilities are described in [ITU-T Y.3500].

3.1.5 distributed cloud [ITU-T Y.3508]: A paradigm for enabling interworking among a pool of distributed resources used for a cloud service with low latency, limited bandwidth and real time processing performing in a massively distributed environment.

3.1.6 edge cloud [ITU-T Y.3508]: A part of the distributed cloud, which is located as close as possible to the cloud service customer (CSC) to support fast response time, latency sensitive and low computing density services.

3.1.7 regional cloud [ITU-T Y.3508]: A part of the distributed cloud, which provides cloud computing capabilities and is deployed for efficient configuration between the core cloud and the edge cloud optionally.

3.2 Terms defined in this Recommendation

None.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

AI	Artificial Intelligence
AR	Augmented Reality
API	Application Programming Interface
CC	Core Cloud
cc-NUMA	cache coherent Non-Uniform Memory Access
CSC	Cloud Service Customer
CSP	Cloud Service Provider
DNS	Domain Name System
EC	Edge Cloud
FaaS	Function as a Service
GPU	Graphics Processing Unit
GUI	Graphical User Interface
HDD	Hard Disk Driver
ID	Identification
I/O	Input/Output
IoT	Internet of Things
IP	Internet Protocol
NUMA	Non-Uniform Memory Access
PCIe	Peripheral Component Interconnect Express
QoS	Quality of Service
RC	Regional Cloud
SSD	Solid State Driver
VR	Virtual Reality

5 Conventions

In this Recommendation:

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

The keywords "**is recommended**" indicate a requirement which is recommended but which is not absolutely required. Thus, this requirement need not be present to claim conformance.

6 Overview of global management of distributed cloud

For the need of the Internet of things (IoT), augmented reality (AR) / virtual reality (VR), artificial intelligence (AI), etc., the service close to users having low latency and real time processing is built with massively distributed nodes to improve user experience. Cloud computing is extended across multiple nodes, as a new paradigm – distributed cloud defined in [ITU-T Y.3508]. Distributed cloud covers the whole set of cloud computing and consists of three types of cloud: core cloud (CC), regional cloud (RC) and the edge cloud (EC).

The core cloud generally consists of a large-scale resource and the edge cloud consists of a small-scale resource close to the user. Also, the regional cloud has a middle scale resource between the core cloud and the edge cloud.

The distributed cloud needs a management for core, regional and edge cloud locally. At the same time, cloud service providers (CSPs) need the global management to provide cloud services according to the requests of the cloud service customers (CSCs), where global management has an important role in the distributed cloud in which the core, regional, and edge cloud interact with each other. It is also important to define a common framework and functional requirements for global management.

6.1 Global management framework of distributed clouds

The global management framework of distributed cloud provides the management capabilities and includes the following aspects as shown in Figure 6-1.

- Resource management: management capabilities for resources of distributed cloud including virtual machine, container, application software, platform software, storage, network and other infrastructure resources.
- Data management: the management capability of data processing and transmission in a distributed cloud.

NOTE 1 – Data processing and transmission includes filtering data, encrypting the filtered data (capabilities located at EC), analysing data and storing data (capabilities located at the core cloud).

- Platform service management: management capabilities of platform services, interfaces and functions.
- Application service management: management capabilities of service for application (e.g., End-to-end product) to a CSC.
- Operation and maintenance management: management capabilities of the operation and maintenance of the distributed cloud, such as scheduling, deployment, allocation and monitoring of resources, etc.
- Risk management: management capabilities of the risk and faults.

NOTE 2 – Risk management includes but is not limited to assigning different roles and authorities.

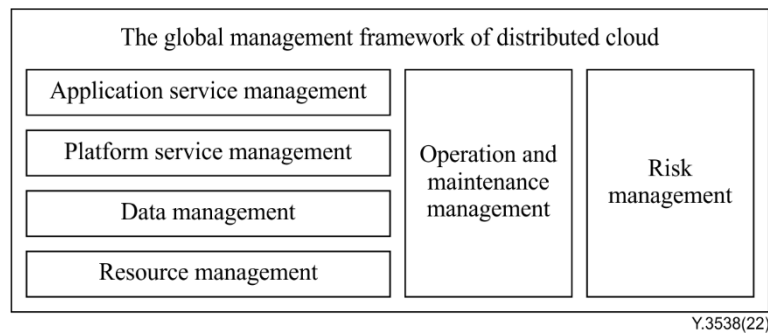


Figure 6-1 – Global management framework of distributed cloud

6.2 Operation model using global scheduling for distributed cloud

In a distributed cloud, the global scheduler performs the overall role of the distributed system management from managing CSC's requests to executing the actual jobs. Figure 6-2 illustrates the operation model using global scheduling for distributed cloud.

NOTE 1 – A distributed cloud includes all cloud deployment models such as public, private, hybrid cloud, etc.

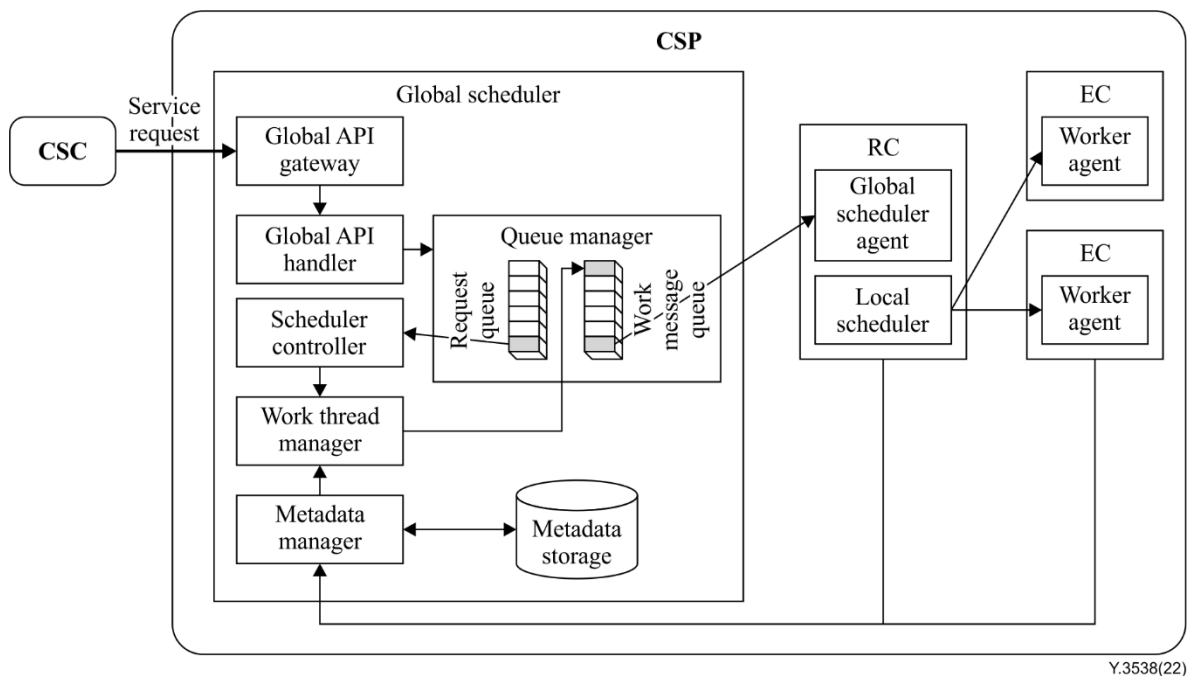


Figure 6-2 – Operation model using global scheduling for distributed cloud

The global scheduler controls the distributed cloud infrastructure according to the CSC's request data which is the CSC's requirements or policies for operation or maintenance, stores the requested data in the queue, processes the job based on the enqueued data, processes the job according to the CSC's request data and collects the results.

NOTE 2 – CSC's requirements or policies are applied to the scheduler for training the historical data provided by distributed cloud infrastructure for intelligent and automatic scheduling.

In Figure 6-2, the global application programming interface (API) gateway receives the CSC's request data and the requested data is processed through the global API handler. The queue manager receives and stores the request data for scheduling (such as service allocation request data and the execution data) from the global API handler. The scheduler controller gets back the data for scheduling from the queue and creates and runs job threads.

Also, the work thread manager converts message types for delivery to global scheduler agents in the distributed cloud and stores them back in the work message queue for scheduling across the core, regional and edge clouds and the metadata repository stores the metadata of the distributed cloud.

7 Functional requirements of global management for distributed cloud

7.1 Resource management

It is recommended that the CSP provides load balancing to deploy the application at the various types of platforms in the distributed cloud.

NOTE 1 – The types of platforms in the distributed clouds include virtual machine-based platforms, container-based platforms, function as a service (FaaS) platforms, etc.

NOTE 2 – A FaaS platform can deploy and run applications by allocating resources whenever an event occurs.

It is recommended that a CSP provides auto-scaling that expands and contracts the distributed cloud automatically with load balancing.

It is recommended that a CSP provides scaling policy setting and scaling group for auto-scaling and reflecting CSC's scaling policy in the distributed cloud.

NOTE 1 – Scaling policy includes scaling metrics and service information.

NOTE 2 – Service information in the scaling policy includes service ID for auto-scaling, minimum or maximum number of replications, support of the type of platform, etc.

NOTE 3 – Scaling metrics in a scaling policy determine how to execute and trigger auto-scaling and include a policy for a resource (such as min/max/average resource utilization), policy for the type of platform (such as the number of requests), policy for customization (such as user defined triggering value), etc.

It is recommended that the CSP monitors and stores auto-scaling information in the distributed cloud for auto scaling.

NOTE 1 – Auto-scaling information includes service specification, resource information, scaling specification, etc.

NOTE 2 – Service specification includes service identification (ID), CSC's internet protocol (IP), type of platform, required resources, auto-scaling enabled information, quality of service (QoS) such as latency and performance, etc.

NOTE 3 – Resource information includes utilization for hardware and requests number or customized metrics of software resources, etc.

NOTE 4 – The scaling specification is aligned with scaling metrics in the scaling policy.

It is recommended that CSP scales in and out with consideration of the QoS according to the scaling policy automatically.

NOTE – If the application is sensitive to latency, CSC sets the scaling policy with the CSC's QoS, and the platform is automatically scaled according to the QoS so that applications can run on the lowest latency platform.

It is recommended that the CSP provides the monitor of connection information between the distributed cloud.

NOTE 1 – The connection information between distributed cloud includes connection topology and device connection information (e.g., peripheral component interconnect express (PCIe), NVLink, etc.) between distributed cloud, and their connection state is monitored.

NOTE 2 – The connection topology (e.g., non-uniform memory access (NUMA), cache coherent non-uniform memory access (cc-NUMA), etc.) describes how devices or nodes are connected.

It is recommended that CSP provides the resource allocation according to the resource affinity.

NOTE 1 – The resource affinity is the distance between resources which is calculated through hop, based on the connection information between the corresponding devices and nodes.

NOTE 2 – A hop is a path located between a source and a destination in distributed cloud.

NOTE 3 – The resources are allocated according to the best resources' affinity (minimum distance between resources).

It is recommended that the CSP provides the resource allocation according to the input/output (I/O) congestion.

NOTE 1 – The I/O congestion is calculated as the average value of how much is used against the maximum bandwidth by measuring the bandwidth of I/O devices (network and storage, etc.) during a specific time slot.

NOTE 2 – The resources are allocated according to the lowest I/O congestion.

It is recommended that the CSP provides the management of the connection topologies in the distributed cloud.

NOTE – The connection topology management informs the available resources according to the resource affinity and I/O congestion to local and global schedulers based on the CSC's request.

It is recommended that CSP provides the interface of the CSC's resource allocation request.

NOTE 1 – Interface includes CLI, graphical user interface (GUI) and specific file format.

NOTE 2 – Resources to be requested by the CSC include the number of CPUs, the amount of memory and storage, I/O hardware device, network device, accelerator information (Graphics processing unit (GPU), TPU and other accelerator resources), etc.

NOTE 3 – The resources from the CSC's requests are reflected by the global scheduler.

It is recommended that CSP provides the ability to monitor the global resource.

NOTE 1 – The global resource management monitors the already allocated resource, the available resources, connection information and I/O congestion of resources periodically in distributed cloud.

NOTE 2 – The global scheduler requests and receives the information on the available resources, I/O congestion, and resource affinity from monitoring the global resource.

It is recommended that the CSP requests the resource allocation from the global scheduler to the local scheduler.

NOTE – The global scheduler selects the resources (nodes or devices) with information from the global resource management according to the CSC's request and sends the request to the local scheduler to deploy the application.

It is recommended that the CSP updates the resource information after the resource allocation.

NOTE – The global scheduler updates the available resource information through the global resource management when the service deployment of the local scheduler is successful.

It is recommended that the CSP provides the network gateway to access the distributed cloud for migration.

NOTE – Network gateway is the access method between distributed cloud which is connected by a dedicated and direct network using tunnelling such as IPsec tunnelling or a route bypass through a network proxy.

7.2 Data management

It is recommended that the CSP provides the interface of storage provisioning.

NOTE – The interface of storage provisioning includes APIs or device drivers. This interface is provided in form of a repository (such as memory, NVMe, solid state driver (SSD), hard disk driver (HDD), data storage federation architecture, etc.) using a container or a virtual machine.

It is recommended that the CSP provides the deployment for data storage.

NOTE – By integrating with the data storage (such as a data storage federation), it is possible to provide both high performance of main memory and large capacity of disk storage for distributed cloud.

It is recommended that CSP provides the minimization of data transmission delay and processing time.

NOTE – The minimization of data transmission can be realized by collecting and processing data from a location close to CSC.

It is recommended that the CSP provides high scalability and error-resistance for data storage.

NOTE – The error resistance includes the iteration process in which jobs restart due to an error. If the number of iterations is exceeded by a certain number of times, it is considered as a job failure.

It is recommended that the CSP provides the distribution of applications according to the system load and network delay to maximize the computing resources between ECs.

NOTE – The examples of distributed applications for data processing to deploy in EC are as follows. Each processing is deployed as a container or a virtual machine so that it could operate independently in a different node.

- A pre-processing: filtering the input data, adjusting the input size of data, or performing data processing for better performance.
- A main processing: treating the main job or task of processing such as big data analyses, ML inference and training, VR/AR engine, etc.
- A post-processing: visualizing results from the main processing or performing additional data processing.

It is recommended that the CSP provides a high-speed network for sharing storage.

NOTE 1 - The shared storage communicates the job information and status using the database for distributed applications.

NOTE 2 – The sharing job information includes the registered number in a database, user identification (ID), data path where the data was uploaded to the repository, data type, data status, job execution time, etc.

NOTE 3 – The job status includes enqueued processing, pre-processing execution, pre-processing completion, main processing execution, main processing completion, post-processing execution, post-processing completion, exit, etc.

7.3 Platform service management

It is recommended that the CSP provides the interface for service management.

NOTE – The interface for service management includes APIs with an API server.

It is recommended that the CSP provides the service discovery.

It is recommended that the CSP provides a load balancer so that the service is distributed according to CSC's policy.

It is recommended that the CSP provides service templates for the CSCs to design services.

NOTE – Service template is the technical description including service name, service type, resource information for service, CSC's request, etc.

It is recommended that the CSP provides the configuration for service deployment.

It is recommended that the CSP provides the service monitoring during service operation.

7.4 Application service management

It is recommended that the CSP provides the development of the CSC's own functions for a CSC's application.

It is recommended that the CSP provides domain name system (DNS) service so that CSC connects with the application.

It is recommended that the CSP provides application templates for CSCs to design applications.

It is recommended that the CSP provides a connection for an application service, including network IP address, port, etc.

It is recommended that the CSP provides the CSC with the interface for application monitoring.

7.5 Operation and maintenance management

It is recommended that the CSP provides global scheduling for the applications to be executed.

It is recommended that the CSP provides the API management to handle API.

NOTE – For example, CSC requests the API for application service and the global API handler manages and stores the request message in a queue for fast response.

It is recommended that the CSP provides the management of policies for global scheduling.

NOTE – Policy for the global scheduling includes the work priority, affinity, resource usage, etc.

It is recommended that the CSP provides and manages the metadata repository storing cloud information from the regional cloud and edge cloud.

NOTE – The cloud information includes the type of cloud service, usage, etc. and this information is collected by monitoring the regional cloud and edge cloud.

It is recommended that the CSP provides the queue management for work priority.

NOTE – The queue for the global scheduling includes the CSC's request message queue, work message queue, etc.

It is recommended that the CSP provides the connection management for global scheduling.

NOTE 1 – Connection management for global scheduling manages the connection protocol or agent to utilize the distributed resources of the regional cloud or edge cloud.

NOTE 2 – Agents in the regional cloud or edge cloud receive messages from the work message queue in the global scheduler.

7.6 Risk management

It is recommended that the CSP provides the replica management to create and manage replicas as a replacement in case the system breaks down.

It is recommended that the CSP provides backup copies to keep data in distributed cloud.

NOTE – Global management provides storage tiering, archiving, and distributed share management functions to manage backup copies of data.

It is recommended that the CSP provides the network connections to the detour path in case of network failure.

It is recommended that the CSP provides the restriction of network access for unauthenticated devices or users.

It is recommended that the CSP provides encryption / decryption for data protection to transfer data between clouds.

It is recommended that the CSP provides the migration in case of system failure.

NOTE 1– Migration includes a control process with checking the status in EC, checking the network of destination nodes, saving the current status of a snapshot image and restoring the image.

NOTE 2 - The data management for migration utilizes the mechanism of shared storage or cache to reduce snapshot (such as checkpoint) transfer between nodes.

NOTE 3 – The data management for migration includes the data management using micro-services and virtual machines and the deployment for data storage.

It is recommended that the CSP provides migrating the CSC's application and data to the available resources in case of load balancing, system failure, and the CSC's policy.

8 Security considerations

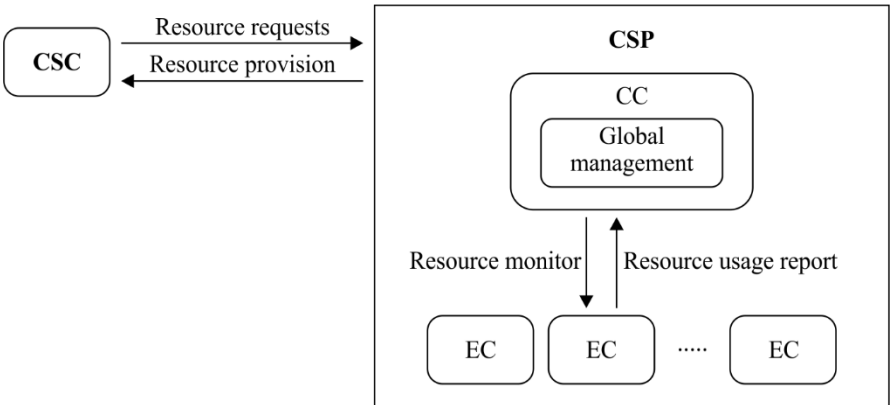
Security aspects for consideration within the cloud computing environment, including inter-cloud computing, are described in [ITU-T X.1601], which analyses the security threats and challenges, and describes the security capabilities that could mitigate these threats and meet the security challenges.

Appendix I

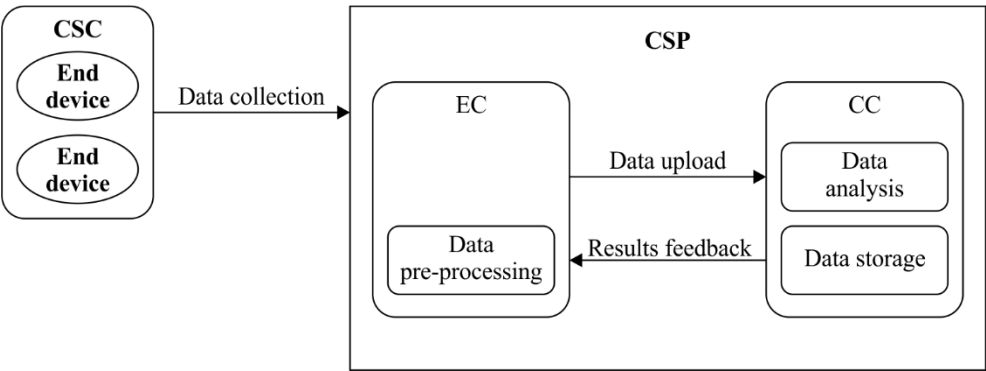
Use case of global management for distributed cloud

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

I.1 Use case of resource management

Title	Resource management
Description	According to the resource requirements of the CSC, global management on the core cloud (CC) could flexibly schedule the resources of the EC nodes and could monitor the usage of the resource. The EC nodes feed the resource usage status to the global management.
Roles	CSC, CSP
Figure (optional)	 <p>The diagram illustrates the resource management process. On the left, a box labeled 'CSC' sends 'Resource requests' to a larger box labeled 'CSP'. The 'CSP' box contains a sub-box 'CC' which includes 'Global management'. Below the 'CC' box, there are several 'EC' nodes. A 'Resource monitor' arrow points from the EC nodes up to the 'Global management' box, and a 'Resource usage report' arrow points from the 'Global management' box down to the EC nodes. The 'CSP' box also sends 'Resource provision' back to the 'CSC' box. The reference 'Y.3538(22)' is located at the bottom right of the diagram area.</p>
Pre-conditions (optional)	
Post-conditions (optional)	
Derived requirements	Resource management (see clause 7.1)

I.2 Use case of data management

Title	Data management
Description	Under global management, the EC nodes collect data from the edge device of the CSC. The EC nodes pre-process the data and upload it to the CC node. The CC node performs the data analysis and stores the large-scale data. The results of the data analysis, such as models, are fed back to the EC nodes by the CC node.
Roles	CSC, CSP
Figure (optional)	 <p>The diagram illustrates the data management process. On the left, a box labeled 'CSC' contains two 'End device' boxes. An arrow labeled 'Data collection' points from the 'End device' boxes to a larger box labeled 'CSP'. The 'CSP' box contains an 'EC' box with 'Data pre-processing' and a 'CC' box with 'Data analysis' and 'Data storage'. An arrow labeled 'Data upload' points from the 'EC' box to the 'CC' box, and an arrow labeled 'Results feedback' points from the 'CC' box back to the 'EC' box. The reference 'Y.3538(22)' is located at the bottom right of the diagram area.</p>

Title	Data management
Pre-conditions (optional)	
Post-conditions (optional)	
Derived requirements	Data management (see clause 7.2)

I.3 Use case of platform service management

Title	Platform service management
Description	CSP provides an interface for platform service management. Interfaces for platform service management include the APIs and API servers. CSP provides a load balancer to deploy the CSC's application to a distributed cloud. It also provides platform service discovery and policy for deployment. Platform service management can help the CSC by using various service templates.
Roles	CSC, CSP
Figure (optional)	<p style="text-align: right; font-size: small;">Y.3538(22)</p>
Pre-conditions (optional)	
Post-conditions (optional)	
Derived requirements	Platform service management (see clause 7.3)

I.4 Use case of application service management

Title	Application service management
Description	<p>According to the request of the CSC, the global management on CC manages the applications deployed in the EC, including deploying application templates to EC, monitoring the application running status, etc.</p> <p>There are two situations for deploying applications on EC nodes:</p> <ol style="list-style-type: none"> 1. An application is only deployed on one EC. For example, App1 and App2 are deployed in different ECs respectively. 2. An application is deployed in multiple ECs. For example, App3 is deployed in two different ECs.
Roles	CSC, CSP

Title	Application service management
Figure (optional)	
Pre-conditions (optional)	
Post-conditions (optional)	
Derived requirements	Application service management (see clause 7.4)

I.5 Use case of operation and maintenance management

Title	Operation and maintenance management
Description	According to the requests of the CSC, the CSP monitors the operation of the EC resources through global management, schedules and allocates the EC resources, and finally feeds back the results to the CSC.
Roles	CSP
Figure (optional)	
Pre-conditions (optional)	
Post-conditions (optional)	
Derived requirements	Operation and maintenance management (see clause 7.5)

I.6 Use case of risk management

Title	Risk management
Description	CSP manages the risks of all the ECs through global management, including assigning different roles and permissions, authentication, etc.
Roles	CSP, CSC
Figure (optional)	
Pre-conditions (optional)	
Post-conditions (optional)	
Derived requirements	Risk management (see clause 7.6)

I.7 Use case of auto-scaling for distributed cloud

Title	Auto-scaling for distributed cloud
Description	<p>This use case is an auto-scaling technology for providing flexible services in a distributed cloud as shown in the following figure. This auto-scaling technology is applied to a distributed cloud to scale in and out services. CSC sets the scaling policy, which is information such as the service to be scaled, the minimum / maximum number of scaling instances, resource utilization, and request limits by a supported platform type.</p> <p>Global management monitors the information for auto-scaling including service specifications (service name, execution time, required resources, service QoS including latency), system resources (CPU, memory, network, etc.), and system expansion specifications (expansion group, number of services), and saves it to the auto-scaling repository.</p> <p>When the load increases due to a large number of service requests from the CSC, the resource is automatically adjusted according to the auto-scaling policy and the information for auto-scaling. Auto-scaling can effectively respond to a service load and reduce costs.</p> <p>In a distributed cloud, the characteristics of platforms such as virtual machine-based platforms, container-based platforms and PaaS (including FaaS or serverless) provided by the CSP are very diverse, and the CSC can use various applications on these various platforms.</p> <p>At this time, when the CSC request increases and the load increases, an additional application is created on each platform for auto-scaling. If the CSC needs auto-</p>

Title	Auto-scaling for distributed cloud
	scaling for high-quality applications (low latency and high performance) on a variety of platforms, applications can be added to platforms that support high-quality services.
Roles	CSC, CSP
Figure (optional)	<p>The diagram illustrates the auto-scaling architecture. On the left, 'N-service requests' from a 'CSC' (Cloud Service Client) are processed through a 'Scaling-policy' to an 'Auto-scaling repository' within 'Global management'. This repository is linked to an 'Auto-scaling information monitor'. The 'Auto-scaler' (containing 'Auto-scaling metric comparison' and 'Application selector') receives input from the repository and sends 'Scaling in/out' commands to a 'Load balancer'. The 'Load balancer' then directs traffic to a 'Scaling group' on the right. This group consists of three 'EC' (Elastic Cloud) instances: 'Scale-out' (VMs), 'Container' (App-A), and 'Scale-in' (PaaS Functions). Each EC instance includes 'MEM/CPU/NET' resources. A feedback loop labeled 'Auto-scaling information' returns from the Scaling group to the monitor and repository. A reference 'Y.3538(22)' is noted at the bottom right.</p>
Pre-conditions (optional)	
Post-conditions (optional)	
Derived requirements	Resource management (see clause 7.1)

I.8 Use case of resource allocation in global management for distributed cloud

Title	The resource allocation in global management for distributed cloud
Description	<p>In general, resource allocation in a distributed cloud environment utilizes the nearest resource in order to minimize service delay. This is a key feature of the distributed cloud. In order to solve the service delay, it is important to optimally allocate resources (CPU, memory, GPU, TPU, network, storage, other accelerated HW, etc.) of the distributed cloud.</p> <p>In order to provide a distributed cloud service with low latency, most computing servers use a resource connection topology (e.g., NUMA cc-NUMA, etc.) and device connection (e.g., PCIe, NVLink, etc.), and their connection state is monitored and resources are allocated according to the distance between the resources (resource affinity) and the input/output congestion.</p> <p>Resource affinity is calculated through hops based on the connection information between corresponding devices. The I/O congestion is calculated as the average value by monitoring the bandwidth of the I/O devices for a specific time.</p> <p>These values are based on the CSC's resource requirements when requesting a service. The CSC requests resources (CPU, memory, network, storage, and accelerator device) to deploy an application on a container and virtual machine. Based on this requested resource, the distance between the resources described above and the input/output congestion is measured.</p> <p>The global resource manager manages the allocated resource information, available resource information, and input/output congestion information of each cloud. The global scheduler receives and processes the distributed cloud service execution requests.</p>

Title	The resource allocation in global management for distributed cloud
	<p>The global scheduler updates the available resource information of the node through the global resource manager when the service execution of the local scheduler is successful. The node's I/O congestion is periodically transmitted to the global resource manager by monitoring the previously described I/O congestion (network congestion, storage congestion, etc.).</p> <p>The topology manager manages the allocated resources and the available resources based on the device topology information.</p> <p>The topology manager, who receives the resource requirements from the local scheduler, selects and returns the available resources based on the congestion, distance, and resource connection. The local scheduler receives the execution request of the distributed cloud service from the global scheduler.</p> <p>The local scheduler receives the allocable resource information from the topology manager, configures the virtual environment setting information, makes a request to the virtual machine-based platform or container-based platform and transmits the result to the global scheduler. The virtual machine-based platform or the container-based platform configures the virtual environment and runs the service based on the virtual environment information received from the local scheduler.</p>
Roles	
Figure (optional)	<p>Container: name: example-con image: example:v1.0 resources: cpu: 2 memory: 20G nvidia.com/gpu: 2 inputDev: NIC</p> <p style="text-align: right;">Y.3538(22)</p>
Pre-conditions (optional)	
Post-conditions (optional)	
Derived requirements	Resource management (see clause 7.1)

I.9 Use case for an AI application in global management of distributed cloud

Title	AI application in global management for distributed cloud
Description	<p>This use case introduces an AI inference service that improves the resource utilization through a micro-service structure so that the video data can be inferred quickly and flexibly in a distributed cloud. Through this, it is possible to minimize data transmission delay and execution time by collecting and processing video data from a</p>

Title	AI application in global management for distributed cloud
	<p>location close to the CSC. In addition, it is expected that a highly scalable and error-resistant inference service can be provided to the CSCs through a deep learning model which is specialized for flexible and rapid resource usage between edge clouds. By integrating with the deep learning model, it is possible to provide both high performance of the main memory and large capacity of the disk storage. Therefore, data required for inference can be quickly shared for high performance.</p> <p>In the figure, deep learning models for image inference can be divided into three main parts:</p> <ul style="list-style-type: none"> – A pre-processing process that adjusts the input size of the image or performs an image processing for better inference performance; – An inference process that extracts and recognizes object features through a deep learning network; – A post-processing process in which the results from the inference process are displayed on the image. <p>Among them, the pre-processing process and the post-processing process are processes that mainly utilize the CPU among computing resources, and the inference uses a large amount of computation including GPU to maximize the usage of the flexible computing resources in the edge cloud.</p> <p>For the existing deep learning model each process is deployed in a container so that it operates independently.</p> <p>In addition, when creating an edge cloud-based environment using a container-based orchestration platform, one thing to consider is that communication between each service must occur smoothly without conflicts.</p> <p>To this end, as the video data is uploaded the job information is stored in the shared storage and service containers communicate with the database and share the job status.</p>
Role/Sub-role	CSC, CSP
Figure (optional)	<p style="text-align: right; font-size: small;">Y.3538(22)</p>
Pre-conditions (optional)	
Post-conditions (optional)	
Derived requirements	Data management (see clause 7.2)

I.10 Use case for migration in a distributed cloud

Title	Migration in a distributed cloud
Description	<p>The migration service in the distributed cloud is provided in order to re-deploy services in real time and maintain continuous services. Migration is performed between edge clouds, or between edge clouds and other clouds in consideration of resource scarcity, failure, and cost-effectiveness.</p> <p>To perform migration, it is necessary to manage the migration in the distributed cloud for service transfer between clouds.</p> <p>To do this, it is necessary to have a reliable and efficient way to capture the state of the application in real time in a transferable format and to save it in the form of a snapshot.</p> <p>Basically, in a distributed cloud global management, the scheduler assigns the applications to the nodes, and the controller communicates with the nodes through gateways to create, delete or update applications. This gateway is responsible for the control function that accesses the network between clouds.</p> <p>In each node, only the control tasks that maintains the applications and creates containers or virtual machines are performed by individual agents residing in the target node itself. Migration is provided through migration controllers from the migration brokerage function in global management.</p> <p>If a CSC has a snapshot (checkpoint image) that maintains the state of an application, CSC also needs a mechanism to transfer this data between clouds. Therefore, to reduce data transfer between nodes and containers, it utilizes the mechanism of shared high-speed storage and utilizes high-speed network connectivity.</p>
Role/Sub-role	<i>CSC, CSP</i>
Figure (optional)	<p>The diagram illustrates the migration architecture. At the top, a 'Global management' block contains a 'Global scheduler' and a 'Global resource manager', which are connected to a 'Migration brokerage'. To the left, a 'CSC End device' is connected to the 'Global management' block. Below, two 'EC' (Edge Cloud) blocks are shown, connected to a central 'Public NET'. Each EC contains an 'Application', a 'Migration controller', and a 'Gateway'. The left EC's 'Application' is crossed out with a red 'X', and its 'Snapshot' is highlighted in grey. The right EC's 'Application' is active, and its 'Snapshot' is also highlighted in grey. A large curved arrow indicates the migration of the application state from the left EC to the right EC via the 'Public NET' and 'Migration brokerage'. The 'Migration controller' in each EC manages the application and its snapshot. The 'Gateway' in each EC connects to the 'Public NET'.</p> <p style="text-align: right;">Y.3538(22)</p>
Pre-conditions (optional)	
Post-conditions (optional)	
Derived requirements	Risk management (see clause 7.6)

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