ICS 35.020
CCS L70

Social Organization Standard

T/GBA 019-2024

General technical requirements of computing network for the Greater Bay Area

大湾区算力网络总体技术要求

(English Translation)

Issue date:2024-07-25 Implementation date:2024-08-01

Issued by Guangdong-Hong Kong-Macao Greater
Bay Area Standards Innovation Alliance

Contents

Foreword

This document is drafted in accordance with the rules given in the GB/T 1.1-2020 *Directives for Standardization -- Part 1: Rules for the Structure and Drafting of Standardizing Documents.*

This document is proposed by Peng Cheng Laboratory.

This document was prepared by the Guangdong-Hong Kong-Macao Greater Bay Area Standards Innovation Alliance.

This document is authorized for use by organizational partners and all members of the Guangdong-Hong Kong-Macao Greater Bay Area Standards Innovation Alliance, who are required to adopt the same transformation into their own group standards, and publicize the standard basic information on the National Group Standards Information Platform.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. The issuing body of this document shall not be held responsible for identifying any or all such patent rights.

This is the first release.

General technical requirements of computing network for the Greater Bay Area

1 Scope

This document specifies the overall architecture, functional requirements and interface requirements of the computing network for the Greater Bay Area.

This document is applicable to the overall design and construction of the computing network for the Greater Bay Area.

2 Normative references

The contents in the following documents constitute necessary clauses of this document by normative references. Meanwhile, for dated references, only the version of the corresponding date applies to this document; for undated references, the latest version (including all the revisions) applies to this document.

GB/T 41867-2022 *Information technology—Artificial intelligence—Terminology*
YD/T 4255-2023 *General technical requirements of Computing and Network Convergence*

3 Terms and definitions

The terms and definitions defined in GB/T 41867-2022 and the following apply to this document.

3.1
data center
a structure or group of structures, dedicated to the centralized accommodation, interconnection and operation of information technology and network telecommunications equipment providing data storage, processing and transport services together with all the facilities and infrastructures for power distribution and environmental control together with the necessary levels of resilience and security required to provide the desired service availability.

Note 1: A structure can consist of multiple buildings and/or spaces with specific functions to support the primary function.

Note 2: The boundaries of the structure or space considered the data center, which includes the information and communication technology equipment and supporting environmental controls, can be defined within a larger structure or building.

[Source: ISO/IEC 22237-1:2021, 3.1.8]

3.2
computing center
facility designed to provide computing services to a variety of users through the operation of computers and auxiliary hardware and through services provided by the facility’s staff.

[Source: ISO/IEC/IEEE 24765:2017, 3.741]

artificial intelligence computing center
a structure or group of structures, dedicated to provide artificial intelligence computing services and data accommodation for a variety of users. It can provide data storage, processing, transport services and artificial intelligence computing acceleration together with all the facilities and infrastructures for power distribution and environmental control together with the necessary levels of resilience and security required to provide the desired service availability.

Note 1: A structure can consist of multiple buildings and/or spaces with specific artificial intelligence functions to support the primary functions of the artificial intelligence computing center.

Note 2: The servers in the artificial intelligence computing center can consist of artificial intelligence servers, general servers, etc. The server can be called "node".

[Source: ISO/IEC 22237-1:2021, 3.1.8 and ISO/IEC/IEEE 24765:2017 3.741, modified]

3.3
computing network
a facility that provides computing resources to users by connecting the computing centers across various places through network technology such that the computing tasks can be allocated and scheduled across the computing centers.
[Source: ITU-T Y.2501, modified]

3.4
computing awareness
a capability of the network to perceive the computing resources and computing services from multiple dimensions such as deployment location, real-time status, load information, and service requirements.
[Source: YD/T 4255-2023]

3.5
computing & network orchestration and management
the unified management and scheduling of computing resources and network resources, including registration, OAM, etc.
[Source: YD/T 4255-2023]

4 Abbreviation

The following abbreviations apply to this document.
AI (Artificial Intelligence)
CPU (Central Processing Unit)
DetNet (Deterministic Networking)
DNS RR (Domain Name System Resource Record)
FCFS (First Come First Served)
IP (Internet Protocol)
I/O (Input/Output)
OTDR (Optical Time Domain Reflectometer)
OTN (Optical Transport Network)
OXC (Optical Cross Connect)
OMSP (Optical Multiplex Section Protect)
OSU (Optical Service Unit)

QKD (Quantum Key Distribution)
QoE (Quality of Experience)
QoS (Quality of Service)
QUIC (Quick UDP Internet Connection)
RDMA (Remote Direct Memory Access)
ROADM (Reconfigurable Optical Add-Drop Multiplexer)
SDWAN (Software Defined Wide-Area Network)
SRV6 (Segment Routing over IPv6)
TCP (Transmission Control Protocol)
UDP (User Datagram Protocol)
WDM (Wavelength Division Multiplexing)
WSON (Wavelength Switched Optical Network)

5 Overall architecture

5.1 General Framework

The computing network for the Greater Bay Area connects AI computing centers, supercomputing centers, and cloud computing centers scattered throughout the Greater Bay Area to gather and share computing resources, data, models, and applications. Various types of computing centers can share resources by joining the computing network for the Greater Bay Area. Through the unified scheduling of computing resources, the resource utilization rate of the whole network can be improved and the huge computing demands of the Greater Bay Area can be met more efficiently.

The framework design and layering of the computing network for the Greater Bay Area are shown in Figure 1.

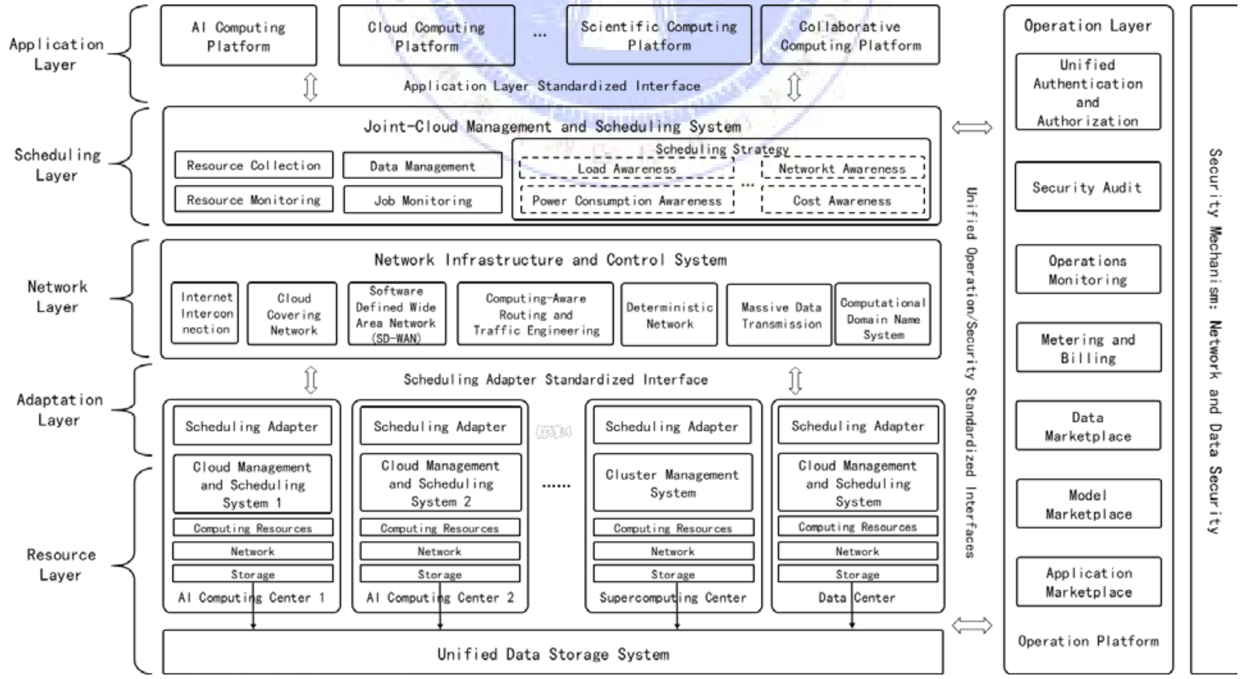

Figure 1. General framework of the computing network for the Greater Bay Area

5.2 Deployment Framework

The interconnection of the computing centers and hub nodes in the Greater Bay Area Computing Network is as follows:

a) Computing centers can be interconnected through the following three types of networks:

1) IP private network: It is mainly used for routing control plane signaling, low-throughput data interchange between computing centers, and transmission scenarios with specific requirements for security and network quality;

2) OTN high-speed network: It is mainly used for high-throughput data interchange between computing centers and transmission scenarios with specific requirements for security and network quality;

3) Internet: It is mainly used for low-throughput data interchange between computing centers and transmission scenarios with little requirements for security and network quality.

b) The backbone node is chosen as the large/important AI computing center, supercomputing centers, or cloud computing center in a certain region. The priority (from high to low) of the interconnection modes between the backbone nodes is: OTN high-speed network, IP private network, and Internet;

c) The priority (from high to low) of the interconnection modes between other computing centers is: IP private network, OTN high-speed network, Internet;

d) Computing network platform for the Greater Bay Area includes a joint-cloud management and scheduling platform and an operation platform, which can be deployed in a backbone node or in an independent computing center.

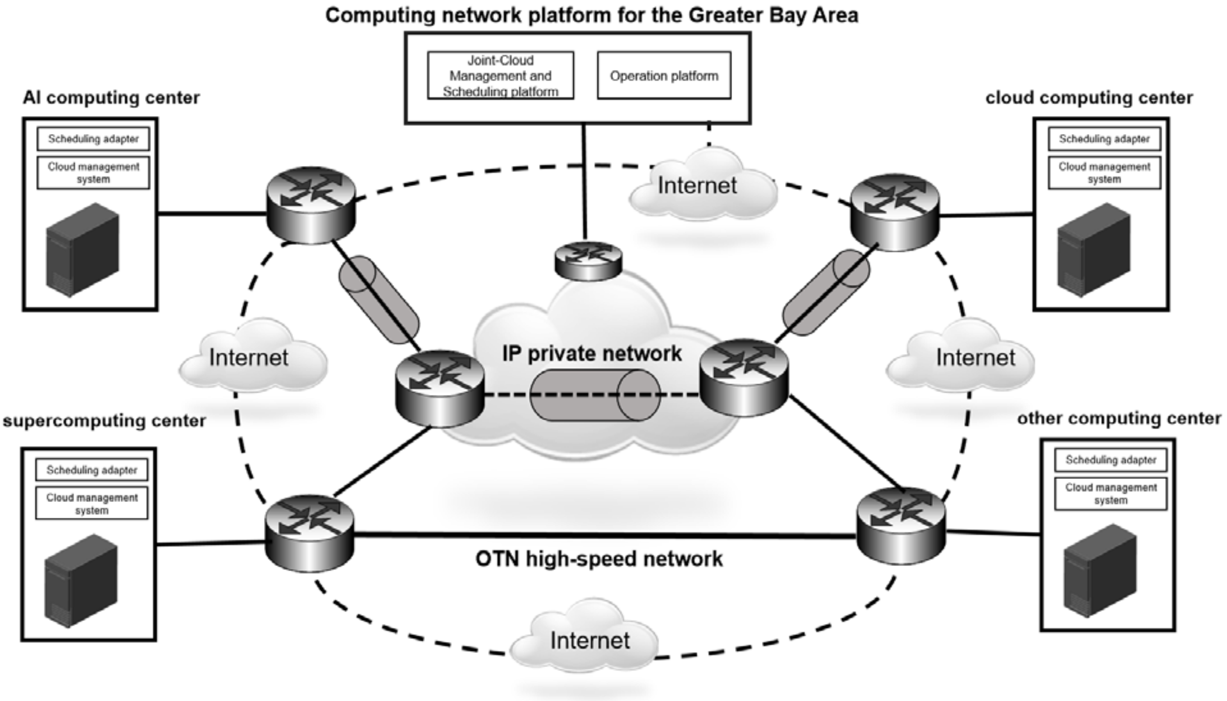

Figure 2. Deployment framework of computing network for the Greater Bay Area

6 Functional requirements

6.1 Overview

The general framework of the computing network for the Greater Bay Area can be divided into resource layer, adaptation layer, network layer, scheduling layer, application layer, operation layer, and security mechanism. The specific functional requirements of each layer are as follows.

6.2 Resource Layer

The resource layer of computing network for the Greater Bay Area contains the computing, storage, network, data, and other resources of each computing center, and should have the following capabilities:

a) Providing computing resources: Should provide computing, storage, and other resources and support on-demand resource expansion and contraction;
b) Interconnection and openness of computing centers: Should support the interconnection and opening-up of computing and data resources of each computing center;
c) Unified data storage: Should support constructing a unified converged storage service of block, file, and object storages based on the storage resources of the computing centers to support unified data management and migration across computing centers;
d) User authentication: Should support user creation, deletion, modification, authentication, and other functions to ensure that authorized users of the computing network can log in to the computing centers to use various resources;
e) Job management: Should support submission, cancellation, inquiry, and modification of jobs in each computing center;
f) Resource management in the computing center: Should support real-time monitoring of computing center resources (such as CPU, GPU, memory, etc.), allocation of resources for jobs, and resource isolation to ensure the security of sensitive data and isolation between different jobs;
g) Account management: Should support account creation, deletion, and modification via the management and scheduling system of the computing center;
h) Scheduling strategies within the computing center: the management and scheduling system of the computing center should support various scheduling strategies to meet the complex computing requirements of the computing network. Should support selecting or customizing the scheduling strategy as required;
i) Stability: Each computing center should have continuous and stable operation of the user authentication system and the management and scheduling system for providing stable and reliable computing resources for the computing network.

6.3 Adaptation Layer

The adaptation layer of the computing network for the Greater Bay Area realizes interaction between the scheduling layer and the resource layer through a scheduling adapter. The scheduling adapter is deployed in each computing center and interconnects with the heterogeneous management and scheduling systems in each computing center. It collects and reports information of heterogeneous cluster resources as well as realizing the forwarding and management of jobs. The scheduling adapter can collect cluster resources, jobs, energy consumption, and fees from each computing center and report the information to the scheduling layer for scheduling jobs. The adaptation layer should have the following capabilities:

a) Should support the unified adaptation of the interfaces of AI computing center, supercomputing center, and data center for shielding the differences in technology stacks of the heterogeneous clusters;

b) Should support interface expansion to meet the requirements for other types of computing centers to access the computing network for the Greater Bay Area;

c) Job agent: Should support sending the job delivered by the scheduling layer to the local scheduling system of the computing center, periodically collecting the resource usage information of the job status, and reporting it to the scheduling layer;

d) Computing resource agent: Should support periodically collecting resource information of the computing cluster to which it belongs and reporting it to the scheduling layer;

e) Data management and caching: Should support the data management between the adaptation layer and the scheduling layer, including data upload, download, breakpoint continuation, etc. Should support data delivery during job scheduling and data retrieval after job execution. The data should be cached to avoid the re-transmission when the data is re-used;

f) Account mapping: Should support the mapping between the unified user account of the computing network and the user account of the local management and scheduling system of each computing center.

6.4 Network Layer

6.4.1 Overview

The network layer is mainly responsible for the network access and interconnection of multiple heterogeneous computing centers, achieving routing control and high-speed forwarding of data for computing. The network layer mainly includes the optical layer and the IP layer.

6.4.2 Optical Layer

The computing network for the Greater Bay Area should support an all-optical network as the base, providing low latency and high-throughput transmission channels for data exchange between important hub nodes. The optical layer specifically includes the following capabilities:

a) Should support long-distance and high-capacity transmission capabilities of WDM/OTN, such as 100G/400G/800G, to achieve interconnection of computing resources and high-speed transmission of massive data;

b) Should support all-optical switching and scheduling capabilities such as ROADM/OXC, including 9 dimensions, 20 dimensions, and 32 dimensions;

c) Should support full optical computing power access capabilities through various methods such as OTN/OSU/WDM;

d) Should support multiple protection and collaboration mechanisms based on WSON/OMSP and other methods to improve the survivability and reliability of the network layer;

e) It is advisable to support cross layer collaboration between optical and IP layers, including collaboration in multiple aspects such as business, path, operation and maintenance, and protection.

6.4.3 IP Layer

The computing network for the Greater Bay Area needs to build an IP private network that supports comprehensive coverage of major computing backbone nodes in the Greater Bay Area, supports various networking methods such as squared-shape and crossover, and supports IPv6+, soft and hard slicing, SDWAN and other technologies. The IP layer specifically includes the following capabilities:

a) User access: Should support individual users or government and enterprise users to access computing network resources through VPN dedicated lines and other methods;

b) Computing awareness: Should support selecting the optimal path by monitoring the usage of computing resources (such as CPU, memory, storage, etc.) in the network and performing routing calculations based on this information; It should support the awareness and understanding of current computing and network resources, and be able to identify system bottlenecks and hotspots by real-time monitoring and analyzing performance indicators of computing resources (such as CPU utilization, memory usage, disk I/O, etc.) and network resources (such as bandwidth, latency, jitter, etc.), and dynamically adjust resource allocation according to user needs and priorities to maximize the performance and efficiency of computing resources;

c) Application awareness: It is advisable to support deep analysis and identification of network traffic, achieve awareness and optimization of network applications, and carry out targeted network optimization and adjustment, including adjustment of traffic paths and adjustment of priority, etc; It is advisable to support the use of programmable space in network protocol packets (such as SRv6 packets) to carry application information (identification and/or network performance requirements) into the network, enabling the network to aware applications and their needs, providing them with fine network services and precise network operation and maintenance; It supports differentiated service capabilities for computing, guiding traffic into corresponding SRv6 Policy tunnels, network slicing, DetNet (Deterministic Network) path, service function chain path, etc., to achieve application diversion and flexible routing;

d) Service awareness: It is advisable to support the representation of specific services through service identification, and can be used for application of task invocation, end-to-end connection, network service routing, addressing and scheduling of computing; It is advisable to support the awareness of SLA requirements at the service level granularity through the awareness of service identification; Support the association of corresponding forwarding table entries through service identification indexes to achieve fine-grained awareness of services and traffic forwarding;

e) Software defined wide-area network: Should support automated management and control of wide area networks, and achieve network programmability; Should support flexible deployment and management of network connections between multiple branch offices; Should support the ability for rapid business deployment; Should support technologies such as dynamic path selection and failover to improve the reliability of the network;

f) Network quality assurance: Should support the demand for computing resources as the guide, to slice and carry different business through software/hard slicing capabilities to achieve the separation of latency and bandwidth; Should support end-to-end rapid activation of business based on SRV6 technology; Should support various data transmission protocols, including UDP TCP, QUIC, etc. to ensure the quality and reliability of data transmission;

g) Massive data transmission: It is advisable to support data compression and sharding technology to reduce the size of data and reduce the occupation of network bandwidth; It is advisable to support data express delivery services, fully utilize the redundant bandwidth of existing networks, and achieve efficient on-demand elastic data cloud services through controllers.

6.5 Scheduling Layer

Scheduling layer of the computing network for the Greater Bay Area mainly implements resource management and global job scheduling among computing centers. The scheduling layer should have the following capabilities:

a) Cross-center computing resource management and monitoring: Should support real-time information collection and monitoring of computing, storage, and network resources in each computing center, and support the unified measurement and management of various computing resources;

b) Computing and network orchestration: Should support the orchestration of computing, network, and other resources on demand;

c) Data resource management: Should support the monitoring and management of data resources on the computing network, and support data upload, download, breakpoint transmission, and cache;

d) Job management: Should support job submission, job list viewing, and job management operations;

e) Computing resource management: Should support the management of computing, network, storage, and other resources of the whole network, including resource and job information reported by the scheduling adapter;

f) Cross-center scheduling policy: Should support selecting appropriate computing centers to run computing jobs according to job requirements and the available resources of the computing network. At least one of the following scheduling strategies should be supported:

   1) Manual scheduling strategy: Manually specify the computing center;
   2) FCFS scheduling strategy: First-come-first-served scheduling strategy;
   3) Load-aware scheduling strategy: Scheduling jobs based on the load status of each cluster. Clusters with low load and available resources are preferentially selected;
   4) Energy-aware scheduling strategy: Scheduling jobs based on the overall energy consumption level of each cluster. Clusters with low energy consumption are preferentially selected;
   5) Price-aware scheduling strategy: Scheduling jobs based on the fee of each cluster. Clusters with lower prices are preferentially selected;
   6) Network-aware scheduling strategy: Scheduling jobs based on network parameters such as bandwidth, delay, jitter, and packet loss rate of each cluster. Clusters with better network performance are preferentially selected;
   7) Data-aware scheduling strategy: For scenarios with a small amount of data, select the most appropriate computing center to run jobs and move data to the computing center through scheduling; For scenarios with big data or privacy data that are not suitable for data migration, should sense the storage location of the data and select the computing center where the data resides to run the job;
   8) Computing services QoS/QoE scheduling strategy: Take the QoS/QoE of the service as the highest priority decision-making basis. Clusters with balanced and optimized network paths are preferentially selected.

g) Computing job scheduling: Should support distributing the computing jobs to the selected computing center according to the scheduling strategy;

h) Computing domain name system: Shall support the publication and subscription of computing network resources based on the domain name resolution mechanism, and expressing the information of computing network resources based on DNS RR to realize the publication and subscription of computing network resources.

6.6 Application Layer

The application layer is responsible for the information interaction between the computing network for the Greater Bay Area and the application platforms. The application platforms include AI computing platforms, general computing platforms, scientific computing platforms, etc. The application layer of the computing network for the Greater Bay Area should have the following capabilities:

a) Should support users to query the available resources (e.g., computing nodes, computing resources, network resources, open datasets, etc.) of the computing network for the Greater Bay Area;

b) Should support users to initiate resource application requests and service access requests to the computing network for the Greater Bay Area;

c) Should support the creation, query, termination, deletion, and obtaining of calculation results of jobs.

6.7 Operation Layer

The operation layer should achieve unified operation of computing, network, data and other resources of multiple computing centers (including supercomputing, intelligent computing, and general cloud computing centers), and should include the following functions:

a) Unified authentication and authorization: Should support account registration, authentication, and authorization to ensure that the account is globally unified and can be used in multiple computing centers;

b) Unified measurement and billing: Should support statistics on the usage of computing resources and services, and generate orders for cost settlement;

c) Resource operation visualization: Should support the visualization management of computing resources, network resources, and infrastructure;

d) Resource monitoring and alarm: Should support real-time information collection and fault detection of the network, computing, and other resources of the computing center, and can generate alarm logs;

e) Unified computing console: Should support the unified registration and access of computing for major computing centers, and support the configuration, joining, and exiting of the computing network for computing centers and network resources;

f) computing Trading Center: Should support the search and viewing of the supply of computing resources across the entire network, and support users to subscribe to computing services through methods such as annual and monthly packages, metering;

g) Product management: Should support the release, removal, labeling, and other capabilities of computing network products and solutions;

h) Data trading center: Should support data providers to publish data, and support users to subscribe to transactions and use data services;

i) Computing network measurement: Should support the evaluation and level definition of computing, storage capacity, and transportation capacity for basic computing equipment, storage devices, and network conditions with different specifications, models, and brands of computing centers, and provide users with unified abstract descriptions to achieve standardization of computing, storage capacity, and transportation logic unit products for user services.

6.8 Security Mechanism

The security mechanism of the computing network for the Greater Bay Area needs to provide security protection for the computing network from multiple aspects such as data security, network security, transmission security, security assessment, and security management. It should include the following functions:

a) Data security: Data security capabilities can provide data privacy protection, critical data authentication, and calculation result verification for the computing network services. Specifically, the following capabilities should be supported:

1) Should suppourt important data backup and recovery to avoid data loss or damage, ensuring

the integrity and reliability of system data.

2) Should support data storage encryption to prevent important data from being tampered with;

3) Should sensitive data isolation storage , and should adopt access control technology to only allow authorized users or nodes to access, avoiding illegal access and data leakage;

4) Should support data traceability function, which can trace the source, flow path, and processing of data, to improve the credibility of data;

5) Should support the introduction of computing power for confidential computing, build a highly trusted computing environment, and serve high security business needs;

6) It is advisable to support the introduction of secure multi-party computation to ensure the confidentiality of data at the computing node;

7) It is advisable to support blockchain, and through blockchain reconciliation and proof preservation, billing data can be guaranteed to be trustworthy and tamper proof, ensuring the credibility of all parties involved in the computing network business transactions.

b) Network security: Network security capabilities support the detection of network attacks, timely response to attacks, and other functions to ensure the normal operation of computing networks. Specifically, the following capabilities should be supported:

1) Should support access control at the boundaries of computing networks and check the business flow of computing networks to prevent unauthorized malicious access;

2) Should support network boundary intrusion detection and defense, and monitor attacks on computing nodes and businesses;

3) Should support situational awareness, display real-time security risk status and development trends in the computing network, and be able to respond to security risks in a timely manner.

c) Transmission security: The transmission security capability provides encryption and authentication for data transmission between major computing centers, and can support multiple encryption transmission methods. Specifically, it should support the following capabilities:

1) Should support the secure transmission of physical layer information through chaotic algorithms, noise disturbances, and other methods;

2) Should support encrypted transmission based on OTNsec method;

3) It is advisable to support quantum encryption transmission based on QKD;

4) It is advisable to support the use of OTDR and other methods to achieve state awareness of the physical layer of the optical network, prevent leakage and eavesdropping, etc.

d) Cross border security: Cross border security capabilities provide data governance and security protection capabilities for data exchange and transmission within the province, Hong Kong and Macau, ensuring that cross-border data circulation complies with relevant national laws and regulations;

e) Security assessment: The security assessment capability provides security assessment, baseline verification, vulnerability scanning, and other capabilities for computing nodes, ensuring the security of computing nodes. Specifically, it should support the following capabilities:

1) Should support security evaluation to detect and evaluate the security protection mechanisms and configurations of computing nodes, such as host security, virtualization security, and network security, as well as whether the computing nodes have confidential computing, privacy computing, or other security service capabilities;

2) Should support baseline verification to verify whether the security protection mechanism and security configuration of computing nodes meet security requirements;

3) Should support vulnerability scanning, conduct vulnerability scanning on computing nodes, and timely detect and discover the vulnerabilities of computing nodes.

f) Security management: Security management capabilities should support capabilities such as capability management, security business management views, operational auditing, and be able to present security situations for security administrators to analyze and make decisions. Specifically, the following capabilities should be supported:

1) Should support the registration, management, and configuration of security capabilities;

2) Should support the management and coordination of the computing network and business to call the resources of security capability layer ;

3) Should support operational auditing by analyzing the logs generated by the computing network, recording and monitoring normal processes, abnormal states, and security events, forming operational auditing information and conducting timely security analysis and evaluation;

4) It is advisable to support the management, configuration, and presentation of security requirement analysis and security capability perception functions.

7 Interface requirements

7.1 Overview

Different functional layers of the computing network for the Greater Bay Area are accessible to each other through various interfaces. A complete computing network system should contain the following interfaces.

7.2 Interfaces between the Resource Layer and Adaptation Layer

Between the resource layer and the adaptation layer, the resource and job information of the computer centers can be obtained through the command line, RESTful, or Socket interfaces. The interfaces should support job submission and management.

7.3 Interfaces between the Adaptation Layer and Scheduling Layer

The scheduling layer submits, queries, and manages jobs through the interfaces to the adaptation layer. The adaptation layer reports the resource and job information of the computing centers to the scheduling layer through the interfaces.

7.4 Interfaces between the Adaptation/Scheduling Layer and the Network Layer

Information is transmitted between the adaptation layer and the scheduling layer through the network layer using the existing standard network interface.

The network layer reports the current network status information by calling the scheduling layer interface to support the scheduling layer's network orchestration function.

7.5 Interfaces between the Operation Layer and Scheduling Layer

Through the interfaces, the operations layer obtains the resource and job information of the computing centers in the Greater Bay Area from the scheduling layer, and submits, queries, and manages jobs.

7.6 Interfaces between the Application Layer and Operation Layer

The application layer calls interfaces of the operation layer for unified authentication and authorization.

The operation layer obtains the model and data information by calling the interface of the application layer.

7.7 Interfaces between the Application Layer and Scheduling Layer

The application layer can directly connect to the scheduling layer and submit and manage jobs through the interfaces. Based on the scheduling strategies at the scheduling layer, computing requirements from the application layer can be accurately and quickly routed to the corresponding computing nodes.

Appendix A

(Informative)

Application Case of UAV Supervision in Cities

In this application scenario, the UAV patrols specific areas of the city along a pre planned route and takes aerial photos. Based on the computing network for the Greater Bay Area, the aerial images and video data are quickly transmitted back to the backend deployed in the cloud computing center. The backend selects appropriate reasoning nodes for the intelligent computing center based on the computing network scheduling layer, and quickly transmits relevant image and video data to the intelligent computing center for reasoning analysis. The intelligent computing center then feeds back the analysis results to the backend. The backend will synchronize the analysis results with the relevant platforms of the regulatory authorities. If there are any abnormal situations in the analysis results (such as store fires, debris blocking roads, etc.), it will promptly alert the regulatory authorities.

Due to the large amount of image and video data collected by UAVs, in order to ensure timeliness, it is necessary to provide a high-speed and low latency network environment for image and video data transmission back to the background, as well as cross computing center data interaction scenarios. At the same time, it is necessary to support various scheduling strategies and achieve flexible scheduling of computing power for various AI reasoning and analysis nodes.

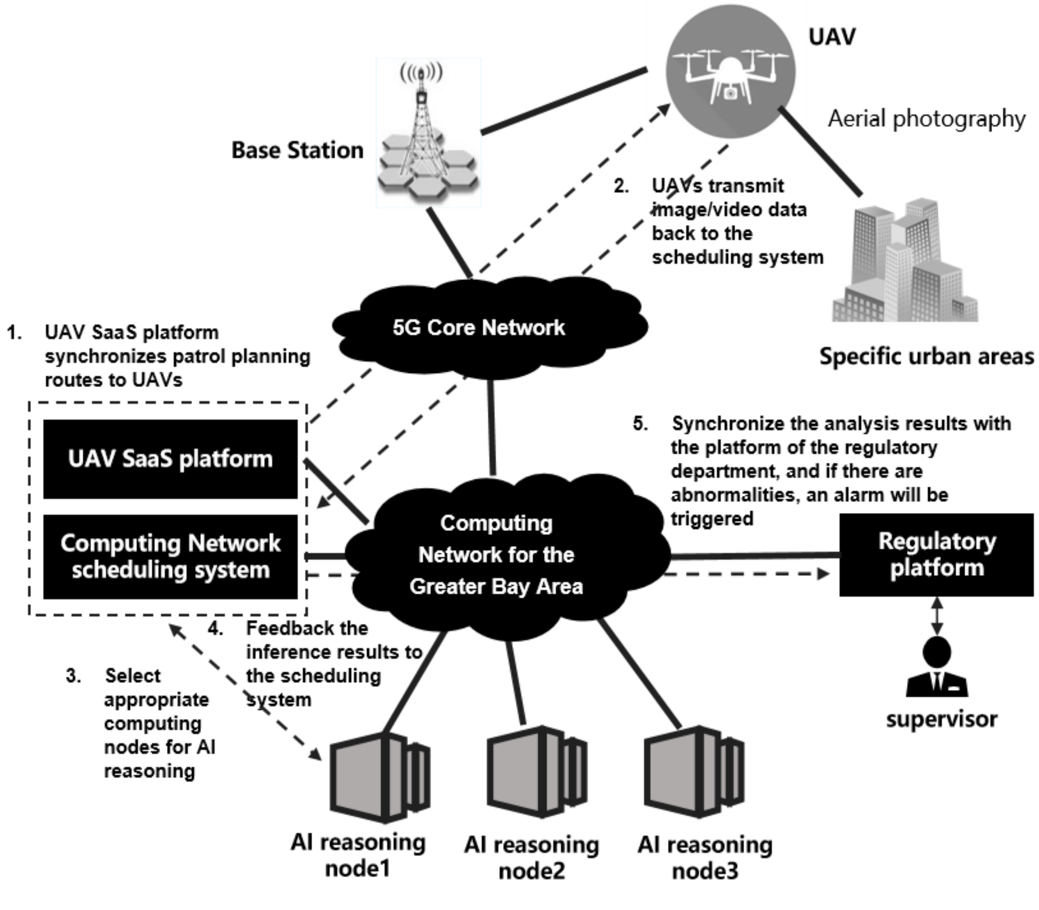
UAV
Aerial photography
Base Station
2. UAVs transmit image/video data back to the scheduling system
Specific urban areas
5G Core Network
1. UAV SaaS platform synchronizes patrol planning routes to UAVs
UAV SaaS platform
Computing Network scheduling system
5. Synchronize the analysis results with the platform of the regulatory department, and if there are abnormalities, an alarm will be triggered
Computing Network for the Greater Bay Area
Regulatory platform
supervisor
4. Feedback the inference results to the scheduling system
3. Select appropriate computing nodes for AI reasoning
AI reasoning node1
AI reasoning node2
AI reasoning node3

Figure A.1 Schematic diagram of UAV supervision in cities

Appendix B

(Informative)

Application Case of Intelligent security

In this application scenario, users access MEC resources through the R1 node, and under normal circumstances, video traffic is connected to the nearest MEC-1 node. However, assuming that the resource utilization rate of MEC-1 is close to 90% at that time point, and it is in a high load state, the video will frequently experience lag. By utilizing the on-demand scheduling capability of the computing network for the Greater Bay Area, video services are automatically redirected to the optimal computing node MEC-2 for processing, achieving uninterrupted video service flow and smoother videos, thereby significantly improving the experience of users.

Due to differences in computing resource allocation and uneven computing load among MEC nodes, local MECs may not be able to fully meet the real-time and deterministic needs of business. Through the computing awareness protocol, the required computing resource requirements are embedded in the control layer information of the network. The computing resource information is distributed through the network control plane, and computing tasks are scheduled to nodes that meet user needs based on the computing resource information, achieving computing resource awareness and optimized scheduling.

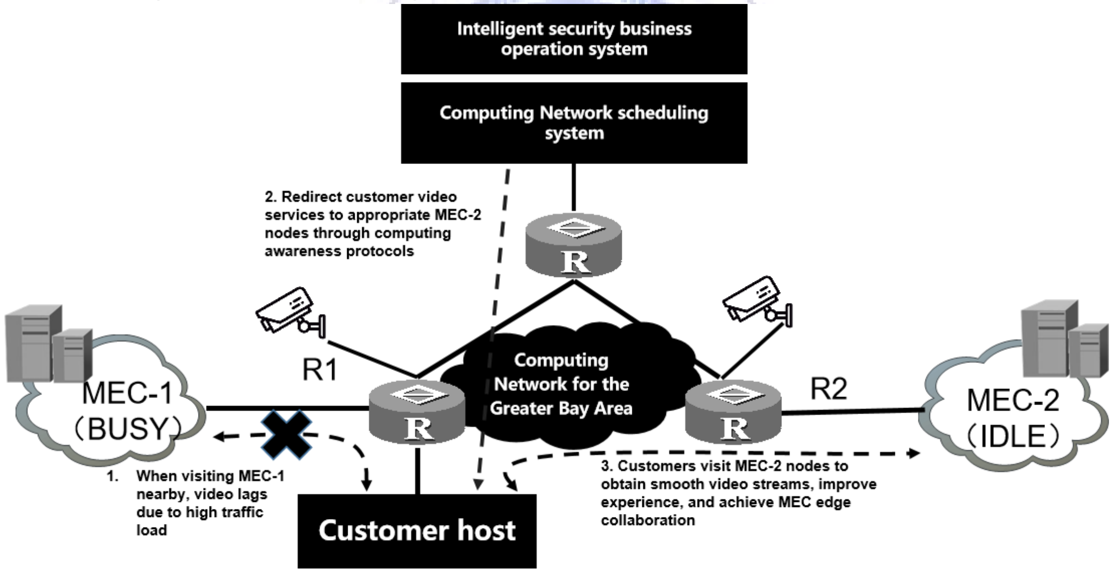
Intelligent security business operation system
Computing Network scheduling system
2. Redirect customer video services to appropriate MEC-2 nodes through computing awareness protocols
R
R1
MEC-1
（BUSY）
R
Computing Network for the Greater Bay Area
R
R2
MEC-2
（IDLE）
1. When visiting MEC-1 nearby, video lags due to high traffic load
Customer host
3. Customers visit MEC-2 nodes to obtain smooth video streams, improve experience, and achieve MEC edge collaboration

Figure B.1 Schematic diagram of intelligent security application

Appendix C

(Informative)

Application case of 3D real-time rendering

In this application scenario, users log in to the real-time cloud rendering portal, select computing specifications, submit their designed 3D models, and the computing network scheduling system selects rendering edge nodes nearby to complete cloud rendering based on the user's computing requirements and access location. After rendering is completed, the portal will return a publishing URL to the user. Users can access this URL to obtain real-time 3D rendering effects and interact with the model in real-time (such as scale out, scale out, zooming in, zooming out, etc.).

Due to the need for real-time interactive operation of the model by users, the transmission delay of the network is required to reach the millisecond level in order for users to obtain a better experience. Therefore, based on the perception ability of the computing network, it is necessary to select rendering edge nodes with corresponding computing resources and close to users for 3D real-time rendering applications to provide services, ensuring that users do not encounter problems such as frame drops and lagging during real-time interactive operations with the model.

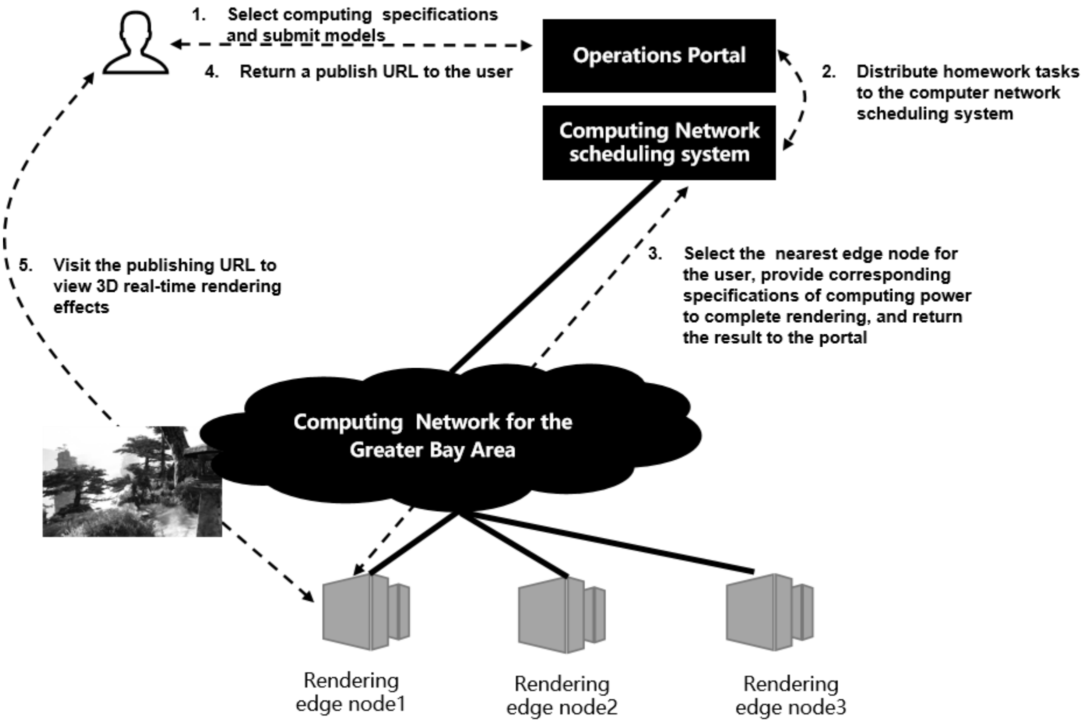
1. Select computing specifications and submit models
4. Return a publish URL to the user
Operations Portal
2. Distribute homework tasks to the computer network scheduling system
Computing Network scheduling system
3. Select the nearest edge node for the user, provide corresponding specifications of computing power to complete rendering, and return the result to the portal
5. Visit the publishing URL to view 3D real-time rendering effects
Computing Network for the Greater Bay Area
Rendering edge node1
Rendering edge node2
Rendering edge node3

Figure C.1 Schematic diagram of 3D real-time rendering application

Appendix D

(Informative)

Application Cases of AI computing Product Standardization

In the traditional way, when users choose to buy AI arithmetic products, they need to choose the form of physical GPUs based on the floating-point arithmetic values provided by GPU vendors, which requires high expertise from users.The standardized application solution for AI arithmetic products is to customize an AI arithmetic benchmark that hides the real GPU device model, and provides a virtual GPU for sale based on the reference arithmetic benchmark only. Users can intuitively judge the choice based on the reference performance value of the standard AI computing power model application, making it easier for users to judge. The advantage of this solution is that cloud vendors can use GPU virtualization software to customize GPU devices, and can make reasonable use of pooling new and old different GPU devices to greatly increase the utilization rate of the entire GPU resource cluster by a number of times, resulting in the lowest cost per unit of computing power. The following is a practical example of the standardized application of AI arithmetic products:

Based on the typical pervasive AI model Resnet50 as a benchmark, the GPU devices in the resource pool for testing, and based on the test data will be divided into different specifications of the GPU devices of different arithmetic levels, for example, into the basic version of the B (Basic), the standard version of the S (Standard), the advanced version of the P (Premium) of the three major levels, and support in the B / S / P Inside the three major levels, there are smaller granularity of arithmetic level instances based on operational requirements and user needs, and the final AI arithmetic product specifications are shown as follows (taking the Basic Edition as an example):

Table D.1 Example diagram of AI computing product specifications

First level classification	Second level classification	Third level classification	Instance specifications	Computing specifications
Basic version - B (Basic)	B1	small	#Number of GPU cards: 1 #GPU graphics memory per card: 6 GB #CPU cores: 4 #Memory: 12 GB	B1.small : The throughput is approximately 145 images/sec #ResNet50 model with training accuracy of fp16 #ImageNet2012 dataset #BatchSize=112
		medium	# Number of GPU cards: 1 # GPU graphics memory per card: 12 GB # CPU cores: 4 # Memory:12 GB	B1.medium: The throughput is approximately 390 images/sec #ResNet50 model with training accuracy of fp16 #ImageNet2012 dataset # BatchSize = 224

Table D.1 Example diagram of AI computing product specifications（*Continued*）

First level classification	Second level classification	Third level classification	Instance specifications	Computing specifications
Basic version - B (Basic)	B1	large	# Number of GPU cards: 1 # GPU graphics memory per card: 24 GB # CPU cores: 8 # Memory: 24 GB	B1.large: The throughput is approximately 801 images/sec #ResNet50 model with training accuracy of fp16 #ImageNet2012 dataset # BatchSize = 448

Appendix E

(Informative)

The Application Case of Computing Network-Native Large Language Model

In this application, the computing network for the Greater Bay Area deploys *large language models* (LLMs) in large computing centers (i.e., backbone nodes) in the region, and supports the whole development process of an LLM, including fine-tuning, distillation, inference deployment, and other applications for various industries by invoking computing and data resources in multiple computing centers.

For example, users log in to the model development platform, select the base model, downstream task data, and computing power specifications, and then submit the model fine-tuning request. The scheduling system of the computing network selects the appropriate node to complete the model fine-tuning according to the computing resource demand and data location. Once the fine-tuning is completed, users can download the fine-tuned model or submit a model deployment request through the model development platform. The computing network scheduling system selects one or more appropriate computing centers according to the model scale and user requirements, completes the inference service deployment of the model, and releases the model call API. Then, model users can submit model inference requests through the API. The computing network schedules services according to the deployment location of the model and the resource status of each computing center, and then selects appropriate deployment nodes to complete model inference services. At the same time, with user permission, the model development platform can collect user feedback through API and feed relevant data to the model provider and the base model for continuously updating the LLM.

Due to the various requirements of users, the continuous learning, fine-tuning, and inference of LLMs may be completed in different computing centers. Therefore, the computing network is required to effectively shield the heterogeneity of different computing centers from the computing chip, storage, network, as well as the system interface such that the same model can be quickly transferred to different computing centers to execute the corresponding training and inference operations for supporting the ecological construction of LLM applications.

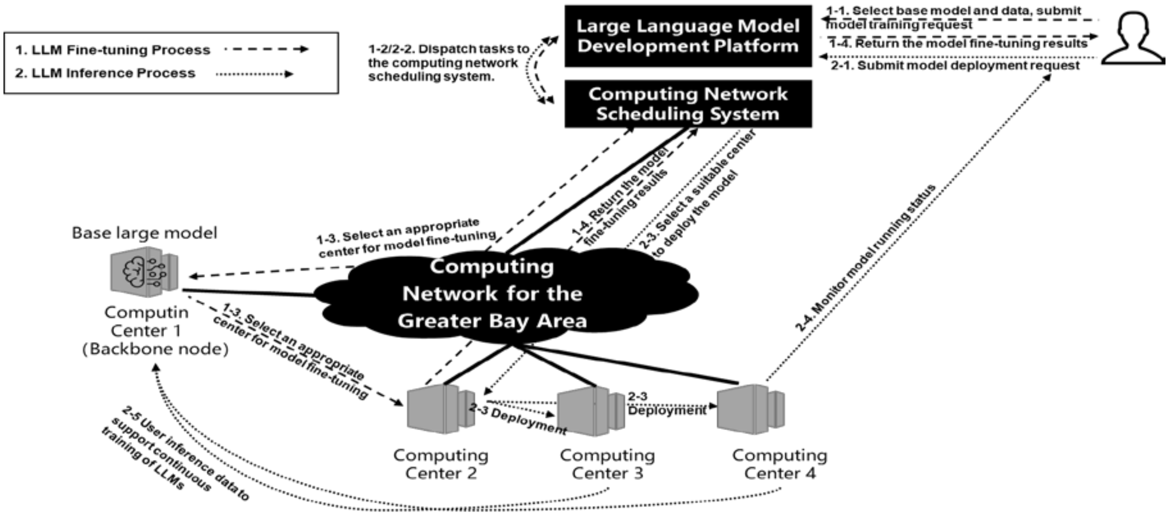
1. LLM Fine-tuning Process
2. LLM Inference Process
1-2/2-2. Dispatch tasks to the computing network scheduling system.
Large Language Model Development Platform
Computing Network Scheduling System
1-1. Select base model and data, submit model training request
1-4. Return the model fine-tuning results
2-1. Submit model deployment request
Base large model
1-3. Select an appropriate center for model fine-tuning
1-4. Return the model fine-tuning results
2-3. Select a suitable center to deploy the model
2-4. Monitor model running status
Computing Network for the Greater Bay Area
Computin Center 1 (Backbone node)
1-3. Select an appropriate center for model fine-tuning
2-3 Deployment
2-3 Deployment
2-5 User inference data to support continuous training of LLMs
Computing Center 2
Computing Center 3
Computing Center 4

Figure E.1. Application diagram of the Computing Network-Native LLMs

Bibliography

[1] *ISO/IEC 22237-1:2021 Information technology – Data center facilities and infrastructures – Part 1: General concepts*

[2] *ISO/IEC/IEEE 24765:2017 Systems and software engineering – Vocabulary*

[3] *ITU-T Y.2501 Computing power network – Framework and architecture*

[4] *YD/T 4255-2023 General technical requirements of Computing and Network Convergence*