NIDA-PD-003-2025

NIDA/TC/WG01/TF03

Date: 2025-09-26

Next Generation Network for 5.5G Era (NET5.5G) Deployment Guideline

Contents

Foreword

The Network Innovation and Development Alliance (NIDA) is a voluntary international, industry-specific, non-profit social organization comprised of industry organizations, universities, research institutes, companies, and other entities from around the world, dedicated to promoting fixed network technology innovation and industrial upgrading.

The work of preparing NIDA Standards is normally carried out through the NIDA Technical Committee (TC), its Working Groups (WGs), and Task Forces (TFs). The procedures used to develop this document and those intended for its further maintenance are described in the **NIDA Standards Development Guidelines**.

This document is drafted by: China Unicom, China Academy of Information and Communications Technology, China Telecommunications Corporation, Internet Association of Kazakhstan, Huawei Technologies Co., Ltd.

Main drafters of this document: Chang Cao, Ran Pang, Shuai Zhang, Wei Gao, Yongqing Zhu, Zehua Hu, Shavkat Sabirov, Li Zhang, Shuanglong Chen.

Copyright Notice:

This document was developed and is maintained by the **NIDA**. The content of this document is made available solely for use by authorized members and users for purposes consistent with the goals of the NIDA. For authorization, contact: contact@nida-alliance.com.

Patent Statement:

NIDA draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). NIDA takes no position concerning the existence, evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, NIDA may receive notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the relevant patent database. NIDA shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of NIDA specific terms and expressions related to conformity assessment, please refer to the relevant NIDA IPR documents.

This document was prepared by the Technical Committee (TC), Working Group WG 01, Network Evolution working group, Task Force TF 03, *Next Generation Network for 5.5G Era (NET5.5G) Deployment Guideline.*

Any feedback or questions on this document should be directed to the NIDA TC Secretariat.

Next Generation Network for 5.5G Era (NET5.5G) Deployment Guideline

1 Scope

This Recommendation describes general principles of next generation network for 5.5G era (Net5.5G) deployment, then specifies the architecture and key technologies for overall new three connection scenarios, including computing connections, intelligence connections, and data connections.

This Recommendation provides key technologies for support of various applications in Net5.5G. Following the network deployment guideline with architecture and key technologies for the three scenarios are provided.

2 Normative references

[1]. ITU-T Y.LDT IMT-2020 Networks and Beyond: Requirements and Framework for applications demanding large data transmissions—*Part 6: Overview of applications demanding large data transmission*

[2]. WBBA White paper (2024), Network Evolution For the 5.5G And 6G Era—*NET5.5G NETWORK ARCHITECTURE AND*
KEY TECHNICAL FEATURES

3 Terms and definitions

3.1

Net5.5G

Net5.5G refers to the next generation network era to support computing connection, intelligence connection, data connection and space connection. It is divided into two phases, this Recommendation focus on phase Ⅰ, including the scenarios and key technologies of computing connection, intelligence connection and data connection.

3.2

Computing connection

Computing connection refers to the scenarios support AI training and inference, e.g., network in a single data network (intra–DC), network across multiple data centers (inter–DC) and network for data transmission to data centers (data–transmission–to –DC).

3.3

Intelligence connection

Intelligence connection refers to the scenarios support intelligent terminals access and reference, e.g., personal services (e.g., autonomous driving and XR), home services (e.g., intelligent robotic vacuum cleaner and intelligent camera) and industry services (e.g., industry robots and intelligent manufacturing).

3.4

Data connection

Data connection refers to the scenarios support data transfer between data suppliers and consumers for different data types (e.g., private and public) and different scopes (e.g., across–domains and across–countries.).

4 Abbreviations

ADN: Autonomous Driving Network

AI: Artificial Intelligent

AOI: Automatic Optical Inspection

APN: Application-aware Network

APT: Advanced Persistent Threat

DC: Data Center

DT: Digital Twin

ECMP: Equal-Cost Multi-Path

ECN: Explicit Congestion Notification

EDR: Endpoint Detection and Response

HBF: Hybrid Beamforming

LLM: Large Language Model

MAC: Media Access Control

MACsec: Media Access Control Security

M&C Management and Control

NSLB: Network Scale Load balancing

PFC: Priority-based Flow Control

PQC: Post-Quantum Cryptography

RDMA: Remote Direct Memory Access

rPFC: Remote PFC

SLA: Service Level Agreement

SASE: Special Application Service Element

SPFC: Subscriber PFC

SRv6: Segment Routing IPv6

XR: Extended Reality

WAN: Wide Area Network

5 Overview of next generation network for 5.5G Era (Net5.5G)

Next generation network for 5.5G Era， abbreviated as Net5.5G, refers to the network evolution for the emerging network connection scenarios, including computing connection, intelligence connection, data connection and airspace connection.

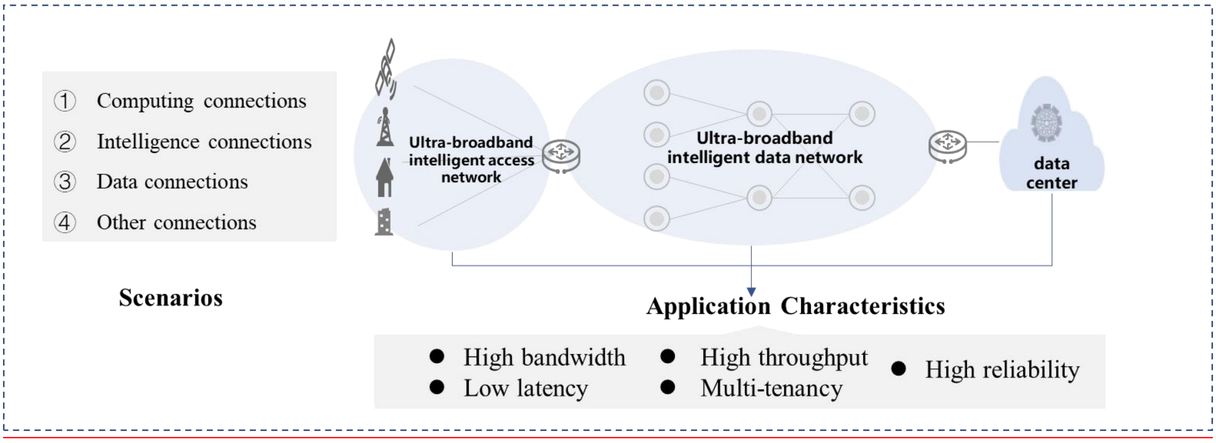

Figure 1 — Overview of Net5.5G

Some organizations have already done some research. World Broadband Association (WBBA) has published a white paper [1] about Net5.5G, which focus on a sustainable evolution of data communication network infrastructure, providing enhanced capabilities in order to meet the demands of new intelligent application trends. ITU-T also proposed a Recommendation [2] about requirements and functions for applications demanding large data transmissions, which defines the enhanced and new functions in IMT-2020 networks and beyond to meet the requirements of applications driven by the new technologies (e.g., AI, cloud, big data, 5G and 5G-A).

— Computing connection scenario

Computing connection refers to the connection between computing power suppliers and consumers. Net5.5G for computing connection is used to interlink the computing power and computing chips. It’s required to support three types of networks, including intra-data center, inter-data center, and data-transmission-to-data center

As far as intra-DC network, the interconnection of computing chips needs to accommodate the ultra-large-scale cluster connections ranging from hundreds to tens, or even hundreds of thousands of chips in one single data center. The network is required to support ultra-large-scale networking, high throughput with no data losses and intelligent fault tolerance.

As far as inter-DC network, the multi data center cooperation for AI training is needed due to computing power and electric limits of single data center. The network of interconnection of multi data centers needs to support high speed, high throughput and long-distance lossless data transmission in order to support high-efficient ultra-large-scale model training.

The network of data-to-DC, serving as a pipeline for computing power, connects a large number of enterprises, research institutions and data centers. It's required to have the capabilities for differentiated scheduling and optimization. AI large model training incurs large data transmission requirements. The model data set usually require ten of GB to hundreds of TB data. The challenges of large data transmission to data center are includes high bandwidth low cost, high network resource utilization and high data security.

— Intelligence connection scenario

With the developments of Generative AI, multimodal understanding and embodied intelligence, the real-time interaction between virtual and the real world is deepening. AI agent are deeply integrated with various intelligent terminals to create more innovative application scenarios. AI terminals, represented by AI assistants, embodied robots and immersive devices, have had a significant impact on life and production. Net5.5G for intelligence connection is required to support the individual, home, enterprise and industry scenarios.

For individual scenario, AI terminals and digital humans, as the entries for connecting people and AI, reshaping the new man-machine relationship, and drive the network architecture transformation of multi-level inference. The network between terminals and cloud is required to meet the high bandwidth and low latency requirements.

For home scenario, an interconnected home intelligent ecosystem with companion robots as the core has become the mainstream. In addition of low-latency experience, the network in home scenario is also required to provide security authentication and data encryption functions.

For enterprise scenario, new technologies such as immersive collaborative office, AI assistant and intelligent manufacturing robots have been applied rapidly, with stronger interaction, higher production efficiency, and the man-man, man-machine, and machine-machine relationship closely implemented. The network is required to ensure the virtual and reality convergence experience, low latency, centimeter-level positioning, application-level policy control, and dimensional data protection.

— Data connection scenario

Data has become the important factor of production now. The data element needs to be circulated across different industries, regions and subjects. Net5.5G for data connection is a new generation of secure and trustworthy network infrastructure to support data element circulation.

Data connections network is required to sensor data identities, types and sensitivity levels and then implement the security control policies, in order to implement manageable and controllable data element circulation paths.

Data connection network should have the capabilities of differentiated data transmission trusted paths and higher security data transmission encryption, in order to meet the security requirements of data element circulation in different industries and at different data levels.

Data connection network is required to provide data access nodes, and should has the all–in–one capabilities of network, data and security. Based on the zero–trust architecture, the network should support the data security enhancement capabilities such as secure data storage, refined data use control, and secure data transmission, to implement flexible and secure data access and lowering the threshold for customer’s data connection.

6 Architecture of next generation network for 5.5G Era (Net5.5G)

6.1 Architecture of Net5.5G

The next generation network for 5.5G era (Net5.5G) is shown in the following figure:

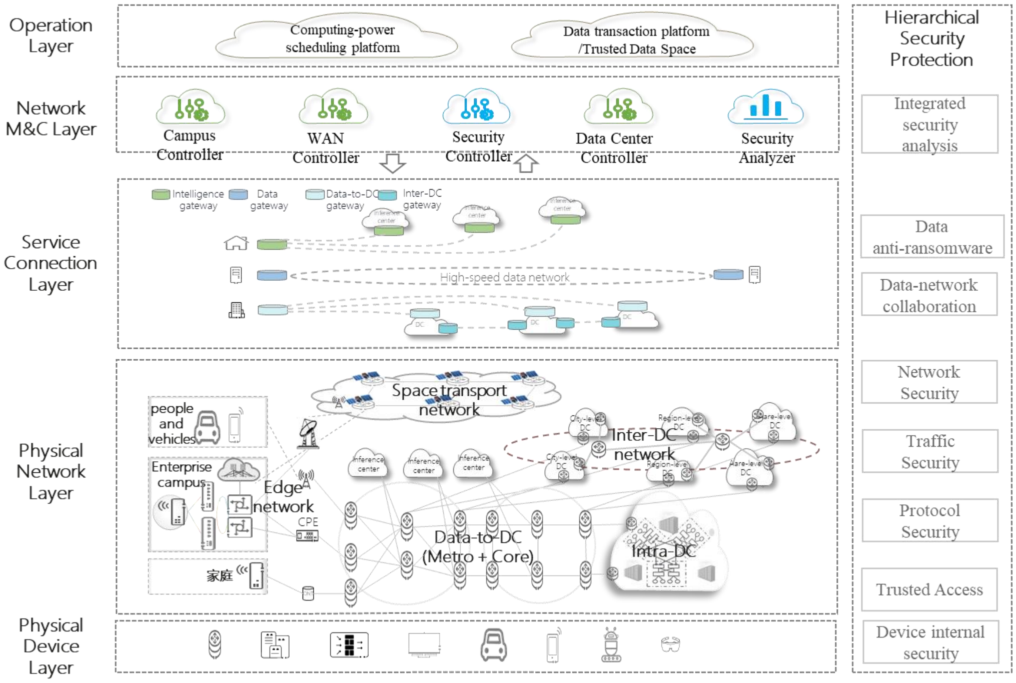

Figure 2 — Architecture of Next Generation Network for 5.5G Era (Net5.5G)

The architecture of next generation network for 5.5G Era (Net5.5G) is divided into five layers, physical device layer, physical network layer, service connection layer, network management and control layer, and operation layer.

In terms of physical device layer, a variety of devices should be supported, including routers, switches, terminal devices, and so on.

In terms of physical network layer, in order to connect the personal smart terminals, homes and enterprises and multi–level inference and data centers, which are used to support AI large model training and inference and new AI-driven applications, Net5.5G needs to build high-quality edge networks, intra-DC network, inter-DC network and data-to-DC network. At the same time, space and terrestrial integration networks is also needed.

In terms of service connection layer, the service gateways (e.g., intelligence gateway, data gateway, data-to-DC gateway and inter-DC gateway) are needed to connect the users, computing power and data.

In terms of network management and control layer, Net5.5G should be composed of multi-region controllers as network brain to provide various of network management, operations and maintenance capabilities to improve the network autonomous level.

In terms of operation layer, which usually works as a coordinator between business services and network, is also required to be improved.

In addition, as the foundational attribute, hierarchical security protection is required in different layers.

6.2 Features and capabilities of Net5.5G

To support new connection scenarios, next generation network for 5.5G Era (Net5.5G) is required to achieve new network objectives in the following figure:

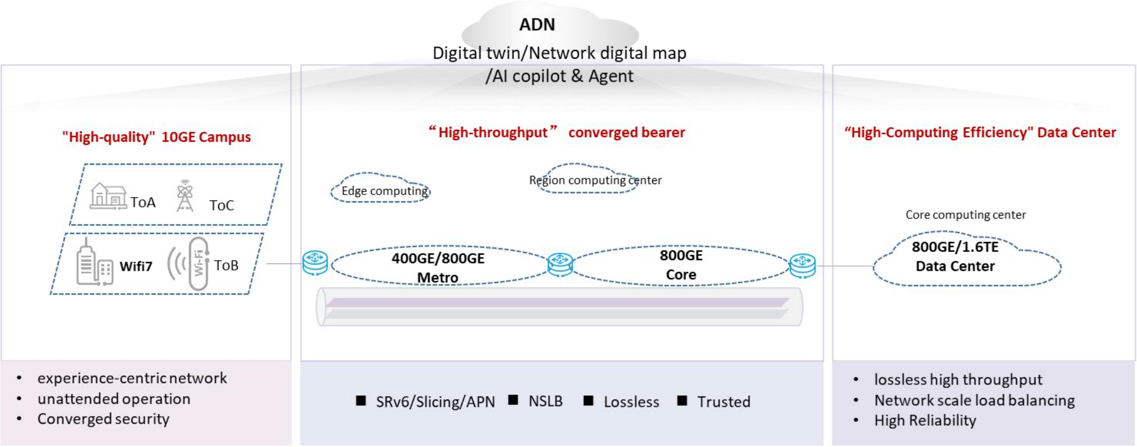

Figure 3 — Objectives of Next Generation Network for 5.5G Era (Net5.5G)

To achieve the goal of high-quality Campus, high-throughput converged bearer and high-computing efficiency data center, the following key features and capabilities are required:

— Extreme simplification and ultra-bandwidth: The network should be extremely simplified. The network devices should support high-density ports and ultra-bandwidth, based on the continually bandwidth volume improvement of the physical layer protocols (e.g., Wi-Fi 7 and 400GE/800GE/1.6TE).

— Intelligence native: The service connection layer is required to support fine application identification, and the physical network layer is required to support lossless high throughput. Both physical device layer and network management and control layer should evolve towards intelligent in order to support autonomous driving network (ADN) to level 4 or even more advanced.

— Integrated security: The devices should have native security, and the physical network layer should support secure transport tunnels. In service connection layer, the service connection security should be guaranteed through different methods such as data and network cooperation, data anti–ransomware and so on. The network management and control layer support integrated

— Green and energy saving: The physical devices should support hibernation on demand, and the physical network layer is required to support low–carbon paths. The network management and control layer should support visibility and controllability.

7 Advanced technologies of next generation network for 5.5G era (Net5.5G)

7.1 Large data transmission-driven physical device improvement

Key technologies in physical device layer refers to the capabilities to enhance bandwidth, reduce power consumption, increase port and bandwidth volume to meet the requirement of large data transmission. The following technologies should be considered:

— High data rate of 400GE/800GE/1.6TE: In network system, high data rate Ethernet ports with low power per bit design and advanced signal processing technologies should be used to support high–speed, low–latency and high–reliability data transmission.

— Evolution of Wi–Fi7[3]: Millimeter wave technology promotes the access bandwidth over air interface from 30Gbps to 100Gbps. Technologies such as high–gain algorithm and dielectric HBF antennas are used to continuously challenge the coverage distance of 10m to 15m under power consumption constraints. This is urgently needed in large uplink scenarios, such as industrial AOI backhaul and XR video local generation.

— Switches with high–density ports: These switches, such as subrack–shaped switches should be used in large scale networks (e.g., tens of, even hundreds of thousands of chips) to support AI large model training and inference.

7.2 High efficient transmission–driven network enhancement

Key technologies of protocols are used to implement large–scale network, simplify network interconnections, support lossless high throughput and enhance network security. The following technologies should be considered:

— Ethernet enhancement: Converged Ethernet is the optimal protocol in AI training and inference. Enhanced Ethernet technologies should be considered, e.g., head suppression, link layer retransmission.

— Segment Routing IPv6 (SRv6) [4]: SRv6 is the last evolution of the source–routing technology. It's compatible with IPv6, and can be used to simplify network protocols, specify explicit forwarding paths, and to support cross–domain network path programming. In addition, SRv6 is based on 128 bits address space, which could be used to identify various network information, locations, VPNs, value–added applications, etc. The programmability of SRv6 is useful to implement flexible network functions. Some enhance capabilities of SRv6 are as follows:

- ● Application-aware network (APN) [5]is used to identify application. This capability will promote the network service system from IP address only to IP address and APNID, to provide more refine (e.g., application-level) network service. It also supports the collaboration of clouds, networks, edges and endpoints to guarantee the application experience.
- ● APN is used to identify data attribute types and data space to ensure that data is recognizable, manageable, and visible during data transfer, ensure data security, and promote data development and utilization.

— Accurate traffic data measurement: Advanced traffic data measurement methods are required to detect the real flow status instead of the traditional probe packet-based measurement methods. The advanced traffic data measurement methods, such as in-situ flow information telemetry (IFIT) [6], encapsulate the measurement information in the traffic data packets, which can provide high accurately end-to-end and hop-by-hop measurement for a flow with respect to packet loss, latency and jitter.

— Network slicing: IPv6-enhanced-based network slices can be used to implement resource and security isolation, differentiated service quality assurance, flexible and customizable topology connection.

— Technologies for network load balance: To address disadvantages of the traditional equal-cost multi-path (ECMP) in network load balance, enhanced technologies, such as packet spraying, should be used.

— Precise flow control: Flow control technologies, e.g., priority-based flow control (PFC) [7] without deadlock and head-of-line blocking, should be used to prevent packet loss in data network. To support intelligent computing across wide area network (WAN) with long distance between data networks, precise control technologies are required. Subscriber PFC (SPFC) is a technology designed to enhanced subscriber-level flow control based on SRv6 network. The collaboration of SPFC on WAN and PFC or rPFC in data center could enable the lossless high throughput for the AI training.

— Congestion control technologies: Congestion control refers to the congestion detection, avoidance and processing. In AI training scenarios, congestion control technologies for RDMA flows should be used to reduce the network congestion. Technologies such as ECN, fast CNP, and AI enabled ECN should be taken into considered. Besides, adaptive routing should be introduced to improve the congestion detection, distribution and re-routing efficiency.

— Big flow recognition: This technology is required to recognize the large flows in the network approximately real-time.

7.3 Emerging service connection-driven network new gateways

Work as the service gateways of intelligence connection, data connection and transmission to DC, new technologies are required to connect the users, computing power and data:

— Computing power gateway: These gateways are required to connect the users and DCs.

- Gateway for data-to-DC are required to be deployed on both enterprise sides and computing power centers. These gateways provide protocol transform at transport layer, allocate identification and select proper tunnels and paths to data transmission to DCs, and support accounting and reconciliation capabilities.
- Gateways for inter-DC connections are required to be deployed in computing power centers. These gateways support protocol connections to provide ultra-high bandwidth and long-distance lossless forwarding paths for collaborative training of multiple DCs.

— Intelligence gateways: These gateways refer to various types of integrated service gateways that establish connections with multi-level inference centers to dynamically select proper inference mode to ensure smooth running of different services.

- As the service entrances, gateways identify and control applications, and work with multi-layer reference to meet the differentiated requirements of various applications.
- The gateways serve as the control point for data encryption and security policies to ensure data security.
- These gateways should be deployed in different levels, e.g., campus, carrier edge network and data center.
- These gateways can work as an edge inference node.

— Data gateways: These gateways provide multi-scenario capabilities for data proximity and trusted access, APN network collaboration, and data fencing for both data suppliers and consumers.

- Through the network and data security convergence access node, the identity, device, data of the data suppliers and consumers can be trusted to access the connecting data networks, and the data can be encrypted stored and used control policies can be specified.
- The data gateways use network labels, e.g., APN6, to carry data information, user identity, data types, sensitivity levels, etc., to ensure that the data are strictly, accurately used as the usage control policies formulated by data suppliers, and to protect data sovereignty.
- The data gateway, by recognizing the authorized data usage scope (e.g., countries and industries) carried by network labels, apply policy control such as blocking and in-depth analysis of data flow beyond the authorized scope, to implement data fence during data transfer.

7.4 Large scale network-driven enhanced network management and control

Work as the network smart brain, network M&C layer require advanced key technologies to promote the network autonomous to level 4 or advanced. The technologies of M&C layer should collaborate with the technologies of device to realize network self-perception, self-management, self-optimization and self-healing. The following technologies should be considered:

— Centralized path computing, re–optimization and scheduling: This technology calculates the optimal traffic distribution by drawing a global traffic matrix and automatically diverts traffic through network devices. Based on the multi–dimensional data of the networks, applications, AI tasks, implement intelligent algorithms, large flow recognition, e.g., network–scale load balancing (NSLB), cloud–map, maximum flux algorithm.

- The NSLB technology takes advantage of the global perspective of both network and AI training tasks, achieving 100% network wide traffic balancing and improving performance in AI training scenarios.
- Cloud–map algorithm take all the network and cloud resources as the compute factors to calculate the most proper path and recommend proper cloud.
- Maximum–throughput multi–path algorithm focus on computing flexible multiple paths to carry the large data transmission to the cloud site.
- Based on the big flow recognition, flow–level scheduling is used to match the flow to the optimal paths, in order to realize high network throughput.

— Application–level experience guarantee is required to identify the applications, perceive the application status and implement close–loop policies to guarantee the application experience.

7.5 AI–driven network integrated intelligence

According to the development of AI, next generation network (Net5.5G) is required to support integrated hierarchical AI ranging from physical device to network M&C. Work as the network smart brain, network M&C layer require advanced key technologies to promote network ADN to level 4 or even more advanced. The AI technologies of M&C layer should collaborate with the AI technologies of device to realize network self–perception, self–management, self–optimization and self–healing. The following technologies should be considered:

— Device–level native AI: AI in devices such as AI main control board, AI line card and AI computing card should be improvement to support multi–level interference capabilities range from light fast level to high performance and real–time level. These inference functions are used to support various applications, intelligent energy saving, attach detection, refine flow and convergence perception, etc.

— Digital twin: The multi–dimensional and high–precision network digital twins provide inference acknowledges and decision–making basis for the network new smart brain. The main capabilities include:

- Collect various network data to build a network data lack that associates and models multi–dimensional data, network, security, application, user, experience, etc.
- Transform the historical, real–time and in future communication data to associated acknowledges, in order to support the decision–making of the network new smart brain
- Provide predictive network change simulation evaluation for complex networks, to ensure the decision–making accuracy of large models.

— Network digital map: It's a solution applying the digital twin concept to provide agile service provisioning, automatic traffic optimization, intelligent fault analysis, and potential risk prediction. It is a virtual twin that digitally maps physical network entities and performs real-time interactive alignment with physical networks. It provides multidimensional dynamic topology, localization, network configuration verification, and service experience assurance capabilities through a map-based experience. It also supports network plug-and-play.

— Generative AI: Both large and small models should collaborate to implement decision-making and task-close-loop. It covers different technologies and features:

- Copilot technologies provide person in different positions with personalized assistant support with multi-modal intent understanding capabilities. It could simplify the complex network operations, and increase work efficiency.
- AI agent has excellent logical reasoning abilities, and can independently detect and solve the network operation and management problems, can provide precious data analysis to assist the decision maker to make more reasonable judgement.

7.6 high-security technologies

Network security is a systematic project, which consist of various security technologies covering different dimensions, trustworthiness, resilience, encryption, etc. The following security technologies should be considered:

— Device internal security: A set of internal security mechanisms are required to protect devices against network attacks. Abnormal scenarios are automatically restored to ensure service running. Device internal security includes two aspects:

- As far as device self-protection, capabilities, e.g., anti-penetration, anti-removal and anti-vandalism should be built to redefend against APT attack.
- As far as the device resilience protection, functions and service continuity must be maintained in the face of attacks or other disturbances.

— High security: The network integrates data space, privacy computing, and federated learning technologies to provide comprehensive secure pipeline services for data transmission. In addition, post-quantum cryptography (PQC) and quantum security network are the key security technologies of the next generation network (Net5.5G). Specifically,

- The network use IPSec technology to construct end-to-end line-speed encryption secure transmission channels.
- The network use MACsec and PHYSEC technologies to establish secure interconnections between devices.
- The network use quantum key agreement (QKA) and quantum key distribution (QKD) against quantum computing attacks to ensure zero data leakage.

— Integrated cloud-network-edge-device security: This security defense system is required to deal with more systematic and covert security attacks by technologies such as generative AI.

- On the cloud side, the security agents have capabilities of semantic understanding and rich context awareness, to trace threat sources to hunt high-risk intrusions, and build red-blue games to warn and eliminate risks in advance.
- On the network and edge sides, the next generation special application service element (SASE) security service based on edge gateway and edge cloud should provide zero-trust network access, high-performance integrated security engine, real-time AI detection for known, variant, and unknown threat traffic detection.
- On the device side, the next generation endpoint detection and response (EDR) use technologies, e.g., dynamic and static confrontation, traceability graph analysis and malicious instruction detection to quickly detect unknown threat and 0-day vulnerability attacks and synchronize network-wide data against ransomware crime, financial fraud, and sensitive information theft.

— Trusted data transfer: This refers to build data identifiable, visible and manageable capabilities, in order to identify data, ensure that data does not leave trusted area, and implement supervision.

- In terms of data identification, large language model (LLM) could be used for semantic understanding and data annotation, and to specify the security level requirements for identifying data.
- In terms of data visibility, multi-dimensional digital twin enables 100% visualized supervision of data transfer paths, and the centralized or distributed routing algorithms based on security factors is used to avoid the sensitive areas and low-security devices during path computation.
- In terms of data management, the network policies based on IPv6+ accurately control the data transfer path, and the path consistency management is ensured by path verification algorithms. These features will ensure that the abnormal data transmission could be locked in seconds, and the important data dose not leave the boundary of the trusted zone.

— Security analyse: This technology is used to provide security analysis capabilities, e.g., traceable compliance audit and accurate threat detection.

— Security control: This technology is required to implement unified security authentication, intelligent security policy control, e.g., trusted path, encryption and data fence.

8 Next Generation Network (Net5.5G) Deployment Specification

8.1 Computing connections

8.1.1 Intra-DC intelligence computing

Intra-DC network consists of intelligent computing zone, storage zone, general calculation zone and management zone. The following figure shows intra-DC network architecture.

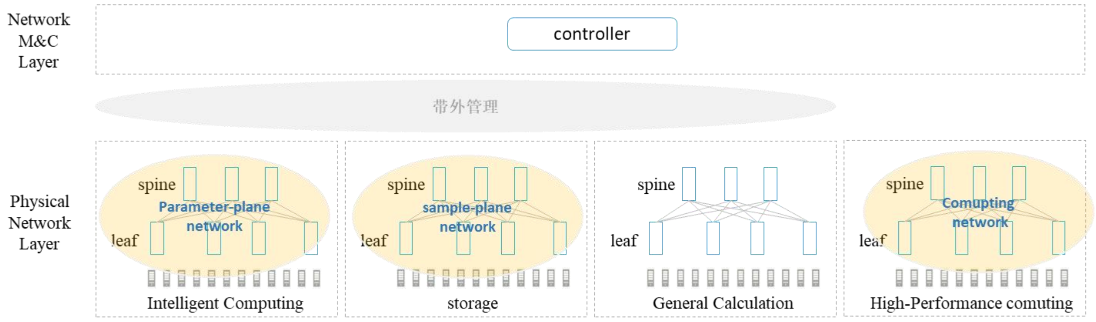

Figure 4 — Architecture of intra-DC

With the increasing scale of AI large model training and inference, the intelligent computing data center networks is also growing larger, requiring support for scale ranging from thousands to hundreds of thousands of cards. As the scale of networks and network communication efficiency increasingly impact the computing power of AI clusters, large-scale intelligent computing networks that provide lossless forwarding, high throughput, low latency, high reliability and virtualization management are becoming more and more important.

To support networks scale of ten thousand and hundreds thousand cards, and considering the reduction of latency, a multi-plane networking based on two-layer CLOS is recommended. This requires further enhancement of the port density of switches and the port bandwidth. The specific requirements are as follows:

— Switches with high-density ports. For example, 51.2T switches can support 64*800GE, 128*400GE, 256*200GE and 512*100GE.

— High port bandwidth of 400GE/800GE/1.6TE

RoCEv2 is most used in intelligent computing data center networks, and RDMA is highly sensitive to network packet loss. The efficiency of large model training depends on the inter-card communication with the lowest possible latency, which make network scale load balancing and non-blocking forwarding of the network more and more crucial. The following key technologies should be deployed:

— Ethernet enhancement, such as head suppression and link layer retransmission which can improve the transmission efficiency and reliability.

— Technologies for network load balance, such as NSLB which perceive the AI model training tasks and calculates the optimal traffic distribution by drawing a global traffic matrix and automatically diverts traffic through network devices. In addition, packet spraying can be used to provide packet-level load balancing with capability to solve the packet disorder problem.

— Congestion control technologies, such as ECN, CNP and adaptive routing techniques should be used.

— Fine flow control, such as PFC with enhanced capabilities of preventing and resolving dead-lock problem automatically. Besides, rPFC could be used to provide flow-level flow control.

As the scale of intelligent computing data center networks expands, the difficulty of manual network fault localization increases while efficiency gradually declines. The primary demand is to provide visualized and intelligent network fault demarcation and locating capabilities. In network M&C layer, controller, the following key technologies should be deployed:

— Digital twin, especially the specific capabilities, including:

- In-situ flow information telemetry (IFIT), which support precise E2E and hop-by-hop application SLA measurement, such as packet loss, latency and jitter, which are important for the AI task visualization and fault demarcation and location.
- AI task-level virtualization: The controller in intra-DC is required to visible the forwarding path and status between any two chips of an AI task, in order to reduce the fault locating time
- RDMA communication performance monitoring: The controller should support monitor indicators such as RDMA communication flow completion time and effective flow throughput
- Optical module health monitoring: The controller in intra-DC should support to detect the optical module health (e.g., smudge and loosing) before and during the AI training, in order enhanced the high reliability of the network
- High-precision packet loss detection and locating: The controller is required to support high-precision (e.g., 1‰) packet loss detection and direct query and locate the packet loss statistics caused by port convergence, ACL policy, route table query failure, etc.

8.1.2Inter-DC intelligence computing

Inter-DC refers to the interconnections of multiple data centers. The following figure shows inter-DC network architecture.

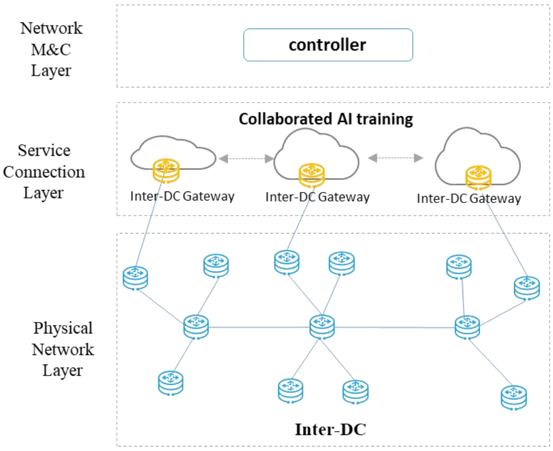

Figure 5 — Architecture of intra-DC

Inter-DC AI training is required to address the problems in single data center located in a specific area, e.g., the cards and network scope limits and the lack of electric energy. To support the maximum computing power usage rate, the network between distant data centers should also take high-rate bandwidth, lossless throughput, high reliability, and virtualization management into consider.

The traffic for AI large model training is typically characterized by fewer instances but larger data volumes. The inter-DC network should provide high-bandwidth, including:

- High bandwidth of 400GE/800GE

To support the inter-DC AI large model training, the network is also request to support lossless forwarding. The network load balancing and high throughput are also crucial. The following key technologies should be deployed:

- Segment Routing IPv6 (SRv6), especially programming capability to provide APN identifier for AI training tasks and other applications, then implement flexible tunnel scheduling for different services;
- Network slicing, which should be introduced to guarantee network resources for different network application if the applications require resource exclusivity;
- Technologies for network load balance, such as NSLB, which is also important in inter-DC network.
- Fine flow control, such as SPFC, which provides flow-level flow control in SRv6 networks for the AI large model training flows without impact on other active flows.

When AI training traffic is conducted across data centers, it requires the interaction of traffic awareness and control both within and between DCs. This functionality needs to be implemented by deploying a computing power gateway:

- Computing power gateway, especially gateways for inter-DC connections, which should be introduced to provide ultra-high bandwidth and long-distance lossless forwarding paths for collaborative training of multiple DCs.

The autonomous and intelligent management and operate of inter-DC network is also requested. In network M&C layer, controller, the following key technologies should be deployed:

- In-situ flow information telemetry (IFIT), which support precise E2E and hop-by-hop application SLA measurement, such as packet loss, latency and jitter, which are important for the AI task visualization and fault demarcation and location.
- Digital twin, especially the specific capabilities, including:
  - AI task-level virtualization: The controller in intra-DC is required to visible the forwarding path and status between any two chips of an AI task, in order to reduce the fault locating time
  - RDMA communication performance monitoring: The controller should support monitor indicators such as RDMA communication flow completion time and effective flow throughput

— Network digital map, which provides multidimensional dynamic topology, localization, network configuration verification, and service experience assurance capabilities through a map-based experience.

— Centralized path computing and re-optimization, especially NSLB for network scale load balancing, maximum-throughput multi-path algorithm to maximize the use of network bandwidth to transmit big flows traffic and flow-level scheduling used to match the flow to the optimal paths.

— Generative AI for AI copilot and AI agent for network change simulation, network re-optimization and network intelligent fault process.

The AI large model training traffic across different DCs have a long distant, the security capabilities of the devices and networks are crucial to protect the data privacy and security. The following key technologies in security should be deployed:

— Device internal security, to protect devices against network attacks.

— High security, especially IPsec, MACsec and PHYSEC to provide multi-layer data protection.

8.1.3Data-to-DC

Data-to-DC refers to the sample uploading and decoupled storage and compute between the user sites and cloud sites. The following figure shows data-to-DC network architecture.

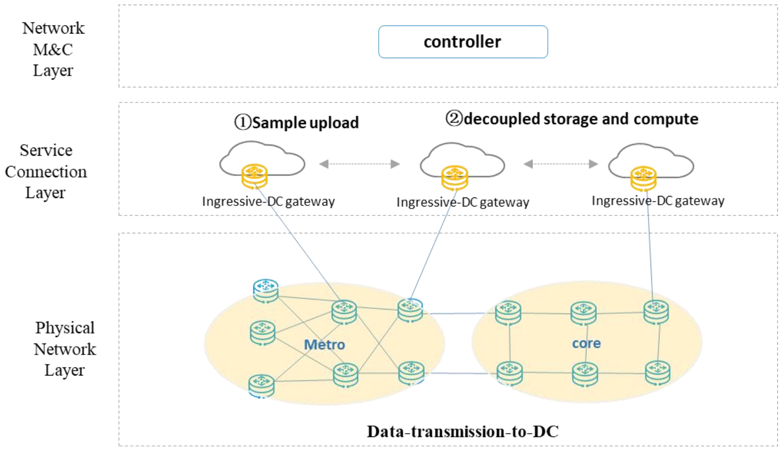

Figure 6 — Architecture of data-to-DC

The networks for data transmission to data center cover the metro and the core networks. The traffic volume of sample upload and decoupled storage and compute for AI training or inference could reach hundreds of GBs or even TBs, the networks are required to provide large bandwidth, low latency, high throughput, network resource isolation and high security.

The traffic volume in data-to-DC could be hundreds of GBs or even TBs, the network device should provide high bandwidth, including:

— High bandwidth of 400GE/800GE

These large data coming from different users and applications are usually require to be transmitted within specific time periods. The following key technologies in data-to-DC network should be deployed:

— Segment Routing IPv6 (SRv6), especially programming capability to provide APN identifier for applications demanding large data transmission and other applications, then implement flexible tunnel scheduling for different services;

— Network slicing, which should be introduced to guarantee network resources for different network application if the applications require resource exclusivity;

— Big flow recognition, which could be used to recognize the flows with large data transmission in order to implement specific flow scheduling.

The large data is required to be transmitted from the user side to the computing power centers, gateways for data-to-DC in both enterprise sides and computing power centers are necessary. In service connection layer, the following key technologies should be deployed:

— Computing power gateway, especially gateways for data-to-DCs, which provide protocol transform at transport layer, allocate identification and select proper tunnels and paths to data transmission to DCs, and support accounting and reconciliation capabilities.

Network of Data-to-DCs is also require autonomous and intelligent management and operation. In network M&C layer, controller, the following key technologies should be deployed:

— In-situ flow information telemetry (IFIT), which support precise E2E and hop-by-hop application SLA measurement, such as packet loss, latency and jitter, which are important for the AI task visualization and fault demarcation and location;

— Network digital map, which provides multidimensional dynamic topology, localization, network configuration verification, and service experience assurance capabilities through a map-based experience;

— Centralized path computing, re-optimization and scheduling, especially maximum-throughput multi-path algorithm to maximize the use of network bandwidth to transmit big flows traffic and flow-level scheduling used to match the flow to the optimal paths;

— Generative AI for AI copilot and AI agent for network change simulation, network re-optimization and network intelligent fault process.

The data for AI training usually need high security. Within a long distant from enterprise sides to cloud center, the security capabilities of the devices and networks are crucial to protect the data privacy and security. The following key technologies in security should be deployed:

— Device internal security to protect devices against network attacks;

— High security, especially IPsec, MACsec and PHYSEC to provide multi-layer data protection.

8.2 Intelligence connections

Intelligence connection refer to the scenarios of smart home, smart office and intelligent manufacturing. The following figure shows intelligence connection network architecture.

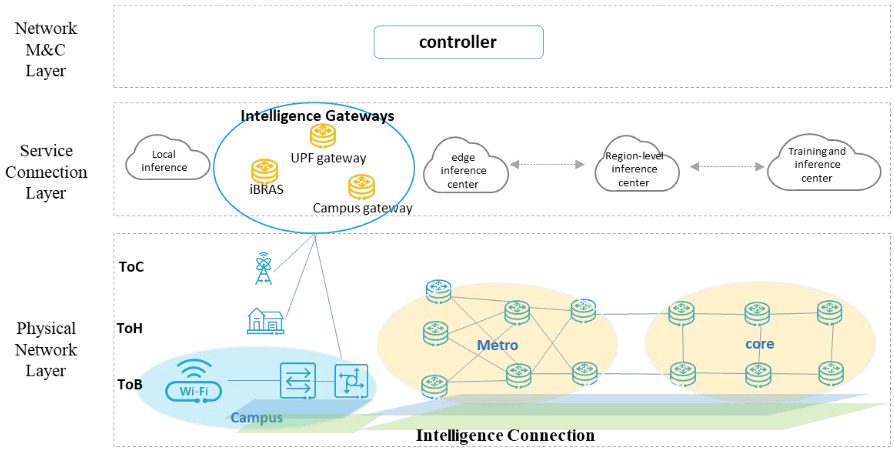

Figure 7 — Architecture of intelligence connection

The networks for connecting intelligence are used to connect various intelligent terminals, to meet the interaction requirements between intelligent terminals and inference centers, and implement the intelligent functions such as real-time decision-making. Whether personal, home or enterprise services, the intelligent terminals and device-cloud collaboration bring new requirements on networks, e.g., high-rate bandwidth, low latency, secure authentication, data encryption and application-level policy control. To ensure the virtual-reality convergence experience of services, the inference centers need to be deployed hierarchically. The intelligent gateways are required to have flexible orchestrated capabilities to meet various demands of intelligent agents, and implement quickly deployments and adjustment through management and control layer.

In order to transmit the increasing traffic volume generated by the intelligent terminal devices, the network should support high bandwidth. The following key technologies should be deployed:

- High-rate bandwidth of 400GE/800GE
- Evolution of Wi-Fi7, which improve the air bandwidth from 30G to 100G, and the technologies to continuously challenge the coverage distance of 10m to 15m under power consumption constraints. This is urgently needed in large uplink scenarios, such as industrial AOI backhaul and XR video local generation.

Different applications have different SLA requirements in bandwidth or latency. In order to protect the user experience, the network should identify the applications and give specific transmission policies. The following key technologies should be deployed:

- Segment Routing IPv6 (SRv6), especially programming capability to provide APN identifier for applications demanding different transmission requitements, then implement flexible tunnel scheduling for different applications to meet the requirements of latency or bandwidth;
- Network slicing, which should be introduced to guarantee network resources for different network application if the applications require resource exclusivity;
- Device-level native AI, which provide AI-based capabilities, including advanced precise application identification even encrypted application streams, energy saving, attack detection and so on.

To acquire the deterministic transmission for the different applications, an intelligent gateway is necessary, including:

- Intelligent gateways, which are various types of integrated service gateways tant can be deployed in campus, carrier edge network and data center to identify and control applications, serve as the control point for data encryption and security policies, and work as an edge inference node.

Network of intelligent connection also requires autonomous and intelligent management and operation. In network M&C layer, controller, the following key technologies should be deployed:

- In-situ flow information telemetry (IFIT), which support precise E2E and hop-by-hop application SLA measurement, such as packet loss, latency and jitter, which are important for the AI task visualization and fault demarcation and location;
- Digital twin, which is compose of multi-dimensional and high-precision network data to provide inference acknowledges and decision-making basis for the network new smart brain.
- Network digital map, which provides multidimensional dynamic topology, localization, network configuration verification, and service experience assurance capabilities through a map-based experience;
- Application-level experience guarantee, which is required to identify the applications, perceive the application status and implement close-loop policies to guarantee the application experience.
- Generative AI for AI copilot and AI agent for network change simulation, network re-optimization and network intelligent fault process.

The intelligence connection is usually composed of important and privacy data, which require high security. The key technologies in security should be deployed:

- Device internal security to protect devices against network attacks;
- High security, especially IPsec, MACsec and PHYSEC to provide multi-layer data protection;
- Cloud-network-edge-device integrated security, which is required to deal with more systematic and covert security attacks by technologies such as generative AI.

8.3 Data connections

Data connection refer to the network between data suppliers and data consumers in both private and public industries. The following figure shows data connection network architecture.

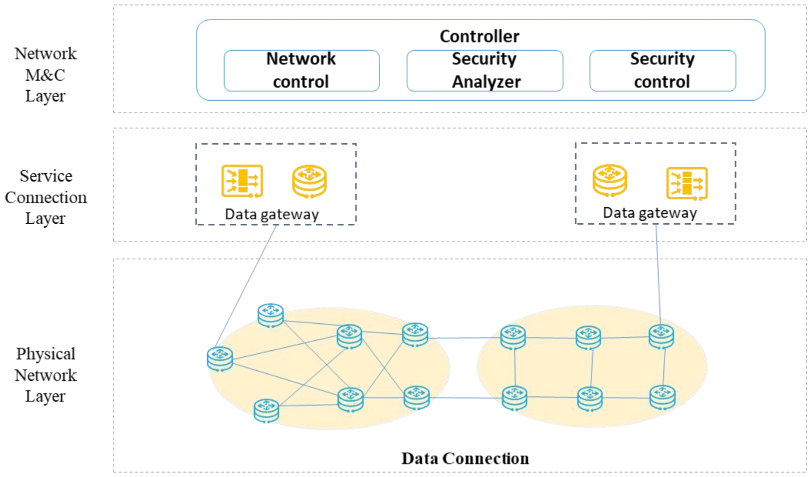

Figure 8 — Architecture of data connection

Data connection network is a network infrastructure that provides data suppliers and consumers with capabilities of trusted access, trusted transfer, alone-path usage control and cross-region/cross-country transfer control during the circulation of data elements. The data connection network is required to support end-to-end data element integrity, confidentiality and availability.

In physical device layer, the following key technologies should be deployed:

- — High-rate bandwidth of 400GE/800GE;
- — Evolution from Wi-Fi7, which improve the air bandwidth from 30G to 100G, and the technologies to continuously challenge the coverage distance of 10m to 15m under power consumption constraints. This is urgently needed in large uplink scenarios, such as industrial AOI backhaul and XR video local generation.

In physical network layer, the following key technologies should be deployed:

- — Segment Routing IPv6 (SRv6), especially programming capability to provide APN identifier for applications demanding different transmission requitements, then implement flexible tunnel scheduling for different applications to meet the requirements of latency or bandwidth;
- — Network slicing, which should be introduced to guarantee network resources for different network application if the applications require resource exclusivity;
- — Device-level native AI, especially for security detection and analysis.

In order to provide reliable data access in multiple scenarios, a data gateway is necessary, including:

— Data gateways, which should be used to provide multi–scenario capabilities for data proximity and trusted access, APN network collaboration, and data fencing for both data suppliers and consumers.

Network of data connection also requires autonomous and intelligent management and operation. In network M&C layer, controller, the following key technologies should be deployed:

— In–situ flow information telemetry (IFIT), which support precise E2E and hop–by–hop application SLA measurement, such as packet loss, latency and jitter, which are important for the AI task visualization and fault demarcation and location;

— Digital twin, which is compose of multi–dimensional and high–precision network data to provide inference acknowledges and decision–making basis for the network new smart brain.

— Network digital map, which provides multidimensional dynamic topology, localization, network configuration verification, and service experience assurance capabilities through a map–based experience;

— Centralized path computing, re–optimization and scheduling, especially cloud–map algorithm.

— Generative AI for AI copilot and AI agent for network change simulation, network re–optimization and network intelligent fault process.

The data connection is usually composed of important and well–defined data, which require high security. At the same time, the security analysis and security control capabilities are increasing important in data transmission to reduce risk and enhance control. The key technologies in security should be deployed:

— Device internal security to protect devices against network attacks

— High security, such as IPsec, MACsec and PHYSEC to provide multi–layer data protection, and especially quantum security to provide high–level security protection.

— Trusted data transmission, especially trusted access, trusted paths, encryption transmission;

— Security analysis on controller, which is used to provide security analysis capabilities, e.g., traceable compliance audit and accurate threat detection;

— Security control on controller, which is required to implement unified security authentication, intelligent security policy control, e.g., trusted path, encryption and data fence.

Bibliography

[1] WBBA - White paper (2024), *Network Evolution For the 5.5G And 6G Era*

[2] ITU-T-Y.LDT, *IMT-2020 Networks and Beyond: Requirements and Functions for applications demanding large data transmissions*

[3] IEEE-802.11be, *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment: Enhancements for Extremely High Throughput (EHT)*

[4] IETF-RFC-9256, *Segment Routing Policy Architecture*

[5] IETF-draft-apn-ipv6-encap, *Application-aware IPv6 Networking (APN6) Encapsulation*

[6] IETF-RFC-9341, *Alternate-Marking Method*

[7] IEEE 802.11bb, *Priority-based Flow Contro*