NIDA-TS-004-2025

NIDA/TC/WG03/TF05

Date: 2025–07–17

Technical Requirements for Financial Data Center Network Construction

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**.

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 WG03, *Datacenter Network Work Group*, Task Force TF05.

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

This document is drafted by: Huawei Technologies Co., Ltd, Industrial and Commercial Bank of China, Ping An Technology (Shenzhen) Co., Ltd, Internet Association of Kazakhstan.

Main drafters of this document: Li Zhang, Jiuyong Li, Xueshan Yu, Zurui Meng, Shavkat Sabirov.

Technical Requirements for Financial data center network

1 Scope

This document specifies the technical requirements for financial data center network, including the network architecture, networking scenarios, and key technologies of the data center network.
This document is applicable to the network construction planning and technical evaluation of the financial data center network, and provides technical basis for the network test.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp
— IEC Electropedia: available at https://www.electropedia.org/

3.1

Data plane fast recovery

Data plane fast recovery in this document refers to a sub–millisecond fault recovery technology. It provides data–plane fast fault detection and convergence capabilities for hardware faults, such as optical module faults and looseness of transmission optical cables, improving device–level reliability.

3.2

Data plane crossing fault

Data plane crossing fault in this document refers to an automatic fault recovery technology. It is used to solve session–level service exceptions caused by silent faults, such as link faults, forwarding entry exceptions, forwarding component exceptions, physical ports that are Up but cannot forward traffic, and configuration errors. Provides fast fault detection and convergence capabilities based on the data plane, achieving path switchover within seconds and improving network reliability.

4 Abbreviations

The following abbreviations apply to this document.

BFD: Bidirectional Forwarding Detectio

BGP: Border Gateway Protocol

CI/CD: Continuous Integration and Continuous Delivery

CLOS: Clos Architecture

DC: Data Center

DPCF: Data Plane Crossing Fault

DPFR: Data Plane Fast Recovery

ECMP: Equal–Cost Multi–Path

ECN: Explicit Congestion Notification

FC: Fibre Channel

FCT: Flow Completion Time

FET: Flow Efficient Throughput

FNR: Flow NAK Rate

HDD: Hard Disk Drive

IOPS: Input/output Operations Per Second

LLM: Large Language Model

MDC: Multi–Datacenter–Controller

M–LAG: Multi–chassis link aggregation group

NOF: NVMe over Fabric

NVMe: Non–Volatile Memory express

OSPF: Open Shortest Path First

PFC: Priority–based Flow Control

POD: Point of Delivery

RDMA: Remote Direct Memory Access

RoCE: RDMA over Converged Ethernet

SLA: Service Level Agreement

SSD: Solid–State Drive

SZTP: Secure Zero Touch Provisioning

5 Background of Financial data center network Standards

5.1 Service evolution

In the context of digital transformation, digital upgrades based on technologies such as 5G, big data, artificial intelligence, and the Internet of Things have covered all industries. The new services of Smart Connect of Everything pose new requirements on enterprise data centers. Take the financial industry as an example. Digital finance has been a new business form that uses digital technologies and Internet platforms, as well as the digitalization, intelligence, and online of services and products to provide customers with more convenient, efficient, secure, and personalized financial services and products. With the strengthening of AI technologies, digital finance will provide more accurate financial services, improve operational efficiency and change business models.

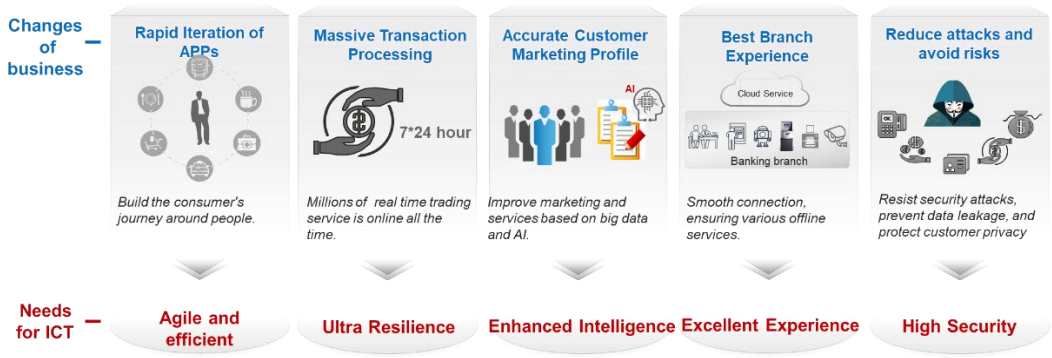

Figure 1 — Digital finance scenarios and requirements

The new requirements of the financial industry data center in the digital finance era include service automation and efficient openness, ultra–reliability, high performance, intelligent O&M, and full–lifecycle security, which have the following connotations:

- The financial industry provides customers with all–round financial services by building platforms and APPs in data center. Apps are updated frequently, and services must be deployed as soon as possible. Traditional manual configuration mode is time–consuming and error–prone. Service deployment across multi–vendor devices and multi–service technologies has high requirements on O&M personnel. The data center network needs to provide efficient and open automatic service deployment capabilities, build an E2E service automation architecture that includes the O&M system, management and control system, and

interconnection of devices from multiple vendors, and realize the end-to-end service deployment automation in service scenarios such as underlay networks, overlay services, and network-security integration.

- In the 5G mobile payment era, the financial industry faces customer access requests 24 hours a day. The number of daily customer visits and the average daily transaction service calls are more than 100 million. This requires data center network to provide 7*24 ultra-reliable service and ultra-fast response rate.
- The financial industry has a wide range of requirements for big data and AI technologies, which are used to enable customer digital profiling and credit risk analysis, financial fraud detection and market risk prediction, product and service model innovation, digital services and operations, etc. Financial data centers are growing in computing power and storage scale. With the evolution of large language models and storage technologies, large-bandwidth and lossless networks are required for data centers to ensure no-packet-loss and low-latency transmission of intelligent computing, model training and inference, and high-speed storage.
- Customers in the financial industry have higher requirements on experience in the digital era. Continuous and rapid service requirements pose challenges to the timeliness of network troubleshooting. In the future, with the further expansion of financial data centers, AI, digital twins, and expert experience technologies are required to implement self-awareness, self-analysis, self-healing, and self-optimization of networks, enabling intelligent collaboration, dynamic optimization, and interactive innovation of data, technologies, processes, and organizations.
- As financial services become increasingly complex in life scenarios, the security protection chain becomes increasingly complex. At the same time, the external information security situation is grim, facing more and more complex security threats. Financial business has natural high security requirements. data center network should build a comprehensive and multi-dimensional security assurance system to prevent data leakage and protect customer assets and privacy.

5.2 Technology evolution

5.2.1 Automatic service deployment

The improvement of service deployment efficiency and accuracy in financial data center effectively enhances customer experience and reduces network O&M OPEX. Build an automatic interactive ICT system that integrates the resource system, service orchestration system, work order system, service collaboration and deployment system, and service detection system to automate the entire

process of service planning, resource allocation, configuration change approval, service deployment, service detection, and service visualization. It has become a key requirement for financial data center network operation and management.

Data center services are usually deployed across PODs or even across DCs, involving devices from multiple vendors and frequent service changes. Service automation must have the capabilities of agile development, continuous evolution, openness and programmability, and usability. With the evolution of CI/CD–related technologies, such as workflow orchestration and plug–and–play (PnP) of multi–vendor devices, it provides technical support for the integration of multiple systems to build an automatic ICT system.

5.2.2 Intelligent computing

With the breakthrough development of AI large language models, training and reinforcement training of large–scale models in the financial industry are also evolving. High–performance computing is the core requirement of intelligent computing data centers. Professional high–performance chips such as GPU, FPGA, and ASIC are used to support a large number of floating–point and parallel computing requirements. Although the improvement of AI training algorithms reduces the dependence on the scale of the intelligent computing cluster to some extent, the computing power improvement rate of single–chip still cannot meet the computing power requirements of the increasing number of parameters of large model training. The scaling–law principle is still applicable, and the scale of the intelligent computing cluster is still expanding.

At the same time, the enhancement of model inference capabilities also drives the deployment of inference–based applications in the financial industry, and the demand for computing resources in model inference is expanding. Improving computing efficiency is the key to intelligent computing clusters with high investment.

RoCEv2 (RDMA over Converged Ethernet Version 2) has become the first choice for intelligent computing networks. It effectively reduces network delay but is sensitive to packet loss, which poses higher requirements on intelligent computing networks, requiring large bandwidth, lossless, stable latency, and higher network throughput to improve communication efficiency.

5.2.3 Storage

The digitalization of the financial industry will make data centers face hundreds of billions of transactions and payment operations performed by tens of millions of users through apps every day.

The digitalization of the financial industry will make data centers face hundreds of billions of transactions and payment operations performed by tens of millions of users through apps every day. This requires parallel and fast data processing capabilities. As the infrastructure of Artificial Intelligent applications and services, storage systems need to provide over 50 GB bandwidth and over 1 million IOPS performance. Due to the maturity of the flash technology, SDDs improve storage density and performance, and gradually reduce costs. Therefore, SDDs gradually replace HDDs and gain wide application. The Non–Volatile Memory express (NVMe) provides an efficient communication mechanism for SSDs to fully utilize the performance advantages of SSDs and improve read/write speed, latency, and more.

The storage network that works with SSDs and NVMe also needs to support higher performance and lossless forwarding. Traditional FC networks face technical bottlenecks of bandwidth growth and are separated from other areas of the data center, which complicates O&M. Ethernet has a good technical accumulation in large bandwidth and lossless transmission, and has a complete supply chain. All–flash storage and NoF storage network based on NVMe over high–speed Ethernet will be the main development direction of financial data centers in the future.

5.2.4 AI for network

The mutual access relationships between online financial services are complex. To complete a service, multiple service system may be involved. E2E real–time awareness of service quality and segment–by–segment demarcation and visualization of different service systems have been new requirement for visualized O&M of financial data centers. With the emergence of distributed and big data services, the proportion of east–west traffic in data centers increases. Distributed traffic is prone to microbursts, which are difficult to detect by traditional monitoring methods. In addition, the distributed architecture and intelligent computing bring multiple devices. Stable service running requires that potential network risks can be identified in advance and faults can be quickly located. This brings new challenges to network O&M.

To address the preceding challenges, data center O&M is in urgent need of transforming from manual traditional O&M to intelligent network O&M. With the development of technical capabilities, such as Telemetry second–level collection, digital twin, knowledge graph, AI, Co–pilot, and Agent, more efficient, automated, and intelligent network O&M is possible.

5.3 Networking Evolution

To meet the service development requirements brought by the digital transformation of the financial industry, multiple data centers are required to deploy services. In addition, data security

and service reliability and continuity are increasingly valued. Backup and disaster recovery are becoming a common requirement. Multiple data centers must be constructed to solve the disaster recovery and backup problem.

With the development of 5G, cloud computing, and big data, virtualization and resource pooling become the mainstream requirements. Cross–region and cross–DC resources need to be integrated to form a unified resource pool. In addition, service systems are deployed in multiple DCs in a distributed manner to provide services nearby to improve user experience. Distributed multiple data centers are the mainstream solution now.

To meet the continuous innovation requirements of financial data centers, the data center architecture must be flexible, scalable, and evolvable. The design of the data center network must meet the development requirements of computing, storage, and large data centers, take into account the technology evolution in a certain period, and support the evolution of distributed and cloud–based data centers.

5.3.1 Multi–POD

Multiple PODs are deployed in a single DC. Proper service POD division improves the computing and storage resource pooling and sharing capabilities, and ensures the scalability of the DC and the elastic scaling capability of each POD. Deploy security policies between PODs to implement security protection and fault isolation at different layers. Services can be flexibly backed up and migrated between PODs to achieve load balancing and high reliability. To meet the continuous improvement of service capacity and performance in the future, the data center must reserve appropriate bandwidth capacity and reliable access settings.

5.3.2 Multi–DC

With the increasing digitalization of financial services, data centers have high security and reliability requirements. Data centers must continuously ensure reliability, meet the requirements for 7 x 24 network operation, and meet the requirements for high reliability of systems and applications deployed in the same or across data centers. Therefore, data center construction must support multi–DC scenarios, including:

- Intra–city active–active DCs

Two service systems running at the same time are deployed in two DCs in the same city. The security policies of the same subsystem are the same and the same services are provided. This scenario provides double service capabilities and supports real–time DR and takeover of each

other. Full redundancy is implemented at the application processing layer, greatly improving service continuity and reliability and preventing users from sensing faults.

- Geo–Redundant DCs

The remote DR center is the backup data center of the two active data centers in the same–city active–active solution. It is used to back up data, configurations, and services of the active data center. When the master active–active data centers are faulty due to natural disasters, the remote disaster recovery data center can quickly recover data and applications to ensure normal service running and minimize losses caused by disasters.

5.4 Policy Constraints

The financial industry is of great significance to the country and plays an important role in promoting economic development, optimizing resource allocation, stabilizing the economy, and promoting innovation and cooperation. Countries have also issued corresponding policies to ensure the construction of data centers in the financial industry. For example, the intelligent O&M mechanism for financial data centers should be established and improved, and multi–scenario collaboration and multi–sites integrated management and control should be strengthened to improve site awareness, exception detection, and fault prediction capabilities and reduce manual operation risks.

6 Financial Data Center Network Construction Roadmap

6.1 Single DC and multiple PODs

6.1.1 Overall Architecture

A single DC is divided into multiple PODs to carry different services. PODs are connected by the switching core POD, as shown in Figure 2.

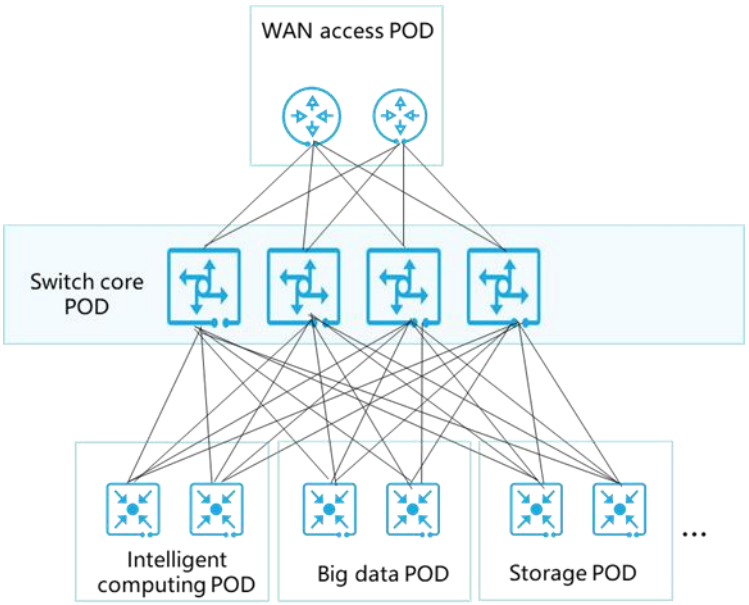

Figure 2 — Single–DC network construction architecture

6.1.2 Switching core POD

The switching core POD is the core of the financial data center network. It connects all the functional PODs in the data center and works as the switching bus of the data center.

- Core devices are deployed in the switching core POD independently. The spine–leaf architecture ensures the scalability of the switching core POD and allows flexible addition or deletion of functional PODs.
- The bandwidth of the switching core POD must be scalable and reserved for future service development. Currently, 200G/400G is the mainstream, and 800G is a new trend in large financial data centers.
- The switching core POD must support the establishment of a Layer 3 routing control plane to ensure network reliability through fast convergence of standard routing protocols.

6.1.3 Intelligent computing POD

An intelligent computing POD is a partition that performs AI model training. It is usually divided into three planes: service plane, parameter plane, and management plane, as shown in Figure 3.

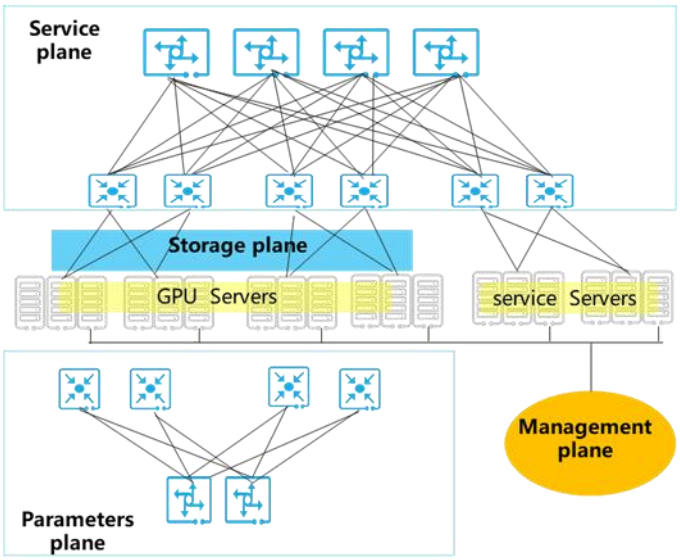

Figure 3 — Intelligent computing POD network construction architecture

In intelligent computing POD, the communication efficiency between GPUs is critical. RDMA is widely used for data transmission and memory access, and the Intelligent computing networks are usually constructed based on the RoCEv2 protocol. RoCEv2–based AI training is sensitive to network transmission indicators such as network delay, packet loss, and jitter. The packet loss rate of 1 per 10,000 decreases the training efficiency by 50%. RoCEv2 does not provide a comprehensive packet loss protection mechanism. To improve the computing power availability of intelligent computing clusters, high–performance, large–bandwidth, and lossless network are required. The requirements for building an intelligent computing POD are as follows:

- The scale of intelligent computing PODs can be flexibly adjusted based on service requirements. High network availability and scalability are required. The CLOS spine–leaf architecture is used.
- The network bandwidth of intelligent computing PODs is increasing driven by the increasing of AI model scales and GPU bandwidth. Currently, 200G/400G is the mainstream, and 800G is a new trend in order to support large scale AI training.
- A non–blocking network in intelligent computing PODs is needed. Traditional ECMP is insufficient to support lossless network transmission. More efficient load balancing is required to steering a small number of high–bandwidth AI flows. Additionally, more efficient flow control and congress control are needed.
- The complexity of network operation and management in the intelligent computing POD is creasing with the network scale. It's required to provide RDMA–based visualization in order to simplify and shorten the fault detection and analysis time.

The key technologies of intelligent computing networks are as follows:

PFC Priority Flow Control
PFC deadlock monitoring
PFC Deadlock Prevention
ECN
Network scale load balancing
RDMA–Based Visualized O&M

6.1.4 Big data POD

The big data POD is used to process large–scale distributed data computing tasks. The big data POD in the financial industry data center is used to calculate and analyze customer information, including financial behavior, preference, and risk analysis for account opening, transaction, and transfer. It can also be used to calculate and analyze financial organization information, such as business processing process improvement points, resource allocation optimization, and financial data. Figure 4 shows the big data POD network architecture.

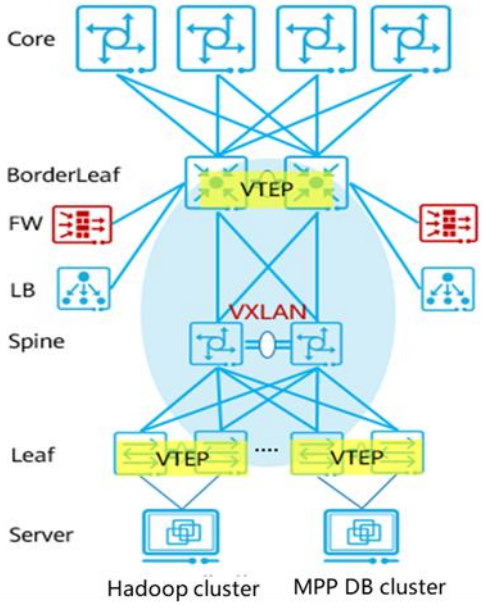

Figure 4 — Big Data POD network architecture

The requirements for building a big data POD are as follows:

- The scale of big data PODs can be flexibly adjusted based on service requirements. High network availability and scalability are required. The spine–leaf architecture is used to support ECMP.

- Flexible server migration is required, and network should support automatic service configuration and deletion.
- The big data POD processes a large amount of data input and output. The network must have sufficient access bandwidth, for example, 10G access bandwidth and 100G/200G aggregation bandwidth.
- Big data computing services, especially services related to customers' mobile applications, require fast response. Therefore, the network must ensure low latency and no congestion.
- Big data computing involves key data processing and requires higher reliability. Therefore, the network must support fast fault switchover, fast fault locating and recovery, and fault prediction. In addition, the network should provide service–level visualized O&M capabilities to display service performance indicators and path changes in real time.

The key technologies of the big data POD include:

Service automation
ECN
Overlay ECN in VXLAN
M–LAG reliability
Intelligent service O&M

6.1.5 Storage POD

Storage PODs are used for data storage and management. The financial data center should provide high–performance, reliable, and available data storage services to ensure data security and persistence and meet data storage requirements of different applications. Figure 5 shows the storage POD network architecture.

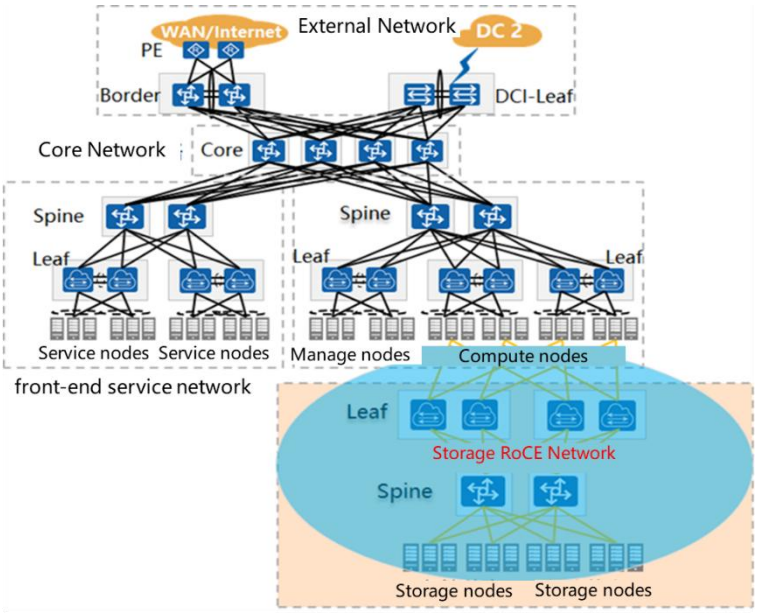

Figure 5 — Network construction architecture of storage POD

A storage network refers to the communication network used by computing servers to access storage data. Generally, the network is an independent physical network. With the popularization of all–flash storage, the NVMe over fabric become the main option. In addition, in order to support mixed application of TCP and RoCE protocol and to meet the requirement of unified data center network technology, the Ethernet–based NOF solution is preferred. NVMe over Ethernet requires high–performance, lossless, reliable, easy–to use, and unified O&M network capabilities. The requirements for building a storage POD are as follows:

- The scale of storage PODs can be flexibly adjusted based on service requirements. High network availability and scalability are required. The spine–leaf architecture is widely used.
- The storage network is used to carry RoCE data, which is sensitive to packet loss. Efficient flow control and congestion control capabilities must be supported to ensure lossless.
- The reliability of the storage network has high requirement according to the importance of key information storage in financial data center. Multi–level reliability should be implemented. In terms of link–level reliability, based on the local redundant links, fast link fault detection and switchover are the key requirements. In terms of device–level reliability, the fast fault detection and switchover both in local redundant links and remote devices are required. In terms of network–level reliability, the silent fault should be detected and support network scale convergence.
- To store a large amount of data, a storage system usually needs to manage a large number of hosts, and allow new hosts to access the network dynamically. The storage network is required

to quickly detect and manage newly added hosts and intelligently adjust storage network configurations.

- As the storage network scale expands, network O&M becomes more complex. Intelligent network O&M becomes necessary.

The key technologies of the storage POD include:

AI–ECN
PFC
M–LAG
Data plane fast recovery
Data plane crossing fault
Storage host plug–and–play
Intelligent O&M

6.2 Multi–DC Deployment

The particularity of the financial industry requires data centers to provide ultra–high reliability. Multi–DC deployment solutions, such as intra–city DC active–active or backup, and geo–redundant three–center deployment solutions, become mandatory. Figure 6 shows the multi–DC deployment architecture.

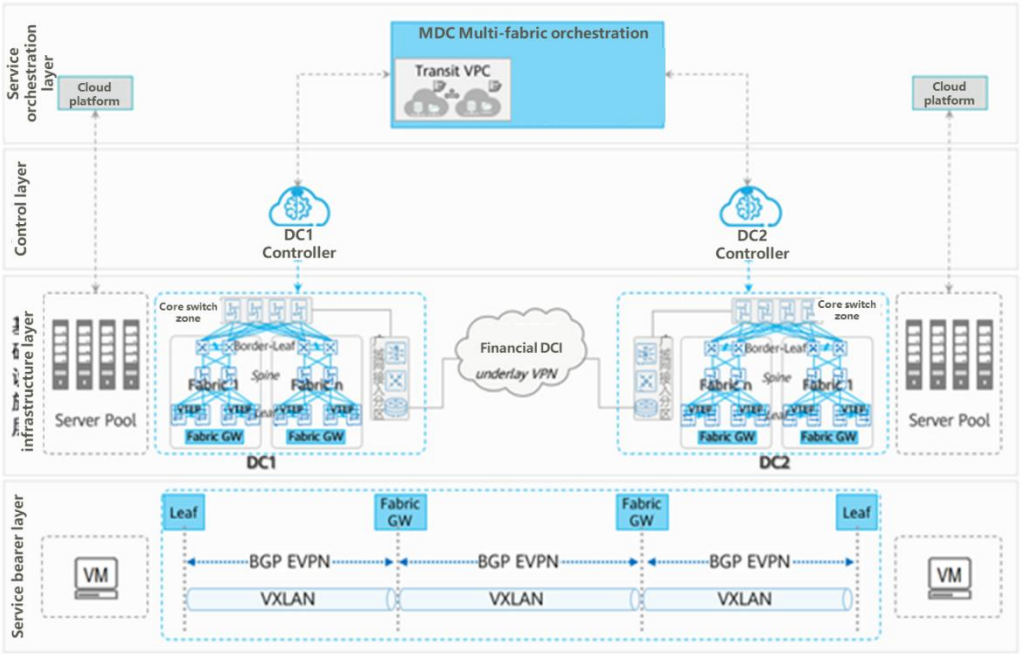

Figure 6 — Multi-DC network architecture

The MDC is connected to the controllers of multiple DCs to implement automatic deployment of cross–DC service interworking, flexible security policy control, and cross–DC network O&M.

- Intra–city active–active DCs: The same service system is usually deployed in the intra–city active–active DC to provide services externally.
- Geo–redundant DCs: The same service systems are deployed in the local and remote DCs. In normal cases, the service system in remote DC does not provide services externally and is used only for backup. The following figure shows the service interconnection requirements of geo–redundant three–centers:

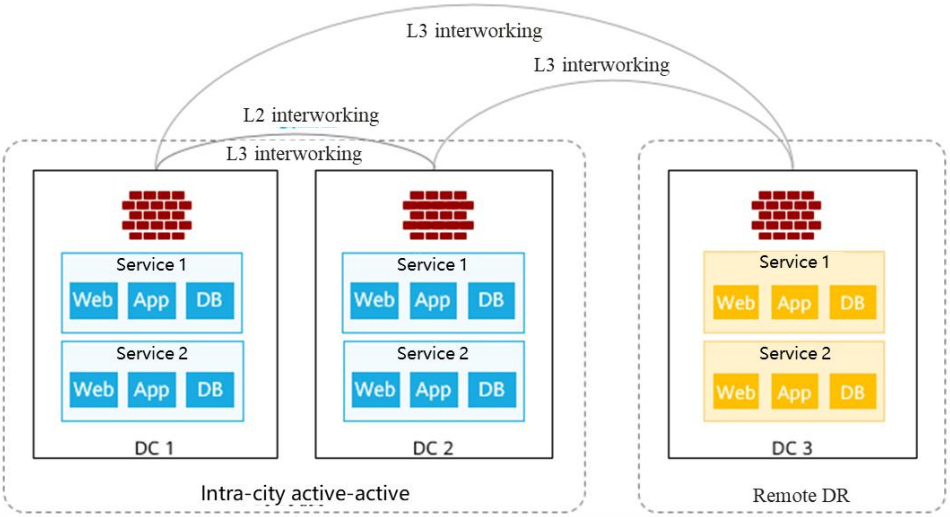

Figure 7 — Multi–DC network architecture

Multi–DC is required to support high–reliability of financial data centers. In order to achieve the goal of application and data backup and recovery, the multi–DC should provide the following capabilities:

- The multi–DC status detection and switchover mechanism needs to provide higher reliability.
- Automatic cross–DC service deployment is required to improve service deployment efficiency and reduce the configuration error probability.
- The storage in different DCs need to replicate data for use of backup. Generally, there is a long distance between DCs, so that the flow control and congestion control capabilities that support long distance need to be implemented to ensure lossless.
- In the scenario of cross–DC big data which need to collaborate with each other to process services, high–bandwidth is required.
- Cross–DC security is required to protect the data transferred between different security zones.

The key technologies of the multi–DC include:

Intra–city active–active DCs
Geo–Redundant DCs
Advanced long–distance PFC
Multi–DC O&M
Multi–DC Security

7 Key Technical Requirements for a Single DC Network

7.1 Network automation

7.1.1Automatic service deployment

The network automation technology is used to provide the automatic deployment capability of network services, improving the service deployment efficiency. The technical requirements for network automation include:

- Support automatic deployment of underlay services, including but not limited to IGP, BGP, Eth–Trunk, DHCP, VLANIF, STP instances;
- Support automatic deployment of overlay services, including but not limited VXLAN, EVPN;
- Support automatic deployment of network–security convergency services, including both network configuration and security configuration;
- Support automatic deployment of heterogeneous network services, including multi–protocol interworking configuration.

7.1.2 Service flow orchestration

Service flow orchestration technology is a network service–oriented capability. Through the service customization programming function, customers who do not have professional development capabilities can quickly and flexibly orchestrate network operations to apply service knowledge to networks to adapt to rapid service changes. Service flow orchestration needs to streamline the boundaries of systems and allow service objects to flow between systems. The technical requirements for service flow orchestration include:

- Service flow orchestration supports the design–state functionalities, including:
  - Support action management: An action is a basic element of a process. It contains the action model description and action design description, which describe the input and

output attribute data models required by an action during execution and describe how an action is executed, such as do, retry, rollback, and dry–run;

- Support process design and management: Multiple atomic interfaces can be orchestrated into a process, and processes can be used independently or in nested mode;
- Support the drag–and–drop UI capability to simplify the design;
- Support modeling of E2E network capabilities instead of single–site functions and features, and orchestrates model–based network capabilities into service workflows;
- Support model–driven abstract modeling of network capabilities, shields NE implementation, and adapts to differences between forwarding protocols and vendor implementations;
- Support necessary service resource management, automatic allocation, and recycling.

– Service flow orchestration supports the running state functionalities, including:
  - Support automatic generation of APIs and integration with other ITSM platforms by describing network service functions instead of specific implementation;
  - Support workflow statistics management, including execution records and status maintenance;
  - Support automatic service processing, including decomposition of service configurations, dry–run after decomposition, and simulation verification of configuration changes;
  - Support automatic delivery of service configurations and consistency check;
  - Support automatic service rollback when service deployment fails and automatically check of configuration consistency after rollback.

7.1.3 Configuration Simulation Verification

The simulation verification technology is used to evaluate and verify the impact of configuration changes in advance, intercept incorrect network configurations, and help experts verify the effectiveness of policies and reduce implementation risks on the live network. The technical requirements for simulation verification include:

– Support multi–branch simulation verification. The optimal solution can be selected based on the impact of different configuration changes in each branch;
– Support configuration change risk analysis and provides risk assessment from multiple perspectives, including but not limited to routing, tunnel, forwarding;
– Support multi–scenario simulation, including but not limited to protocol simulation, network connectivity simulation, and reliability simulation;
– Support configuration optimization suggestions based on risk assessment results.

7.2 Openness

Financial data center services involve inheritance and interworking between multiple systems, and are usually constructed across devices from multiple vendors. The data center network management and control system must be open and programmable in both northbound and southbound.

7.2.1 Northbound Openness

The financial data center network management and control system supports northbound openness to implement interconnection between multiple systems. The technical requirements for northbound openness include:

- Support northbound GUI capability to support human–machine interaction;
- Support northbound API gateway to support machine–machine interaction and invoking between different systems;
- Support northbound interface protocols, including but not limit to RESTFUL, SNMP, FTP.

7.2.2 Southbound Openness

The financial data center network management and control system supports southbound openness capabilities to interconnect devices from multiple vendors and implement cross–device network and service management and O&M. The technical requirements for southbound openness include:

- Support simplified adaptation of multi–vendor devices to quickly support multi–vendor device management and O&M.
- Support out–of–the–box use of multi–vendor devices.
- Support southbound interface protocols, including but not limit to RESTCONF, SNMP, Telemetry, FTP.

7.3 High network performance

Based on the RoCEv2 protocol, intelligent lossless network, implementing the convergence of intelligent computing and storage networks, should integrate advanced flow control, congestion control and load balancing technologies to transmit traffic over Ethernet with zero packet loss, low latency, and high throughput.

Flow Control and congestion control must be used together to solve network congestion. The difference between flow control and congestion control is: Flow control is end–to–end, and the transmission rate of the sending endpoint needs to be suppressed so that the receiving endpoint

can receive the data in time. Congestion control is a global process, involving all hosts, network devices, and all factors related to reducing network transmission performance.

7.3.1 Flow control technologies

The flow control technologies work through suppressing the sending rate at the sending endpoint so that the receiving endpoint can receive the data in time. The technical requirements of flow control include:

- support PFC (Priority–based Flow Control) technology, including:
  - Support priority–based flow control. PFC is a flow control mechanism defined in the IEEE 802.1Qbb standard, designed to prevent frame loss in Ethernet networks by enabling per–priority pause control. The loss–sensitive traffic like RDMA in AI or storage scenarios, which has specific service classes indicating specific service priorities, can be individually controlled;
  - Support PFC deadlock prevention to reduce the probability of PFC storms, which will block the traffic;
  - Support automatic recovery of PFC deadlocks, eliminates PFC deadlock loops, releases cache dependency, and resolves PFC storm problems;
  - Support the setting of the PFC threshold.
- support advanced long–distance PFC technology, including:
  - Support short–period, high–frequency, and a small amount of traffic adjustment and pause mechanism to implement long–distance lossless transmission compared with PFC when the buffer size and bandwidth are fixed;
  - Support periodically scans the buffer usage of interface priority queues. By sending PFC backpressure frames to the upstream device to control the duration for the upstream device to stop traffic in each period. In this way, the traffic transmission and suspension can be adjusted continuously;

7.3.2 Congestion technologies

Congestion control is a global process. It aims to ensure that the network can bear the existing network load. Generally, the forwarding device, traffic sending endpoint, and traffic receiving endpoint need to cooperate with each other. In addition, the congestion feedback mechanism in the network is used to adjust the traffic on the entire network to relieve congestion. The technical requirements of congestion control include:

- Support ECN (Explicit Congestion Notification) technology, including:

- Support ECN work together with PFC. The ECN threshold should be properly set together with the PFC threshold, so that the buffer space between the ECN threshold and the PFC threshold can accommodate the traffic sent by the application in the period from the ECN congestion mark moment to the application rate reduction moment in order to avoid triggering network PFC flow control;
- Support ECN work for both delay–sensitive small flows and throughput–sensitive large flows on the network. The ECN threshold should be properly set in order to meet the bandwidth requirements for throughput–sensitive large flows with high traffic sending rate and buffer space for absorbing burst traffic in the queue, at the same time meet the delay requirements for delay–sensitive small flows with instantly triggering of ECN congestion flag to notify applications to reduce the rate;
- Support ECN mechanism. Forwarding devices can mark ECN congestion in IP packets. The traffic receiving endpoint detects network congestion based on the ECN field in IP packets. If the network congestion is detected, packets with ECN congestion information are sent to notify the traffic sending endpoint to reduces the traffic sending rate.

– Support AI ECN (Artificial Intelligence Explicit Congestion Notification) technology, including:
  - Support intelligently adjusts the ECN threshold for lossless queues based on the traffic model on the live network, ensuring low latency and high throughput in the case of zero packet loss and achieving optimal performance for lossless services;
  - Support the AI ECN intelligent algorithm, performs AI training based on the live network traffic model, predicts network traffic changes, and infers the optimal ECN threshold in a timely manner. In addition, the ECN threshold can be adjusted in real time based on traffic changes on the live network, and the lossless queue buffer can be accurately managed and controlled to ensure the optimal performance of the entire network;
  - Support devices to collect traffic characteristics on the live network and send the collected data to the AI ECN component;
  - Support AI ECN to be used together with the queue scheduling technology. The AI ECN function of lossless queues can implement hybrid scheduling of TCP traffic and RoCEv2 traffic on the network, ensuring lossless transmission of RoCEv2 traffic and achieving low latency and high throughput.
– Support overlay ECN technology, including:
  - Support the overlay ECN technology in VXLAN networks to map the ECN field in IP packets to VXLAN packets and transmit the congestion status to the traffic receiving end.

In this way, VXLAN network congestion is relieved in a timely manner and network performance is maximized;

- Support AI ECN in hardware network overlay, software vSwitch overlay, and hybrid software and hardware overlay scenarios;
- Support overlay ECN processing when congestion occurs on devices entering a VXLAN tunnel;
- Support overlay ECN processing when congestion occurs on the forwarding device of a VXLAN tunnel;
- Support overlay ECN processing when congestion occurs on the device that exits the VXLAN tunnel.

7.3.3 Load balancing

Load balancing is the prerequisite for achieving high network throughput. The traditional static load balancing mode based on 5–tuple–based (source IP address, destination IP address, source port, destination port, and protocol) has disadvantages, which will lead to unevenly traffic distribution among links and decreasing of network throughput. Advanced load balancing should support network–scale load balancing and network throughput of more than 90%. The technical requirements of load balancing include:

- Support the network–scale load balancing function, including:
  - Support perception of service traffic models, such as access devices and traffic models involved in AI training tasks on intelligent computing networks；
  - Support unified path computation based on the traffic model and network resource status to solve traffic congestion caused by unbalanced local and global traffic load and improve network throughput.
- Support packet–based load balancing technology, which implements packet–spray to the network links and solves the problem of out–of–order packet reassembly.

7.4 High reliability

The RPO (Recovery Point Objective) and RTO (Recovery Time Objectives) have strict requirements in finical data center. The rate and type of network fault detection and rapid network fault switching are critical to improving network reliability. Multi–level reliability technologies should be implemented to build a high–reliability network system.

7.4.1 Link-level reliability

Link-level reliability aims to implement link-level protection and fast link fault detection and switchover. The technical requirements of link-level reliability include:

- support M-LAG (Multi-chassis Link Aggregation Group) technology, including:
  - Support M-LAG, which aggregates ports on different devices into a logical interface. Even if a device fails or one of the aggregated links fails, the aggregated links will not fail completely. This ensures reliable data transmission;
  - Support M-LAG active-active mode, which implements load balancing traffic forwarding and backup protection between two links in active-active mode. M-LAG devices can be paired to negotiate and manage the master/backup status of M-LAG devices and M-LAG member interfaces;
  - Support information synchronization through M-LAG synchronization packets, such as MAC address entries, ARP entries, and ND entries, and the status of M-LAG member interfaces ;
  - Support M-LAG active-standby mode, which implements two links for backup protection while only the master M-LAG device sends and receives traffic. The master/backup status of M-LAG member interfaces can be elected based on protocol packets, in terms of ARP, ND, IGMP, DHCP, and MLD message, sent by the access devices, covering initial scenarios, failover scenarios, and failover scenarios;
  - Support fast fail switchover and switchback of M-LAG in both active-active mode and active-standby mode;
  - Support heartbeat detection and dual-active detection. The dual-active detection failure must not affect the normal operation of M-LAG.

7.4.2 Device-level reliability

When a link fault occurs, the traditional route convergence the information exchange and recalculation of dynamic routing protocols (such as OSPF and BGP) on the control plane. Even if the BFD technology is used, the fault detection speed is accelerated, but the route convergence time is still hundreds of milliseconds. In a large-scale data center network, route convergence takes even seconds, which cannot meet the delay requirements of high-performance storage services or high-performance database access services. Advanced technologies based on data plane fast recovery which support sub-millisecond convergence are required to provide high reliability and stability for high-performance database, storage, and intelligent computing applications. The technical requirements of data plane fast recovery include:

- Support detection of faulty ports through the forwarding plane. The faulty ports may be caused by faulty optical modules or loose transmission cables;
- Support forwarding plane path switchover based on detected faults when local backup paths are available;
- Support fault detection propagation to the remote network;
- Support that the remote device can quickly switchover the backup path based on fault notification.

7.4.3 Network–level reliability

Network–level reliability refers to data plane crossing fault capability, which enables fast fault identification and second–level fault convergence when a silent fault causes service session–level exceptions on the network. The technical requirements of data plane crossing fault include:

- Support silent fault detection, such as link faults, abnormal forwarding entries, abnormal forwarding components, physical ports that are Up but cannot forward traffic, and configuration errors;
- Support network fault recovery. After the fault flow on the network is identified, the hash of route selection is driven to performed again and the flow can quickly switchover.

7.5 Intelligent O&M

Intelligent O&M is the key to improving the service capabilities of financial data centers. Based on the O&M big data platform and intelligent algorithms, intelligent O&M builds O&M capabilities in multiple scenarios, such as pre–event prediction, in–event monitoring, post–event analysis, on–demand capacity expansion, and timely emergency response, to improving data center service capabilities. After continuous polishing and optimization of intelligent O&M algorithms and models, the financial data center will become a "digital intelligence center" and one of the core competitiveness of future financial data centers.

7.5.1 Network digital map

Based on the digital twin technology, the network digital map forms digital copies through the big data platform and supports data–based analysis, computing, and presentation capabilities.

7.5.1.1 Multidimensional visibility and analysis

- Support multidimensional data visualization capabilities, including:

- Support real–time data collection of the network, including but not limited NE configuration, network service configuration, network and service status, and performance data;
- Support associated storage of multiple types of data and provides intelligent search capabilities;
- Support multi–dimensional data visualization and drill–down viewing;
- Support associated playback of historical data, data change trend analysis based on the specified time period, and multi–dimensional data drill–down visualization.

– Support multi–dimensional data analysis capabilities, including:
  - Support collection, cleaning of O&M data and storage of the O&M data in a unified manner;
  - Support correlation analysis on multi–source data for assist fault location ;
  - Support correlation analysis result of multi–sources of data for model training, converts O&M data into knowledge and insights, and provides support for intelligent O&M.

7.5.1.2 Network Path Navigation

Differentiated and deterministic service quality assurance is implemented by the bandwidth, latency, and availability of services. IP networks must accurately perceive service quality changes and network quality deterioration quickly and navigate through network paths. Service flows are groomed in a timely manner to avoid network congestion. The technical requirements of network path navigation include:

– Support network topology and status awareness, which quickly detects network topology changes based on BGP–LS, including node and link faults, link bandwidth, and delay changes;
– Support service SLA awareness through IFIT (In–situ Flow Information Telemetry) technology for flow detection, and telemetry technology to report service SLA data in real time;
– Support service SLA poor–quality demarcation and locating: Service quality deterioration automatically triggers hop–by–hop IFIT detection, detects poor quality based on service forwarding paths, and combines with the network topology to visually locate and locate the faul;
– Support service SLA poor–quality recovery: Based on the poor–quality location result, network paths are re–optimized in order to guide traffic to avoid poor–quality links and nodes to ensure the service SLA.

7.5.1.3 Storage host plug–and–play

This technology supports quick management and control of access hosts and intelligent adjustment of intelligent lossless network configurations to achieve low latency, no packet loss, and high throughput. The technical requirements of storage host plug–and–play include:

- Support host access and leave perception and spread of host access and leave information on the network;
- Support intelligent adjustment of intelligent lossless network configurations based on host access and leave scenarios to support effective host transmission;
- Support spreading error–down information about an interface on a network device to other devices after the interface enters the error–down state due to PFC deadlock or CRC error packets reaching the alarm threshold. After receiving the error–down information, the device adjusts the path information in a timely manner.

7.5.2 Network fault demarcation and locating

7.5.2.1 Fault detection and prediction

Fast fault detection is the basis of network O& M capabilities. Fault detection requires the network status and service status to be detected, reported in real time, and analyzed in a centralized manner, improving real–time and accurate. The technical requirements of fault detection include:

- Support data collection in seconds. For example, the Telemetry technology is used to collect data of multiple types of indicators, such as services, devices, links, ports, and optical modules, in seconds, and monitor the running status of network devices in real time;
- Support to Capture network traffic data through port mirroring, obtain complete service flow data in real time from traditional, virtual, cloud, and container environments, and parse and build a real–time unified service view. The service logic, dependency, service running quality, and alarms of application services are intuitively displayed. The automatic fault analysis function is provided. The service path of each transaction is displayed and the time consumption on each application node is distinguished. The performance bottleneck is found and O&M assurance is performed during the entire service provisioning process;
- Support lossless forwarding in the intelligent computing scenario. A small number of packets will greatly reduce the aggregate communication and slow down the training speed of the entire cluster. Therefore, the intelligent computing network must be able to detect the packet loss rate of the communication traffic between NPUs and GPUs with a precision of more than 1‰ and be capable of identifying the location of packet loss. In this way, the intelligent

computing network can take timely intervention when packet loss occurs to prevent single–point packet loss from slowing down the entire intelligent computing cluster.

7.5.2.2 Fault demarcation and locating

Fault demarcation and locating refers to analyzing and confirming the fault location and causes based on network and service deterioration and faults. The technical requirements of fault demarcation and locating include:

- Support correlation analysis of multiple network indicators based on network status data, alarm data, performance data, and log data. For example, based on the knowledge graph or AI capability, it identifies potential network reliability, capacity, performance, and stability risks in advance, and evaluates potential network risks in a unified manner to reduce the fault occurrence probability;
- Optical module stain detection is supported. If the fiber endface is dusty or dirty, the link may be intermittently disconnected, which affects continuous cluster training or reduces network communication efficiency;
- Support optical module looseness detection, which automatically detects optical module looseness after the cluster runs for a long time;
- Support for locating packet loss causes on the network. When packet loss occurs on the intelligent computing network, the intelligent computing network needs to quickly solve the problem based on the drop cause to restore training services. Therefore, the intelligent computing network needs to be able to directly query and locate the packet loss count caused by port congestion, ACL drop, and route query failure;
- Support monitoring of RDMA communication performance. During AI training, inter–card communication is usually burst traffic in milliseconds. The port bandwidth usage cannot effectively monitor the RDMA flow–level communication performance. Therefore, the intelligent computing network needs to be able to monitor indicators of the RDMA communication flow including FCT (Flow Completion Time), FET (Flow Efficient Throughput) and FNR (Flow NAK Rate);
- Support network congestion monitoring: When network congestion occurs, the aggregate communication performance will be affected. However, the number of PFC packets on ports or queues cannot be quantified due to the impact of the threshold and traffic model. Therefore, the intelligent computing network must be capable of monitoring and collecting statistics on the backpressure duration and pause of ports/queues to quantify the congestion degree.

7.5.2.3 Hitless upgrade

Support hitless upgrade of devices in order to reduce the service interruption time. The technologies such as M-LAG-based device upgrade can be used to upgrade devices independently to achieve high efficiency and second-level service interruption.

7.5.3 Application-level fault demarcation and locating

7.5.3.1 Application fault detection

Support application-level flow detection is supported. Technologies of flow detection, such as IFIT, should be used to insert detection parameters into service packets in order to get the accurate SLA of the application. The data then be sent to the controller for centralized analysis.

7.5.3.2 Application fault demarcation and locating

The technical requirements of application fault demarcation and locating include:

- Support automatic demarcation and locating of application-level faults. Technologies, such as IFIT, should be used to support hop-by-hop detection to identify the NE and link where services deteriorate ;
- Support monitoring and automatic fault analysis and troubleshooting for AI training or inference tasks in the intelligent computing POD.

7.5.4 Full-flow analysis

As the number of devices on an SDN network increases and more services are carried, users have higher requirements on SDN network O&M. Comprehensive monitoring of network-wide traffic helps detect and analyze exceptions in a timely manner, which is critical to network O&M. Traditional network traffic monitoring methods have their own characteristics, but they cannot monitor the traffic information of the entire network.

The full flow analysis technology can analyze network-wide traffic. With the help of the built-in chip, the device supports 1:1 sampling and does not affect forwarding performance. In addition, it provides the capability of reporting key events, reducing the processing burden of the remote analyzer.

- Support flow collection and creation of full flow analysis flow tables to monitor typical TCP exceptions. When the buffer space of the flow table is full or the flow expression reaches the aging time;
- Support flexible flow information uploading to the analyzer, such as NetStream V9 ;

- Support the creation of full flow analysis flow tables based on collected flow fields to implement traffic statistics, path visualization, application access visualization, TCP application-side packet loss detection, and TCP anomaly detection ;
- Support original flow output mode. When the flow aging time expires or the reporting period expires, the statistics of each flow are exported to the analyzer. The advantage of the original flow output mode is that the analyzer can obtain the detailed statistics of each flow ;
- Support the aggregation flow output mode. In the aggregation flow output mode, the built-in chip of the CPU summarizes the original flow statistics that are the same as the aggregation key items to obtain the corresponding aggregation flow statistics. The aggregation of original flows and the output of the aggregated flows significantly reduces the network bandwidth usage. Key aggregation items include the source IP address, destination IP address, destination port number, and protocol type. Only the flows whose session status is Normal are aggregated.

7.6 Security

7.6.1 Device Layer Security

The technical requirements of device layer security include:

- Support factory trustworthiness. The factory software package does not have hard-coded authentication credentials. The internal account and communication matrix provide a list, which is open and transparent and secure without backdoors ;
- Support deployment security. Implement SZTP (Secure Zero Touch Provisioning) based on security technologies such as certificates and signatures to ensure secure network access of devices ;
- Support host-based Intrusion Prevention through monitoring whether the local system is intruded or infected. Once a suspected intrusion or infection event is detected, the system sends logs to prompt the administrator to isolate and protect the system, preventing further intrusion or even endangering the security of other devices ;
- Support for network exit security. All devices that need to be removed from the network can be released only after they pass the security check.

7.6.2 Network Layer Security

The technical requirements of network layer security include:

- Support protocol security. All network protocols use high-strength security cryptographic algorithm by default, which is compatible with insecure algorithms of protocols and prompts risks;

- Support link security. Hardware–based MacSEC encryption is supported to prevent data from being stolen or modified during transmission and ensure data integrity. In addition, encrypted packets can be forwarded at line speed, implementing secure transmission of encrypted data.

7.6.3 Management and control layer security

The technical requirements of management and control layer security include:

- Supporting security configuration check, such as insecure protocols and algorithms, abnormal ports, account and password policies, and password storage modes, to identify insecure configuration risks in a timely manner and mitigate risks on the live network in advance;
- Support for management security. A security management system is built through first login, identity authentication, privilege minimization, external intrusion detection, and data encryption to prevent security attacks such as identity spoofing, repudiation, tampering, information leakage, DoS, and privilege escalation. Support log security. Independent storage space for security logs, preventing tampering.

8 Key Technical Requirements for Multi-–DC Data Center Networks

In the multi–DC scenarios, besides the technical requirements of a single DC, the specific requirements of multi–DC should be meet.

8.1 Multi–DC reliability

8.1.1 Intra–city active–active DC reliability

In intra–city active–active DC scenario, two DCs should be deployed and managed in active–active mode. The technical requirements of intra–city active–active DC include:

- Support active–active status detection of DCs. In the normal, both DCs process the services. If the active–active detection fails, services are not affected;
- Support fast DC switchover when a single DC is faulty in order to reduce the service interruption time;
- Support quick active–active DC recovery and service switchback when a single DC recovers from a fault;
- Data backup in intra–city active–active scenarios is supported. The link RTT and bandwidth should meet the requirements.

8.1.2 Geo-Redundant DCs Reliability

Geo-redundant DCs means that DCs are deployed in different regions to further improve the DR capability of DCs. The technical requirements of Geo-Redundant DCs reliability include:

- Support geo-redundancy with two active-active sites and a DR site. In addition to intra-city active-active DCs, remote DC backup is added;
- Support heartbeat detection among DCs. In the normal, all the DC perform their own duties. If the heartbeat detection fails, services are not affected;
- Support Geo-Redundant DCs switchover and switch back mechanism. When one of the active DC failed, fast active-active DC switchover. When the intra-city active-active DCs are failed at the same time, the services should be switchover to the backup DC. When the failure of DC recovered, the intra-city active-active DC should switchback;
- Support data backup between DCs. The link RTT and bandwidth should meet the requirements.

8.2 Multi-DC O&M

The technical requirements of multi-DC O&M include:

- Support One controller to manage intra-city active-active DCs;
- Support multi-DC management. Each DC can be managed by an independent controller, and cross-DC management is implemented through cross-DC controllers.
- Support cross-DC automatic network security convergence service deployment, in order to keep the consistent of the security polices and reduce configuration errors;
- Support cross-DC fault demarcation and locating and E2E fault locating result in a visualized manner in order to reduce the fault demarcation and locating time;
- Support cross-DC configuration simulation to improvement the cross-DC service deployment reliability.

8.3 Multi-DC security

The technical requirements of multi-DC security include:

- Support data security between DCs. Technologies like MacSEC should be used for data encryption;
- support management security among different DC controllers. Technologies like SSH, TLS should be used for management channel encryption.

Annex A

Bibliography