HUAWEI

Striding Towards the
Intelligent World White Paper

Digital Finance

Navigate Change,
Shaping Smarter,
Greener Finance Together

Building a Fully, Intelligent World

Catalogue

Executive Summary

Many new uncertainties threaten global financial institutions, including declining exports, wars, geopolitics, sluggish economies, and increasing interest rates. To rise to these challenges, they need agility and the capacity to quickly adapt their operations.

At the same time, finance companies are advancing along their digital journeys, with AI integrated into financial business processes. Intelligent applications have brought infinite possibilities for their development.

The financial industry is developing following these trends:

Trend 1: Emerging customer groups accelerate transformation from digital connections to intelligent emotional interaction

The digital natives who were born after 1995 gradually dominate consumer spending, and the financial interaction models should be changed to meet their consumption needs. ChatGPT still takes the world by storm in 2023, foregrounding the wide applications of artificial intelligence (AI). Leading banks have begun to use AI customer service robot assistants and branch robot assistants to identify customer emotions by analyzing their tone and micro-expressions, while providing caring services for customers with the help of AI assistants.

Trend 2: Global central banks are transforming digital currencies, bringing digital payment back to the banking system

Digital currencies are getting more mature and have been officially issued by seven countries globally. China has also piloted digital currency in a large scale and in many batches. Digital currency will bring payments back to the banking system. Therefore, banks need to be prepared for digital currency transactions and intelligent regulatory systems, so as to handle massive volumes of encrypted transactions smoothly.

Trend 3: Platform economy activates micro finance, with financing powered by AI

Scenario-specific finance and micro finance are among the top two critical fintech application domains in the Chinese Banker Surveys Report (2022) published by the China Banking Association. However, financial institutions with poor digital capabilities can hardly support scenario access. For example, it remains difficult to ensure trusted ownership and supervision for inventory financing, so small and micro enterprises cannot obtain needed financial services. This can be solved by building a scenario-based ecosystem and digitalizing open banking. In addition, IoT and AI can be synergized to make inventories trustworthy.

Trend 4: New transaction frauds and intelligent network attacks threaten fund security

Illegal transactions such as fraud and money laundering are becoming increasingly covert, which cannot be detected instantly using traditional risk control means and may lead to fund losses. As technology is evolving, network viruses are getting more intelligent, clustered, and lingering longer, resulting in frequent information leakage and ransomware incidents. At the same time, financial institutions are burdening huge regulatory compliance pressure and they can no longer rely on traditional passive compliance methods. Many of them turn to a convergence of real-time data and AI, alongside global collaboration to improve the efficiency and effectiveness of risk control, security, and compliance audits.

Trend 5: Accelerated transformation and large-scale growth of IT drive systematic construction of business resilience

Cloud-native platforms can vastly unleash digital productivity. More banks are migrating their core transaction systems to open architectures. However, it is difficult for an open system to reach the same level of latency and reliability as mainframes and midrange computers. Cross-domain collaboration, distributed optimization, and intelligent O&M can effectively improve the performance and availability of open systems.

Trend 6: With asset scale growth slows down, banks are moving towards refined operations

Affected by the pandemic, import and export, and the overall economic situation, financial institutions have seen heavier pressure from revenue and slow asset scale growth. Banks start to control operating expenditures and IT investment to reduce costs. The best way to achieve this goal is to build a green and low-carbon cloud infrastructure, adopt a proper IT architecture, and use green energy-saving algorithms for automatic O&M.

Trend 1 .

Emerging customer groups accelerate transformation from digital connections to intelligent emotional interaction

The popularity of mobile and Internet finance, alongside aggressive over-the-top (OTT) services, has made financial institutions shift their focus from ensuring stable transactions to improving user experience. Centering on digital interaction, financial institutions are combining online with offline to reach external and internal users through multiple channels. They are also planning and guiding user journeys to reconstruct business and operational models. Financial services are moving from transactions to intelligent interactions.

Boston Consulting Group (BCG) divides the current financial customer groups into four generations: Generation X born from 1965 to 1979, Generation Y from 1980 to 1994, Generation Z from 1995 to 2009, and post-2010 Generation alpha. Each generation has different group memories, value orientations and consumer preferences. Generation Z, the digital natives, is gradually becoming the main consumer force, and the meta-universe generation alpha is about to emerge, requiring financial interaction models to keep evolving. (Figure 1.1 Heart model for user experience)

China Merchants Bank (CMB) is committed to providing users with a premium experience. It uses the HEART model to evaluate the interaction between users and financial products or services. The HEART model was first introduced by Google in a paper.

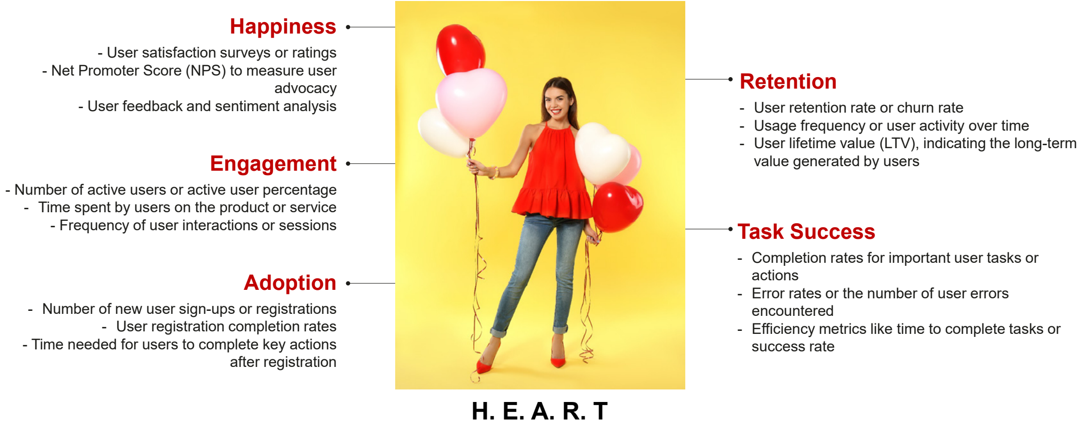

Happiness
- User satisfaction surveys or ratings
- Net Promoter Score (NPS) to measure user advocacy
- User feedback and sentiment analysis
Engagement
- Number of active users or active user percentage
- Time spent by users on the product or service
- Frequency of user interactions or sessions
Adoption
- Number of new user sign-ups or registrations
- User registration completion rates
- Time needed for users to complete key actions after registration
Retention
- User retention rate or churn rate
- Usage frequency or user activity over time
- User lifetime value (LTV), indicating the long-term value generated by users
Task Success
- Completion rates for important user tasks or actions
- Error rates or the number of user errors encountered
- Efficiency metrics like time to complete tasks or success rate
H. E. A. R. T

Figure 1.1 Heart model for user experience

It measures user experience from five dimensions: happiness, engagement, adoption, retention, and task success.

Leading financial institutions have been endeavoring to provide online financial institutions through livestreaming and achieve real-time interactive marketing. In 2022, CMB hosted more than 2,000 sessions of livestreaming, with each attracting more than 3 million customers.

A top bank in Brazil has set up tens of thousands of customer service agents to improve service quality. The bank can now communicate with customers through third-party platforms, including mobile apps, virtual teller machines (VTMs), and What's App, besides traditional phone calls and text messages. In this way, it can reach customers through various channels and handle various actions like remote customer acquisition, loan collection, investment advisory, claims, training, and after-sales services. The booming business volume, an increasing number of agents, and the introduction of HD videos have increased multimedia traffic tenfold in the last five years.(Figure 1.2 AI contact center)

The boom of ChatGPT in 2023 pushed AI to the foreground. Leading banks have started to use AI customer service assistants and branch assistants, from semantic to tone recognition, from image to micro-expression recognition, to perform emotional care for their customers. Virtual humans are used to improve marketing reach rate online and offline.

Challenge 1 Insufficient computing power

- In the training phase, the virtual human training needs more than 100,000 facial expressions, costumes, and prop models, consuming massive computing power. Insufficient computing power prolongs the

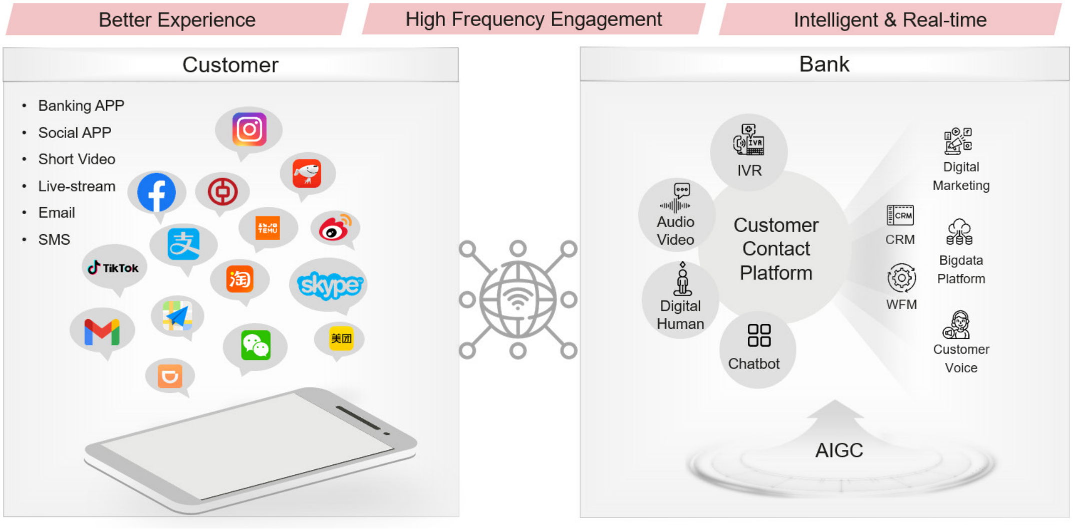


Figure 1.2 AI contact center

training period to more than one week.

- In the application phase, a virtual human with smooth communication capabilities needs an AI inference GPU, which is costly. Ordinary banks are hard to support the concurrent online services for massive users.

Challenge 2 Rapid traffic growth and high video quality requirements

- From audio and image to video and micro expression recognition, HD video transmission requires a 50-fold increase in network bandwidth. The growth of customer service agents further enhances bandwidth requirements, posing great challenges to video quality assurance and bandwidth leasing cost.

Challenge 3: Hard collection of real-time feedback

- Product design and optimization require real-time customer response and feedback. However, technically collected data cannot reproduce customers' real journey, especially their emotional journey.

Suggested actions

1. Purchase virtual human services on the public cloud to train virtual human models, reducing the training time and costs.

2. Optimize computing resources. Slice resources and time for multiple tenants in the cloud-based service mode to support multiple virtual humans with one inference GPU and reduce application costs.

3. Establish a user experience monitoring and troubleshooting system. Enhance the usability of data systems to facilitate service personnel to record customer journey information. Multi-dimensional data collection enables continuous monitoring of key parameters such as service availability, throughput, latency, saturation, and detailed parameters such as emotional response.

4. Build a network quality assurance system. Provide one network for the entire bank to enable ubiquitous and multi-channel service access, support high-quality multimedia interaction involving audio, video, and text through high bandwidth and E2E quality assurance measures, and reduce network leasing costs through data compression.

1.1 Data Intelligence

1.1.1 Financial institutions' self-built AI model training environments

ChatGPT has ignited the spark of intelligence in the financial industry. Using the model training service from public cloud service providers is a shortcut for financial AI training. However, the financial industry has strict regulatory requirements, and uploading data related to key banking services to the public cloud is difficult.

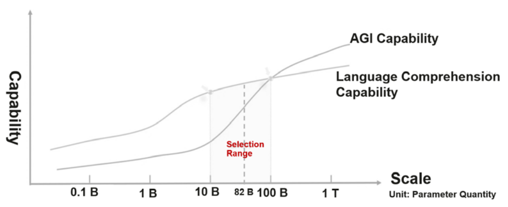


Figure 1.3 AI training capability surge

AI's emergent abilities will take effect when the number of large model parameters reaches about 62 billion. Currently, many open-source large models support 10 billion or even 62 billion parameter sets, significantly lowering the threshold for AI training. As a result, various financial institutions have built their own AI infrastructure and joined the competition in financial modeling.(Figure 1.4 Three-layer architecture of financial AI models)

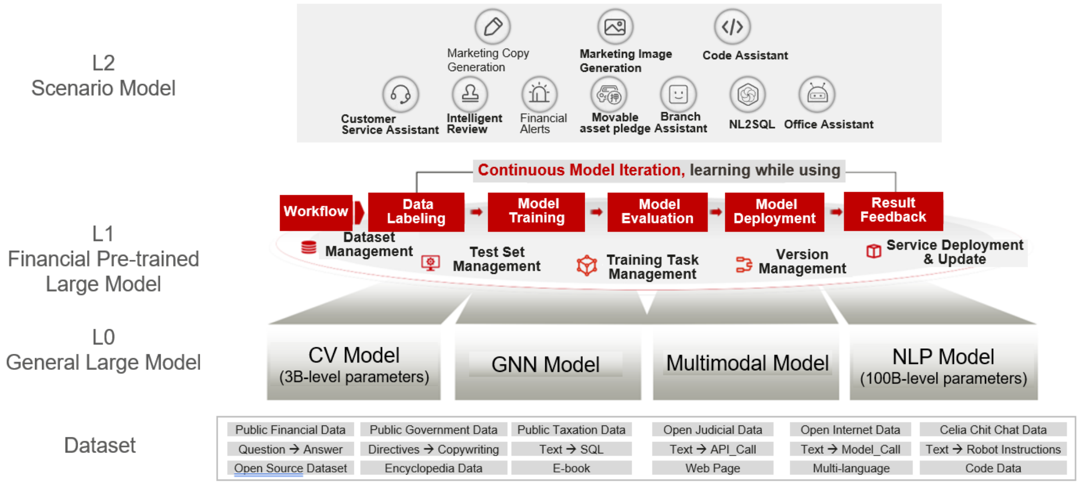
The development of AI models for the financial industry is divided into three layers.

L0 is a general pre-trained foundation model offered by AI training service providers. Top public cloud providers offer AI training services.

L1 is a pre-trained foundation model for the financial industry generated based on L0 and trained by industry-specific datasets. They are usually jointly developed by AI training service providers, industry organizations, or industry-leading enterprises.

Based on L0 and L1, L2 is a scenario-based model trained for specific financial scenarios. These models can help financial institutions generate intelligent applications in specific scenarios such as customer service, code generation, and business review.

For example, banks use NL2SQL to generate over 300 types of reports each year, totalling tens of thousands. These reports include weekly, quarterly, and annual reports. Through NL2SQL, users can input natural language



Public Financial Data	Public Government Data	Public Taxation Data	Open Judicial Data	Open Internet Data	Celia Chit Chat Data
Question → Answer	Directives → Copywriting	Text → SQL	Text → API_Call	Text → Model_Call	Text → Robot Instructions
Open Source Dataset	Encyclopedia Data	E-book	Web Page	Multi-language	Code Data

Figure 1.4 Three-layer architecture of financial AI models

searches like "analyze the online spending habits of women aged 30 to 40 in July and August and compare them with the same period last year" and obtain the analysis report they need. This makes data usage easier and accessible to everyone.

The effectiveness of AI training depends on computing power, algorithms, and data. Financial institutions must prepare data and initial models to build an AI training environment, establish the training infrastructure, and deploy and integrate output models. In the process, the following challenges are faced:

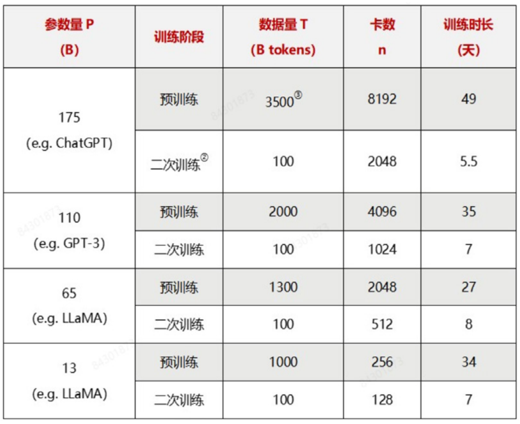
- Financial institutions do not lack a massive amount of data, but lack high-quality data for training.
- Complex model selection, architecture adjustment design, and technical verification process result in high trial-and-error costs.
- Model development requires the parallel usage of multiple technical paths and simultaneous verification of hundreds of technical points.
- On an unstable computing platform, the average constant training time is about 2.8 days.
- The integration of output models into live-network services is complex and financial institutions lack experience in this process.

Suggested actions

1. Streamline data analysis and AI data flows. Use a data lakehouse to clean training data.
2. Select reliable vendors to deliver one-stop services. Request training and coaching.
3. Build the AI training infrastructure with high performance and reliability.
4. Choose the simplest and most effective scenarios to deploy AI applications and form a positive cycle.

1.1.2 Infrastructure architecture transformation

There are three types of GPU volumes in self-built AI model training environments.



参数量 P (B)	训练阶段	数据量 T (B tokens)	卡数 n	训练时长 (天)
175 (e.g. ChatGPT)	预训练	3500[3]	8192	49
	二次训练[2]	100	2048	5.5
110 (e.g. GPT-3)	预训练	2000	4096	35
	二次训练	100	1024	7
65 (e.g. LLaMA)	预训练	1300	2048	27
	二次训练	100	512	8
13 (e.g. LLaMA)	预训练	1000	256	34
	二次训练	100	128	7

Figure 1.5 Resources required for foundation model training

- Small and medium-sized banks start with about four AI GPUs; typically, they have 16 and 64. The number of parameters ranges from 10 million to 100 million.

- Large banks generally invest between 100 and 1,000 training GPUs. And super-large banks invest thousands of training GPUs, with the number of parameters ranging from one billion to tens of billions.

- Top investment banks and funds that engage in quantitative securities transactions are strong and pursue high returns. They often invest in a thousand or even ten thousand GPUs with the number of parameters close to 100 billion.

Since 2012, the computing power required for global AI model training has been doubling every three to four months, resulting in nearly a ten-fold yearly increase. Meanwhile, the trend of computing performance doubling every two years according to Moore's Law has slowed down. In fact, the growing demands for computing power in AI model training are not aligned with Moore's Law, which requires architectural changes in computing, connectivity, and storage infrastructure. (Figure 1.6 AI model training infrastructure)

The AI model training infrastructure must be centered on computing power. The computing power, connectivity, and storage infrastructure should coordinate with one another. Training tasks should be highly parallel, without performance bottlenecks or bandwidth convergence. Faults can be rectified as soon as possible to continue training.

Diversified computing power (Figure 1.7 Diversified computing power)

X86 is no longer the only choice for computing power in DCs.

X86, which features long pipelines, is good at processing complex instructions, such as office and text applications. However, processing batch and current tasks is very wasteful with

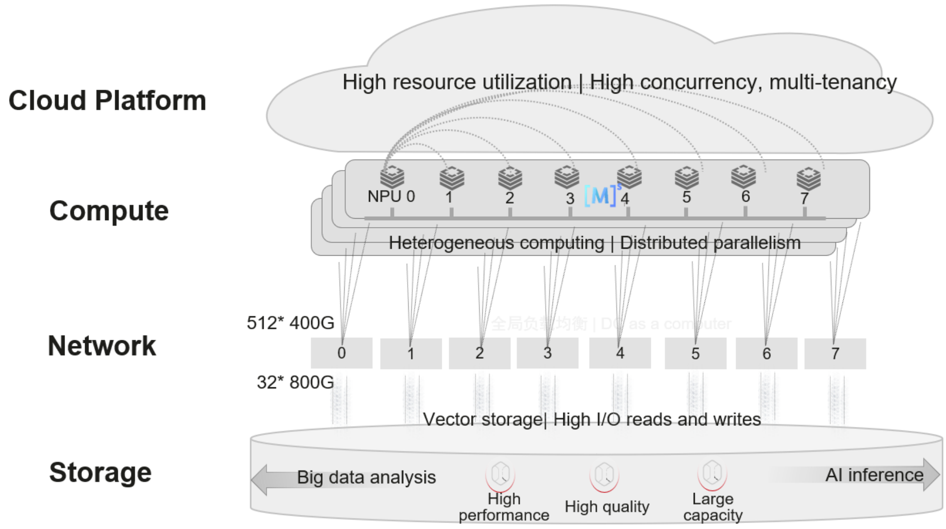

Cloud Platform
High resource utilization | High concurrency, multi-tenancy
Compute
NPU 0
1
2
3
[M]s
4
5
6
7
Heterogeneous computing | Distributed parallelism
Network
512* 400G
全局负载均衡 | DC as a computer
0
1
2
3
4
5
6
7
32* 800G
Vector storage| High I/O reads and writes
Storage
Big data analysis
High performance
High quality
Large capacity
AI inference

Figure 1.6 AI model training infrastructure

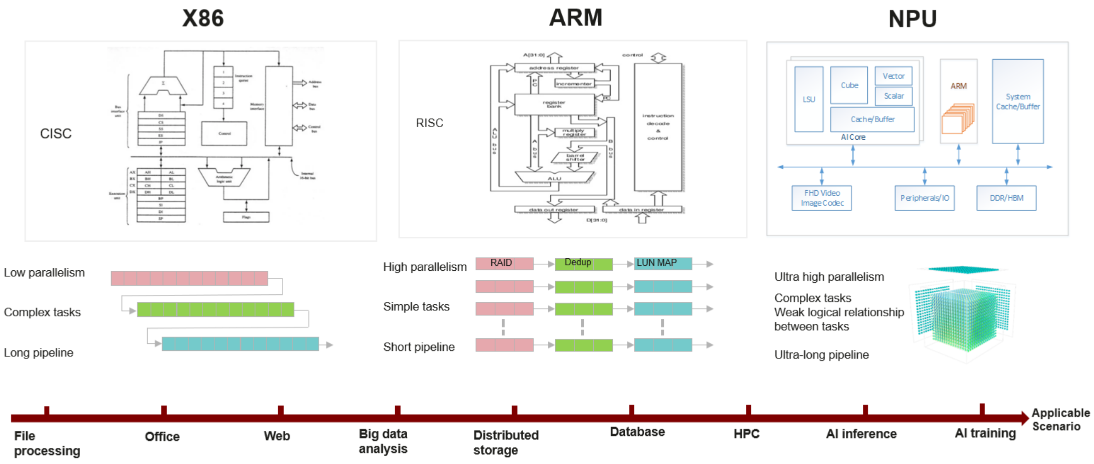


Figure 1.7 Diversified computing power

low CPU usage.

ARM, featuring short pipelines and high parallelism, has an extremely high power consumption/performance ratio for batch and high-concurrency tasks. For example, in terms of storage RAID computing based on bit XOR operation or the read-write of distributed databases, ARM has a higher computing power than x86 servers with the same dominant frequency and core number.

NPU, a dedicated training processor (AI training processors have different names depending on the vendor, such as Google's TPU), is based on multidimensional vector operations. It can output an order of magnitude of computing capability exceeding that of a CPU in a clock cycle. For example, an NPU core with a 3D vector and 16 instruction depth can perform 4096 (16 x 16 x 16 = 4096) operations in a single clock cycle, while a CPU performs only 256 (16 x 16 = 256) operations.

DC-as-a-Computer (Figure 1.8 DC-as-a-Computer)

For the first time, even in DCs, AI training has enabled the NPU to replace the CPU as the core. To match NPU's performance, data traffic does not pass through the CPU, which

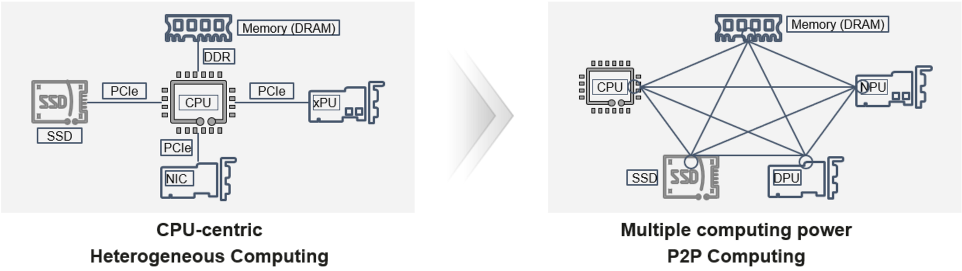


Figure 1.8 DC-as-a-Computer

only plays the role of control. NPU uses a higher-speed bus to access the memory, interface GPU, and storage, evolving from a single-core architecture to a peer-to-peer computing architecture.

Ultra-high parallel network

Generally, each AI training server is configured with four to eight training GPUs, and multiple AI training servers are grouped into a cluster. They form a two-layer high-parallel connection system combining the internal bus and external network. The synchronization of massive data parallelism needs to be implemented between AI chips/AI servers.

1. Parallel data: The training sample set is split into multiple mini-batches to be trained on multiple AI processors in parallel.

2. Parallel models: A model is divided into multiple sub-models and stored in multiple AI processors, supporting large models as a whole.

3. Parallel tensors: A model is split into multiple sub-layers and run on multiple AI chips.

4. Parallel pipelines: The intermediate computing results synchronize with the parameters between training GPUs and server nodes.

The communications of large model training have few data flows but high single-stream bandwidth with a high synchronization burst. There is also a lot of traffic for each iteration, with intra-server traffic reaching 100 GB levels and inter-server traffic reaching GB levels. （Figure 1.9 Parallel connection without convergence）

The training network is required to achieve the following.

- High reliability and zero packet loss: RoCE or InfiniBand networks are used to ensure zero packet loss.

- High bandwidth: Small-scale training networks with fewer than 64 GPUs use 100 GE network connections, and large-scale training networks use 400 Gbit/s or 800 Gbit/s connections. It is expected to reach 1.6 TB in 2024.

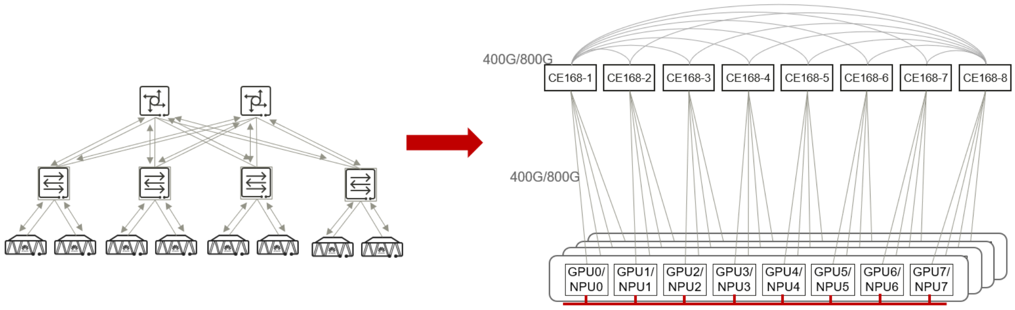


Figure 1.9 Parallel connection without convergence

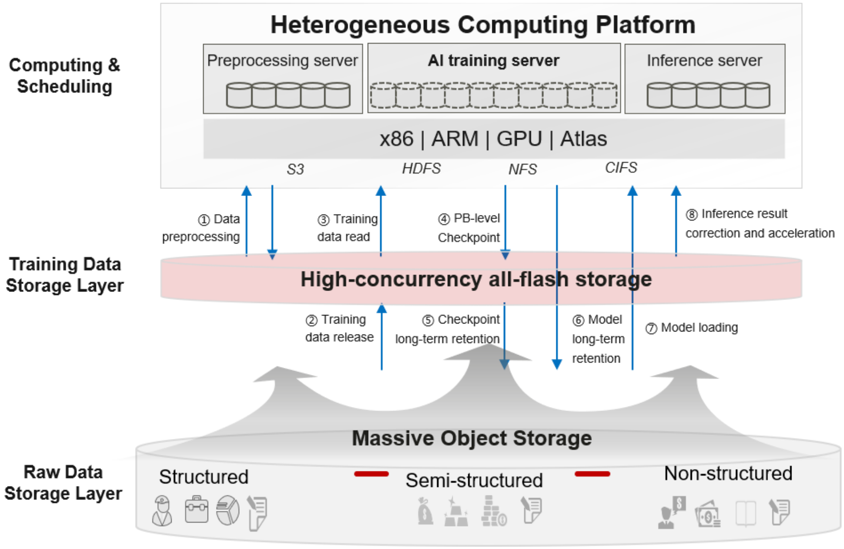


Figure 1.10 Massive training storage resource pool

- No bandwidth convergence: The traditional layer-3 network has many downlink (southbound) outbound interfaces and few uplink (northbound) outbound interfaces. The northbound bandwidth is hard to align with the southbound bandwidth, which requires bandwidth convergence. The traditional layer-3 networking is replaced by the direct switch inter-connection solution shown in the figure. It implements parallel connection without convergence.

- Network-scale global load balancing (NSLB): Based on the congestion status, the adaptive routing algorithm is used to evolve from local to global load balancing, preventing any congestion across the training environment.

- High-reliability multicast: Through ACK high-reliability multicast protocols, all terminals need to confirm the data reception status and ensure that the models and intermediate parameters involved in the training process are delivered to all training GPUs.

High-performance massive storage resource pool

The quality and quantity of data directly impact the level of AI intelligence. As the scale of model parameters increases, large models have higher requirements on the data scale, data retrieval, and read/write speed. (Figure 1.10 Massive training storage resource pool)

Collection of massive amounts of data: The transition from text to multi-modal data from video, audio, text, and images has resulted in a 1000-fold increase in data volume. Additionally, it takes three to five weeks to collect data of different protocols and formats from DCs, edges, and clouds.

- Quick data preprocessing: The collected raw data cannot be directly used for training. It needs to be parsed, cleaned, and deduplicated. This process involves at least three complete data migrations. Generally, this process for PB-level data takes more than 50 days.

- High-speed data retrieval: A GPT model with 6 billion parameters and a high-quality knowledge base for training can provide a higher output precision than the GPT model with 60 billion parameters. Quickly retrieving the knowledge base has become a necessary capability for AI data storage.
- Resumable training: Model training is costly. GPU servers have multiple components and a high failure rate. On average, a fault occurs every 2.8 days in industry's model training. So, in a training process, there are checkpoints to suspend the training task and periodically save the intermediate data. This allows training to be resumed after a fault occurs. The data storage speed determines the duration of the training suspension.

Data storage for model training must have the following features.

- Efficient data provisioning: EB-level scalability supports raw data storage.
- High-performance data acceleration: Hundreds of GB/s of bandwidth and ten million IOPS support fast training data write and collection, helping the training platform read quickly.
- Multi-protocol unified data foundation: Multi-protocol interworking between NFS, CIFS, S3, and HDFS are implemented in one pool to streamline data analysis.AI training data flows and supports different services in multiple phases.
- Efficient tiering of massive amounts of data: Automatic tiering of hot, warm, and cold data lowers the storage costs.

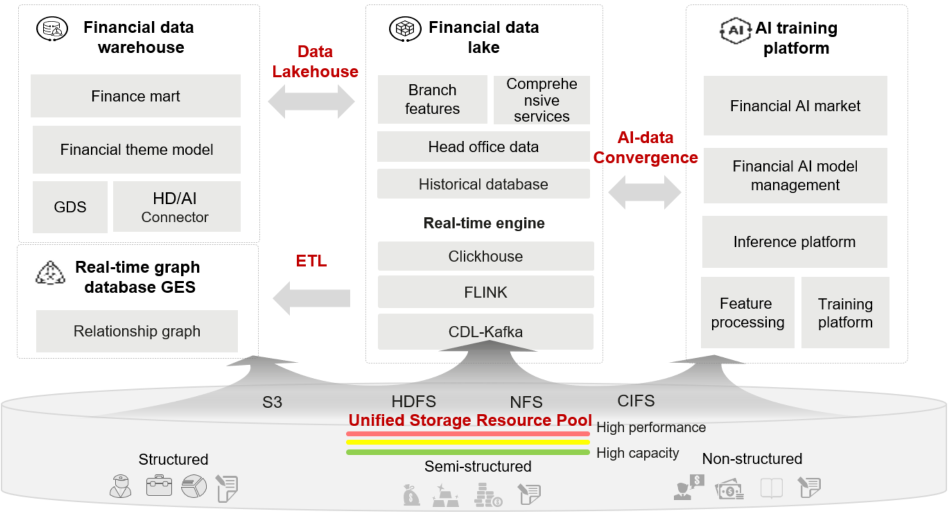


Figure 1.11 Data-AI converged infrastructure

Data-AI converged infrastructure

Build an integrated data-AI platform to streamline data and service processing flows for data analysis and AI training, to significantly improve data utilization efficiency. (Figure 1.11 Data-AI converged infrastructure)

- Unified storage resource pools enable the centralized management of the data lifecycle and ensure the security of the entire bank. They also enable data sharing and flow across the bank.

- The data lakehouse platform governs and cleans raw data, extracts high-quality data for training, and transmits training data to the AI training platform without copying data through the horizontal data flows in unified storage pools.

- The vertical data flows in unified storage pools heat and cool data, improving training performance, and reducing data storage costs.

1.2 High-quality real-time interactions

Real-time interactions by financial institutions include remote outbound calls, livestreaming, and videoconferencing. Take videoconferencing as an example. As multimedia office apps are getting more popular, directors and employees of financial institutions often turn to remote collaboration. Each employee needs to attend four audio or video conferences daily on average and must spend half of the working hours in video conferences when busy. Audio and video conferencing traffic increases by 30% every year. Especially the quality of high-level conferences directly affects enterprise operations.

Network O&M personnel determine the network health status based on network KPIs. The access success rate is an important KPI for measuring Wi-Fi quality and, generally, should reach 95%.

From the perspective of users, key quality indicators (KQIs) can better reflect the actual user experience. Different KQIs are set for different businesses. Web page browsing and video playback are typical scenarios where user experience can be easily affected. So, the browsing latency should be lower than 0.3 seconds, and the percentage of smooth video playback should exceed 99%.

1.2.1 IPv6 Enhanced and WAN

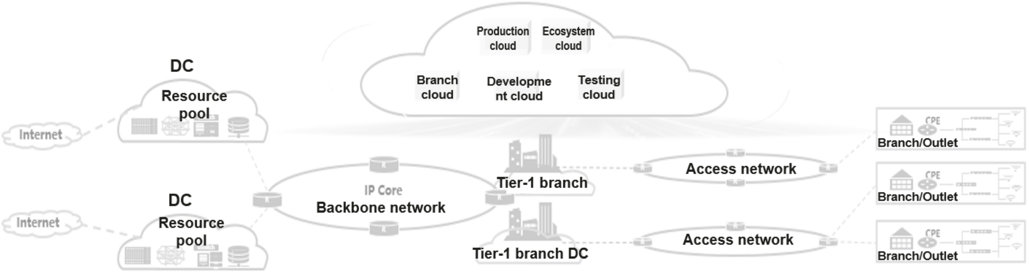


Figure 1.12 Financial WAN

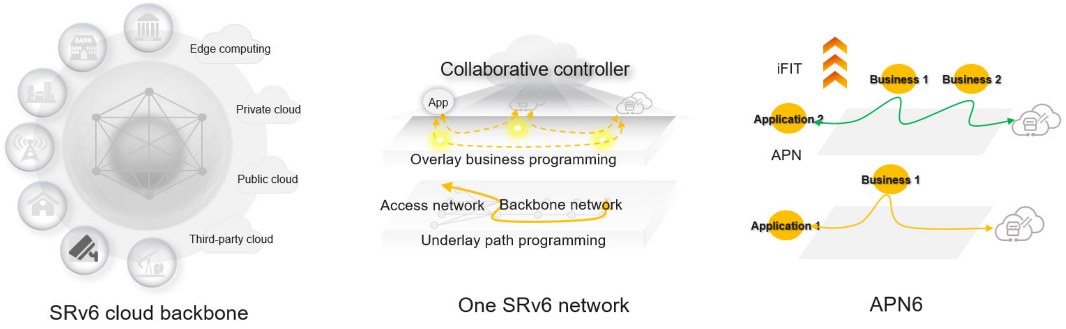


Figure 1.13 IPv6-based network innovation

Financial WANs connect outlets, offices, and DCs in different places to ensure efficient and secure data transmission. The application of IPv6 in financial WANs has many advantages.

IPv6 can improve the performance, security, and scalability of financial WANs. (Figure 1.13 IPv6-based network innovation)

1. IPv6 Enhanced provides massive IP address resources for financial services. All devices can be allocated with a unique IP address. Financial institutions can digitally manage tens of millions of devices with one IP address per device.

2. SRv6 introduces segment routing to the IPv6 network, which offers many benefits.

- Simplified network protocols

  Removes MPLS LDP and RSVP-TE and simplifies the network control plane requiring only IGP and BGP, reducing the workload of network administrators.

- SRv6 path programmability improves private line utilization

  SRv6 paths are automatically optimized to achieve network-wide bandwidth load balancing. This improves link utilization and ensures user experience. SRv6 requires 30% less WAN bandwidth and saves a considerable number of rentals each year.

- Network slicing ensures differentiated quality

  The network-slicing solution can slice one physical network into multiple planes that do not affect each other, comprehensively ensuring business quality.

3. APN6 prioritizes and allocates more network resources to critical applications. With APN 6, network devices can identify different applications, such as latency-sensitive real-time voice and video communications, high-bandwidth file transfer, and financial transactions that require high reliability, and optimize themselves to meet the needs of these applications.

4. iFIT improves management and O&M efficiency. When an exception occurs,

the system automatically collects quality information hop by hop along the business path to demarcate and locate the fault, restore the business, and ensure its continuity. When a fault occurs on the network, the system immediately switches to the backup path through SRv6, quickly rectifying the fault and ensuring business continuity.

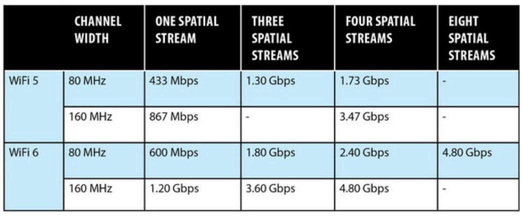
In areas where IPv6 is not available, financial institutions can use IPv6 Ready network devices to smoothly switch to IPv6 if network conditions permit, enhancing network quality.

1.2.2 High-quality 10GE campus network

Financial institutions need to build an integrated 10GE campus network that covers office, IoT, security, and customer services. With centralized access, this network can carry diverse businesses. It should be ultra-broadband, simplified, intelligent, secure, and open.

1. Building all-scenario Wi-Fi 6 for an all-wireless office

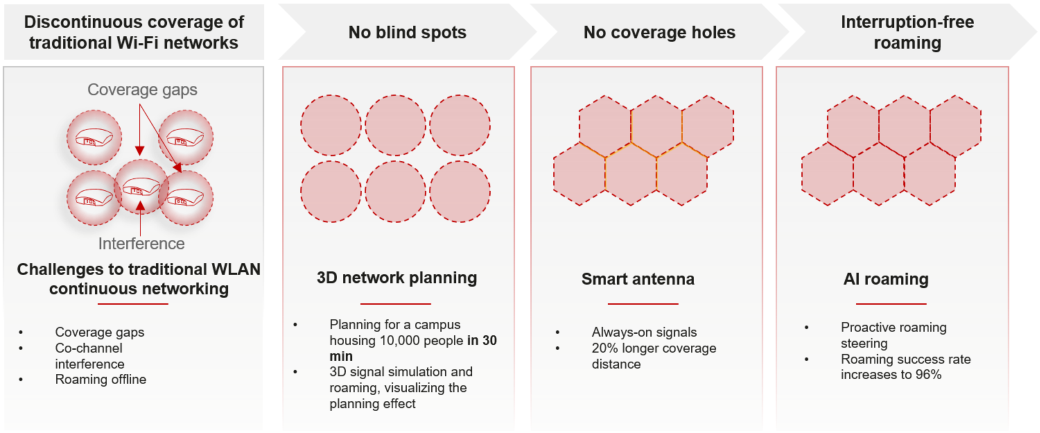
Wireless networks can be upgraded to Wi-Fi 6/7 to achieve full and high-density WLAN coverage and ensure a smooth access experience.



	CHANNEL WIDTH	ONE SPATIAL STREAM	THREE SPATIAL STREAMS	FOUR SPATIAL STREAMS	EIGHT SPATIAL STREAMS
WiFi 5	80 MHz	433 Mbps	1.30 Gbps	1.73 Gbps	-
	160 MHz	867 Mbps	-	3.47 Gbps	-
WiFi 6	80 MHz	600 Mbps	1.80 Gbps	2.40 Gbps	4.80 Gbps
	160 MHz	1.20 Gbps	3.60 Gbps	4.80 Gbps	-

Figure 1.14 Wi-Fi 5 vs. Wi-Fi 6

Wi-Fi 6 is the sixth generation of Wi-Fi and incorporates many key 5G technologies. Compared with Wi-Fi 5, Wi-Fi 6 quadruples the network bandwidth and the number of concurrent users. On average, it reduces the network latency from 30 ms to 20 ms. It also facilitates applications like 4K ultra-HD video conferences (ultra-high bandwidth), high-density (ultra-high concurrency)



Discontinuous coverage of traditional Wi-Fi networks

Challenges to traditional WLAN continuous networking

- Coverage gaps
- Co-channel interference
- Roaming offline

No blind spots

3D network planning

- Planning for a campus housing 10,000 people **in 30 min**
- 3D signal simulation and roaming, visualizing the planning effect

No coverage holes

Smart antenna

- Always-on signals
- 20% longer coverage distance

Interruption-free roaming

AI roaming

- Proactive roaming steering
- Roaming success rate increases to 96%

Figure 1.15 Continuous Wi-Fi coverage

coverage, and virtual reality (ultra-low latency). (Figure 1.15 Continuous Wi-Fi coverage)

Financial institutions need to consider the following to deliver an uninterrupted user experience in wireless systems.

- Network planning: Learning from 5G wireless networks, it is possible to use 3D planning that features a honeycomb-like structure to ensure full signal coverage in office space, including open office areas, corners, aisles, and tea break areas.

- Resource scheduling: Dynamic-zoom smart antennas can perceive the density of access terminals. The angles and RF resources of antennas can be dynamically adjusted to support comprehensive coverage and high-density access, ensuring a smooth access experience for users.

- Interference suppression: Intelligent radio calibration proactively detects environmental changes, predicts future changes based on historical access loads and behaviors, and intelligently optimizes AP channels, frequency bandwidth, and transmit power based on the prediction results. It thereby reduces co-channel and adjacent-channel interference and delivers an optimal network experience.

- Roaming switchover: Terminals dominate roaming switchover on Wi-Fi networks, leading to delayed terminal switchovers or switchovers to non-optimal APs. Relying on the proactive roaming technology, the network intelligently identifies terminal types, proactively learns the roaming behaviors and habits of each terminal, formulates guidance policies and parameters that vary with terminals, and guides roaming. This improves the overall roaming efficiency to over 95%.

- HD videos: Intelligent multimedia scheduling algorithms can identify high-priority multimedia businesses and low-priority ones, such as background downloads. They can also monitor the latency of multimedia businesses. When they find a damaged high-priority business, the congestion control algorithm can precisely suppress greedy business traffic, preventing high-priority businesses from being affected.

2. Free mobility and secure access anytime, anywhere

   Network management software can be used to plan unified user access policies, which can be shared with campuses in other regions. When a user uses different terminals to access the network in different places, the network controller automatically identifies the user's identity and delivers execution policies to the corresponding devices on the network. This ensures that the user obtains a unified policy and consistent experience no matter where he or she accesses the network.

[Case] Bank U in Europe had outdated network devices with an unreasonable design at its HQ. The wireless network coverage was below 50%, and business quality could not be guaranteed. After the pandemic, many employees started using videoconferencing, requiring higher network capacity. Even executives could not access smooth conferencing services, severely affecting the office experience.

Bank U deployed Huawei's high-quality campus switches and Wi-Fi 6 solution to upgrade to a 10GE campus network. The wireless network fully covers the entire campus using 3D cellular-like networking. Both wireless and wired networks feature intelligent application identification. In addition, intelligent HQoS provides QoS policies for different users and applications, ensuring network quality for VIP users and critical services.

The bank now enjoys a better network experience and receives much fewer complaints. The access failure rate of video conferences has been reduced from 10% to almost zero.

1.2.3 Branch network

Typical pain points of financial branches:

- There are many branches, ranging from hundreds to thousands.
- There is a wide range of services, including traditional production office, security, IoT, mixed operations, and public cloud access
- The many access lines are widely dispersed, and frequent network faults are difficult to locate.
- Burst traffic may affect key businesses. Video conferences occupy a lot of bandwidths in a short period.
- Line leasing is costly.

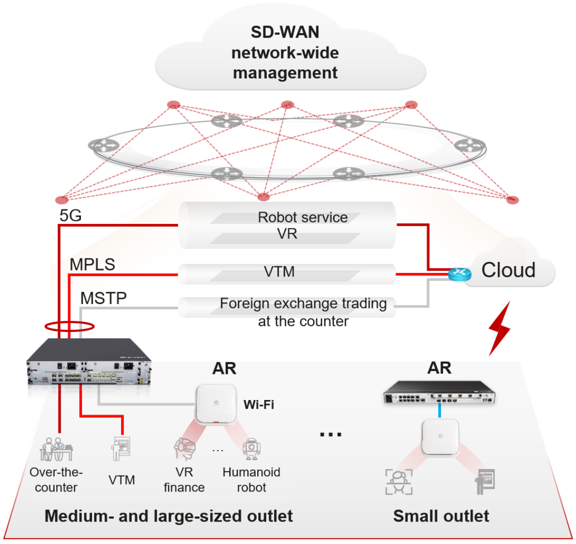


Figure 1.16 Simplified branch network built on SD-WAN

SD-WAN for simplified branches can solve many problems.

Simplified devices: An SD-WAN gateway can support routing, switching, PoE, firewall, IPS/AV, and 5G at the same time, providing hyper-converged access. One device can act as a branch network.

Simplified management: One unified network management platform centrally manages tens of thousands of branches, with only one platform and one network needed.

Simplified deployment: Site templates can be used for batch deployment. With 5G capabilities, 1000 sites are configured within one day, and businesses can be instantly provisioned.

Application assurance: In-depth application identification can distinguish key transaction, video, and other businesses to provide differentiated services. The private network preferentially guarantees transactional businesses and offloads non-key ones to the Internet. This vastly saves the bandwidth of the private network and reduces the lease costs of the private network.

Seamless switchover: The optimal link is selected to carry businesses based on the quality and congestion status of each link. This ensures optimal business quality and zero impact on businesses during link switchover.

[Case] 5G+ smart branches of the China Construction Bank (CCB)

CCB has built smart branches that integrate experience, conversation, and entertainment spaces. It collects a range of data through IoT awareness. Over 20 interactive games facilitate customer acquisition and retention. In addition, smart branches can be combined with auto banking, home banking, space capsules, and humanoid robots to enrich the customer experience.

CCB's branch network uses SD-WAN to centralize the production network, Internet, and IoT. It has built 5G+ MSTP fixed-mobile convergence private lines that improve the branch bandwidth 100-fold, with a measured rate exceeding 1 Gbit/s. The latency of remote verification and AR interaction is reduced by 70%. This facilitates over 300 financial services, including smooth AR and VR experiences. In addition, multi-link resource pooling based on SD-WAN improves bandwidth utilization. Intelligent traffic steering and optimization for critical applications ensure that services are not affected or interrupted during link switchovers.

Trend 2 .

Global central banks are transforming digital currencies, bringing digital payment back to the banking system

Basically, financial institutions need to ensure the accuracy, security, stability, and continuity of transactional systems. China has become a cashless society, with less than 3% offline transactions happening offline. Mobile and Internet payments are explosively increasing, and online promotion activities such as hot events and flash sales activities have triggered transaction peaks, which also pose great challenges to transaction systems. In China, the online transaction traffic during online shopping festivals will be more than 30 times the daily traffic. Booming online interactive transactions driven by influencer-led live shopping make the time and scale of transaction peaks more unpredictable. It remains a new challenge for financial transaction systems to address shocks from massive amounts of transactions.

Internet and mobile payments have disrupted traditional payment methods, giving rise to internet giants that dominate the market and create their own ecosystems. However, as digital currency is put into application, its security and convenience will reshape the payment product experience. And the E2E encryption mechanism avoids the possibility of internet acquiring user information and brings payment back to banks.

It is estimated that companies can save US$100 billion per year by using the digital currencies released by central banks for cross-border transactions. Seventy-eight countries are exploring the application of central bank digital currencies, and seven have officially launched them. More than 20 countries are conducting digital currency pilot projects. Switzerland and Singapore have piloted cross-border payment and settlement. China's pilot projects have covered up to 140 million consumers, opened 261 million personal wallets, and achieved US$9.5 billion in consumption expenditure. China plans to gradually expand pilot scenarios and establish corresponding regulations to apply digital currencies on the market as soon as possible.

Challenge 1. Transaction stability

- It is hard to ensure long-term stable and low-latency transactions due to the rapid growth of digital payments and complex IT systems.
- The traditional scheduling mode of IT resources cannot bear surging traffic caused by massive online transaction of users.

Challenge 2: Massive encrypted transactions supporting digital currency

- To ensure transaction security, a digital currency transaction needs more than 10 rounds of encryption and decryption. The encryption and decryption speed affects customer experience.
- Digital currency brings Internet payment back to banks, but banks' infrastructure cannot support the surging traffic.
- There is no intelligent regulation for encrypted transactions of digital currencies.

Suggested actions

1. Ensure the stability of real-time transactions. Build dedicated, reliable, and highly scalable software and hardware systems to ensure the stability and low latency of transaction systems under extreme conditions.
2. Build E2E trusted infrastructure, build a digital currency transaction system based on trusted computing, network, and storage and isolate the system from other systems. Perform digital currency encryption and decryption through the hardware system certified by the central bank.
3. Use distributed encryptable databases and distributed smart contract systems to ensure the scalability and security of digital transactions.

2.1 Real-time transactions with consistently low latency

The financial core transaction system mainly supports banking deposits, loans, remittance business, insurance purchases or claims, and counter or online securities trading. The following figure shows the positioning

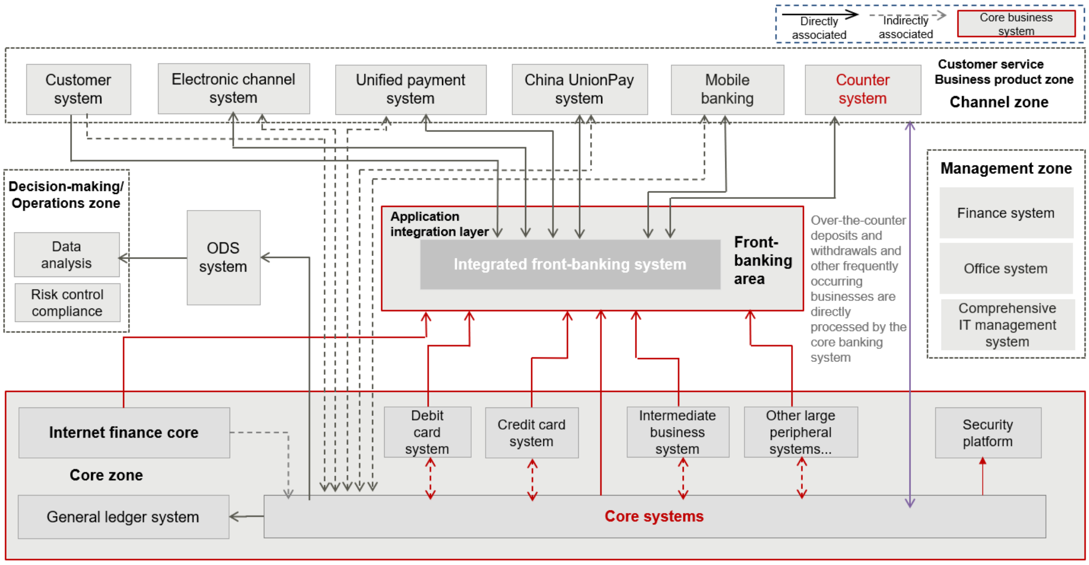


Figure 2.1 Core transaction system of banks

of banks' core transaction system in the overall business architecture(Figure 2.1 Core transaction system of banks)

The foremost requirements of the core transaction system when it comes to the infrastructure are:

1. Consistently low latency: Under the average traffic pressure, the latency of the whole transaction process — from when the front end receives the transaction instruction given by a customer to when the transaction is fully processed — must remain stable and within 120 ms.

2. Supporting traffic bursts: Traffic during peak hours generally exceeds 10 times the average daily traffic and 30 times in extreme situations. The transaction system must be highly redundant to ensure that the transaction system can bear traffic during peak hours.

3. Remarkable business continuity: Mobile finance, online transactions, and transactions across time zones are getting more popular, meaning transactions may take place anytime throughout the whole year. Therefore, the core transaction system should maintain a constant 99.99% robustness without fail.

2.1.1. Consistently low latency

（Figure 2.2 Transaction chain of digital payment）OLTP database is a key component of the core transaction system. One typical payment transaction chain conducts database reading and writing over ten times, and each database operation leads to reading data from or writing data into storage around 40 times, meaning this number would reach 500 to 800 during each typical transaction.

Each time the storage data is read or written with any delay, subsequent reading or writing will be postponed, resulting in wider and higher latency. It's like the phantom traffic jam — all it takes is for one car to touch its brakes and subsequent cars all have to brake harder, a chain reaction that slows traffic to a crawl. When this effect happens during reading/writing, it only congests more businesses, and

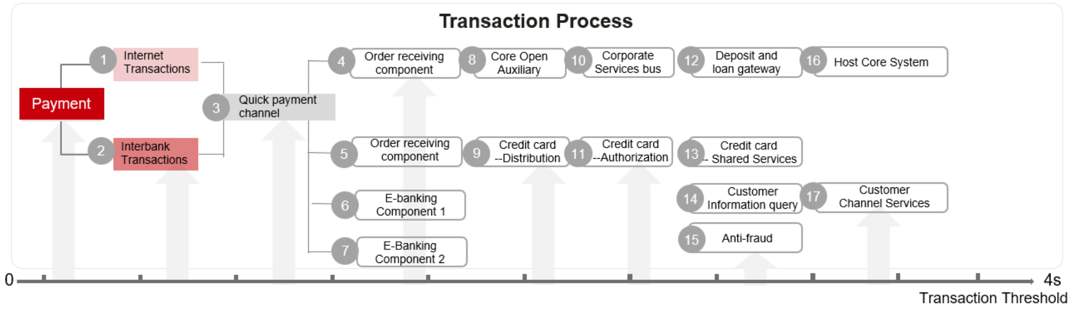

Transaction Process
Payment
1 Internet Transactions
2 Interbank Transactions
3 Quick payment channel
4 Order receiving component
5 Order receiving component
6 E-banking Component 1
7 E-Banking Component 2
8 Core Open Auxiliary
9 Credit card --Distribution
10 Corporate Services bus
11 Credit card --Authorization
12 Deposit and loan gateway
13 Credit card - Shared Services
14 Customer Information query
15 Anti-fraud
16 Host Core System
17 Customer Channel Services
0
4s
Transaction Threshold

Figure 2.2 Transaction chain of digital payment

they can only gradually get back to speed. This situation worsens during peak hours.

Therefore, to ensure E2E consistent transaction latency, the response latency of each data reading/writing must be stable. The core transaction IT system needs to guarantee the high performance and availability of the OLTP database. Therefore, we should configure dedicated compute, network, and storage devices that are high-performance and reliable to ensure business is not interrupted because of a component fault.

In the compute domain, mainframes and mid-range computers have recently gone into a decline after more than 40 years of development. Financial institutions are abandoning them and moving towards an open architecture. As a result, there are fewer Common Business-Oriented Language (COBOL) programmers, and many banks can no longer use COBOL to develop applications deployed in mainframes or mid-range computers.

X86 and ARM servers can be clustered to provide the same or even higher performance with flexible scalability compared to mainframes or mid-range computers. Based on open operating systems such as Linux, multiple languages can be used to develop the system, ensuring sustainable development and maintenance.

In the storage domain, all-flash provides over 100 times higher reading/writing performance and more stable latency than traditional hard disk drives (HDD). It has become a consensus in the industry to replace traditional HDDs with all-flash storage for core transactions. If we replace traditional fiber channels (FC) or SAS SSDs with NVMe SSDs, the number of data path hops will be reduced from four to two and the initiator and receiver need to interact with each other twice (seven times required when using FCs or SAS SSDs) during any single reading/writing process. This reduces the E2E latency by more than half.

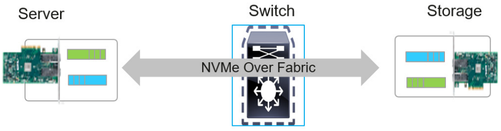


Figure2.3 NOF (NVMe over Fabric)

In the network layer, after upgrading to all-flash storage, compute and storage devices will suffer from performance bottlenecks. We can apply the NVMe protocol to replace the traditional FC protocol, which is the implementation of the NoF and can reduce the latency by more than 50%.

NoF can be carried on three types of underlying networks, including the FC network, traditional IP network, and zero-packet-loss Ethernet network (RoCE). However, the traditional IP network can only be used in non-critical scenarios since it cannot avoid congestion and packet loss. Therefore, RoCE Ethernet with zero packet loss technology has reached the same reliability level as FC, with 20% lower latency than it. The bandwidth of RoCE has increased to 400

Gbit/s/800 Gbit/s, far exceeding the 64 Gbit/s bandwidth of FC. Therefore, RoCE will become the main bearer network for NoFs.

Currently, new versions of key ecosystems such as Linux OS are now compatible with NoF, and have been applied by some banks.

2.1.2. Smooth failover

The financial IT systems are comprised of application and basic software, as well as hardware provided by various vendors. Among all these components, any fault will affect business and lower transaction performance, which can cause a ripple of delayed and unsuccessful transactions. Multi-layer protection is necessary to minimize the impact of faulty components on business, including resource redundancy, smooth business switchover, and E2E monitoring of the transaction train.

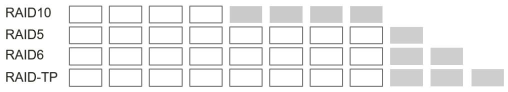


Figure 2.4 RAID resource redundancy

1. Resource redundancy: Take the infrastructure layer as an example. Redundant resources should be reserved at the storage, compute, and network layers to ensure that enough resources are available to protect the system when a fault occurs.

   Typically, the RAID technologies for storage, such as RAID5, RAID 6, and RAID 2.0, adopt the N+M (N copies of data and M copies of redundancy check) mode to protect data.

2. Smooth business switchover: Switchover takes place when a business is transferred from the faulty end to the redundancy protection end. An unsmooth switchover would interrupt business. To avoid this, we must achieve predictable faults, load balancing between the working end and protection, and automatic and fast switchover.

At the compute layer, owing to SLB load balancing and compute clusters, when a server is faulty, the business load must be shared by other servers in the cluster so as not to affect the business.

At the network layer, the active and standby paths of traditional DC networks use software handshakes to detect faults and complete switchovers. The detection process takes between 1 to 5 seconds. The hardware must be integrated with the Bidirectional Forwarding Detection (BFD) technology to shorten the detection time to milliseconds and cut the link switchover duration by 10 times. (Figure 2.5 Global load balancing of storage resources)

At the storage layer, the active-active architecture of storage controllers makes for global load balancing of the access path. The business of storage controllers and hard disks should be taken over if they are predicted to fail based on the monitoring of their performance indicators.

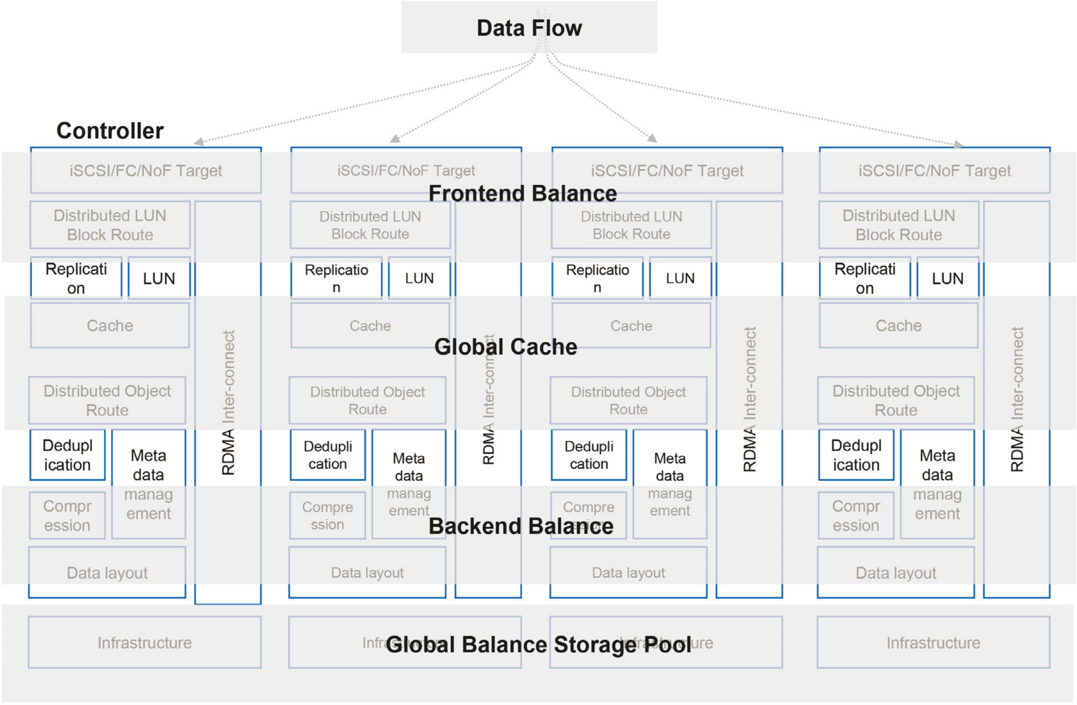


Figure 2.5 Global load balancing of storage resources

The RAID 2.0 technology can randomly distribute data to disk pools by data blocks (such as 64 MB) to achieve global load balancing. This ensures smooth failover and no impact on business when components like controllers and disks are faulty. The storage system monitors disks from multiple dimensions and automatically starts switchover in advance when detecting a potential failure of the disk, ensuring businesses are not affected.

2.1.3 Global monitoring and analysis

Traditional IT monitoring is implemented horizontally. This means the business system, application system, and infrastructure are separately monitored. When the indicator implies a potential fault and its source cannot be located instantly, such horizontal monitoring will lengthen the time of fault location and business interruption. (Figure 2.6 Cross-layer proactive monitoring and analysis)

We can resolve this problem using multi-layer associated O&M. We need to centrally manage the performance of businesses, applications, networks, and devices and seamlessly combine path analysis in order to quickly narrow down the scope of faults and precisely locate them.

[Case] Bank Z, a top bank in China, used to see an increased latency during demand peaks. Transactions frequently failed. This resulted from sub-healthy components with a high failure rate and the inability to quickly

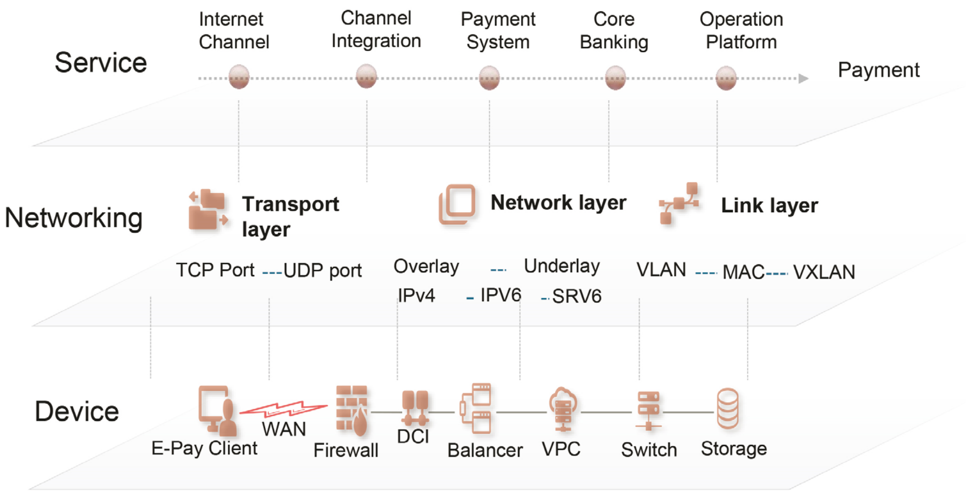


Figure 2.6 Cross-layer proactive monitoring and analysis

locate faults. To address these problems, Bank Z rebuilt and upgraded its systems.

1. Replaced traditional storage on transaction links with high-end all-flash storage and reached a consistently low latency of read/write responses (within 0.5 ms).

2. Replaced traditional FC connections with NoF, which has reduced the average response latency by 28%.

3. Allocated dedicated storage, compute, and network resources for systems on transaction links.

4. Monitors the entire transaction link using AIOps for quick fault locating and troubleshooting.

After system reconstruction, Bank Z has seen a tenfold decrease in disk failures and 80% fewer failed transactions. It can now locate faults within minutes.

2.2 Build E2E Trustworthy System

Trustworthy network:

Through the combination of "forward construction" and "reverse check" security concept and the combination of three elements of network protection, network uncertainty is continuously eliminated, and device, network, and management and control are trusted.

- Forward construction: Build internal security capabilities of devices and networks in the planning, design, development, and deployment phases to ensure that devices are secure and reliable when they are born, ensure that cyber resilience is available when services are rolled out, and build a deterministic trust chain transfer mechanism.

- Reverse check: During network operation, network traffic and log monitoring technologies are used to continuously monitor service changes and abnormal behaviors, monitor network security status in real time, and prevent risks and losses in a timely manner. The security brain is deployed on the network layer by layer to monitor security in all domains, implement security posture visualization, and implement threat correlation and network security collaboration and intelligent collaborative defense based on the security brain

Trustworthy data flow (Figure 2.7 Trustworthy Data Space)

Data may be forged, tampered with, and replayed by internal personnel or external hackers during circulation and processing, and may be obtained, disclosed, or abused by unauthorized personnel or organizations. Financial institutions rely heavily on internal and external data elements to support their business activities. The trusted data circulation mechanism is adopted to trace the whole process of data circulation, resolve security concerns of multiple parties, promote orderly sharing, exchange, and transaction of data elements between different subjects and boundaries, and fully release the value of data elements.

Trusted data flows are implemented in a unified manner. Trusted data flows are managed and controlled. Secure and trusted hardware capabilities (such as TEE and TPM) are used to build a trusted computing environment. Secure and encrypted networks are used for data transmission. In the trusted data space with a secure and trusted execution environment, all departments of the bank and data inside and outside the bank are shared and exchanged. Data is stored in a secure and encrypted storage resource pool to achieve E2E data trust.

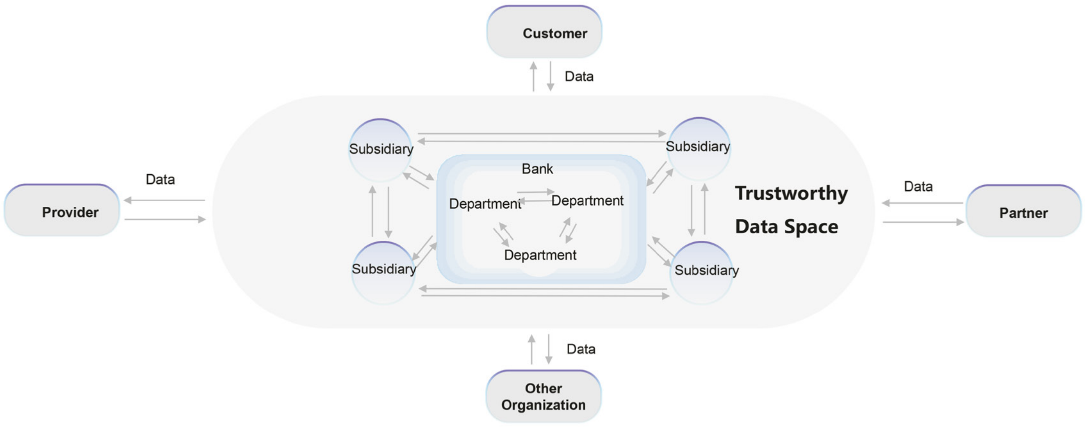


Figure 2.7 Trustworthy Data Space

Trend 3 .

Platform economy activates micro finance, with financing powered by AI

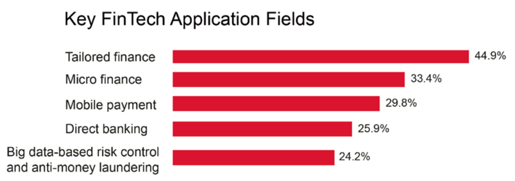


Figure 3.1 Tailored finance and micro finance ranked top two in the FinTech application survey

According to the Chinese Bankers Survey Report (2022), released by the China Banking Association, tailored finance and micro finance ranked the top two among 20 fintech applications.

China Merchants Bank (CMB) is an example of building tailored finance. CMB has built a Super App-centric ecosystem, bridging SMEs and individual users. CMB's Super APP has more than 100 million monthly active users. A leading regional bank in Myanmar grew its customer base to 12 million and increased its transaction volume by 280 times in just four years by developing a collaborative ecosystem.

However, a large number of banks are difficult to build a complete ecosystem in tailored finance, nor gain benefits from it.

In addition, many SMEs are generally less digital, lack of connectivity with industries and supply chain management capabilities. They are also difficult to integrate into the industry chain and banking ecosystem.

According to the data in the same report, SME loans and supply chain financing rank the top two among financial services that banks focus on. However, SMEs have urgent capital demand, fast turnover and huge industry differences, resulting in difficult financing, high costs and low efficiency. Moreover, the difficulty in evaluating SMEs' operating assets and the lack of credit systems make it hard for financial institutions to serve them through digital risk control and other technologies.

Traditional finance cannot meet SMEs' capital needs. In 2021, SMEs needed CNY104.3 trillion in loans, but only 48.5% of this demand was met. SMEs have an urgent need for inventory financing.

Statistics from the World Bank show that China's movable property, such as inventory, was worth between CNY50 trillion and CNY70 trillion in 2020, and financial institutions' annual balance of short-term loans is around CNY30 trillion, of which only CNY5-10 trillion was issued based on movable property. This is far lower than the 60–70% share in developed countries. China's inventory financing market has huge potential for growth and innovation.

However, there are no specialized inspection methods and no real-time monitoring for inventory financing. Some enterprises exploit these vulnerabilities to obtain loans or insurance benefits by deception and repeatedly mortgage movable property. So, banks are quite cautious about movable property financing. For example, Kingold Jewelry, one of China's largest gold jewelry manufacturers, has mortgaged 83 tons of gold to several trust companies since 2019 for financing. When the loan was overdue, these trust companies found that all the gold bars were gold-plated and made of copper alloy inside.

Challenge 1 Difficult integration of SMEs into the ecosystem of financial scenarios

- SMEs are less digital and lack of interconnection methods and capabilities.
- SMEs have low expectations and confidence in digital connection with banks.

Challenge 2 Difficult loaning and financing for SMEs

- SMEs are lack of connectivity with industries. There are no effective methods to evaluate their operating assets and no credit systems.
- The vulnerabilities to obtain loans or insurance benefits by deception raises banks' cautiousness about movable property financing.

Suggested actions (Figure 3.2 Building a B2B and B2C ecosystem through open banking)

1. Banks should increase the investment in open banking and leverage digitalization to enable SMEs to access the banking ecosystem through APIs. They should also operate platforms based on Super App or Market Place, forming a positive cycle of B2B and B2C partner ecosystem.

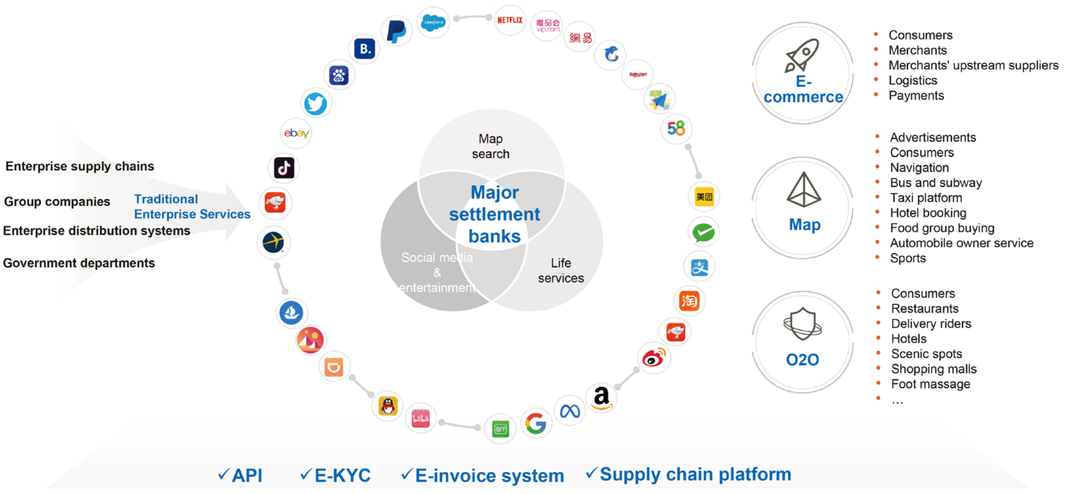


Figure 3.2 Building a B2B and B2C ecosystem through open banking

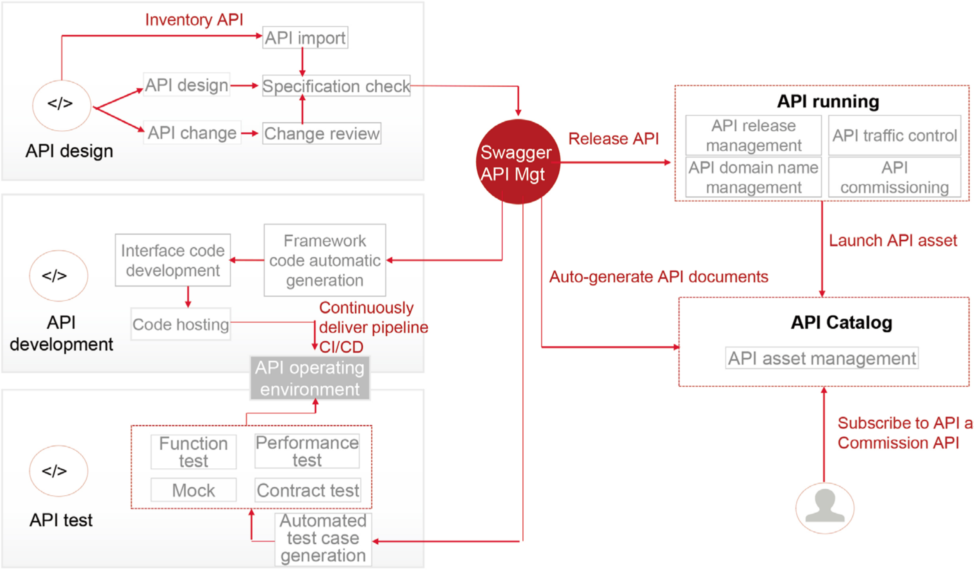


Figure 3.3 Unified API governance

2. Use IoT and AI technologies to achieve always-online real-time intelligent monitoring of pledged inventories. Identify pledged goods through AI and ensure its physical and ownership trustworthiness through automatic monitoring and analysis.

3.1 Open APIs for service innovation

APIs have become the core carrier for enterprises' connection business and internal and external service provision. With the fast development of open banking and enterprise internal applications, the scale and total calls of APIs in financial institutions increase rapidly. An API gateway aggregates APIs and provide full-lifecycle governance capabilities, centralizing API governance. (Figure 3.3 Unified API governance)

1. Provide internal and external services through APIs.

Interfaces are the only channel between services and external entities. Service providers cannot rely on service consumers nor expose internal technical implementation details. Services can interact with each other only through service interfaces. Released service interfaces cannot be changed. Service upgrades must be compatible with earlier versions.

2. Define service interfaces in the RESTful style.

REST is an architectural style proposed by Roy Thomas Fielding in 2000. Featuring universal and easy-to-use interfaces, it loosely couples components and enables high scalability in component interactions, making it easy for developers to invoke components.

3. Perform full-lifecycle API management through API gateway

API gateway aggregates APIs for enterprise systems providing external services and provides full-lifecycle API governance capabilities. It is the general entry for north-south traffic in the IT architecture. Full-lifecycle API operations and management are oriented for security, flexibility, and unified O&M:

Security: Multiple methods, such as security encryption, identity authentication, permission management, and traffic control, to ensure API security and reduce API openness risks.

Flexibility: Full-lifecycle management, including API design, creation, test, deployment, O&M, and removal, is provided and SDK API description documents are generated to improving API management and iteration efficiency.

Easy O&M: More convenient O&M tools, such as monitoring, alarm triggering, analysis, and API marketplace are provided to improve API O&M efficiency.

4. Develop open banking, effectively connect third-party ecosystems through APIs and cloud markets, accelerate scenario innovation, and form a win-win and sustainable financial ecosystem model.

[Case] On the Super App platform, a top bank in China onboards B2B and B2C merchants through APIs, building a complete ecosystem. It builds different life service zones in each city, and each zone provides hundreds of financial and life services. The bank combines offline marketing and live finance, attracting over 50 million monthly live users.

3.2 Cloud-pipe-edge-device collaborative inference

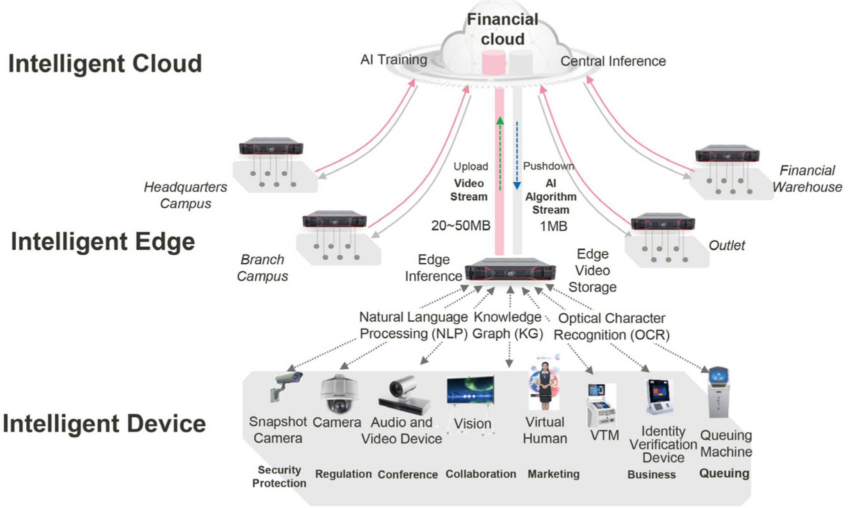


Figure 3.4 Cloud-pipe-edge-device collaborative inference

After training models are output, they need to be pushed to the AI inference server. Financial institutions need to build an AI inference collaboration mechanism covering clouds, pipes, edges, and devices to accelerate intelligent application deployment in specific scenarios.

- Devices with limited computing power have two main tasks: scenario awareness and raw data collection. They can perform primary AI inference where conditions permit. For example, some intelligent cameras have some AI-based recognition capability for videos and images and can be federated with nearby cameras. In this way, non-intelligent cameras can also implement image recognition.

- Intelligent edge devices have dedicated inference GPUs to complete most inference tasks, reducing the processing time and data transmission volume on WANs.

- On the cloud, high-computing servers are used to build a central inference cluster to process the high-performance inference tasks of banks.

- For WANs, SD-WAN is used for in-depth application identification. Through high-priority QoS service channels, AI models can be quickly pushed to edge AI servers in campuses and branches, ensuring efficient backhaul of edge data.

Through cloud-pipe-edge-device collaboration, intelligent applications can be deployed in intelligent branches, intelligent security, and inventory financing.

[Case] Inventory financing (Figure 3.5 Inventory financing)

China's inventory financing market space exceeds CNY20 trillion. Still, financial institutions cannot solve the issues of the physical, ownership, and value trustworthiness of inventories, and fraud occurs occasionally. Financial institutions face a dilemma.

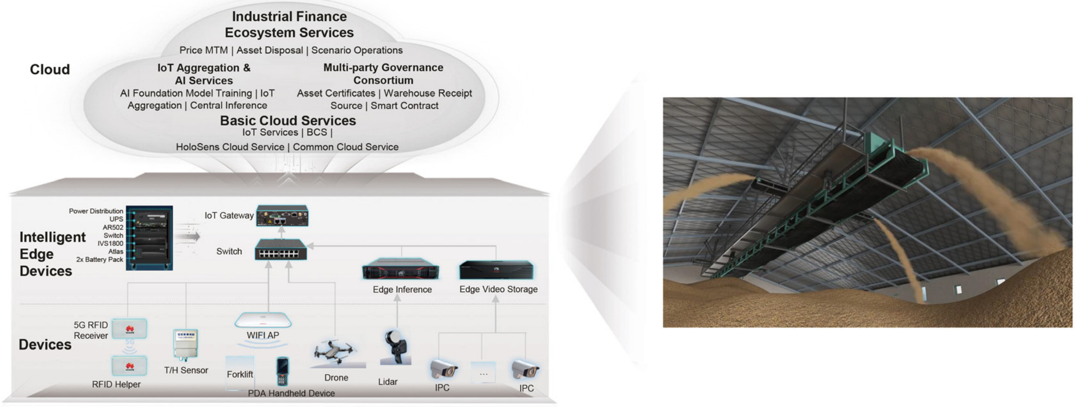

Cloud
Industrial Finance Ecosystem Services
Price MTM | Asset Disposal | Scenario Operations
IoT Aggregation & AI Services
AI Foundation Model Training | IoT Aggregation | Central Inference
Multi-party Governance Consortium
Asset Certificates | Warehouse Receipt Source | Smart Contract
Basic Cloud Services
IoT Services | BCS |
HoloSens Cloud Service | Common Cloud Service
Intelligent Edge Devices
Power Distribution
UPS
AR502
Switch
IVS1800
Atlas
2x Battery Pack
IoT Gateway
Switch
Edge Inference
Edge Video Storage
Devices
5G RFID Receiver
5G
RFID Helper
T/H Sensor
WIFI AP
Forklift
PDA Handheld Device
Drone
Lidar
IPC
...
IPC

Figure 3.5 Inventory financing

Shanghai Pudong Development Bank (SPDB) worked with Huawei to launch SPDB Finwarehouse. The solution uses cloud-based large models for training. One AI model covers nine logistics scenarios, including forklift, personnel, goods, receiving, inbound, storage, and outbound. Intelligent sensing and IoT devices, such as RFID and AI cameras, are deployed in standard warehouses through device sensing and edge-cloud AI collaborative inference. These devices transform typical warehouses into digital financial warehouses, extending financial services to onsite warehouse operations. The solution monitors cold chain and bulk dry goods 24/7. Inbound and outbound risk control provides real-time warnings. The solution also ensures that asset quantity is verified by multiple parties, enables objective and authentic asset assessment, and creates consensus and mutual trust in pre-loan rights confirmation.

Trend 4 .

New transaction frauds and intelligent network attacks threaten fund security

1. Banks always find fraudulent transactions difficult to prevent and control. Criminals often make fraudulent transactions through identity theft, account hijacking, cognitive vulnerabilities in new payment methods, and illegally acquired account and identity information. A bank in Southern Africa had to replace 12 million credit cards after employees printed and then stole its master key to access accounts and made more than 25,000 fraudulent transactions, stealing more than US$3.2 million from customer balances . Driven by technical progress, big data and AI have become powerful tools for professional criminals, who can use less covert means to swindle more people at lower costs, such as AI face swap and voice cloning, and virtual scenes.

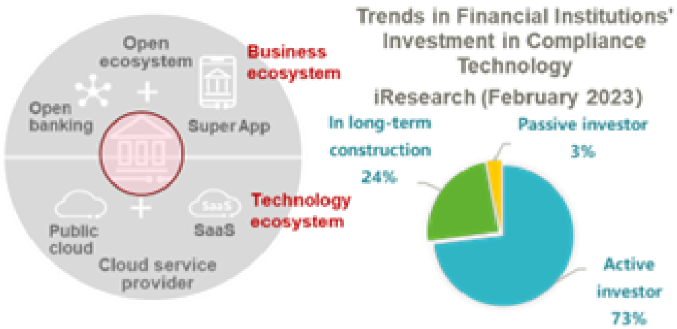
2. With the development of technology, network viruses are becoming more and more intelligent, clustered, and lingering longer. The pattern and intensity of cyber-attacks are increasing rapidly, resulting in higher data leakage costs. In 2021, banks saw 520% more virus attacks, and each virus breach caused a loss of US$7.83 million on average. It takes 16.3 days on average for an institution to recover from being attacked by a ransomware virus. Since 2021, global financial institutions have suffered more than US$1 billion in losses due to cyber-attacks.

   In March 2020, the target recognition system of a Chinese bank's app was hacked. Hackers used fake identities to register multiple accounts and resold them for profit.

   A top European bank was hacked three times within two years. Information about 4 million credit cards was leaked, and the bank lost over US$10 million.

   Some banks have begun to build forward-looking 3D in-depth defense systems to avoid problems. For example, J. P. Morgan has included "protecting the company and customers" as one of four major technology strategies, aiming to prevent cyber threats through proactive defense.

3. The financial industry has strict regulatory requirements, and regulators will heavily penalize financial institutions for money laundering, tax evasion, market manipulation, and illegal lending. In addition, enterprises in this space also need to reach certain business continuity and network security standards as required



Security compliance of external suppliers such as cloud service providers are not controlled. Third-party data compliance issues in the financial ecosystem are prominent.

73% of financial institutions plan to continue to invest in compliance technology in the next 3 years.

Figure 4.1 Compliance challenges faced by financial institutions

by regulators. If not, they will be facing financial or reputational losses.

（Figure 4.1 Compliance challenges faced by financial institutions）Financial institutions are also carrying heavy burdens on compliance. According to GlobalData, 50% of banks' spending is in response to regulations. All staff of a bank in China spent 14 days checking the compliance status due to a slight change in regulations. Traditional passive compliance mainly relies on manual work, so financial institutions cannot respond to regulatory policies instantly and find it challenging to define responsibility boundaries. Such a manual method is not sustainable.

Citibank defines compliance as one of the four major areas of transformation (data, risk and control, financial infrastructure, and compliance). It controls compliance through automatic evaluation and predicts the impact of policy changes to make contingency plans.

Challenge 1 Fraudulent transactions made via covert means

- Difficult and slow to identify identity forgery and false information, resulting in fund loss risks.

Challenge 2: Information leakage caused by attacks and ransomware

- The leakage of key data causes property and legal risks.

- Cyber-attacks and ransomware encrypt and steal critical data, causing large-scale unavailability of financial services and lasting trust crisis.

Challenge 3: Inefficient compliance audit and difficult responsibility definition

- Traditional compliance is inefficient. Data is required to be archived for more than a decade, and usually, a large amount of data needs auditing. Experience-based passive compliance cannot adapt to changes in regulatory policies and businesses immediately.

- Open banking and super apps blur the boundaries between financial institutions and third-party institutions, making it difficult to define compliance responsibilities

Suggested actions

1. Establish an integrated data and AI platform and introduce data from various departments of banks and third-party institutions with public trust through data lakes; let AI learn new fraud models to update risk control models promptly, and achieve T+0 real-time risk control through big data and AI collaboration.

2. Develop a protection platform featuring confrontation between the red team and blue team, self-checks, and intelligence sharing to enhance security measures. Build a zero-trust security authentication and management mechanism, as well as an in-depth security protection system based on the security operations center.

3. Improve data security resilience while strengthening network protection. Use timely and effective data backup to quickly restore business systems when production data is encrypted by ransomware.

4. Build a regulatory framework using machine learning and AI. Introduce AI supervision to identify policy semantics and replace manual retrieval and audit with RPA (Robotic Process Automation). This significantly improves efficiency and prevents human errors.

5. Build a centralized log archiving platform to provide content management capabilities and facilitate quick searches. This will enable high scalability to support long-term storage (over a decade) and unified DR capabilities to prevent log losses.

6. Purchase cloud services and software and hardware systems with compliance qualifications. Build business systems using compliant products and ensure quick system rollout in compliance with regulations.

4.1 Build Real-time Risk Control Platform

Traditional data analysis usually lacks unified planning, which has the following disadvantages:

- BI (operation report) and DI (big data analysis) are constructed independently. BI uses a data warehouse to collect structured data in batches, while DI uses a data lake to analyze unstructured file data. They are not connected to each other.

- Isolated data and siloed construction cannot describe the complete journey of users, nor provide accurate suggestions on user behavior.

- Data migration from other clusters delays data analysis by T+1 days and prevents real-time or near-real-time decision-making.(Figure 4.2 Data analysis of data lakehouse)

Suggested actions

- Adopt a decoupled storage and compute architecture. A unified storage resource pool stores various data from multiple data sources, providing data resource sharing.

- Build a data lakehouse platform to support BI and DI with a unified data management engine. The platform simultaneously supports data batch and stream processing, enabling multi-dimensional user profiles.

- Use a real-time decision engine (RTD) to speed up analysis and build second-level data analysis capabilities. This will speed up analysis result output from T + 1 day to T + 1 hour or even T + 10 minutes.

- [Case] The Bank of Communications (BOCOM) has built a comprehensive data analysis platform based on a data lakehouse for data governance. The customer conversion rate increased by 164%. Real-time fraud identification reduced the number of risk incidents by 52%. The real-time T+0 report search and analysis of banking services quickly support decision-making.

- [Case] China Merchants Bank has replaced its traditional data warehouse with a new data lakehouse platform. The platform has over 4,000 nodes, supports customer journey analysis from 4,000 dimensions, and provides real-time data services for thousands of market development and operations personnel. Also, the platform analyzes over 1.5 million transactions daily, reduces fraudulent transactions by 82%, and prevents tens of millions of dollars in fraud losses each year.

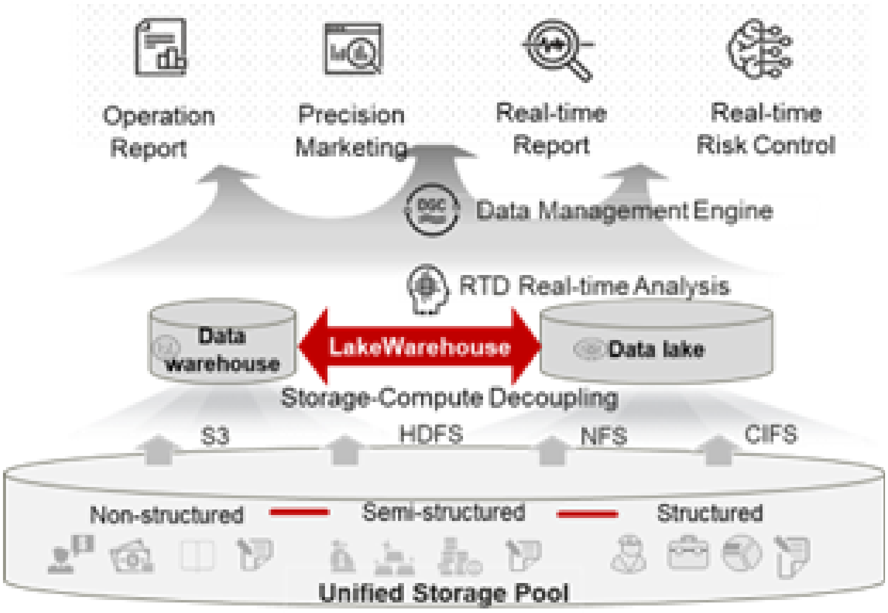


Figure 4.2 Data analysis of data lakehouse

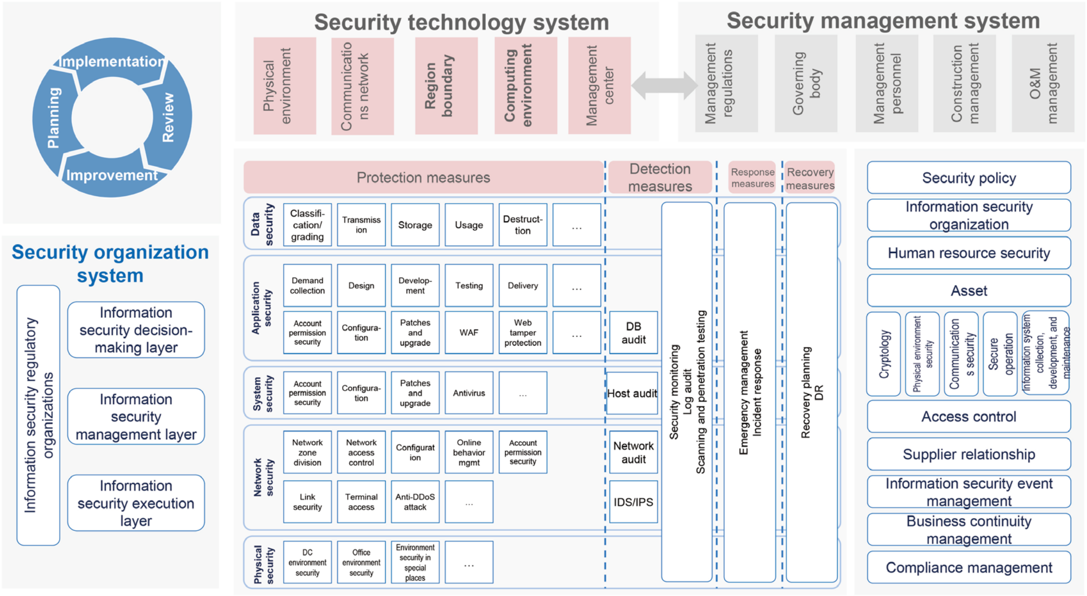


Figure 4.3 Financial security organization system

4.2 Overview of financial security management

(Figure 4.3 Financial security organization system)Security governance in the financial industry involves organization, technical, and management systems. Financial institutions should manage security from many aspects, including organization, technology, process, and culture. The technical system alone concerns the physical environment, network, system, applications, and data.

(Figure 4.4 Security defense in-depth model) The figure above is the security defense-in-depth model. It shows the route of a typical virus attack, which starts at the perimeter and then affects the network, application, host, and finally, storage. Financial enterprises must build five security protection levels along the attack route.

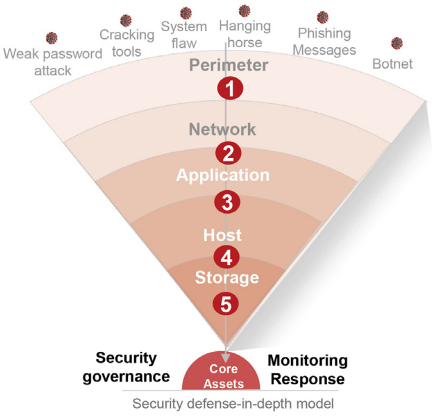


Figure 4.4 Security defense in-depth model

4.2.1 Network perimeter security

The secure access service edge (SASE) provides various converged networks and SaaS (security as a service) functions based on cloudified services, including SD-WAN, SWG, CASB, NGFW, and zero-trust network access (ZTNA).

SASE helps branches, remote staff, and enterprises securely use the Internet. It is delivered as a service and supports ZTNA based on the identity of the device or entity. This is combined with real-time context as well as security and compliance policies.

There are three typical external access scenarios in branches, including:

- Branch Internet Access (BIA)
- Branch Private Access (BPA)
- Branch SaaS Access (BSA)

Following a radical cloud migration policy, a bank migrates 60% of its businesses to the public cloud. With a complex traffic access model, point-to-point fully connected networking among branches, campuses, and cloud service providers is required. This exponentially increases the number of network and security policies. However, traffic models and policies change rapidly, making policy O&M difficult for the bank.

Traditionally, the network is the core security concern for branches, so banks focus on protecting the border since DCs feature radial networks. Now, identity-centric near-source protection and refined management and control should replace the traditional protection method. This is possible if banks use a full-mesh cloud network supporting point-to-point mutual access at any two ends. (Figure 4.5 SASE security solution)

SASE consists of four components: SASE cloud service platform, security management engine, security gateway, and terminal security client.

1. The cloud service platform centrally analyzes network and security data, features centralized network and security operations, and centrally manages security

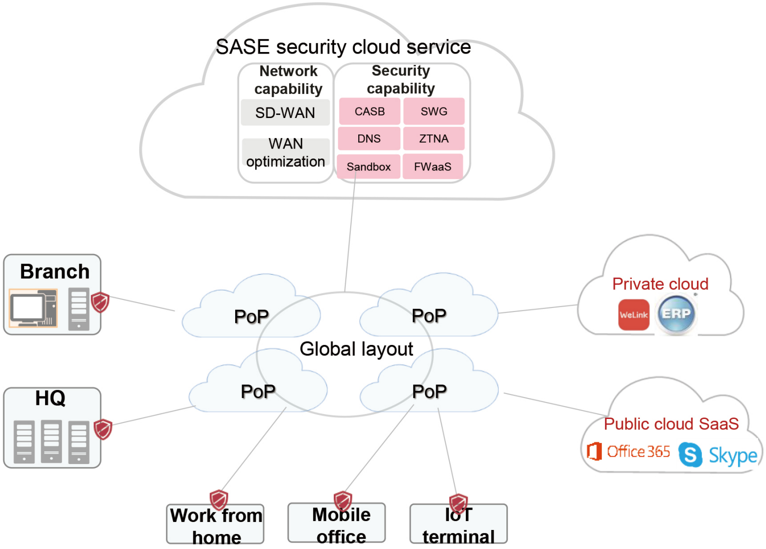


Figure 4.5 SASE security solution

policies. It is deployed on the cloud of carriers or MSPs.

2. Banks can create PoP access based on the backbone network of carriers or MSPs. They can deploy value-added security services at POPs to provide software firewalls, IPS, IDS, VPN, and ZTNA, ensuring centralized and trusted edge access.

3. There are two ways to deploy the security gateway.

   Mode 1: Distributed SASE solution

   The security gateway is directly deployed at the access edge of branches and outlets. This mode is ideal for headquarters and multiple medium-sized and large branches requiring high bandwidth.

   Mode 2: Centralized PoP SASE solution

   Security gateways are deployed at PoPs. This mode is ideal for small branches and mobile offices that require low bandwidth. Multiple branches share the same security gateway based on public PoP.

4. A security terminal includes endpoint detection and response (EDR) and the endpoint protection platform (Evolved EPP). Terminal security is moving towards converged EPP and EDR. Converged security terminals empower banks with basic capabilities such as EPP access control, ZTNA, and compliance checks.

An in-depth defense that features device-edge-cloud synergy ensures E2E anomaly detection, access blocking, network-wide monitoring and analysis, and security policy updates, making branches and outlets safe and secure.

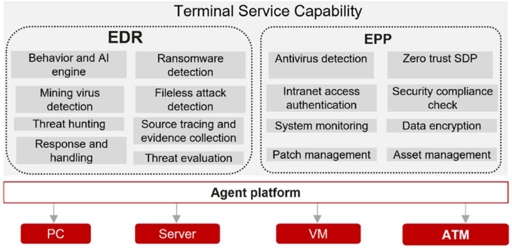


Figure 4.6 Security terminal

4.2.2 Zero trust protection

Zero trust represents a next-generation cyber security protection concept. It prevents visitors from accessing resources without authorization, dynamically evaluates the security status of network access, and dynamically grants access rights to visitors.

Under this concept, the identity of all visitors must be authenticated before they are authorized to access resources. Also, multi-dimensional and associated identity identification and authentication are required for terminals and application software, and their identities may be authenticated repeatedly during access.

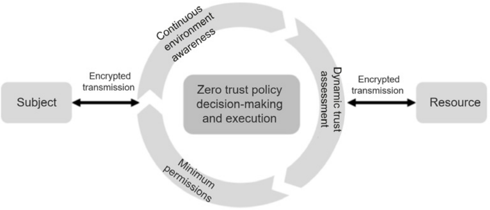


Figure 4.7 Zero trust access model

According to the National Institute of Standards and Technology (NIST), there are three ways to build a zero-trust architecture (ZTA). These are software-defined perimeter (SDP), identity and access management (IAM), and micro-segmentation (MSG).

- SDP enables fast, on-demand security boundary deployment, which isolates services from insecure networks.

- IAM can transfer identity information (revocation and resignation, expiration, abnormality, etc.) to the zero trust system. Then, the system assigns permissions to users, establishes trust relationships between terminals and resources for user identities based on a unique identifier, and blocks the risks it detects.

- MSG is a cyber security isolation technique that logically divides a DC into different security segments and defines access control policies for each independent security segment.

Typical Business Scenario	Category
Bank employees access office applications such as office automation (OA) through the intranet.	Internet access
Bank employees access the bank's data via the Internet and remotely access office applications such as OA.	Remote access
O&M personnel access the bank's data via the Internet for remote O&M.	
Financial institutions need crowdsourced testing for their new businesses. Testing personnel need to access business applications remotely.	
Regulators or business partners access business applications within the organization through the extranet.	
Terminals connected to the production network	IoT access
Terminals connected to the office network	
Monitoring devices connected to the network	
Banks cooperate with third parties to build open banking for a better user experience.	API security
Access to DC applications and devices	DC security

- **Zero trust intranet access:** It can eliminate the security risks multi-purpose office terminals face. The network isolation mechanism automatically disconnects terminals from the Internet and decreases the attack surface of applications and the data isolation mechanism prevents data leakages.

- Zero trust remote access: When a remote user initiates an access request, the convergent client first checks the health status of the terminal. If it is health-compliant, the SDP gateway will open the access paths to the user. This allows the remote terminal to establish a connection with the internal application, ensuring "authentication before access" for businesses. In the financial industry, converged clients usually have security capabilities such as antivirus and terminal sandbox.

- Zero trust IoT access: On-campus terminals access the network through campus access switches. Access switches authenticate the MAC address of terminals, including dumb terminals, to ensure that only authorized terminals can access the network. Unauthenticated terminals will be rejected.

- Zero trust API security: The API gateway provides secure business access capabilities, including channel encryption, traffic limiting, circuit breaker, and API security protection.

- Zero trust DC: A typical financial DC consists of the central business zone, user access zone, extranet access zone, and O&M management zone. Internal and external users access the network and DC resources from the user and external access zones, respectively. The daily operations of a DC cover four typical scenarios: application access, service invoking, cross-network data exchange, and O&M management. Financial institutions need to manage zero trust access in these four zones and scenarios.

4.2.3 DC anti-ransomware

Ransomware attacks core financial data in four phases: intrusion, spread, encryption, and self-destruction. Ransomware has the following four principal risks.

1. Difficult to crack the encryption: Currently, the length of a ransomware key is 2048 bits. Once a file is encrypted by ransomware, it is impossible to crack the encryption.

2. Data leakage: When a user refuses to pay a ransom, the attacker will publicize the data.

3. Horizontal transmission: When ransomware enters the production system, it will sweep the system horizontally to infect files in the local and remote backup centers.

4. Copy pollution: When the source file is infected with the virus, the backup file also carries it.

Network anti-ransomware

Network security devices constitute the first protection wall of the DC.

1. Anti-intrusion at the network border: Firewall IPS features, threat intelligence,

virus sample detection, and detailed ransomware monitoring and identification can block ransomware threats in seconds, preventing vulnerability exploitation and blocking hacker attacks. Combined with an AI detection engine and machine learning classification algorithm, they prevent slow brute-force cracking. The sandbox can detect malicious files to prevent hackers from implanting malicious programs on the network.

2. No spreading across the network: Terminals, security devices, and networks are associated to stop viruses from spreading to the production environment. Situational awareness intelligent analysis is combined with the security manager, network controller, and EDR, helping quickly deliver handling policies and block installation, horizontal spread, and data backhaul of malicious programs.

3. Situational awareness allows threats to be globally visualized. Based on this, operators can view the status of threats, assets, intranet threats, and website security, and manage threat events on one screen.

Storage anti-ransomware

Ransomware attacks storage last. When ransomware enters the data storage layer, it has already passed the interception, detection, and protection measures at the network, application, and computing layers, which have been invalidated. As such, storage anti-ransomware is critical.

E2E anti-ransomware prevents viruses from entering the storage system or tampering with data, completely protects isolation zones, and ensures data can be restored quickly.

Blocked ransomware

When known ransomware starts to attack, the detection and analysis blocklist of the storage device can identify its characteristics and block it in advance. This prevents the virus-infected files from being written into the storage system.

Untamperable data

WORM (write once read many) can be set for key files to prevent data from being tampered with. Ransomware and unauthenticated users cannot modify files with a WORM attribute.

Strong isolation

Data security isolation zones are set in DCs. AIR GAP technology can physically isolate the storage in the production zone from the storage in the isolation zone; it periodically establishes transmission channels to replicate data from the former to the latter. When data is not being transmitted, the transmission links will be completely disconnected to ensure that these two zones are completely isolated. In this way, they are only connected for 1% of the time. This vastly reduces the probability of an attack on the data in the isolation zone.

Rapid recovery

After the ransomware is cleared, the

administrator can restore the data through a secure snapshot or the storage system in the isolation zone. However, it is important to ensure that the data at the recovery source is clean.

Secure snapshots record data changes at a given point in time and can restore the data from that period within minutes or even seconds. However, if all of the data is damaged, secure snapshots cannot restore the data.

Production and backup storage can be used for storage in the isolation zone. Production storage features fast data recovery but could incur high costs. Meanwhile, backup storage features low costs but slow recovery. An all-flash backup solution would be the most desirable. With all-flash backup, the data recovery speed can reach 50–100 TB/hour, which is three to five times faster than traditional HDD backup. It improves the recovery speed and reduces costs.

[Case] In May 2021, the Hong Kong Monetary Authority (HKMA) invited the Hong Kong Association of Banks (HKAB) to develop Secure Tertiary Data Backup (STDB) guidelines for the Hong Kong banking industry, requiring all financial institutions to back up more than three copies of data. This ensures that businesses can quickly recover after cyber-attacks. A local bank combines the all-flash backup solution with backup software and uses Air Gap to isolate data, which enables quick data restoration. Testing shows that it takes less than 30 minutes to recover each TB of data.

Storage-network collaborative anti-ransomware

With technologies evolving, network hackers keep developing new ransomware, finding new ways to attack networks. Even if financial institutions build multi-layer protection walls and can identify 99.9% of viruses, they still fail to identify the remaining 0.1%. Therefore, protection against ransomware requires financial institutions to break the barriers between network, compute, storage, and data protection processes to create complete anti-intrusion, anti-proliferation, virus detection, data encryption, security isolation, and rapid recovery processes. (Figure 4.8 Storage-network collaborative anti-ransomware)

Financial institutions should also set up a handshake mechanism between the security situational awareness engine on the network side and the data management engine on the storage side. This will enable network-storage collaboration.

- When the network-side antivirus database is updated, it is synchronized to the storage side; then, the storage updates it to the virus detection database.

- When a new virus is detected on the storage side through abnormal behavior detection, the new virus will be synchronized to the network side, adding the new virus to the firewall's antivirus database.

- When detecting a virus attack, the network side immediately notifies the storage side. Then, the latter immediately performs

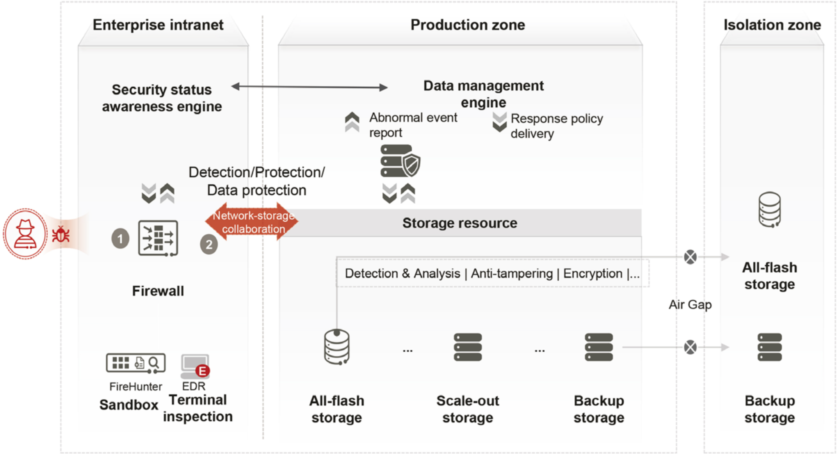


Figure 4.8 Storage-network collaborative anti-ransomware

protective operations, such as secure snapshots and AIR GAP isolation, to ensure clean data is stored.

Efficient network-storage handshake and collaboration combine the first and last step of DC security protection, eliminating all risks.

4.3 Compliance Infrastructure

4.3.1 Archiving infrastructure

Regulators have specific requirements on the data retention period and data type, which may vary from region to region. The Data Governance Measures for Commercial Banks published by the People's Bank of China approached the issue by stipulating different retention periods for different types of data, for example.

- Business voucher: Refers to important banking records, including certificates of deposits, withdrawals, transfers, and remittance, which need to be kept for five years.
- Business ledger: Records the process and results of banking business, including the opening or closing of accounts, deposits and withdrawals, transfers, and loans, which need to be kept for a decade.
- Account archive: Refers to the basis of banking business, including basic customer information, ID card copies, and signing agreements, which need to be kept for a decade.
- Risk management files: Includes risk assessment reports, internal control systems, anti-money laundering measures, etc., which need to be kept for 15 years.

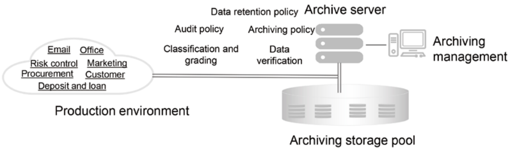


Figure 4.9 Typical archiving solution

Data must not be altered or lost when archived. It should be possible to quickly search archived files.

Archiving devices should be WORM compliant. Data can be read multiple times but cannot be rewritten. The WORM file system can be divided into regulatory compliance WORM (WORM-C) and enterprise WORM (WORM-E).

WORM-C is used for legal compliance audits. The written data cannot be modified. This ensures that complete and accurate records can be provided to regulatory auditors and prevents transactional and operational records from being tampered with.

Financial institutions are responsible for setting permissions for WORM-E, and enterprise security administrators are responsible for setting key records to prevent them from being tampered with.

Typically, text logs and audio or video recordings are archived. Europe has detailed requirements for General Data Protection Regulation (GDPR) compliance, and China requires data-related operations to be audio- and video-recorded. Therefore, financial institutions in these two regions often need to archive over 100 PB of data.

Archived data can only be accessed a limited number of times. Accordingly, there are no high requirements for the archiving system's performance. Therefore, data is usually archived using low-cost media. Tape is the cheapest media, followed by Blu-ray storage and super-large-capacity HDDs. Currently, only one supplier provides archiving drives, with upgrades required every 5 to 7 years. Generations are not compatible with one another, so dumps are required every 5 to 7 years. In addition, tapes are difficult to maintain and are vulnerable to damage; they also deliver poor access performance.

At present, many banks have adopted object-based storage, mainly large-capacity HDDs. If regulatory policies allow data archived for a long time to be migrated to the cloud, the public cloud would be a desirable choice. If not, data can be archived on Blu-ray storage or tapes.

Hundreds of millions of historical files are usually archived for a long time. It takes several days to find the required files if they are stored on traditional tapes or HDDs. SSDs can store the metadata (file index) of object files so that searches can be performed in minutes or seconds instead of days.

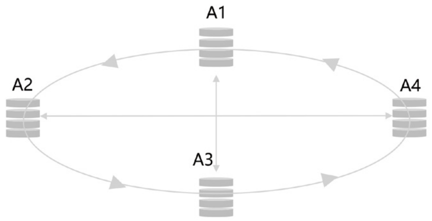


Figure 4.10 Cross-site multi-active objects

However, regulators have strict requirements on the preservation of archiving systems, one of which is zero RPO. Therefore, the archiving system must have active-active or triple-active capabilities. If one DC is faulty, other DCs can take over businesses and data can still be accessed as normal.

4.3.2 Compliance certifications

Global standards organizations such as ISO and regulators in each country or region have formulated certification systems for quality, security, business continuity, and social responsibility. Financial institutions must first obtain industrial and regional compliance certifications to provide regulation-compliant services.

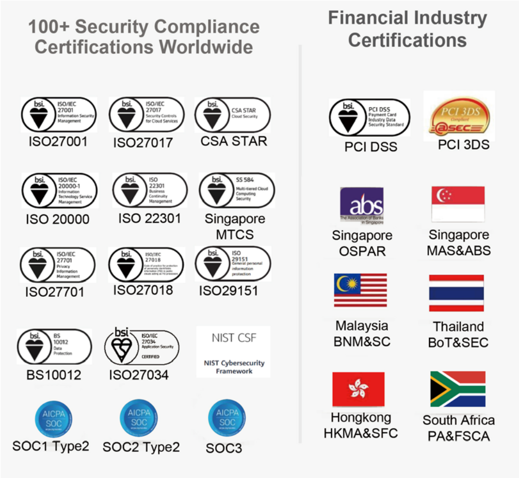


Figure 4.11 Global and regional compliance certifications

Compliance certifications related to the financial industry include the following:

1. Global universal information security standards

ISO27000-series (ISO27K)	ISO27K comprises information security standards published jointly by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC).
Common Criteria (CC) for Information Technology Security Evaluation	CC is an international standard for computer security certification developed based on several information security criteria, including TCSEC and FC of the United States, Europe's ITSEC, and Canada's CTCPEC. CC has two key components, one of which is the Evaluation Assurance Level (EAL). There are seven EALs.
Payment Card Industry Data Security Standard (PCI DSS)	PCI DSS was developed by the Payment Card Industry Security Standards Council, jointly set up by American Express (a founding member of the council), Discover Financial Services, JCB, MasterCard, and Visa. PCI DSS has six goals and 12 requirements concerning information security management systems, network security, physical security, and data encryption.
CSA STAR Certification	Designed to address specific issues related to cloud security, CSA STAR Certification is a joint effort by the British Standards Institution (BSI) – the founder of global standards, and the Cloud Security Alliance (CSA) – an international cloud security authority. CSA STAR comprehensively evaluates the organization's cloud security management and technical capabilities from five dimensions: 1. 1) Communication and stakeholder engagement 2. 2) Policies, plans, and procedures, and a systematic approach 3. 3) Skills and expertise 4. 4) Ownership, leadership, and management 5. 5) Monitoring and measuring
System and Organization Controls (SOC) Reports	SOC is a suite of third-party independent audit reports developed by the American Institute of Certified Public Accountants (AICPA) to audit the internal control of service providers. It is a globally recognized security privacy audit standard. SOC 2 is a dedicated audit standard that defines criteria for managing network security and privacy protection based on five trust service principles: security, availability, confidentiality, process integrity, and privacy.

2. Global DC standards

Uptime Institute Tier Standard

The Uptime Institute is a globally recognized DC standards organization and an independent certifier. It has created the DC Tier classification levels as the international standard for DC performance. It has defined standards for DC electrical parameters, redundancy, floor loading capacity, power supply, cooling equipment, and costs. There are four Tiers for the classification of DC equipment rooms.

	Tier Ⅰ Basic Site Infrastructure	Tier Ⅱ Redundant Site Infrastructure Capacity Components	Tier Ⅲ Concurrently Maintainable Site Infrastructure	Tier Ⅳ Fault-Tolerant Site Infrastructure
Distribution paths	1	1	1 active and 1 alternate	2 simultaneously active
Minimum capacity components to support the IT load	N	N+1	N+1	2(N+1) or (S+S)
Downtime time (per year)	28.8 hours	22 hours	1.6 hours	0.4 hours
Availability	99.671%	99.749%	99.982%	99.995%
Power supply	UPS	UPS and generator	UPS and generator	UPS and generator
Redundant component	Not required	System	System, power	All components
Cost/ft2	US$450	US$600	US$900	US$1100+

TIA-942 Telecommunications Infrastructure Standard for DCs

The TIA-942 Standard was developed by the Telecommunications Industry Association (TIA) under the American National Standards Institute (ANSI). It defines DC availability tiers based on the four tiers from the Uptime Institute and adds definitions for telecommunications infrastructure and architecture. These tiers cover the overall DC (Rate 1–Rate 4), telecommunications (T1–T4), architecture (A1–A4), electric power (E1–E4), and machinery (M1–M4). The standard provides a specific implementation guide for each category, except for the overall DC.

3. National and regional certification standards

In addition to standards widely used across the globe, regulators in each country and region have developed various certification standards or local executive standards. These are usually based on the existing global standards.

China's national standard GB/T 25070-2019 defines cyber security protection levels. Protective measures, security management systems, and emergency response mechanisms are required to ensure physical, network, host, and application security. For example, banking applications should reach Level 3 security protection.

Damaged Object	Degree of Damage Suffered by the Object		
	General	Severe	Extremely Severe
Legitimate interests of citizens, legal persons, and other organizations	Level 1	Level 2	Level 3
Social order and public interests	Level 2	Level 3	Level 4
National security	Level 3	Level 4	Level 5

Commercial cryptography application security evaluation issued by China: China's national standard GB/T 39786-2021 Information Security Technology — Baseline for Information System Cryptography Application sets out the requirements for cryptography application technologies to ensure physical and environmental security, network and communications security, device and computing security, and application and data security. Enterprises must evaluate whether commercial cryptography is used compliantly, correctly, and effectively.

Regulations in Singapore

Technology Risk Management (TRM) 2021 issued by the Monetary Authority of Singapore (MAS)	• Deploy technology risk governance and supervision: Set different roles and responsibilities based on the technology risk governance framework to effectively manage information assets and third-party services. • IT project management and supplier selection process: Establish standards and processes for supplier selection, monitoring, and assurance to avoid the risk that suppliers cannot meet the requirements of financial institutions. • Strengthen cyber security and defense methods: Regularly conduct scenario-based network drills and simulate defense against attacks.
Guidelines on Outsourcing issued by MAS	• Establish a sound outsourcing risk management framework to effectively manage direct third parties, subcontractors, and outsourced infrastructure within the group. • Self-assess all existing outsourcing arrangements based on the Guidelines. • Rectify the deficiencies found no later than 12 months from the effective date of the Guidelines.
ABS Cloud Computing Implementation Guide 2.0	• Ensure cloud computing due diligence framework with KPIs/key risk indicators included. • Materiality outsourcing assessment: financial strength and resources, corporate governance and entity control, DC location, and physical security risk assessment for DCs in Singapore and other countries • Implement baseline controls to cover the entire lifecycle of the cloud. Specific measures should include governance, design, and assurance (pre-deployment) and cloud running (continuing operations).
Multi-Tier Cloud Security (MTCS) Standard for Singapore	MTCS is the world's first multi-tier cloud security standard and provides cloud computing certification for CSPs. MTCS defines three levels of cloud security. • Tier 1: Designed for non-business critical data and systems with basic security control. • Tier 2: Designed for organizations that use cloud services to protect business or personal information. • Tier 3: Designed for companies with specific needs and more stringent security requirements, such as the public cloud service provider.

Financial institutions should use products and solutions with industrial and regional compliance certifications to build financial infrastructure. For example, they can use high-reliability and high-security devices to develop unified DR, security protection, and intelligent O&M solutions for DCs to meet business security and availability certification requirements. Also, they can use resources and services from cloud service providers that have received industrial and local regulatory certifications to achieve compliant cloud migration.

Trend 5 .

Accelerated transformation and large-scale growth of IT drive systematic construction of business resilience

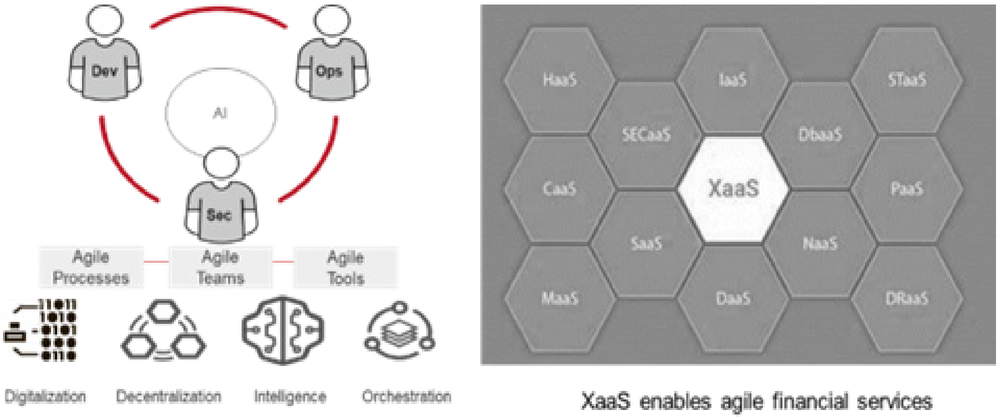
Recently, a leading bank in Thailand launched a commercial loan program in the digital marketplace. It took the bank only six months to acquire 250,000 new customers after building a new digital loan system.

According to Boston Consulting Group, agile transformation has a significantly positive impact on costs, delivery speed, customer satisfaction, and employee engagement:

- Two to four times faster delivery
- A 10% to 20% increase in customer satisfaction and digitalization ROI
- 15% to 25% lower development costs, and two to four times faster production.
- More than 70% to 90% of dedicated employees in agile organizations

Compared with traditional business competition that relies on branch scales, the competitiveness of banking products in the digital economy is determined by the speed of innovation and iteration. Banks' IT architectures are moving from closed to open, enabling banks to achieve shorter time to market, faster user feedback, and swifter product optimization.(Figure 5.1 Agile banking business)

Mainframes and AS series midrange computers would incur high O&M costs, and the number of developers for applications



XaaS enables agile financial services

Figure 5.1 Agile banking business

deployed on them is decreasing, yet the rest developers cannot flexibly adapt to business development. Therefore, banks will inevitably migrate the core transaction system to an open architecture. However, it is difficult for an open system to reach the same level of latency and reliability as mainframes and midrange computers.

With the help of a mainframe supplier, a head bank in Türkiye successfully moved its core system off four mainframes to more than 7000 servers. This means it will have to maintain thousands of times more devices. The integrated architecture of mainframes ensures low system latency while their active-active DR mechanism has been tested by banks for more than 30 years and proven to be feasible. However, the horizontal data synchronization between nodes in a distributed system makes it difficult to achieve a latency that can be delivered by mainframes.

DR of some banks cannot meet 99.999% of system requirements. Especially in third-world countries, more than 50% of banking DR systems do not support switchover drills and a switchover takes more than 4 hours before businesses are restored. Even in Western Europe, many banks are still out of service for more than three hours a month on average, and cannot guarantee 24/7 online services.

Leading banks, however, have successfully migrated core systems off mainframes and midrange computers relying on their own technical strengths. They use multi-site and multi-active IT architecture to keep financial services always online, continuously improve business and operational resilience, and continue to enhance their market leadership. For example, the DBS Bank in Singapore has moved 90% of its host applications to open systems. In the 2022 annual report of DBS Bank, its CEO emphasized that the bank would enhance its technical support for cloud computing, site reliability engineering, and other fields. The goal is to optimize its scalability, automation, speed to market, cost control, and system resilience.

Challenge 1: Smooth evolution

- Migration from mainframes and midrange computers makes it difficult to ensure the stability and continuity of application systems.

- The hardware architecture of open systems cannot ensure stable and low latency.

Challenge 2: Business resilience

- DR switchover is difficult and causes interruptions that can last for hours. After the switchover, businesses may be unstable for a long time.

- In case of DC-level disasters, financial services will likely be interrupted.

- There are numerous components, the faults of which take a long time to locate, severely affecting business running.

Suggested actions

1. Create a bank-wide centralized technical and business platform, and build or reconstruct business systems based on the microservice architecture. Control resource requirements in a fine-grained manner through microservice splitting, and efficiently allocate resources leveraging cloud-native automatic orchestration and service capabilities. Establish a microservice governance system to better cope with business traffic changes and service failures during dynamic running.

2. When planning and designing financial DCs, financial institutions should consider the support for the highly available distributed businesses from multiple dimensions, including the network latency, traffic model, storage synchronization, and backup system. Synergize different products and technologies to improve business availability through system-level optimization.

3. Based on cloud-native reconstruction, evolve from traditional DR to multi-site and multi-active DR and from heavy business switchovers to light access path switchovers. This will make switchovers much easier and safer and reduce RTO to minutes.

4. Build a centralized, intelligent, and automatic O&M platform, and use AI and automation technologies to reduce manual intervention, achieving low-touch O&M management. In addition, the SRE team should quickly carry out quantitative risk management based on error budgets, and actively embrace agile value delivery.

5.1 Modernizing financial applications

Digital transformation has six requirements for financial infrastructure: agility, elasticity,

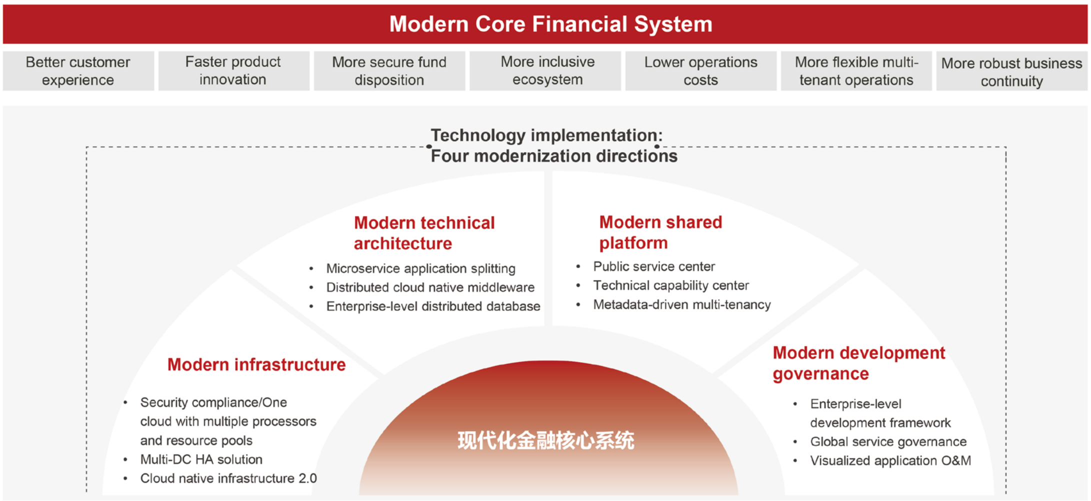

Modern Core Financial System
Better customer experience
Faster product innovation
More secure fund disposition
More inclusive ecosystem
Lower operations costs
More flexible multi-tenant operations
More robust business continuity
Technology implementation: Four modernization directions
Modern technical architecture
Microservice application splitting
Distributed cloud native middleware
Enterprise-level distributed database
Modern shared platform
Public service center
Technical capability center
Metadata-driven multi-tenancy
Modern infrastructure
Security compliance/One cloud with multiple processors and resource pools
Multi-DC HA solution
Cloud native infrastructure 2.0
Modern development governance
Enterprise-level development framework
Global service governance
Visualized application O&M
现代化金融核心系统

Figure 5.2 Four major trends in modern core financial systems

openness, high security, high availability, and unified governance. The finance industry needs to modernize applications through agile service delivery, evolve from centralized IT systems to cloud-native distributed architectures, implement cognitive reshaping, upgrade architecture, and transition technologies from infrastructures to applications. It also needs to provide financial service systems with elastic scaling capabilities, unified service governance and architecture management, agile development, and fast iteration.(Figure 5.2 Four major trends in modern core financial systems)

Financial institutions should modernize the infrastructure, architecture design, sharing platform, and development governance on the comprehensive cloud-native platform. They can adapt new and legacy applications to the modernized environment and achieve full-lifecycle application agility.

Financial institutions should evolve from a service-oriented architecture to a cloud-native architecture to effectively support innovation in core financial services and quickly respond to service changes. To do so, they need to upgrade their technologies by introducing key functions such as containers, service mesh, microservices, and declarative APIs. This will help them fully migrate applications to the cloud. For services, these companies can use a metadata-driven multi-tenancy architecture to quickly build and assemble applications like they would using SaaS. This will allow them to quickly roll out services to efficiently support the rapid development of their headquarters, branches, and ecosystem.

5.1.1 Infrastructure modernization

The cloud-native transformation of traditional infrastructure enables high elasticity and availability of storage, compute, and network resources. It also reduces O&M costs and frees development and O&M personnel from repeated and complicated resource allocation.

1. One cloud for the entire bank: Different subsidiaries and departments of an African bank purchased cloud services independently based on their business needs. As a result, the bank acquired the same services from multiple public clouds and built several independent private clouds. This wasted resources and reduced the effective utilization rate. It is best for financial institutions to formulate unified cloud usage standards and specifications, purchase cloud resources collectively, and develop all clouds for different purposes, including R&D, testing, production, ecosystem, and branch. This will ensure that all cloud resources evolve based on a unified framework and the same architecture.

2. Multi-center DR: Develop DR solutions applicable to different business scenarios, considering factors such as DC locations, resource allocation, and RPO/RTO requirements. These solutions may include intra-city active-active DCs, geo-redundant three DCs, and multi-site multi-active DCs.

3. Seamlessly scalable platform: The platform not only considers the cost-effectiveness of investment to allocate resources on demand, but also supports the scalability required by rapid development in the future.

4. One cloud with multiple processors and resource pools: Tenants are isolated for different applications, implementing "one cloud for multiple purposes" and "one cloud for multiple pools". The multiple pools can be flexibly implemented in batches following the overall planning. In addition, full-stack IaaS, PaaS, and SaaS services must have high compatibility to flexibly match multiple vendors' heterogeneous storage, compute, and network components.

5. Security compliance: The security service system needs to address challenges regarding open-source software security to comply with local and regional regulations. These include vulnerability defense, supply chain security, media trustworthiness, and license changes.

6. Software and hardware synergy: To maximize the performance of cloud-native infrastructure, financial institutions can use some technologies that combine software and hardware, such as offloading network capabilities to dedicated devices or leveraging the pass-through capability of container networks. This will allow banks to build a flat, high-performance, and secure infrastructure resource platform that integrates software and hardware.

5.1.2 Architecture design modernization

Application architecture modernization and the microservice and serverless architectures

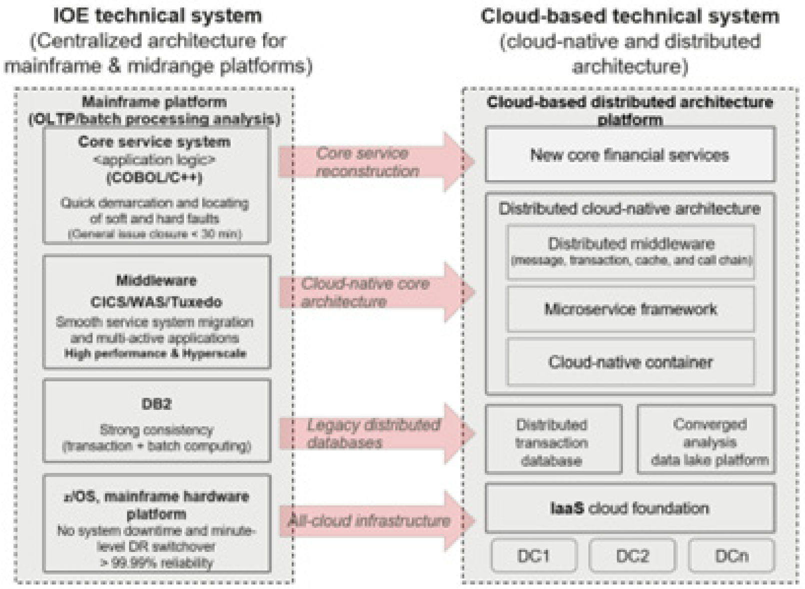

IOE technical system
(Centralized architecture for mainframe & midrange platforms)
Mainframe platform
(OLTP/batch processing analysis)
Core service system
<application logic>
(COBOL/C++)
Quick demarcation and locating of soft and hard faults
(General issue closure < 30 min)
Middleware
CICS/WAS/Tuxedo
Smooth service system migration and multi-active applications
High performance & Hyperscale
DB2
Strong consistency
(transaction + batch computing)
z/OS, mainframe hardware platform
No system downtime and minute-level DR switchover
> 99.99% reliability
Core service reconstruction
Cloud-native core architecture
Legacy distributed databases
All-cloud infrastructure
Cloud-based technical system
(cloud-native and distributed architecture)
Cloud-based distributed architecture platform
New core financial services
Distributed cloud-native architecture
Distributed middleware
(message, transaction, cache, and call chain)
Microservice framework
Cloud-native container
Distributed transaction database
Converged analysis data lake platform
IaaS cloud foundation
DC1
DC2
DCn

Figure 5.3 Architecture design modernization

can help split applications into different modules that can be released independently and quickly. This achieves high cohesion and low coupling.(Figure 5.3 Architecture design modernization)

A microservice architecture can be implemented through cloud-based infrastructure upgrade, distributed database transformation, and core architecture migration to cloud native.

A container microservice platform is built using a distributed DB, distributed cache, distributed transactions, distributed messages, and a microservice governance platform. Heavyweight applications can be split and run on the microservice platform.

5.1.3 Sharing platform modernization

A modern core financial service platform can quickly generate new products/services through service orchestration. It can boost the efficiency of an organization and its business operations. It does so by abstracting, accumulating, integrating and sharing the general capabilities of core services.(Figure 5.4 Shared financial service platform)

This platform will include unified technology stacks, data models, and business models:

1. Core financial service center: Accumulate the capabilities of financial business assets such as the centers for users, accounts, products, limits, credit, deposits, parameters, payments, and authentication center. Help users quickly build financial products through processes/models.

2. Technical component and competence center: Identify and accumulate key technology capabilities such as multi-tenancy, elasticity, and distributed

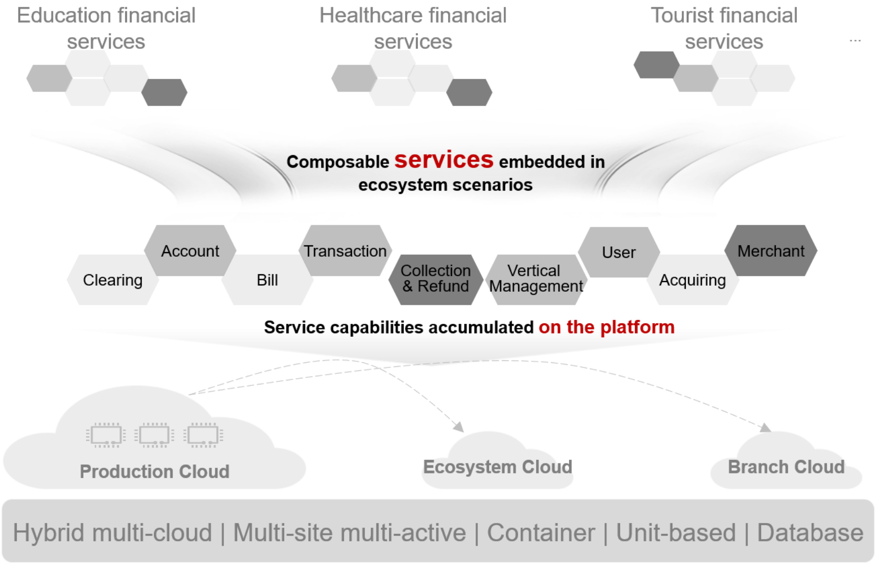

Education financial services
Healthcare financial services
Tourist financial services
...
Composable services embedded in ecosystem scenarios
Clearing
Account
Bill
Transaction
Collection & Refund
Vertical Management
User
Acquiring
Merchant
Service capabilities accumulated on the platform
Production Cloud
Ecosystem Cloud
Branch Cloud
Hybrid multi-cloud | Multi-site multi-active | Container | Unit-based | Database

Figure 5.4 Shared financial service platform

middleware (distributed cache, message, transaction, scheduling, and databases).

3. Secure core system O&M platform: Enable comprehensive observability across applications, platforms, and hardware. This happens after cloud migration or cloud-native modernization of core applications. Prepare DR plans for core systems based on chaos engineering to ensure secure production; Use automatic O&M tools to achieve intelligent O&M.

4. Metadata multi-tenancy agility platform: Use low-code development tools to develop multi-tenancy software products that consider service customization, isolation, and sharing; base these on the standard models for financial digital products and services. Provide personalized services for internal and external users by classifying legal entities, customers, scenarios, and channels. Projects can go live through the subscription of the product generic layer and industry suite layer after simple asset configuration and customization.

5.1.4 Development and governance modernization

Traditional core applications have a slow design and production cycle. At the same time, there is an increasing workload in configurations and development. As a result, agile service transformation cannot support the modernizing core financial applications.

Enterprises need to build an integrated development and governance platform to systematically solve problems, roll out core applications in days or weeks, and ensure intrinsic security and trustworthiness. The platform should cover development, operations, and O&M, achieving DevOps/DevSecOps.(Figure 5.5 DevSecOps platform)

1. An enterprise-level development framework: Preset enterprise-level R&D specifications, built-in standard scaffold code and application configuration in the development framework to significantly reduce workloads associated with repeated adaptation and development. Adopt continuous integration/

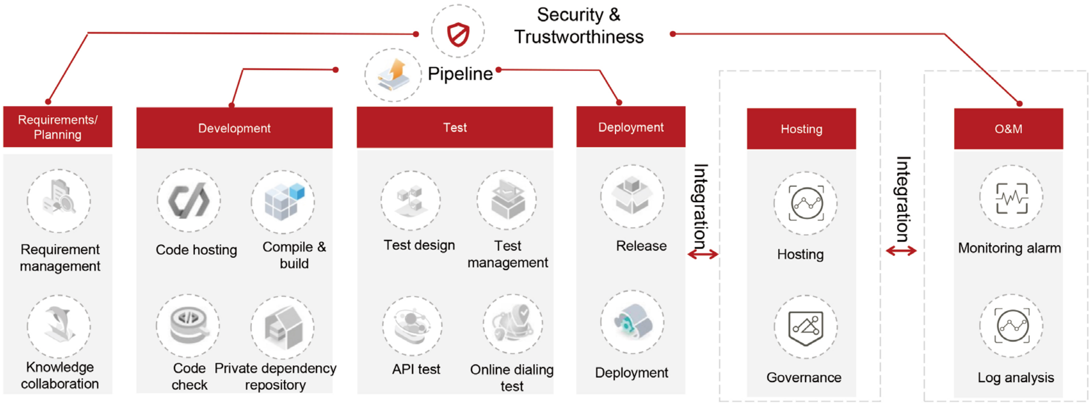


Figure 5.5 DevSecOps platform

continuous delivery (CI/CD), application hosting, and O&M services to achieve efficient application GTM.

High automation is achieved based on the application model, abstract declarative definitions such as Infrastructure as Code (IaC), and the CI/CD that seamlessly integrates with the cloud native infrastructure. Users can submit "code" to deploy and roll out applications.

Developers can use page components, process orchestrator BPM, model orchestrator, and baseline application templates provided by cloud service providers. They also adopt a low-code/no-code composable development approach. This enables them to rapidly create the desired application system, reduce uncertainties in software development, and significantly improve development efficiency.

2. Global service governance: Use the traditional SDK microservice framework, service grid, and more flexible and innovative dual-mode governance system for overall service governance. Support the coexistence, transition, and generational evolution of governance.

3. Visualized O&M of applications: Provide multi-layer, multi-dimensional, and in-depth intelligent O&M for the running environment. This includes infrastructure, the middleware platform, and microservices. Aggregate O&M data to enhance O&M automation and intelligence.

Modernizing core financial systems is a systematic project. To ensure smooth modernization, financial institutions should work with professional vendors that can provide technical and engineering capabilities. They should follow an agile approach, and establish standardized key actions in each phase, including planning and solution design, platform building, core system development, and system O&M. This will help them achieve a successful transformation.

[Case] Postal Savings Bank of China (PSBC) is using agile technologies to drive service agility and modernize its IT platform through a cloud-native platform and distributed database. The bank employs enterprise-level service modeling to create and assemble over 5000 composable parts. These parts can be flexibly combined by scenario to quickly meet the market and customer requirements for personalization, differentiation, and customization. This approach also facilitates a shared and reusable platform. Through cloud-native transformation, PSBC has achieved high performance, scalability, and availability of its core transaction system, supporting 650 million customers and 1.8 billion accounts. The daily transaction volume reaches 529 million during peak hours, with a success rate of 99.99%. "The distributed technology platform ensures secure and stable core system operations and provides a continuously scalable and innovative enterprise engine for PSBC's future IT systems." (Niu Xinzhuang, Vice President of PSBC)

5.2 Quick development of digital core banking

(Figure 5.6 Dual-core digital operations of traditional banks)The existing core systems of traditional banks need to ensure that mission-critical services run stably but cannot provide the agility required for digital innovation. Most banks find it difficult to overhaul existing core systems on a large scale quickly, or to build new core systems entirely. To accelerate the rollout of digital services while preserving the existing service systems, banks can introduce an independent agile service platform and build a digital core banking system based on digital accounts to coexist with the traditional core system.

After obtaining virtual banking licenses from their central banks, emerging financial institutions need to quickly provide financial services to attract customers and generate revenues. However, they need to gain more experience in building core banking systems. The systems must complete many certifications and meet regulatory compliance requirements before operating, which takes a long time.

(Figure 5.7 SaaS-based digital core banking solution)Public cloud SaaS enables public cloud providers to offer cloud resources. It aggregates ecosystem partners across digital banking channels, core accounting, digital loans, micro loans, and open banking systems on the public cloud through APIs. It also builds a complete and integrated business chain covering channel, product, core, operations, and regulation. Thus, a cloud digital core banking solution is formed, which can implement system pre-integration and unified supervision certification. This accelerates the rollout of digital banking services.

[Case] Thailand's second largest bank deployed a next-generation digital loan application based on Huawei Cloud. Service rollout took just three months instead of nine. Meanwhile, automatic loan approvals now take just five minutes instead of several weeks, significantly improving efficiency. The transaction performance of the containerized loan system is three times higher than legacy systems.

[Case] Green Link Digital Bank (GLDB) is one of the first four financial institutions to obtain

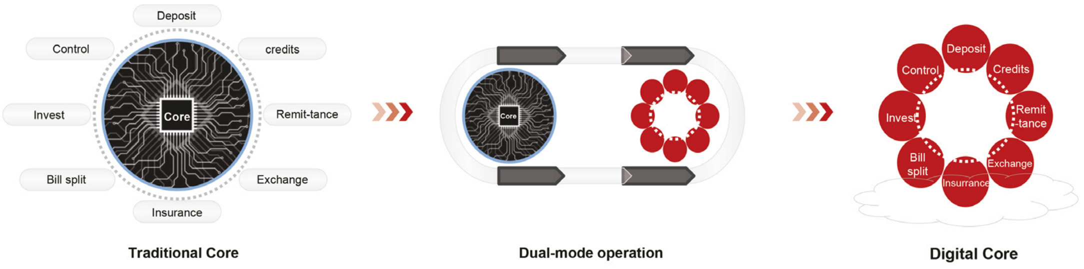


Figure 5.6 Dual-core digital operations of traditional banks

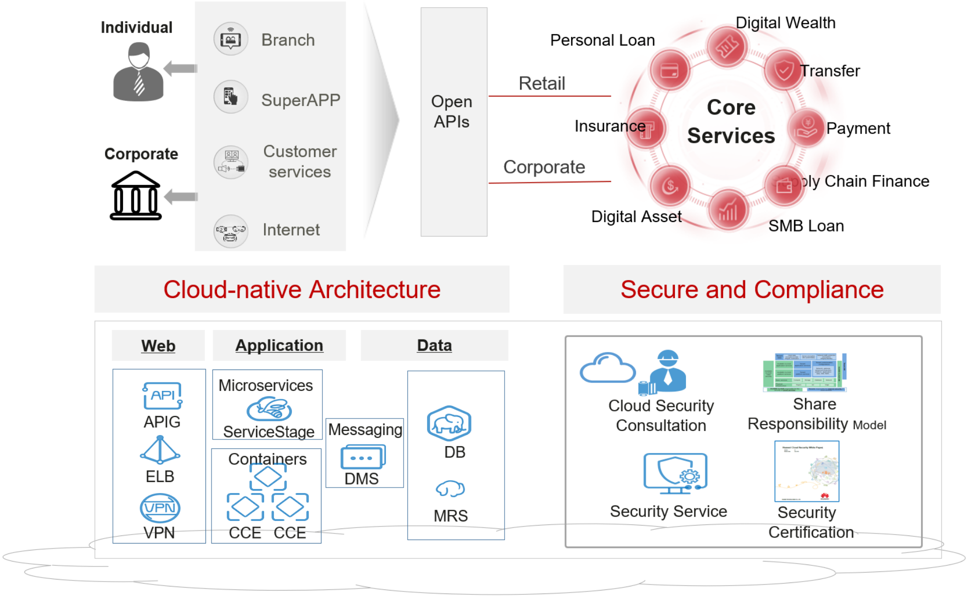


Figure 5.7 SaaS-based digital core banking solution

virtual banking licenses from the Monetary Authority of Singapore and is also the fastest bank to roll out services. GLDB adopted Huawei's digital banking solution on the public cloud to build a core banking system. Based on APIs, GLDB has quickly integrated multiple service ecosystems and completed strict local regulatory certification. It took only eight months to launch services from scratch and release financial products every week, quickly obtaining revenue.

5.3 Business Continuty

5.3.1 Development history of DR

(Figure 5.8 Evolution of DR technologies) The development history of financial DCs represents that of financial DR.

- Since financial institutions have centralized DCs by incorporating multiple dispersed DCs into a larger one, they tended to

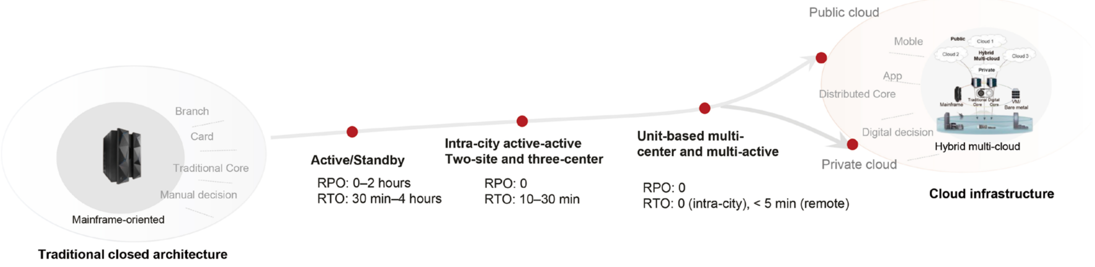


Figure 5.8 Evolution of DR technologies

deploy one active DC and one standby DC. If the former was faulty, business would be switched to the latter.(Figure 5.9 Two-site and three-center deployment)

- Later, three centers in two sites became the typical layout of financial DCs to defend against geographical disasters like floods and earthquakes. Financial institutions deployed active-active or active/standby dual DCs in the same city to achieve manual or automatic failover and business takeover. They also deployed a standby DC remotely to provide data backup. Business takeover started only when the dual DCs in the same city were faulty.

- As we enter the cloud era, leading banks are learning from Internet companies — they pursue the best user experience and business availability in extreme situations across the entire domain by building multi-active DCs on various sites. They usually adopt a business-centric and unit-based architecture. DR is achieved based on mutual backup of units and intra-city or remote multi-active is available according to where these units are deployed.

5.3.2 Business continuity indicators

Financial institutions' overarching concern is business continuity. Generally, their transaction systems should possess availability above 99.99%, which means their businesses cannot be interrupted for 52 minutes or longer in total across the span of an entire year.(Figure 5.10 Business continuity tiers)

The SHARE 78 international standards first defined the eight tiers of availability in 1992, from no off-site data at tier 0 to automatic switchover at tier 7. The classification is still in use today.

The basic indicators of business availability

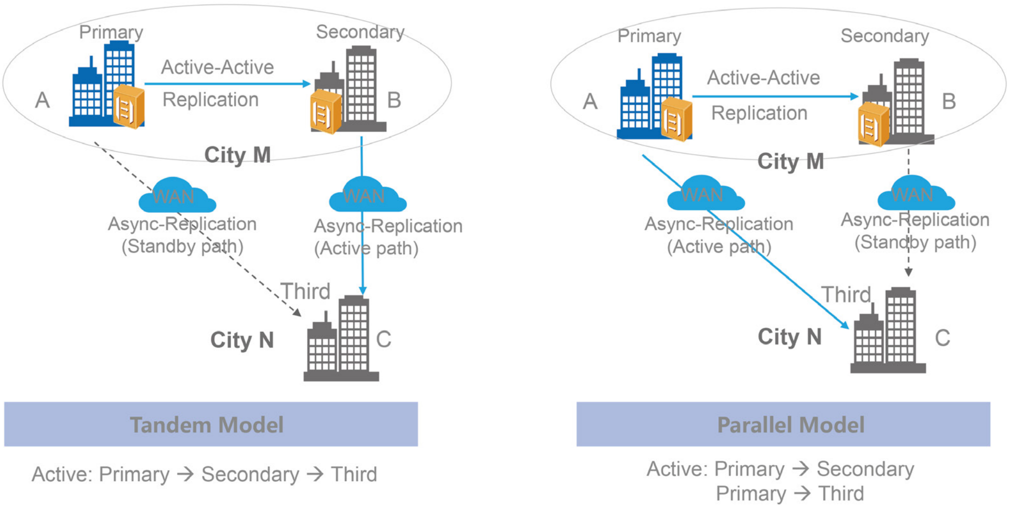
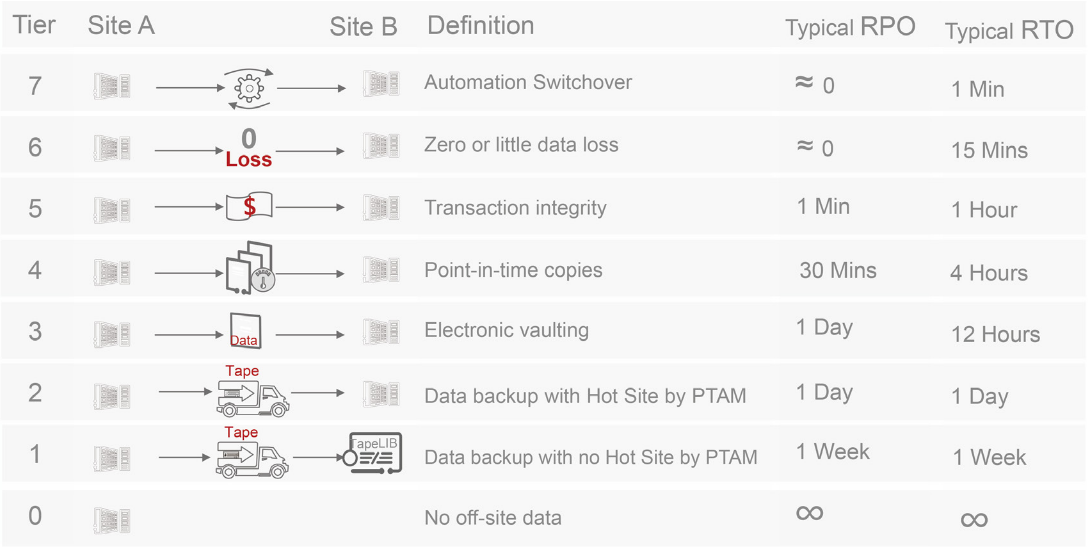


Figure 5.9 Two-site and three-center deployment


0
Loss
$
Data
Tape
Tape
TapeLIB

Tier	Site A	Site B	Definition	Typical RPO	Typical RTO
7			Automation Switchover	≈ 0	1 Min
6			Zero or little data loss	≈ 0	15 Mins
5			Transaction integrity	1 Min	1 Hour
4			Point-in-time copies	30 Mins	4 Hours
3			Electronic vaulting	1 Day	12 Hours
2			Data backup with Hot Site by PTAM	1 Day	1 Day
1			Data backup with no Hot Site by PTAM	1 Week	1 Week
0			No off-site data	∞	∞

Figure 5.10 Business continuity tiers

are RPO and RTO. Financial regulators in most countries have strictly defined RPO and RTO. Financial institutions that fail to keep to regulatory standards must face repercussions.

However, different assurance for indicators means varied costs across equipment, O&M, and management. Financial institutions must classify business into different levels, and invest in the highest-level DR protection measures possible for key transactions. For more common businesses, they can take relevant measures based on how much they can tolerate interruptions and data losses.

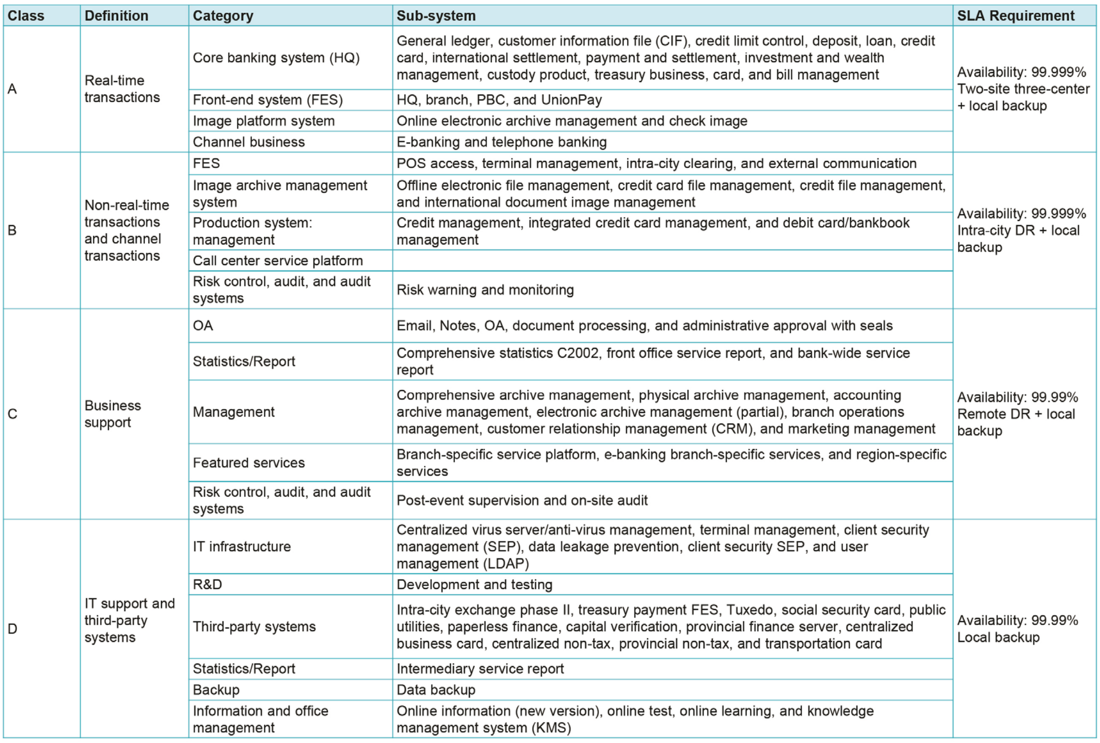
Before DR implementation, financial institutions first need to review and analyze their business and application systems.

Step 1: Straighten out business processes, and survey each department's key business processes, KPIs, and key application systems that carry the processes. Grade business systems into levels, like A+, A, B, and C as shown in the following figure.(Figure 5.11 Tiering Service)

Step 2: Analyze the impact of business failures on financial institutions from the perspectives of finance, contracts, supervision, customers, partners, enterprises and social stability, operations, and brands to prioritize businesses. Analyze and order application systems according to protective measures, business volume, solution substitutability, remedial measures, and service duration requirements.

Step 3: Design RTO and RPO indicators for different systems according to the ranking.

Step 4: Scientifically plan and implement the DR solution based on the RPO/RTO requirements of each business and application system, as well as the budget, current IT architecture, and distance between DCs.



Class	Definition	Category	Sub-system	SLA Requirement
A	Real-time transactions	Core banking system (HQ)	General ledger, customer information file (CIF), credit limit control, deposit, loan, credit card, international settlement, payment and settlement, investment and wealth management, custody product, treasury business, card, and bill management	Availability: 99.999% Two-site three-center + local backup
		Front-end system (FES)	HQ, branch, PBC, and UnionPay	
		Image platform system	Online electronic archive management and check image	
		Channel business	E-banking and telephone banking	
B	Non-real-time transactions and channel transactions	FES	POS access, terminal management, intra-city clearing, and external communication	Availability: 99.999% Intra-city DR + local backup
		Image archive management system	Offline electronic file management, credit card file management, credit file management, and international document image management	
		Production system: management	Credit management, integrated credit card management, and debit card/bankbook management	
		Call center service platform		
		Risk control, audit, and audit systems	Risk warning and monitoring	
C	Business support	OA	Email, Notes, OA, document processing, and administrative approval with seals	Availability: 99.99% Remote DR + local backup
		Statistics/Report	Comprehensive statistics C2002, front office service report, and bank-wide service report	
		Management	Comprehensive archive management, physical archive management, accounting archive management, electronic archive management (partial), branch operations management, customer relationship management (CRM), and marketing management	
		Featured services	Branch-specific service platform, e-banking branch-specific services, and region-specific services	
		Risk control, audit, and audit systems	Post-event supervision and on-site audit	
D	IT support and third-party systems	IT infrastructure	Centralized virus server/anti-virus management, terminal management, client security management (SEP), data leakage prevention, client security SEP, and user management (LDAP)	Availability: 99.99% Local backup
		R&D	Development and testing	
		Third-party systems	Intra-city exchange phase II, treasury payment FES, Tuxedo, social security card, public utilities, paperless finance, capital verification, provincial finance server, centralized business card, centralized non-tax, provincial non-tax, and transportation card	
		Statistics/Report	Intermediary service report	
		Backup	Data backup	
		Information and office management	Online information (new version), online test, online learning, and knowledge management system (KMS)	

Figure 5.11 Tiering Service

5.3.3 Technical DR solutions

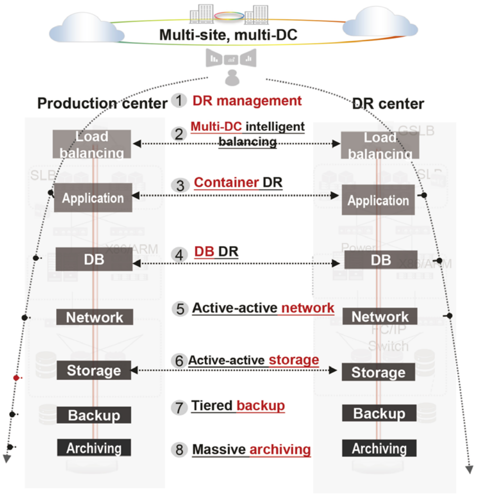


Figure 5.12 DR technical stacks

The goal of DR is to ensure that RPO and RTO reach the optimal value. RPO measures the duration of data losses. DR based on dedicated storage can guarantee the completeness of data and 0 RPO. RTO measures the business interruption duration. Failover and business takeover involve so many systems that some financial institutions can only achieve hour-level RTO. Even only a few leading institutions can achieve an RTO that is no more than ten minutes. The following are the key elements of financial DR.

Storage DR

Professional storage systems usually have complete DR capabilities and provide three basic DR modes: active-active, synchronous replication (sync-replication), and asynchronous replication (async-replication).

Storage DR can be combined with transmission and network to enhance DR capabilities.

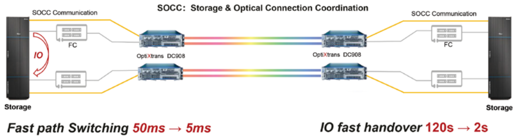

SOCC: Storage & Optical Connection Coordination
SOCC Communication
SOCC Communication
FC
FC
IO
OptiXtrans DC908
OptiXtrans DC908
Storage
Storage
Fast path Switching 50ms → 5ms
IO fast handover 120s → 2s

Figure 5.13 SOCC-based DR

As shown in the above figure, there are four pairs of optical fibers for transmission between two DCs. When the primary and protection links are connected to the same transmission device, it can start a switchover that lasts for no more than 5 ms and does not affect business.

However, when they are connected to different transmission devices, business can be switched only through storage devices. It takes 30 to 120 seconds for storage devices to detect a transmission link fault, which means that businesses will be interrupted for 30 to 120 seconds. If a communications channel is established between the transmission device and storage device, the former will immediately notify the latter of the path interruption, reducing the E2E business interruption time to less than 2 seconds. This technology is called storage-optical connection coordination (SOCC).

Database DR

As key components of business, transactional databases have complete DR capabilities, including active-active, sync-replication, and async-replication. However, the active-active mode requires that the databases at both sites be completely synchronized. The transmission and data links between DCs are not under the control of financial institutions. Aging lines, road works, pipe digging, foundation building, line maintenance, and other operations may cause faults in transmission links to emerge, resulting in link bit errors or even interruptions. Cross-DC active-active databases are not recommended because they will affect two sets of business systems. However, if financial institutions replicate data between active and standby databases, manual intervention will be necessary, and they cannot control the business switchover time.

[Case] Bank G, a top financial institution in China, has achieved high availability of business systems through storage and database collaboration.

It combines active-active storage (or sync-replication) with fast database mounting. Incremental database logs are transmitted to the peer end through storage synchronization, ensuring that the RPO is 0. The active and standby databases are not directly associated with each other and are almost entirely uncoupled because data is transmitted through storage. Even if the transmission link is faulty, only the storage system is affected, and the primary database remains unaltered. In addition, when it is faulty, the standby database at the peer end can be quickly mounted to storage devices via automation scripts and take over the business within 2 minutes. In this way, the RTO is less than 120 seconds.

Container DR

Containers have two defining characteristics. First, data is easy to lose. The moment a container becomes faulty, the data in it will disappear. Second, the container can easily drift to another server and continue to provide container services when a server fails.

With the file sharing provided by professional network-attached storage (NAS), containers can be easily mounted to storage devices. When a container is faulty, RPO can be zero and no data will be lost since all data in the container is stored on professional NAS. If NAS has the active-active capability of automatic switchover, it will support cross-DC active-active containers, which is mandatory in production transaction scenarios.

Cell-based multi-center and multi-active

(Figure 5.14 Unit-based multi-center and multi-active)In the cloud-native era, leading financial institutions have learned from Internet providers. They are building multiple availability zones (AZ) and multi-active DCs to protect data using a containerized microservice platform.

When a business grows to a certain scale, cloudified infrastructure needs to be sliced into different units. The business will also be divided into multiple business units from a certain dimension, preventing system faults from affecting the entire business. The unit contains all the services the business requires and can independently process the complete business cycle. Each unit stores only partial data and processes some businesses. It protects businesses as a relatively independent entity.

For example, a bank with 40 million users can divide them into eight units for deployment in multiple AZs. Each processing unit will have 5 million users. Each unit has a complete business

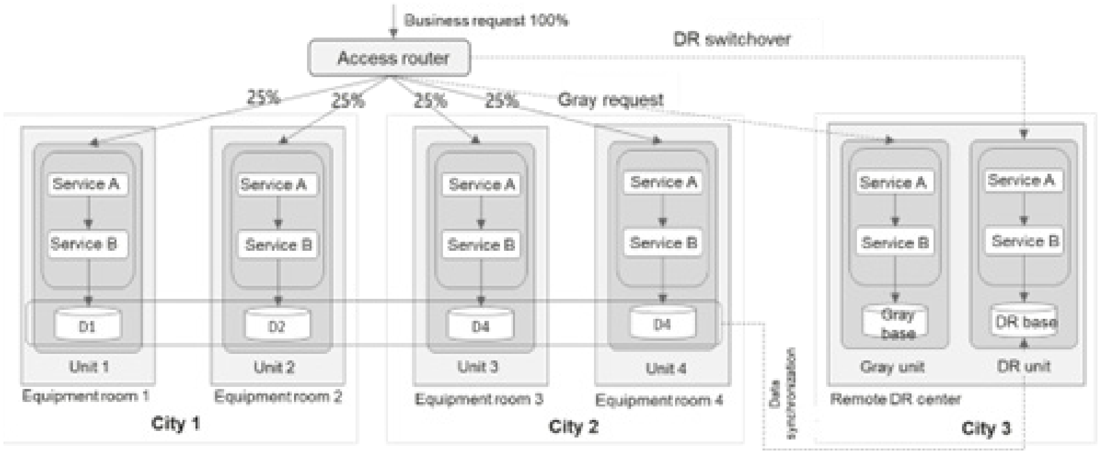


Figure 5.14 Unit-based multi-center and multi-active

system, application system, and hardware systems such as compute, network, and storage. It is completely isolated from other units' resources, and they do not affect each other.

Business units deployed in different AZs run businesses at the same time. For example, business unit 1 is deployed in AZ 1, AZ 2, and AZ 3 to bear the access of 5 million users. Upon user access, the system selects the best routing, and one of the units is responsible for business processing. The changed data, which has been processed, is replicated to other AZs so that business unit 1 can still process businesses even if the local AZ is faulty.

Data backup

As a safety net for financial data and business, backups are not responsible for restoring business immediately, but for point-in-time recovery, which refers to the recovery of data up to a given point in time. Backup involves three key elements: the deduplication and compression ratio, backup time window, and recovery time window.

Financial institutions usually need to perform daily incremental backups and weekly, monthly, or yearly full backups as required by regulators and businesses. Therefore, there is bound to be a large amount of duplicated data in the backup devices. Such data can be deduplicated or compressed to achieve a high data reduction ratio. A typical reduction ratio ranges between 10:1 and 30:1. Some leading vendors can reach 50:1 by using state-of-the-art deduplication and compression algorithms.

Backups have a great impact on the business system's performance. Therefore, backup implementation windows usually open during off-peak hours and close before the institution opens on the next day. In addition, adequate redundancy time must be reserved since business needs to be backed up again if the first backup fails. Therefore, the backup window must be as short as possible, and this imposes high requirements on the backup bandwidth.

Backups are usually only needed in emergencies, like when the business system is destroyed and DR is of no help, or when it is attacked by a virus and business needs to be recovered to the status before the attack. The recovery time window needs to be as short as

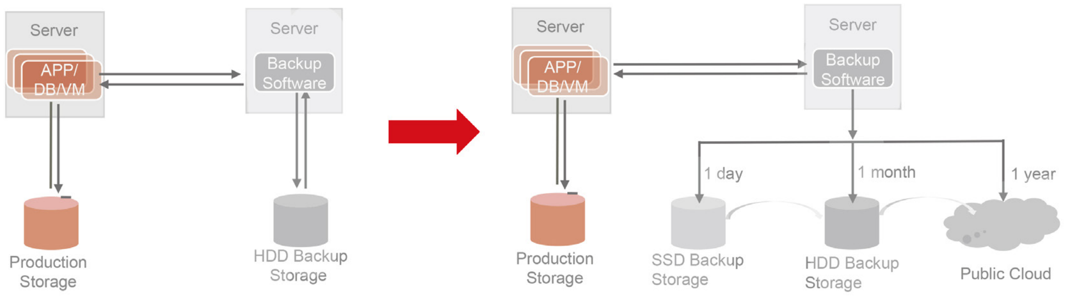


Figure 5.15 Tiered backup

possible.(Figure 5.15 Tiered backup)

The industry has responded to this need by replacing single HDD backup with multi-level backup. Backups can be divided into three types as needed: hot backup, cold backup, and extremely cold backup. Hot backup uses all-slash media. All-flash backup shortens the data backup and recovery time window by more than three times. All-flash backup data will be migrated to HDD backup after a certain period of time (such as 3 months). After more than one year, data can be migrated to the public cloud and stored there for a long time.

It is possible to migrate data between local all-flash backup and HDD backup through backup software, but this would consume a great amount of limited bandwidth resources. Direct data migration between them is the most desirable and backup software is used only for policy delivery and monitoring. This can vastly save bandwidth resources.

On-cloud business and off-cloud backup

(Figure 5.16 On-cloud business and off-cloud backup)If both public and private clouds are involved, banks are advised to deploy business on the cloud and store data within the bank considering data sovereignty. They can deploy ecosystem and channel businesses on the public cloud to increase the business scope and volume. In addition, they can back up and store key customer data and that generated from key business processes in the DC. Local object-based storage is the optimal storage method.

This brings four major benefits, including the:

1. Protection of core data assets: Ensures key data stored in the bank's DC will not be leaked or lost due to faulty third-party clouds. Uploads data backed up in the DC to the cloud to recover business even when faults of third-party clouds lead to data loss.

2. Facilitation of regulatory audits: Centrally stores business data scattered on multiple clouds, and allows it to be centrally

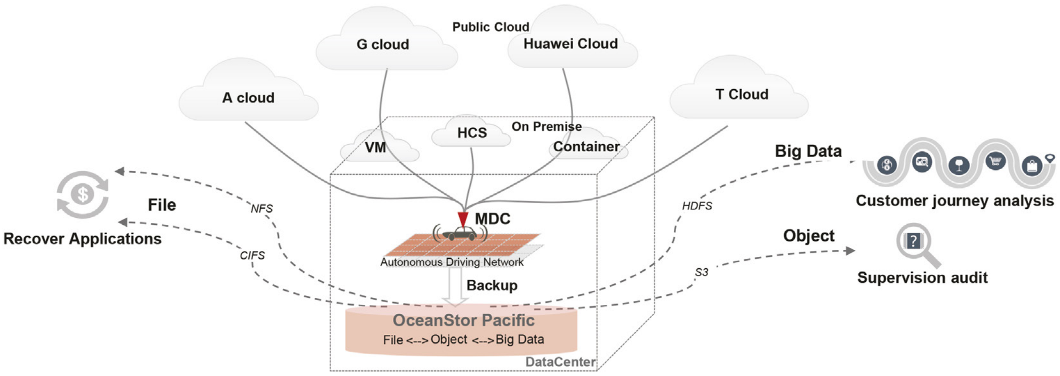


Figure 5.16 On-cloud business and off-cloud backup

managed and read through content management software so as to make regulatory audits easier.

3. Accurate mapping of user journeys: A customer may choose multiple financial services on different clouds. With data from multiple public and private clouds gathered, data analysis software can be used to describe customer behaviors throughout the financial transaction process.

4. Reduction of data storage costs: With gathered data, one storage system supports backup, archiving, and data analysis, reducing data storage costs by two-thirds.

Trend 6 .

With asset scale growth slows down, banks are moving towards refined operations

Affected by the pandemic, import and export, and the overall economic situation, financial institutions have seen heavier pressure from revenue and slow asset scale growth. Collapse frequently occurring in the trust and real estate domains and enterprises with difficulties in running as normal lead to an increase in non-performing loans (NPLs). The Bank of China's housing-related NPLs increased by 20% in 2022, seriously affecting its asset quality. Financial institutions are expanding 2B and 2C businesses to prevent operational risks while making more profits. In addition, they are refining internal operations to control costs. In 2022, a large Chinese insurance institution cut its IT expenditure to 0 considering the tough operations.

Banks' operations and O&M labor force usually accounts for a large proportion of costs. If each person maintains 10 sets of IT devices in a five-year lifecycle, the O&M labor costs in Europe and America will exceed half of the IT system purchase costs.

During the energy crisis in 2022, the commercial electricity price peaked at EUR0.83 per kilowatt-hour in Germany, three times the price in 2021. On this condition, power consumption of IT devices will be more than 1/3 of their purchase costs, and the equipment room space and cooling costs these devices incur will be about 1/2 of their purchase costs.

Challenge 1 High operations and O&M costs

- Cloud-based infrastructure requires complex technology stacks and more IT O&M investment.
- Multi-vendor heterogeneous structure makes management more complex and problem location more difficult.
- Frequent business changes increase IT O&M workload.

Challenge 2 Rapidly rising energy consumption and costs

- Business expansion is accompanied by a sharp increase in hardware systems, yet equipment room space, power supply, and cooling resources are insufficient to carry all of them.
- The hardware resource utilization is low. For example, the average CPU utilization is below 12%. In some banks, it's even below 6%.

- There is a sharp increase in WAN traffic between DCs and campus branches, significantly increasing bandwidth leasing costs.

Suggested actions

1. Focus on TCO throughout device lifecycle instead of one-time purchase costs

   Banks often focus on procurement costs of devices. However, the total cost of operation (TCO) throughout their lifecycle and losses caused by failures often far exceed their purchase costs.

   A top bank in China adopted a cloud-based architecture. However, its universal servers carried its hardware, leading to frequent faults. Slow disks and controllers often became sub-healthy, deteriorating performance and hindering fault locating. In addition, due to business expansion, the bank needed to purchase tens of thousands of servers every year. However, the average resource utilization of these servers was below 10%. There was no more space in the bank's old equipment room, and even the new one was running out of space.

2. Build a hybrid multi-cloud architecture (Figure 6.1 Hybrid cloud)

- Use the advanced technologies and services of cloud service providers to achieve faster business innovation.

- Select the most cost-effective cloud service portfolios for given scenarios to reduce total operational costs.

- A hybrid multi-cloud architecture can effectively reduce the risk exposure of financial data centers (DCs). By distributing data and services across multiple cloud service providers and on-premises DCs, financial institutions can reduce the dependence on a single vendor and the risk of business interruptions due to vendor failures and network issues.

  WeBank in China has migrated all its businesses to the public cloud. It purchases needed services from different public cloud providers with average annual costs per account of only US$0.5, 1/10 of the industry average.

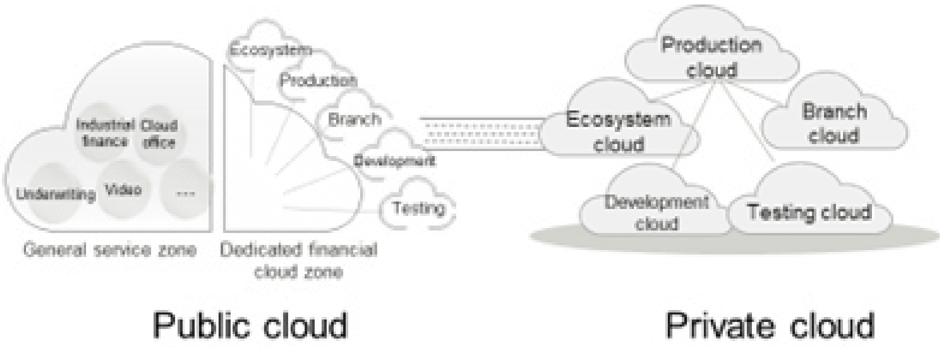


Figure 6.1 Hybrid cloud

3. Develop business process automation

   Based on AI and automation technologies, Robotic Process Automation (RPA) can interact with the existing system as a virtual employee according to the preset process, implementing automatic task completion without manual intervention and without the need to reconstruct the existing system. The pandemic has brought hyper automation that incorporates machine learning and automation to the forefront and made it an alternative for the financial industry to maintain business continuity during times of workforce shortage

   From 2020 to 2022, Gartner has listed hyper automation as one of the annual top technology trends. Based on automation and intelligent O&M, it configured businesses with zero operations and effectively reduced the time for responding to and locating faults.

   Capitalizing on automated and on-demand infrastructure orchestration, J. P. Morgan has sped up hardware provisioning by over 95% and achieved an infrastructure cost efficiency of 15% to 20%.

4. Adopt reasonable architecture and algorithms.

   Adopt the most appropriate infrastructure architecture for different services to improve resource utilization. Leverage green algorithms such as data compression to reduce costs and resource consumption while improving efficiency. Especially, maximize the utilization of WAN bandwidth that costs millions of US dollars annually.

6.1 Building a hybrid multi-cloud framework

Taking business innovation, ecosystem connectivity, costs, risk dispersion, and security compliance into account, financial institutions can choose different public cloud services and build private clouds internally. Hybrid multi-cloud is now a must for financial IT architecture.

To build a hybrid multi-cloud architecture, financial institutions need to consider how they will manage multi-cloud resources and connections as well as data flows among multiple clouds.

1. Infrastructure resource sharing: An infrastructure resource pool can be set up in a DC to support multiple private cloud platforms through a unified compute storage and network platform. For example, some banks have multiple platforms, including K8S, which can be supported by a unified infrastructure resource pool to reduce resource silos and costs.

2. Multi-cloud connections: Financial institutions need to manage public cloud, private cloud, and traditional IT resources at the same time.(Figure 6.2 Connections and flow among multiple clouds)

   [Case] A bank changes its network in the hybrid multi-cloud architecture. The O&M

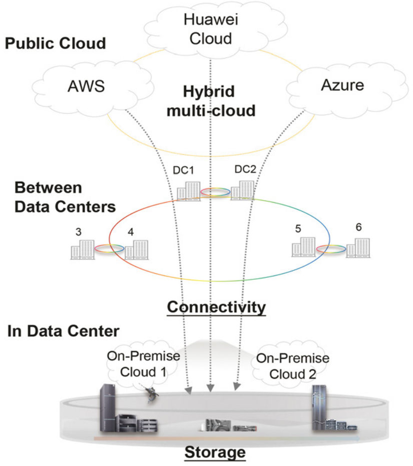


Figure 6.2 Connections and flow among multiple clouds

personnel for public cloud management, DC networks, and security networks need to perform the change at the same time. This costs three times more than a network change in a single environment.

3. Inter-cloud application flow: Build a multi-cloud container management platform. Cloud native technologies such as Docker and Kubernetes can be used to deploy container clusters across clouds. In this way, Kubernetes container services that are the same as those on the central cloud can run in the local DC.

The multi-cloud container management platform centrally releases and manages application services across clouds, achieving cross-cloud multi-active, migration, and DR. Applications between multiple clusters can be flexibly scaled and scheduled by region, status, and resources, enabling applications to be quickly deployed and managed in multiple clusters. The platform supports quick capacity expansion of container resources on the private cloud using cloud resources, effectively coping with traffic bursts.

6.2 Choosing the right architecture

Performance is not the top concern of production, development, and testing systems commonly used by financial institutions, which also do not have strict requirements on business continuity and can tolerate interruptions to a certain degree. Such systems feature non-key OLTP databases and a mass of file sharing, and consume compute resources such as bare metal servers, VMs, and containers. They have high requirements for resource utilization and system flexibility.

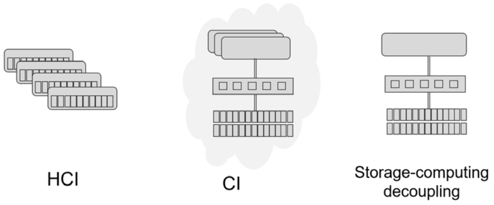


Figure 6.3 Three common IT architectures

These systems mainly use three types of architectures: HCI, CI, and traditional decoupled storage and compute. Financial institutions often choose the one best suited to their usage habits, resource utilization, and how easily the architecture can be expanded and managed.

In the HCI architecture, storage and compute are coupled. The servers function as a single carrier for horizontal expansion and are connected over an IP network at the same rate. All server nodes are completely peer-to-peer. This makes both management and expansion easier. The dedicated storage layer is invisible in this architecture and often appears as a capacity attribute of compute resources.

However, compute and storage have to scramble for resources since they are coupled, including CPU and Cache resources within the node, and network bandwidth resources between nodes. There are no ideal methods to isolate resources and prevent resource contention. This leads to unstable response latency and frequent delays.

In addition, DR mode is often simplified under the HCI architecture, mainly appearing as active/standby DR. Therefore, this architecture is desirable when there are no high requirements for latency and DR. Nevertheless, it should never be used in the channel, business, and core accounting systems of financial transaction chains.

The HCI architecture is expanded through equivalent nodes. Therefore, the compute, storage, and network configurations of all server nodes in a cluster should be consistent to ensure load balancing. Otherwise, many load and traffic balancing operations will be performed between nodes, which severely affects performance. However, HCI often carries complicated applications, each consuming a different amount of compute, storage, and network resources. This limits resource utilization, wasting either compute or storage resources. The latest vSAN Max storage solution has also decoupled storage from compute.

The CI architecture adopts one-stop package delivery. Generally, storage and compute are decoupled for hardware. Integrated management platforms such as VMs and containers are configured and monitored through unified O&M software. This architecture decouples storage, compute, and network from each other to ensure they each have dedicated resources and stable performance. It also simplifies management through integration. However, under the CI architecture, hardware or software products provided by one or several certified vendors are often combined into a few typical configurations for customers, who then have a limited range to choose from.

The traditional decoupled storage and compute architecture ensures resources are dedicated and latency is stable. Financial institutions can choose products from different vendors as needed. However, there are bound to be diverse management platforms provided by these vendors. Financial institutions have to manage two layers whether they use a third-party commercial management platform or adopt an open-source management framework northbound (like the automated configuration of Ansible and Grafana-based graphical monitoring and analysis). Therefore, they need to plan the two-layer management plane in advance to avoid conflicts.

In this architecture, storage and compute resources can be separately expanded as needed to maximize resource utilization, which is a huge advantage.

In addition, the DR capability of dedicated storage devices is much stronger than that of HCI. Multiple DR modes, such as active/standby DR, active-active DR, and three centers in two sites, are available to keep business continuous.

Dedicated storage devices are also superior to HCI when it comes to professional hard disk management.

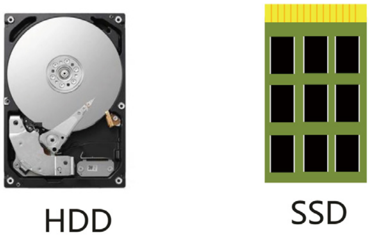


Figure 6.4 HDD and SSD

Traditional HDDs are sensitive to temperature, vibration, and even loud noise. Once, a gas fire suppression system that was set up across multiple financial institutions was set off due to false fire alarms; at the time, entire batches of HDDs were damaged because of the constant high-pitched noise. Data blocks are also often crippled by strong vibrations or clashes between disk heads and disks.

However, SSDs use flash media and have a limited number of times they can perform data block reading and writing. Currently, TLC or QLC media can provide less than 1000 writing times per bit. The IT system needs to take measures to balance the times of reading and writing data to avoid hot disks. When a bad flash data block appears on the SSD, it should be handled specially — data should be restored using redundant data blocks or disks to ensure that it will not be misread.

When complex disk faults occur, professional storage devices usually use bad block preprocessing, background scanning, and DIF (T10) checks to ensure data consistency. However, HCI software vendors often focus on compute resources and treat storage as a capacity element. They cannot process underlying disk faults and fail to cope when data is misread or miswritten, which may shorten the service life of HDDs and SSDs.

6.3 Adopting data reduction technologies

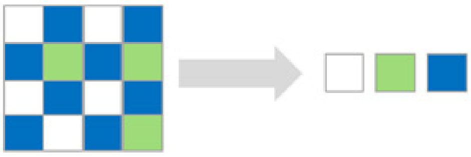


Figure 6.5 Deduplication

Common data reduction technologies include data deduplication and data compression.

When a large number of duplicate data blocks exist in the system, only one may be retained, and the location of the original data block can be recorded with the help of the metadata indexes. However, data compression adopts bit exclusive OR to record the data with certain rules or similar data with fewer bits.

Deduplication and compression technologies are mature. For production storage, most vendors can reach a data reduction ratio of 2:1

or 3:1. When applied to backup storage, the reduction ratio can be 10:1 or even higher.

When data compression technology is applied in a WAN, it will reduce costs more significantly. For the DCI backbone network between DCs and the branch network between campuses and branches, most banks pay carriers more than US$10 million in bandwidth lease fees every year. Large banks even pay more than US$100 million. A 30% reduction in transfers with data compression can save millions of dollars.

6.4 Capitalizing on automatic O&M tools

(Figure 6.6 Digital network map)The network digital map can be used both for offline and real-time simulation. The map simulates device routing protocol behaviors based on NE configuration data, and accurately generates a global routing table of NE protocols in real time, thus completing the analysis and verification of impact on the network.

The digital map supports associated awareness of businesses and networks based on multi-dimensional visualization. For example, a bank's hundreds of applications and tens of thousands of NEs are not associated with each other visually, resulting in low O&M efficiency. Common NMSs are used only to visualize network quality management. When a business fault occurs, however, the network fault point cannot be located.

Digital maps are used to detect the mutual access relationships within and between applications in seconds. They deliver digital insights across services, applications, networks, and devices. As such, a fault can be detected within one minute, located within three minutes, and rectified within five minutes.

6.5 Open-source technology for fair competition

Financial institutions are facing a dilemma in using open-source technology.

On the one hand, open-source platforms can gather global developers, universities, and technology companies to achieve quick iteration, and some techniques even far

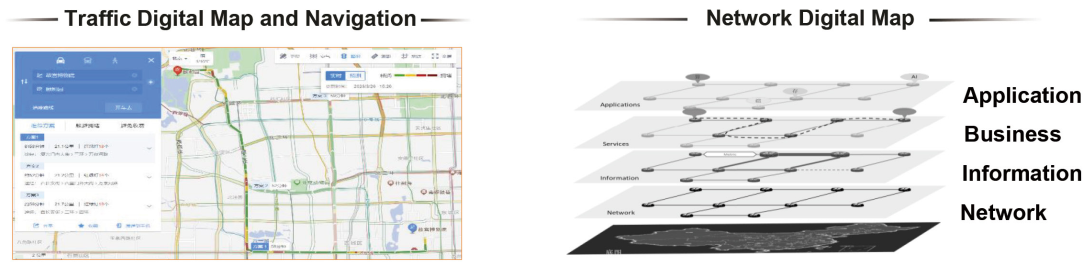


Figure 6.6 Digital network map

exceed the commercial versions of technology companies. Adopting open-source technology can greatly improve efficiency.

On the other hand, open-source platforms are not a complete commercial product and cannot provide comprehensive assurance for reliability, security, testability, maintainability, and manageability. Commercial products will undergo complete processes of design, development, and testing to ensure the integrity of product Design For Everything (DFX). The engineering quality and after-sales service of commercial products are beyond the reach of open-source platforms.

Some leading financial institutions have strong R&D capabilities. After absorbing open-source technologies, they independently develop platform versions to ensure DFX integrity, and perform iteration and maintenance of platforms. For example, large banks invest in dedicated D&R teams to build K8S container microservice platforms and provide unified services for departments and branches of the bank. Most other financial institutions deploy open-source platforms in non-frontline production environments such as development, testing, and office, providing only auxiliary functions.

Financial institutions can also observe the use of open-source technologies by commercial vendors and select vendors with a high contribution rate to open-source communities, fast iteration of commercial versions, and good reputation as providers for commercial versions.(Figure 6.7 Open-source technology promotes fair competition among providers)

The heterogeneous device management of infrastructure is a pain point for O&M. The multifold increase of O&M costs often

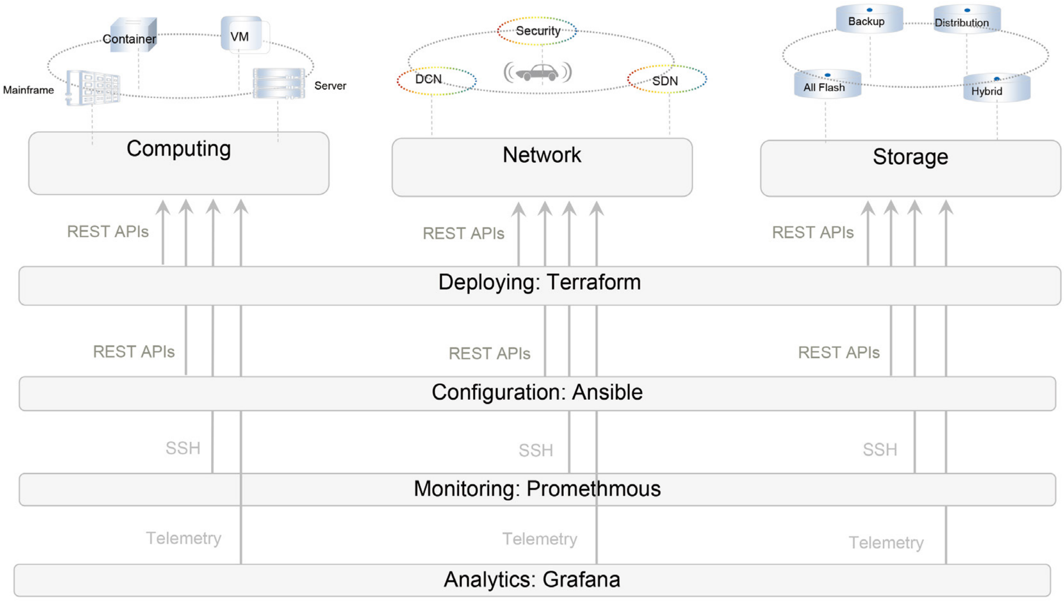


Figure 6.7 Open-source technology promotes fair competition among providers

becomes the root cause of the unavailability of the latest technologies of new vendors, hindering the technological progress. Using open-source tool platforms on the DC's management and O&M plane has no negative impact on the production and transaction system. On the contrary, it enhances O&M automation, and boosts O&M efficiency.

1. Ansible automatic configuration tools: Ansible integrates the advantages of various O&M tools (including Puppet, CFEngine, chef, func, and fabric). It invokes the plug-ins provided by vendors to easily add and delete storage resources in batches and configure functions such as DR and snapshots in batches.

2. Terraform infrastructure as code (IaC) platform: Terraform allows us to deploy infrastructure (compute, network, and storage resources) on the cloud. Without the help of professional O&M personnel, application developers can deploy resources in steps of writing, planning, and application. It is written in a declarative manner, such as "five DB servers, 3 TB storage space", which is easy to use.

3. Prometheus automatic monitoring: Prometheus is an open-source monitoring system based on time series databases. It is ideal for environment indicator collection, service discovery, and alarm management of VMs, containers, and hardware resources. Prometheus periodically captures the status of monitored components through HTTP. Any component can be monitored if the corresponding HTTP interface is provided, without SDK or other integration processes.

4. Grafana visualized display and analysis: Grafana is a trendy data visualization platform. It can easily convert data into charts while analyzing and displaying O&M using engaging visuals, such as dashboards.

Developers have created thousands of colorful dashboards, as well as rich panels and colors. The tool is more stylish than most management software.

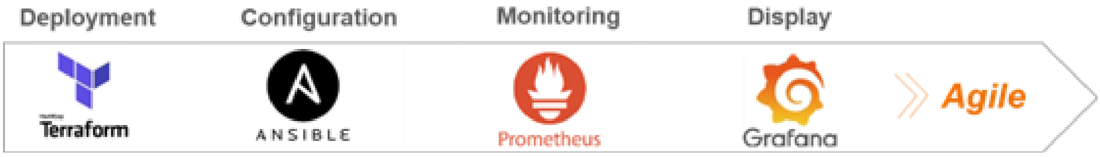


Figure 6.8 Open-source O&M chain in DCs

Using Terraform, Ansible, Prometheus, and Grafana, financial IT personnel can easily create complete agile O&M platforms, significantly improving management and O&M efficiency.

Various software and hardware vendors provide various plug-ins for open-source tools that can be used to deploy, configure, monitor, display, and analyze storage, compute, and network devices and resources of multiple heterogeneous vendors. Therefore, leveraging open-source tools dramatically contributes to fair competition among providers.

Some open-source frameworks can be integrated into vendors' management software. For example, network and storage vendors can invoke the Runbook framework of Ansible to perform configuration, monitoring and other functions on heterogeneous vendors through their own management software.

[Case] A top financial MSP company in Europe provides IT operations for banks in hundreds of cities. For a long time, the company has bound network and security to one vendor, resulting in high costs and no competition. Network providers experienced supply shortages during the pandemic and could not upgrade network devices in time. Now, the company uses Ansible to build a unified network O&M platform. Vendors provide plug-ins and successfully introduce third-party network providers to construct dual-plane networks in DCs, ensuring supply continuity.

Targets and Framework

7.1 Enhance infrastructure resilience and build MEGA infrastructure

(Figure 7.1 MEGA infrastructure)Financial services are evolving to being always online and ubiquitous. In the future, financial institutions need to build a robust and resilient financial digital foundation targeted at "M.E.G.A". This acronym represents the following.

Multi-DC-as-a-Computer:

CPU decentralization is used to implement collaborative scheduling of multiple computing powers and support multiple types of compute units in a system, such as CPU, AI training, and network processing. Architecture innovation enables multiple DCs to run as efficiently as a computer, and improves the performance and efficiency of financial infrastructure.

E2E Experience:

From core transaction to digital interaction, and from DevOps to branches, the collaboration of storage, compute, optical, network, and cloud technologies, provides end-to-end experience

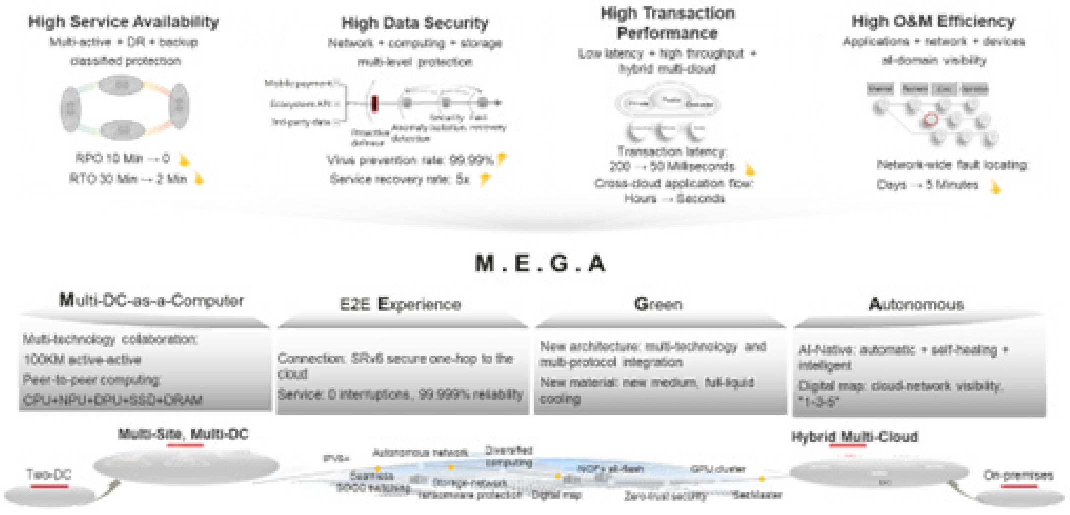

High Service Availability
Multi-active + DR + backup classified protection
RPO 10 Min → 0
RTO 30 Min → 2 Min
High Data Security
Network + computing + storage multi-level protection
Mobile payment
Ecosystem API
3rd-party data
Proactive defense
Anomaly detection
Security isolation
Fast recovery
Virus prevention rate: 99.99%
Service recovery rate: 5x
High Transaction Performance
Low latency + high throughput + hybrid multi-cloud
Transaction latency: 200 → 50 Milliseconds
Cross-cloud application flow: Hours → Seconds
High O&M Efficiency
Applications + network + devices all-domain visibility
Network-wide fault locating: Days → 5 Minutes
M.E.G.A
Multi-DC-as-a-Computer
Multi-technology collaboration: 100KM active-active
Peer-to-peer computing: CPU+NPU+DPU+SSD+DRAM
E2E Experience
Connection: SRv6 secure one-hop to the cloud
Service: 0 interruptions, 99.999% reliability
Green
New architecture: multi-technology and multi-protocol integration
New material: new medium, full-liquid cooling
Autonomous
AI-Native: automatic + self-healing + intelligent
Digital map: cloud-network visibility, "1-3-5"
Multi-Site, Multi-DC
Two-DC
IPV6+
Autonomous network
Diversified computing
Seamless SDCC switching
Storage-network
Ransomware protection
Digital map
NOF+ all-flash
GPU cluster
Zero-trust security
Backmaster
Hybrid Multi-Cloud
On-premises

Figure 7.1 MEGA infrastructure

assurance for key business, making service response more agile, business more secure, and systems more reliable.

Green:

Digital technologies enable green and low-carbon development of the industry, and new algorithms, technologies, and architectures support energy consumption reduction and efficiency improvement of infrastructure.

Autonomous:

Intelligent technologies are widely used in IT infrastructure to make system facilities more efficient, simpler, and more reliable. The infrastructure also supports data analysis, AI training and inference, and intelligent applications such as precision marketing and real-time risk control, enabling smart banking.

Financial infrastructure can be divided into three types based on the characteristics of financial business:

- Transaction infrastructure: Ensures agile response to and stability of financial transactions.

- Interactive infrastructure: Facilitates real-time communication with customers, flexible development and testing, and convenient operations. Interactive infrastructure is a typical representative of general infrastructure.

- Data and AI infrastructure: Provides real-time or quasi-real-time decision-making based on cloud-pipe-edge-device synergy and a bank-wide centralized data and AI training, inference, and application platform.

We need to upgrade these infrastructures towards the following directions to empower business innovation:

1. Digital and intelligent

   Data analysis time: T+1days→T+10 minutes; AI-supported service ratio : 2%→20%

2. Optimal experience

   Transaction response: 150ms→50ms; percentage of smooth interactive video duration: 90%→99%

3. Agile business

   TTM of a new business: 3months→2weeks; resource deployment and configurations: 1 day→10minutes

4. Open and innovative

   The time needed to interconnect with a new ecosystem: 2weeks→1day; the time needed to configure heterogeneous devices: 4hours →5minutes

5. Resilient system

   Data loss rate (RPO): 10minutes→0; business interruption duration (RTO): 2 hours→5minutes

6. Secure business

   Mean time to detect (MTTD): 1hour→1 minute; mean time to repair (MTTR): 1 week →2hours

7. Regulatory compliance

   Data retention period: over a decade; the time needed to retrieve hundreds of millions of files: 1day→1minute

8. Cost-effective

   Automated execution rate: 99%; effective IT resource utilization: 30%→60%

7.2 Target Architecture of Financial Infrastructure

（ Figure 7.2 Target architecture of financial infrastructure ）Financial institutions need to systematically build a future-oriented financial infrastructure architecture, which should be highly available and secure, compliant and trustworthy, and green and low-carbon. In this way, it can deliver a premium financial service experience, agile businesses, universal intelligence, and open innovation.

1. Build a hybrid multi-cloud infrastructure to ensure reliable connections, application flow, and data flow between clouds.

2. Create robust and resilient DCs:

- Hierarchical businesses: Scientifically grade business systems into three to four levels based on their importance and impact.

- Hardware architecture: Use professional and reliable devices and dedicated resources for crucial business systems such as transactions to ensure consistently low latency. For non-critical ones, choose the

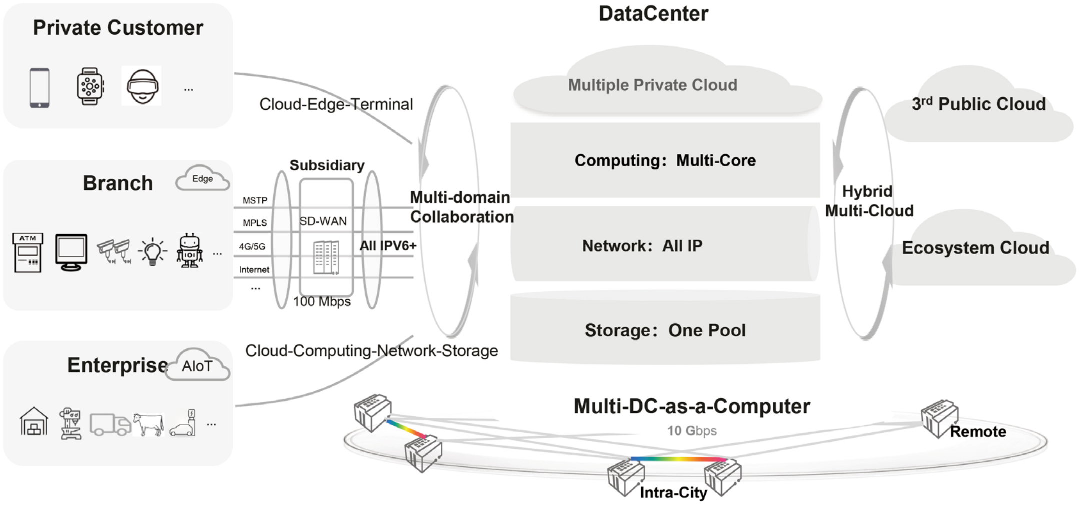


Figure 7.2 Target architecture of financial infrastructure

HCI, CI, or decoupled storage and compute architecture based on the TCO, O&M habits, and system scalability.

- DR: Traditional infrastructure should be based on three or four centers at two sites, and cloud-native systems should be based on unit-based multi-center and multi-active.

- Backup: Create a multi-level backup with all-flash hot backup, mass object storage warm backup, and public cloud cold backup. Centralize data backup of multiple clouds through on-cloud business deployment and off-cloud backup.

- Archiving: Archive massive objects in warm mode and blue-ray/tape/public cloud cold mode.

- Data & AI: Integrate the data warehouse, data lake, and AI training infrastructure, and converge data and AI through decoupled storage and computing and data lakehouse. Build a high-concurrency and high-performance storage and computing network to efficiently train AI models and achieve all-domain data collection and AI inference via cloud-pipe-edge-device synergy.

3. Build secure campus branch networks and WANs to ensure an optimal experience

- WAN: Adopt IPv6 or IPv6 Ready and comprehensively ensure business experience using SRV6, APN6, and iFit. Connect branches based on SD-WAN and compress data to reduce WAN costs.

- Campus network: Use Wi-Fi-6/7 and multi-rate switches to build a high-quality campus network.

- Management: Uses a digital map to gain a complete perspective over the entire network, and detect a fault within one minute, locate it within three minutes, and rectify it within five minutes.

HUAWEI TECHNOLOGIES CO., LTD.
Huawei Industrial Base
Bantian Longgang
Shenzhen 518129, P.R. China
Tel: +86-755-28780808
www.huawei.com

General Disclaimer

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

Editor-in-Chief: Du Yumin, Zhao Junwei, Wang Mingdi

Managing Editors: Cao Chong, Ma Jun, Rao Zengguang, Cui Jie, Chen Zhanwen, Li Jiang

Section Editors: Chen Kunte, Liu Yuanxin, Yu Dawei, Lv Chen, Huang Jin, Lu Zhiqiong, Zeng Lei, Zhang Guobin, Frank.Zhang, Sun Yajun, Li Zhenbin, Li Hongkun, Cui Hongbin, Cheng Xiaoqiang, Wang Lei, Zhou Chunqiang, Li Lin, Gu Qiang, Jiang Xianfeng

Translators: Li Na, Du Xue, Yi Zhuoxuan