HUAWEI

Striding Towards the Intelligent World (2023)

# Data Communication

## Networks Accelerate AI
## and AI Redefines Networks

Building a Fully Connected, Intelligent World

# Preface

It is no exaggeration to say that AI will reshape society in its entirety. From a mere 7% in 2021, the industry penetration rate of AI is expected to hit 30% in 2026. And as foundation models accelerate the intelligent transformation of industries, the penetration rate will exceed 50% by 2030, an increase of 10 times. The rapid development of AI will further promote the digital transformation of industries and provide new opportunities for network innovation.

**WAN agility and quality need to be urgently improved as enterprise cloudification accelerates.**

To date, about 70% of enterprises have migrated to the cloud, with distributed hybrid multi-cloud predominantly being the favored choice for such migration. As a result, there has been a rapid increase in both cloud access and inter-cloud traffic. Enterprises want to use an elastic and agile network to flexibly connect multiple clouds on demand for improved cloud-side efficiency. As traditional industries, such as energy, transportation, and finance, enter a phase of rapid cloud-based transformation, they pose differentiated transport requirements on networks, bringing about the need for networks to provide customized quality assurance capabilities. To effectively support enterprise cloudification, networks must evolve toward better elasticity, agility, security, and reliability.

**Sharp increase in AI computing power drives data center network (DCN) transformation.**

The rapid growth of AI has, in part, been stimulated by applications such as ChatGPT. By 2026, the penetration rate of the AI industry is expected to reach 30%. And from 2023 to 2030, the AI computing power is expected to increase by 500 times. As the computing volume used for AI training increases exponentially, the demand for bandwidth doubles every 3.5 months on average, far exceeding the 18 months defined by Moore's Law. The surge in AI computing power will drive increasing demand for constructing global DCNs and transforming network technologies. Given that a packet loss rate of only 0.1% can deteriorate computing performance by 50%, a DCN with high throughput and zero congestion must be constructed to fully unleash computing power.

**Campus networks are entering the experience-centric era.**

The digital transformation of industries requires a high-speed and stable campus network environment. Campus networks are rapidly expanding from office to production and from connecting people to connecting things. Over the next five years, there is expected to be a 3-fold increase in the number of access terminals on these networks, which need to provide ubiquitous connections and isolate office services from production services. Campus services are undergoing rapid transformation, with mobile office and video conferencing being the two major development trends. What's more, about 80% of traffic on campus networks will come from audio and video applications, meaning that campus networks will enter the experience-centric era. Given these factors, existing networks need to be upgraded, for example, from Wi-Fi 4/5 to Wi-Fi 6/7 and from GE to 10GE access.

**Network complexity increases drastically, and intelligence accelerates network autonomy.**

With the continuous development and application of technologies such as cloud computing and IoT, an intelligent society featuring connectivity of everything, all things sensing, and all things intelligent is gradually materializing. As such, enterprise networks will extend from office to production, shift from static configuration to on-demand adjustment, and transform from single-domain management to network-wide collaboration. Furthermore, network boundaries will expand, network quality attributes will increase, and network O&M will undergo a qualitative change. According to Huawei's Global Industry Vision (GIV) 2025, 97% of large enterprises will be using AI by 2025. Networks that integrate AI capabilities can overcome the efficiency limitations of manual O&M, achieving autonomous driving with high levels of automation and intelligence. Once built, autonomous driving networks (ADNs) can pave the way for enterprises' digital service innovation and agile operations.

**Ubiquitous network attacks require an integrated security protection system.**

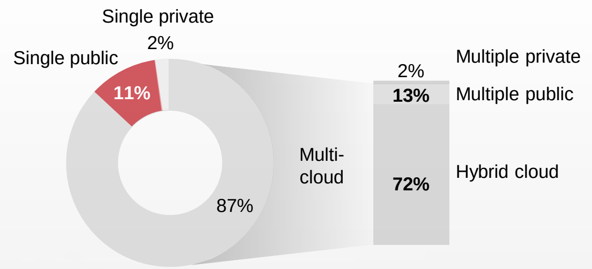
Traditional network boundaries are disappearing as services move to the cloud, making network security far more challenging and uncertain. In 2022, 85% of enterprises experienced network attacks. The number of network attacks launched worldwide increased by 42% over the previous year, and a ransomware attack occurred every 11s on average. Network attacks can interrupt services, leak sensitive data, and even cause huge economic losses. To effectively prevent network attacks, the key is to establish a security defense system that integrates the clouds, networks, edges, and endpoints.

# Contents

# Multi-Cloud Has Become the New Normal for Enterprise Digitalization

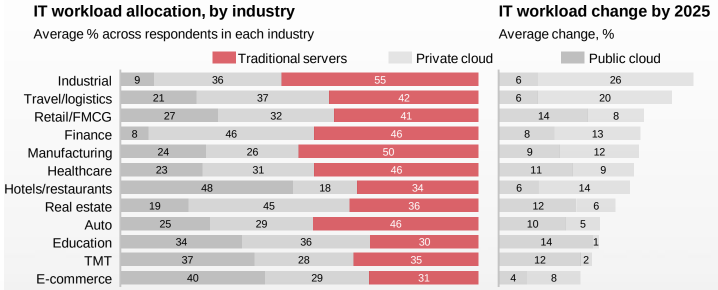
- **Multi-cloud has become the new normal for enterprises.** Enterprises are gradually transitioning from public clouds to hybrid clouds, industry clouds, edge clouds, and distributed clouds in order to achieve cost savings, data security, and cloud technology integration. Multi-cloud allows enterprises to flexibly use different cloud services based on their business needs and distribute multiple workloads across different cloud platforms. According to the Flexera 2023 State of the Cloud Report, 87% of surveyed enterprises have already been using multi-cloud services, and this shows that multi-cloud has become the new normal for enterprise digital transformation.
- **Cloud migration in traditional industries is accelerating.** With advances in cloud computing, big data, and artificial intelligence (AI), cloud services have become a driving force for service innovation and enterprise upgrade in a wider range of fields. More enterprises are embracing the cloud in order to keep pace with technological transformation and seek new development opportunities. In addition to the Internet industry, cloud service practitioners are now also seen in traditional, non-digital-native industries, such as industrial, education, healthcare, government, energy, and finance. Digital transformation is driving cloud migration across industries. Take China as an example. According to the McKinsey 2021 China Cloud Computing Survey, the share of traditional industries' IT workloads running in the cloud will increase significantly by 2025.
- **Cloud migration across industries is dominated by distributed, hybrid multi-cloud.** Cloud migration across industries emerges as enterprises continue to pursue higher efficiency, cost-effectiveness, and business growth. Hybrid multi-cloud combines the advantages of public and private clouds, and not only ensures enterprise data security but also provides a flexible cloud architecture. This makes it a popular choice for enterprises. In addition, a large number of emerging business applications require massive data analysis and computing capabilities. A multi-layer and distributed cloud computing model is going mainstream across industries. Examples include the deployment of "three DCs in two cities" in the financial industry and "distributed DCs + public clouds" in the energy industry.

## 87% of surveyed enterprises are already using multi-cloud in 2023



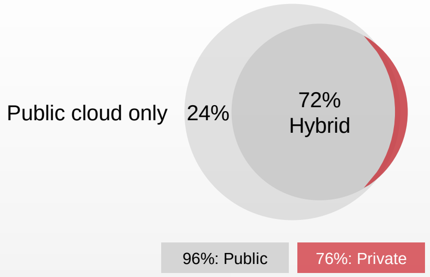
Source: Flexera 2023 State of the Cloud Report

## The share of traditional industries' IT workloads running in the cloud will increase significantly by 2025



Source: McKinsey 2021 China Cloud Computing Survey

## 72% of surveyed customers use hybrid cloud



Source: Flexera 2023 State of the Cloud Report

# Multi-cloud Strategy Drives Multi-cloud Networks Construction Enter Peak Period

**Networks are the cornerstone of a multi-cloud strategy.** Cloud services require powerful network capabilities, while network resource optimization needs to draw inspiration from cloud computing. More and more enterprises are opting for a multi-cloud strategy. Heterogeneous connectivity, complex network management, E2E service experience assurance, and security protection all require a network infrastructure that can better accommodate the requirements of cloud computing applications and optimize network structures to ensure that networks meet the requirements of cloud services across industries.

**Demand for private networks for industry digitalization has soared.**

- **Finance:** The rapid and large-scale growth of financial services and the distributed architecture transformation raise new requirements for the wide-area networks (WANs) on which financial services run. Financial WANs are the channel that connects financial clouds and financial outlets, and this makes them the cornerstone of efficient and stable financial services. Financial institutions should build high-speed, intelligent, and elastic WANs to connect multiple DCs across different cities and also connect those DCs to financial branches. In China, more than 20 financial institutions, including China Construction Bank and Bank of Communications, have engaged in multi-cloud network reconstruction.
- **Energy:** Deploying a distributed multi-cloud architecture is the cloudification strategy of choice in the energy industry. Migrating all production, management, and operations data to the cloud requires efficient and flexible scheduling of networks, computing power, and data. Digital production, front-end data collection, back-end real-time intelligent analysis, and front-end and back-end collaboration for intelligent operations all require energy data networks to provide a deterministic experience. Energy industries such as electric power and oil & gas have started to build multi-cloud networks based on their digital development requirements.
- **Government:** A government cloud is a physically scattered but logically centralized cloud that provides unified cloud services for all government departments across a country. Government multi-cloud networks are expected to break down barriers between government departments to enable resource integration, meet the different needs of different departments and services, and achieve dynamic multi-level collaboration, intelligent services, intensive construction, and across-the-board coverage.

## Between 2019 and 2023, 300+ enterprises in China built multi-cloud networks

招商银行 CHINA MERCHANTS BANK

交通银行 BANK OF COMMUNICATIONS

广东政务服务网

国家管网 PipeChina

中国银行 BANK OF CHINA

中国建设银行 China Construction Bank

PICC 中国人民保险集团公司 THE PEOPLE'S INSURANCE COMPANY (GROUP) OF CHINA

中国农业银行 AGRICULTURAL BANK OF CHINA

国家电网公司 STATE GRID CORPORATION OF CHINA

江苏省大数据管理中心 JIANGSU PROVINCIAL BIG DATA MANAGEMENT CENTER

中信银行 CHINA CITIC BANK

中国光大银行 Bank

中国南方电网 CHINA SOUTHERN POWER GRID

...

## Features of industry multi-cloud networks

**Elasticity and agility:** On-demand multi-cloud connectivity and multi-cloud collaboration

**Service isolation:** Data security and deterministic experience assurance

**Real-time visualization:** Multi-dimensional visualization of networks, services, and experiences

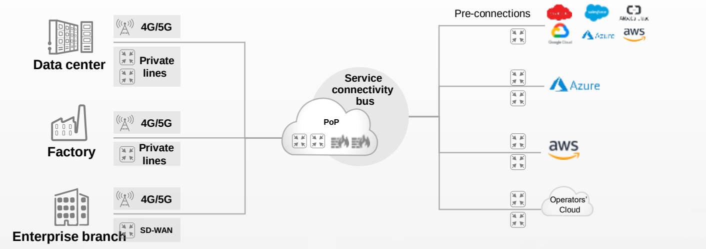
# Carriers Offering Multiple Innovative Multi-Cloud Network Models

**Traditional carrier networks cannot meet industries' multi-cloud service requirements.** Cloud computing—built on networks—and its applications are growing rapidly. Its network requirements are shifting from simple private line access to multi-cloud networks with elasticity, agility, real-time visualization, and reliable experiences. However, carriers traditionally focus on the deployment and O&M aspects of networks, and cannot meet enterprise requirements for business network provisioning speed, adjustment flexibility, and intelligence.

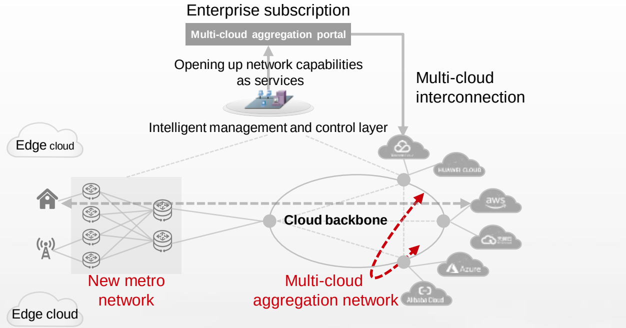
**Carriers have started to innovate in multi-cloud networks.** Carriers have huge network infrastructure, which is advantageous as it allows them to provide various services for enterprises. However, it may also be a burden. When there is a new business need, the sheer volume of network assets means that carriers can only implement network transformation in stages.

- **Cloud Private Line Mode, adding Packages to Increase Profits.** Traditional networking or Internet private lines are replaced by site-to-cloud private lines and multi-cloud connectivity private lines, and SD-WAN is deployed to achieve the agile provisioning of site-to-cloud private lines. Gateway capabilities that support any access methods and pre-connections to multi-cloud resource pools are enabled based on point of presence (POP) resource pools. Site-to-cloud private lines connect enterprises to multi-cloud on demand through POP.
- **Multi-cloud aggregation mode, preempting a unified procurement entry.** An SRv6 cloud backbone is built based on site-to-cloud private lines to achieve pre-connections to local different cloud resources. Carriers offer E2E SRv6 capabilities from enterprises to backbone networks to enable the automated orchestration of enterprise networks in the cloud based on business needs, providing enterprises with access to flexible and high-quality multi-cloud connectivity. Carriers build a multi-cloud aggregation platform with APIs that can be connected to third-party clouds so that they can resell third-party cloud services. This innovative business model fully captures the benefits of accessing multiple clouds with one single line.
- **Industry private network mode, ensuring experience for valued customers.** As office work and production continue their migration to the cloud, high-value sectors like finance, government, and education require isolation from public services for security reasons, and require high-quality networks to deliver a premium experience. To meet these requirements, carriers have started to deploy network slices or industry-specific physical networks based on the SRv6 cloud backbone. In addition, the SLA performance of services running on the cloud is analyzed and displayed in real time through tenant network traffic and performance indicators. This allows tenants to understand the service quality of their private lines in real time and this information can inform SLA monetization efforts. The centralized monitoring of each tenant's SLA performance helps promptly identify incidents like cloud-based network traffic and performance indicator deterioration, so that optimization and targeted maintenance can be performed to enhance cloud service experiences.

## Model 1: Cloud Private Line Mode, Adding Packages to Increase Profits

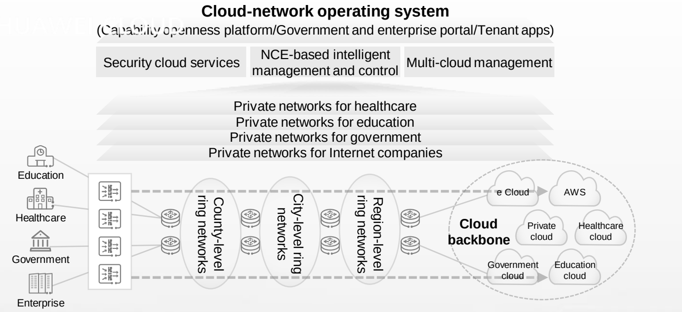


## Model 2: Multi-cloud aggregation mode, preempting a unified procurement entry



Enterprise subscription
Multi-cloud aggregation portal
Opening up network capabilities as services
Multi-cloud interconnection
Intelligent management and control layer
Edge cloud
HUAWEI CLOUD
aws
Cloud backbone
Azure
Alibaba Cloud
New metro network
Multi-cloud aggregation network
Edge cloud

## Model 3: Industry private network mode, ensuring experience for valued customers



Cloud-network operating system
(Capability openness platform/Government and enterprise portal/Tenant apps)
Security cloud services
NCE-based intelligent management and control
Multi-cloud management
Private networks for healthcare
Private networks for education
Private networks for government
Private networks for Internet companies
Education
Healthcare
Government
Enterprise
County-level ring networks
City-level ring networks
Region-level ring networks
Cloud backbone
e Cloud
AWS
Private cloud
Healthcare cloud
Government cloud
Education cloud

# Key Feature 1: Elasticity and Agility to Support the Flexible Scheduling of Computing and Storage Resources

Purchasing private lines with insufficient bandwidth can negatively affect service experiences. Maintaining high-bandwidth private lines long term can be costly. Temporary high-bandwidth services include high-bandwidth real-time communications and periodic data migration. High-bandwidth real-time communications are instantaneous. They are triggered primarily by specific events, and last between several hours and several days. The bandwidth shortage problem cannot be solved by arbitrarily stretching communication time as this compromises the instant communication experience. Periodic data migration does not have high real-time requirements, but does have requirements for total migration time.

**Elastic site-to-cloud private lines meet the bandwidth requirements of enterprise services with significant traffic peaks.** Elastic computing, as a part of cloud computing, has matured. From the perspective of consumers, elastic services fulfill their needs at an optimal cost. These services are designed to meet the requirements across different business scenarios and are offered under the pay-as-you-use (PAYU) model, which makes them highly cost-effective. From the perspective of suppliers, elastic services are essentially efficient management of resources to maximize their effectiveness. Elastic site-to-cloud private lines learn from the concept of elasticity in cloud computing, where network bandwidth resources are pooled and tenant service traffic is detected and predicted in real time to facilitate the flexible scheduling of bandwidth resources across the entire network, ensuring the service experiences of elastic site-to-cloud private lines.

**Elastic private lines guarantee user experiences and help monetize network resources.** Elastic private lines mean enterprises can temporarily increase bandwidth or purchase a traffic package based on their business needs while retaining a fixed-bandwidth private line, with the temporarily increased bandwidth or purchased traffic package offering the same quality as the fixed-bandwidth private line does. This truly allows for on-demand purchase and pay-per-use. Carriers can make full use of idle bandwidth resources to maximize network value.

## Elastic network architecture

**Elastic network service layer**
Elastic network service catalog and open APIs

**Traffic decision-making and scheduling layer**
Balanced traffic scheduling at the network and node levels

**Elastic resource abstraction layer**
Node-level bandwidth granularity abstraction and management

## Four features of elastic networks

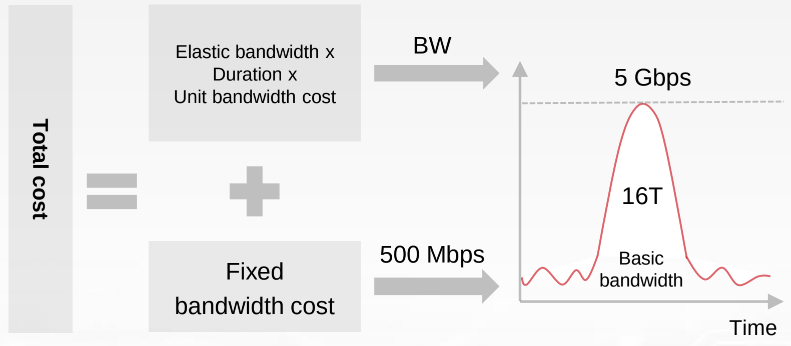
**Mbps-level bandwidth granularity pooling**

**Minute-level tenant traffic prediction**

**Second-level intelligent decision-making and scheduling**

**Second-level elastic bandwidth adjustment**

## Elastic networks enable users to buy elastic traffic packages based on usage

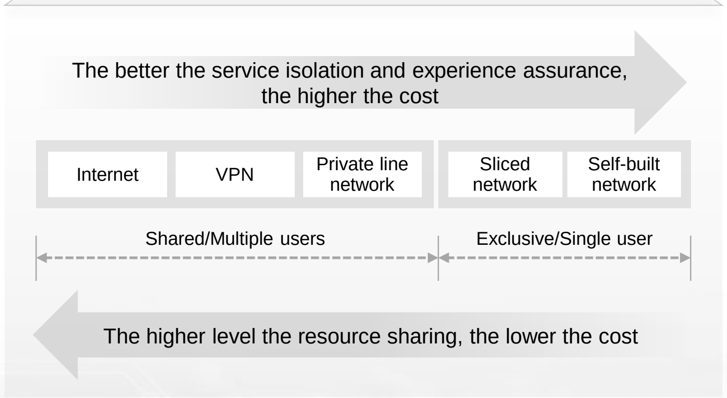


# Key Feature 2: Service Isolation to Guarantee the Quality of Critical Cloud-Based Services

Cloud migration across industries raises higher requirements for security isolation and service experiences. Network slicing can be used to meet the security isolation and differentiated assurance requirements of different services on the same network.

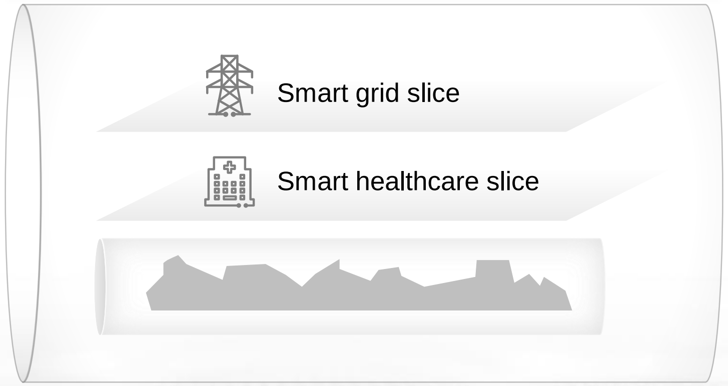
- **Resource and security isolation:** From the perspective of service quality, the purpose of IP network slice isolation is to prevent a service burst or abnormal traffic in a slice from affecting other slices in the same network, which ensures that services in different network slices do not affect each other. This is especially important for vertical industries, such as smart grids, which have strict requirements on latency and jitter and whose performance is highly sensitive to the impact of other services. From the perspective of security, if information about services (private line services, such as finance and government services) in an IP network slice is not expected to be accessed or obtained by users in other network slices, effective security isolation measures need to be taken between different slices.
- **Differentiated SLA assurance:** Network slicing enables carriers to expand beyond selling traffic and provide differentiated services in the form of slices for tenants from different industries. On-demand, customized, and differentiated services will be the main business model for carriers in the future, and will also be a new value-generating area of growth for them.
- **High reliability:** High-value services and ultra-reliable low latency communications (URLLC) require IP networks to provide high availability. Millisecond-level fault recovery has become a basic requirement for IP networks. SRv6-based network slicing provides local protection against any faults on the IP network, such as Topology-Independent Loop-Free Alternate (TI-LFA) and midpoint protection. These protection technologies can help significantly improve the effectiveness of protection and enhance the reliability of IP network slices. In addition, link failover in a network slice can be performed within the slice without affecting other slices.

## Network slices provide the highest possible resource isolation and experience assurance



The better the service isolation and experience assurance, the higher the cost
Internet
VPN
Private line network
Sliced network
Self-built network
Shared/Multiple users
Exclusive/Single user
The higher level the resource sharing, the lower the cost

## Network slices provide private network-style cloud experiences for industries



**Flexible slicing and differentiated SLAs**

- E2E network slicing based on a fixed bandwidth value or bandwidth convergence ratio
- Transparent transmission on a local network that does not support slicing

**Multi-purpose network with a guaranteed experience**

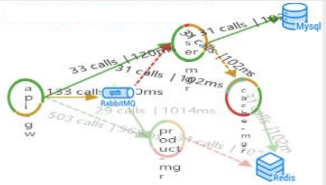
- Slices are subnets which allow for the exclusive use of resources
- SRv6 can flexibly compute the optimal service path within a slice

# Key Feature 3: Real-Time Visualization to Monitor Service Quality from End to End

- **A lack of network visibility leads to inefficient O&M.** The complexity of enterprise networks will grow exponentially. This is reflected in the following aspects: (1) As hybrid office becomes popular, there will be more interconnected branches and access locations; (2) The convergence of office networks and the Internet of Things (IoT) will lead to a surge in connections; (3) Cloudification and new applications will raise higher and more volatile requirements for network performance; (4) A growing number of network equipment varieties and manufacturers can significantly scale up equipment management workloads; and (5) Requirements for network assurance will be higher, and will move from connectivity-oriented to experience-oriented. Meanwhile, the number of O&M assurance engineers will not increase proportionally, if at all, which means that more work will need to be done by fewer people. As a result, the pain points of network O&M will become more evident. There is no unified view to assess the health of enterprise networks. Users may have poor network experiences, resulting in an increased number of fault complaints. Fault recovery can be slow. To sum up, network O&M is not even close to keeping pace with enterprise digital transformation.

- **Network visualization enables the real-time detection of network changes.** Network visualization comprises real-time, dynamic, and HD network-wide resource visualization capabilities. Key technologies like big data computing engines, AI, search algorithms, route simulation, and verification algorithms are used to achieve multi-dimensional visibility, path navigation, searching and locating, and deterministic application experience assurance. Real-time network quality visibility, demarcation and locating, and self-healing capabilities are also provided to help customers switch from the traditional O&M model using static topology to one using dynamic HD electronic maps. In other words, digital maps are used to give an intuitive view of networks and significantly boost network O&M efficiency.

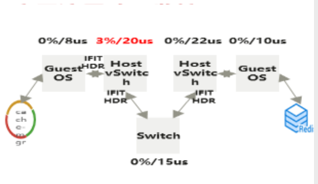
## Integrated monitoring



**Application interaction topology monitoring**

| Interaction topology visualization | Network transmission performance | Host service performance |
|---|---|---|

ACK delay · TCP retransmission · TCP handshake time · Zero window · Number of connections · Connection reset



**Application traffic and road quality monitoring**

| Traffic visualization | Road quality visualization | Node load analysis |
|---|---|---|

Application IFIT · Delay detection · Packet loss detection · Egress/Ingress port visibility · Throughput distribution · Collaborative metric analysis

**Fault diagnosis**

- Traffic dialing test
- Packet capture
- Active link monitoring
- Full-link fault diagnosis
- Audit/ Verification
- Network IFIT

**Intelligent analysis**

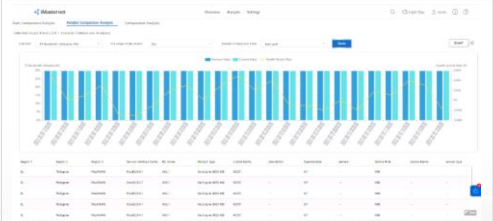
- Log correlation analysis
- Root cause locating
- Network fault prediction
- Network capacity prediction
- Fault impact analysis

**TOP N network issues: Quick identification of network bottlenecks**

**Performance topology: Clear presentation of quality deterioration**

**Regional traffic/quality map: Layer-by-layer drill-down analysis**

**MoM, comparison, and YoY analysis: Early identification of potential network risks**



# Recommendations for Action: Multi-cloud Network with Elasticity, Agility, Service Isolation, and Real-Time Visualization

**Increase cloud-network investment**

Cloud-network deployment has become a trend in many industries. Jump on the bandwagon and develop multi-cloud networks to build a multi-cloud ecosystem. Promote in-depth cloud-network collaboration, make it easier to migrate services to the cloud, and improve cloud-based service experiences.

**Explore business model innovation**

By reselling third-party cloud services and providing elastic private lines, carriers can better meet the network requirements of enterprise customers that are migrating services to the cloud, and better leverage their network resources to achieve revenue growth.

**Actively introduce new technologies**

Network technologies such as SRv6 and network slicing have been effective in simplifying cloud-based networks and guaranteeing cloud-based service experiences. When promoting cloud-network collaboration, enterprises and carriers should consider introducing new technologies that will benefit them.

**Improve network visualization capabilities**

Developing comprehensive network visualization capabilities can help you boost network O&M efficiency, better understand the network status, predict service growth, and effectively scale up networks to prevent network bottlenecks from hindering business development.

# Contents

01 Trend 1: Multi-Cloud Has Become the New Normal, and the Innovation of Elastic, Reliable, and Visualized Networks Is Accelerating

**02 Trend 2: Boosting Foundation Models, Propelling Fundamental Transformation of DCNs**

03 Trend 3: The Sweeping Digital Transformation Propels Campus Networks into an Experience-Centric Future

04 Trend 4: AI Development's Watershed Moment from the Point to System Level

05 Trend 5: Integrated, XaaS, and Intelligent Become New Features of Network Security Construction

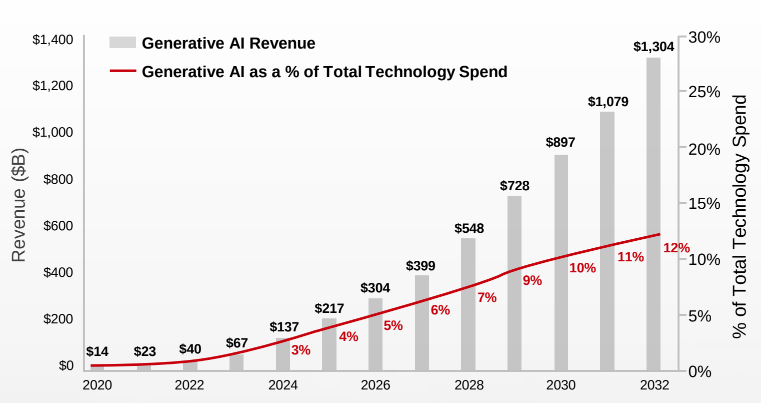
# AIGC Creates Trillion-Dollar Industry Market Spaces and Accelerates Global Computing Infrastructure Construction

- **Foundation models proliferate and the AIGC era arrives.** Sparked by the release of ChatGPT by OpenAI in November 2022, the AI industry has quickly entered a new era of Artificial Intelligence Generated Content (AIGC), creating a new paradigm for human and production interaction. This has also ignited a content productivity revolution in the AI era. According to the latest report from Bloomberg Intelligence, the market space of generative AI will reach US$1.3 trillion by 2032, a significant jump from US$40 billion in 2022, hitting a compound annual growth rate (CAGR) of 42%.

- **AIGC will trigger in-depth transformation across the entire industry.** AIGC is penetrating a wide range of industries at an accelerating rate. Generally, AIGC mainly affects content creation and human-machine interaction. As industries become more online-oriented and the proportion of content in the value chain grows, the more obvious the disruptive effect of AIGC will be. For example, the e-commerce, gaming, and advertising industries are particularly online-oriented, and the content quality directly determines the value created. Therefore, AIGC can maximize its benefits only when it is widely applied across industries.

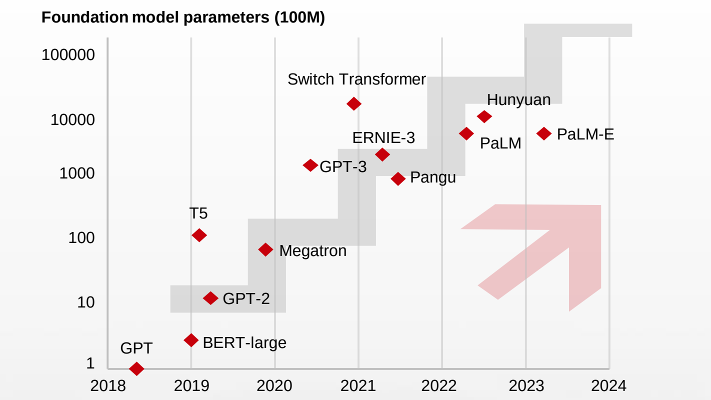
- **Global computing infrastructure construction is speeding up.** The parameter quantity of OpenAI's GPT series models already exceeded 117 million in June 2018. That quantity has since mushroomed, reaching billions or even trillions. On average, the number of model parameters doubles every three to four months, bringing an increasing demand for training computing power. For each 1% increase in the computing power index, the digital economy and Gross Domestic Product (GDP) increases on average by 3.5‰ and 1.8‰, respectively. Computing power is becoming a key factor that affects a nation's comprehensive strength. The construction of computing infrastructure has become a strategic initiative for the high-quality development of a nation's digital economy. According to IDC, global enterprises' investment in AI infrastructure and services is expected to exceed US$200 billion by 2025, far surpassing enterprises' digital transformation (DX) and GDP.

## Generative AI market revenue forecast (2020–2032)



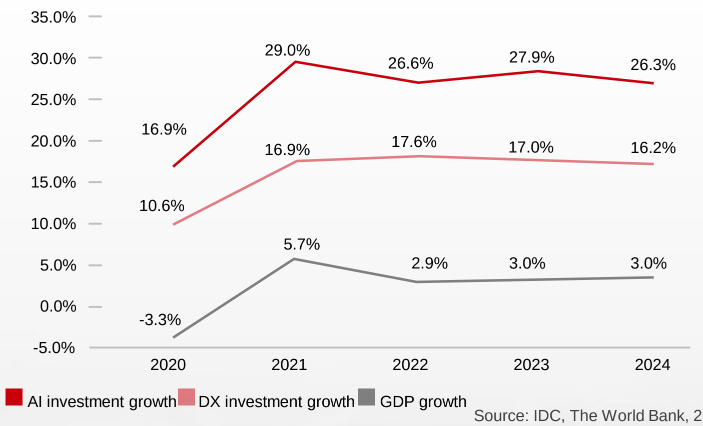
Source: Bloomberg Intelligence

## Parameter quantity change trend of global foundation models



Source: iResearch

## Global AI investment growth far exceeding DX and GDP

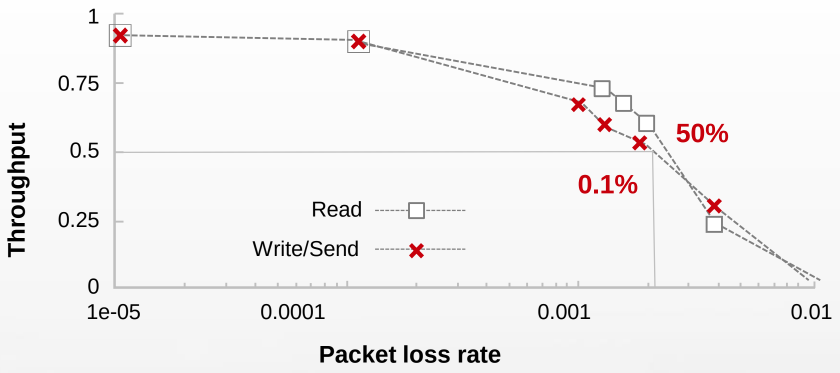


Source: IDC, The World Bank, 2022

# Network Determines Training Efficiency, and Conventional Networks Cannot Meet AI Requirements

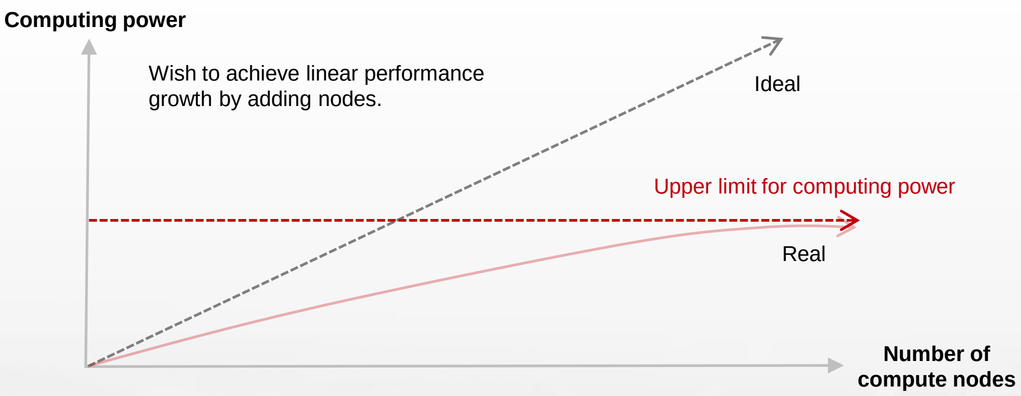
- **Conventional Ethernet cannot meet the requirements of AI DCs**. According to IDC, Ethernet accounts for more than 95% in mainstream DCs. However, in AI training scenarios, conventional Ethernet doesn't perform well in terms of network throughput, latency, and packet loss prevention. The communication mode of AI applications brings a new challenge to CPU and GPU servers as well as the existing underlying network infrastructure. Conventional Ethernet is natively prone to packet loss and cannot meet the requirements of DCs in today's AI era, where data loss cannot be tolerated during AI training.
- **An ultra-large network is required for a compute cluster with 10k GPUs**. To launch foundation models more quickly and meet the growth requirements of model parameter and token quantities, the cluster scale has increased from 1k GPUs to 10k GPUs. Take OpenAI's GPT-4 as an example. It uses thousands of GPUs to train 1.8 trillion parameters. A large-scale training network is required.
- **An ultra-high-throughput network is required for models with trillions of parameters.** Foundation models are trained in distributed mode to improve training quality and speed. In this mode, vast quantities of parameters are distributed to multiple GPUs housed in multiple servers. As such, thousands or even tens of thousands of GPUs are required to train dozens of TBs or more of data. Heavy communication traffic between many GPUs is prone to network load imbalance, which then decreases the network throughput. As a result, the overall AI training performance deteriorates.
- **A highly reliable network is required for long-term stable training.** Foundation model training is a complex project. Ensuring that systems run stably is of vital importance to the entire training process, spanning data preparation, model pre-training, and model training. Network infrastructure is the key to long-term stable training. Take a model with hundreds of billions of parameters as an example. Its total training duration is 65 days, but due to system faults causing the model to restart more than 50 times, the effective training duration is only 33 days. Typically, the training duration of foundation models is long with many interruptions.

## Just 0.1% packet loss rate can reduce network throughput by 50%



Source: Congestion Control for Large-Scale RDMA Deployments

## Network performance causes an upper limit for computing power, leading to serious ROI imbalance



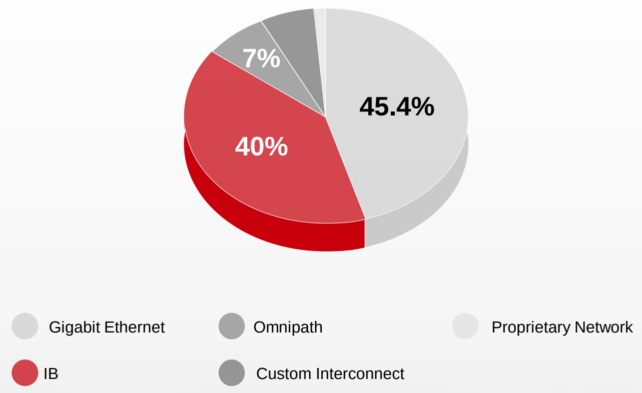
# Continuous Innovation of Ethernet, Promoting AI DCs from Being Closed to Being Open

- **Active industry layout:** In July 2023, the Linux Foundation joined hands with multiple vendors to establish the Ultra Ethernet Consortium (UEC), with the aim of improving the data transmission speed and network performance so as to better adapt to ever-growing AI and HPC workloads. According to the UEC's chairperson, the project is built on Ethernet technology because it is the best representative of durable, flexible, and adaptable basic network technologies in the industry.
- **Wide application**: InfiniBand (IB) features high bandwidth and low latency, and is typically the solution of choice for building conventional HPC networks. However, the IB architecture is closed and has poor scalability, leading to high network deployment and maintenance costs. As Ethernet advances, it is finding wider applications in HPC and AI fields. According to the latest statistics, 45.5% of the world's top 500 HPC systems use Ethernet interconnection, surpassing IB. Remote Direct Memory Access over Converged Ethernet (RoCE) networks have been widely used in the compute clusters for foundation models, such as PCL-G, Huawei PanGu-Σ, and Baidu ERNIE Bot.
- **Continuous implementation of innovative solutions:** Industry players continue to innovate based on Ethernet technology. Since Huawei released the AI Fabric ultra-fast Ethernet solution in October 2018, multiple mainstream vendors have been proactive in promoting technical breakthroughs, and have launched high-speed interconnection products and solutions tailored for the HPC and AI fields one after another.

## Mainstream players release new products based on Ethernet technology innovations one after another

| Time | Vendor | Event |
|---|---|---|
| October 2018 | Huawei | Released the AI Fabric ultra-fast Ethernet solution. |
| August 2020 | HPE | Released Slingshot — Ethernet interconnection technology for HPC. |
| April 2022 | Inspur | Released the RoCE-based lossless Ethernet solution. |
| May 2023 | NVIDIA | Released the high-performance Ethernet architecture — Spectrum-X. |
| July 2023 | Microsoft, Broadcom, AMD, Intel, etc. | Jointly established the UEC. |

## Among global top 500 HPC systems, Ethernet deployment exceeds IB deployment



## Ethernet-based RoCE networks are widely used in foundation models

### Industry applications

| Biomedicine | Electric power | Remote sensing |
|---|---|---|
| Gene research, drug R&D... | Intelligent inspection… | Change monitoring, ground object classification... |
| PCL-G | Pangu Electric Power | LuojiaNET |

### Foundation models

Natural language processing (NLP)

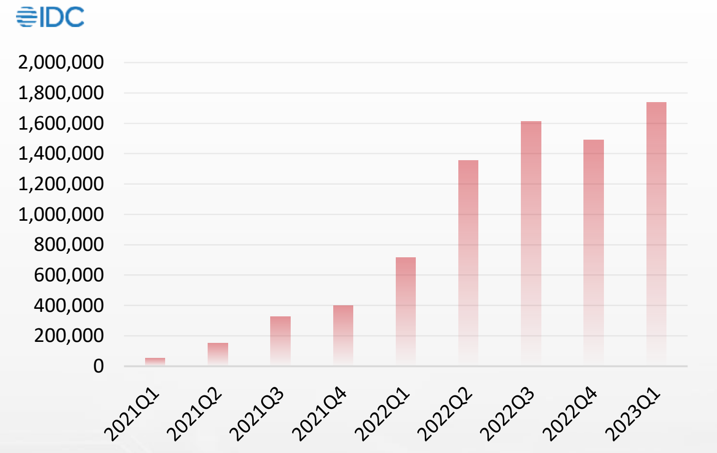
SparkDesk cognition model | ERNIE 3.0

…

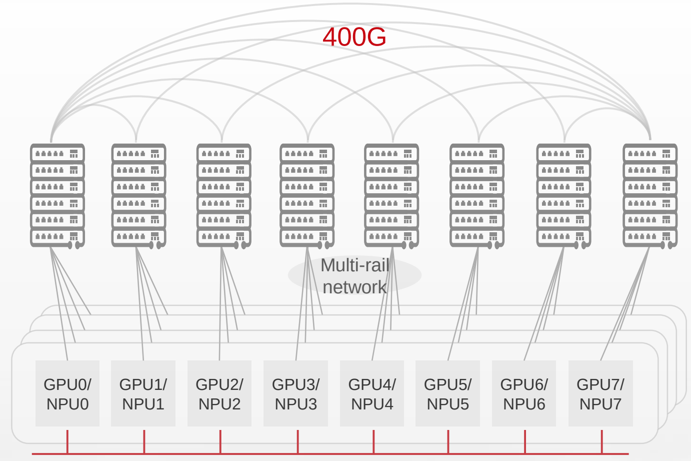
# 400GE Switches Enter the Batch Deployment Phase, Supporting Ultra-Broadband Simplified Architectures

- **Open ecosystem and rapid technology development**: Ethernet has always been an open ecosystem, laying a solid environmental foundation for network technology iteration. After more than 40 years of development, Ethernet has rapidly evolved from offering 10 Mbit/s to now offering 400 Gbit/s, providing ultra-broadband channels for high-speed transmission of massive amounts of data in AI scenarios. According to IDC, from 2021 to 2023, the CAGR of 400GE port shipments was 46%. In the first quarter of 2023, the 400GE port shipments reached 1.73 million.
- **Mature industry spanning standards and products:** In 2013, the 400G Ethernet standards project was officially initiated. In 2017, the Ethernet standards defined by IEEE 802.3bs were approved, indicating the maturity of the 400GE standards. As of today, all mainstream vendors are capable of offering 400GE switches. In 2019, Huawei released the industry's first highest-density 400GE DC switch built for the AI era — CloudEngine 16800.
- **400GE switches are ideal for building an ultra-broadband simplified architecture**: So far, AI clusters generally use high-performance network interface cards (NICs), each of which can provide 200G or even 400G bandwidth. This brings an increasingly urgent need for 400G access and interconnection. By adopting high-bandwidth Ethernet switches, it is possible to build a flexible network architecture to meet networking requirements in different service scenarios. Two important architectures that have attracted wide attention in the industry are the multi-rail network architecture and the Clos network architecture. In the distributed AI training scenario, the multi-rail network architecture only needs to build multiple independent network planes to connect GPUs/NPUs with the same number. This architecture, compared with traditional ones, can effectively reduce the number of network layers, number of data forwarding hops, and network construction costs. The Clos architecture uses a network-wide oversubscription-free design, helping to build a non-blocking large-capacity network featuring on-demand horizontal expansion. This architecture has higher universality and scalability and supports larger-scale networking requirements.

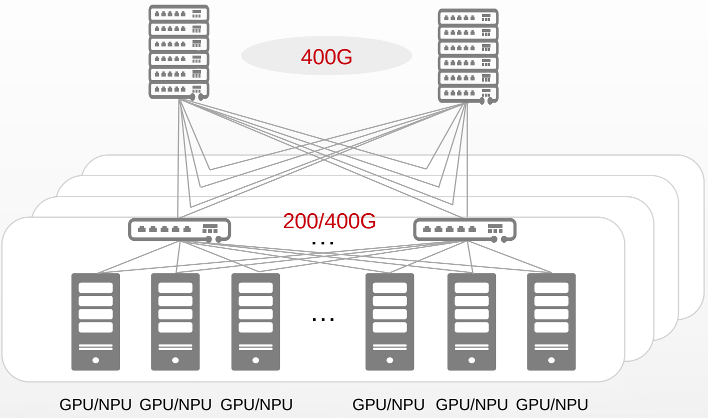
## IDC: 400GE port shipment statistics



## Multi-rail network architecture, reducing network layers and network construction costs



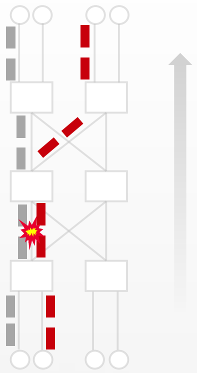
## Clos network architecture, featuring high universality and scalability



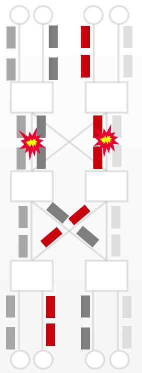
# Zero Packet Loss -> Zero Blocking, Improving AI Training Efficiency

- **Network throughput is the key to AI training efficiency.** Although mainstream vendors already have ways to address Ethernet packet loss, zero packet loss is the minimum network requirement of AI computing. Furthermore, we need to find ways to improve network throughput. There is an inherent problem in network transmission: The receiver will trigger traffic retransmission upon receiving out-of-order traffic, leading to a reduction in speed. AI training mainly produces elephant flows (100 MB to several GB) — while the number of these flows is small, each one involves a large amount of data. If such flows are transmitted in the conventional load balancing mode, network nodes steer traffic only from their own perspectives. As a result, load imbalance can easily occur (the annual network throughput is just 50%). The next round of communication can start only after the slowest flow in this round reaches its destination, meaning that the slowest flow determines the overall network performance. On a network with no global load balancing technique, the overall communication efficiency is between 30% and 50%. This means that half of the network performance is wasted, and the computing power utilization of the entire cluster is only 30% to 50%.
- **Network scale load balancing (NSLB) improves network throughput.** To improve network throughput, mainstream industry players prefer to perform in-depth collaboration and adaptation of devices, networks, and protocols. This allows them to implement network-wide load balancing and achieve over 90% network throughput in addition to facilitating the adaptation of RoCE networks to foundation model training requirements. Huawei's NSLB technology is an ideal choice here. It enables collaboration between the network controller and AI scheduler to perform global path calculation based on the traffic congestion status of network-wide switches and the network topology, and obtains the communication matrix based on the training tasks assigned by the AI scheduler. In addition, NSLB can identify the optimal path based on the communication library, network topology, bandwidth, and congestion status, and automatically deliver the optimal path to network switches. Service flows are then transmitted along the path. All of this helps to improve the network throughput to over 90%.

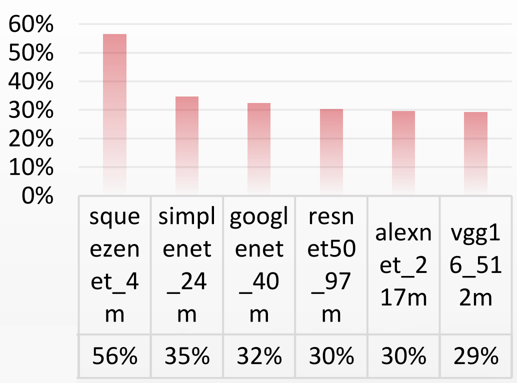
## Conventional solutions are prone to load imbalance, causing network congestion and affecting the training speed



**Local conflict**
Reactive static hash mechanism prone to load imbalance, causing uplink congestion on leaf nodes

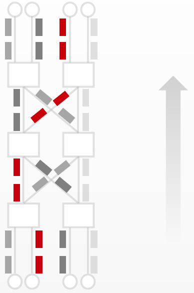


**Global conflict**
Unable to predict next-hop conflicts, causing downlink congestion on spine nodes

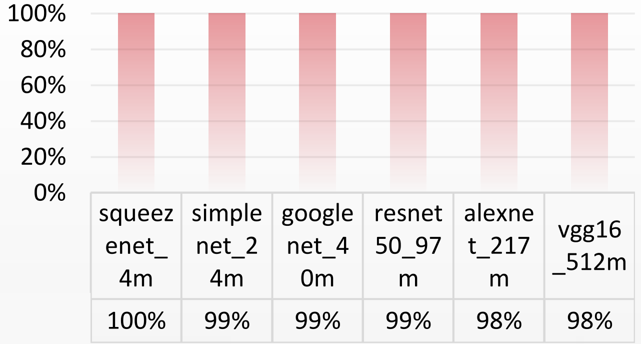


| squeezenet_4m | simplenet_24m | googlenet_40m | resnet50_97m | alexnet_217m | vgg16_512m |
|---|---|---|---|---|---|
| 56% | 35% | 32% | 30% | 30% | 29% |

## Huawei's NSLB technology improves network throughput to over 90%



- Reactive -> proactive
- Static -> dynamic
- Local -> global



| squeezenet_4m | simplenet_24m | googlenet_40m | resnet50_97m | alexnet_217m | vgg16_512m |
|---|---|---|---|---|---|
| 100% | 99% | 99% | 99% | 98% | 98% |

Implement global dynamic load balancing based on AI traffic characteristics to build a non-blocking, full-throughput network in multi-task, multi-tenant scenarios.

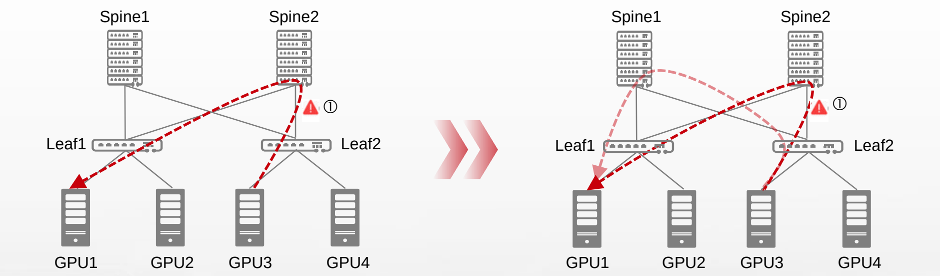
# Fault Rectification Within Milliseconds, Ensuring High Cluster Stability and Reliability

- **High network reliability is fundamental to cluster system reliability.** The AI computing center network on which the foundation model relies serves as the core hub of service traffic. The stability of this network directly affects the stability of the entire cluster system. Because the network fault domain is large, failure of a single network node will affect the connectivity of dozens (or more) of compute nodes. In addition, unlike a single GPU or server that is easy to be isolated, the network shares resources as a cluster. Any performance fluctuation can affect the utilization of all computing resources. As such, continuous network stability is vital to foundation model training. Improving the fault rectification capability and O&M efficiency of the network is therefore an urgent problem that needs to be resolved.
- **Technological innovation directions for high network reliability:**

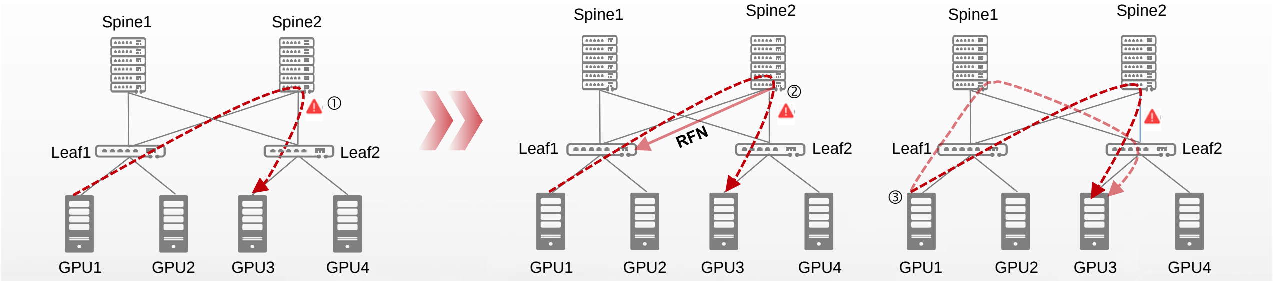
**(1) Fast hardware awareness, fault rectification in sub-milliseconds:** In the AI training scenario, each host communication task takes only a matter of milliseconds to complete. However, the conventional route convergence mechanism takes seconds, potentially interrupting multiple rounds of AI host communications and significantly affecting AI efficiency. One method to solve this problem is to leverage the Data Plane Fast Recovery (DPFR) technology. This technology provides capabilities such as rapid fault detection and rapid fault rectification on the local side and remote side, thereby implementing rapid link failover in sub-milliseconds with no impact on the training task.

**(2) Pre-training intelligent self-check and in-training intelligent operations and maintenance (O&M):** Generally, 90% of faults on high-performance networks are caused by incorrect configurations. And, as the AI training cluster scale grows, the configuration complexity keeps increasing. The computing-network collaboration mechanism — widely considered as an important technology for stable delivery of AI clusters — provides an AI scenario-oriented network model to achieve automatic generation, delivery, and detection of network configurations. Furthermore, foundation models are characterized by heavy traffic and short periods. The conventional polling and packet sampling mechanism cannot support visualization of AI network traffic indicators, and the entire network is regarded as a black box. Through network performance measurement in milliseconds as well as collective communication performance measurement in the computing-network collaboration solution, high service visibility, poor-QoE issue analysis, and rapid fault demarcation are implemented. In addition, the cluster computing O&M platform is used for unified resource scheduling in order to rapidly rectify network faults. This is another exploration direction of industry players.

## Fast fault rectification on the local side



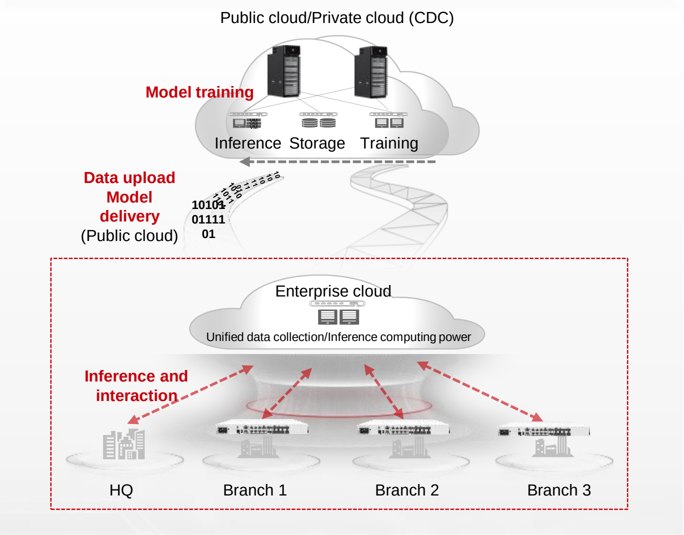
## Fast fault rectification on the remote side



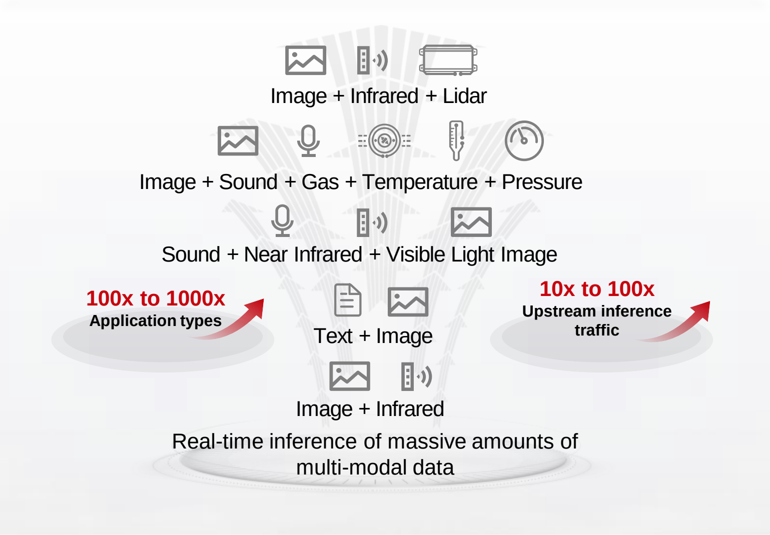
# WANs Are Evolving Towards Elastic and Intelligent, Accelerating AI Inference

- **AI inference brings new network requirements.** As AI technologies and industrial intelligentization gain momentum, the network plays an increasingly important role in connecting human beings, things, and conventional applications, and supporting intelligent applications throughout their lifecycles spanning foundation model training, distribution, inference, and iteration. On the one hand, model training on the cloud and inference off the cloud bring massive amounts of data transfer, requiring high network bandwidth and throughput. On the other hand, as vast numbers of AI inference terminals and applications enter the core production system of enterprises, the number of applications grows 100-fold. Different AI applications have diverse network requirements. Take a typical industrial campus network as an example. AOI machine vision requires real-time inference and interaction, software package download requires high bandwidth, and video conferencing requires stable bandwidth. All of this brings a new challenge: how to enable fine-tuned and differentiated experience assurance on the network.
- **Building 400GE/800GE elastic and intelligent WANs.** The industry is exploring how to use 400GE/800GE devices to build an ultra-broadband network and leverage the network-terminal-computing collaboration technology and intelligent scheduling algorithm to perform intelligence awareness and analysis of applications, accurately predict network traffic change trends, and intelligently optimize network resource allocation by application type so as to remove network congestion in advance, ensure efficient transmission of massive amounts of training data, and meet differentiated service assurance requirements of applications.

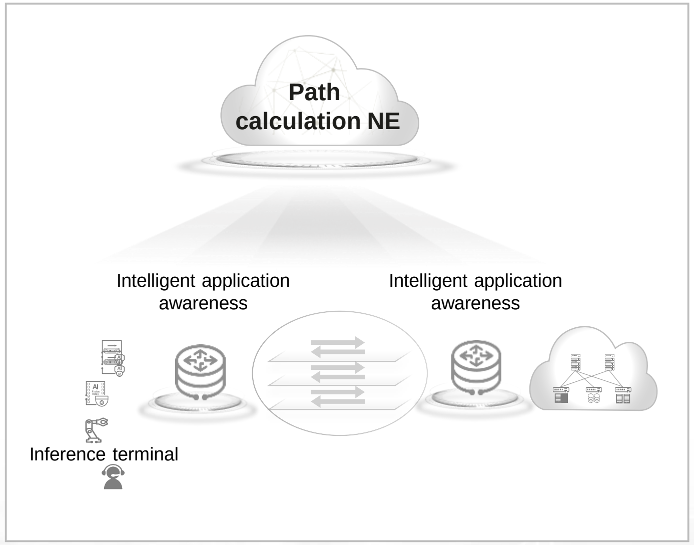
## PB-level training data and TB-level model files delivery, bringing massive data transmission requirements



## AI enters the production system and applications increase 100-fold, networks face challenges



## Elastic and intelligent WAN, accelerating inference and interaction



# Recommendations: Empower DC with AI Networks Featuring Ultra-Broadband, High Throughput, and High Reliability

**Select open Ethernet technology**

As industrial intelligentization gathers pace, AI is penetrating a wider range of industries and domains, not only gaining a larger and larger market space but also bringing a higher demand for computing power. When building a large-scale cluster network, we need to take the scalability of technologies into full consideration in order to better adapt to service growth and changes. The open Ethernet technology enables flexible networking and offers good compatibility with diversified computing power for flexible access based on different service scenarios and computing power requirements. In addition, the use of open technologies can avoid vendor lock-in, increasing the bargaining and selection space.

**Build 400G large-bandwidth networks**

Network performance has become a key factor that determines the training efficiency. Large bandwidth and high throughput are two basic features of AI networks. High bandwidth is defined as at least 200GE, and 400GE/800GE high-speed interconnection needs to be available to meet the transmission requirements of larger-scale AI training data. Furthermore, cutting-edge technologies such as NSLB are recommended for building an ultra-broadband, simplified, and non-blocking network that can improve the effective network throughput and AI training efficiency.

**Proactively promote network automation and intelligence**

As the quantity of foundation model parameters continues to grow, the network scale multiplies. As a result, the complexity of network deployment and O&M increases exponentially. To keep up, the network for AI DCs should be as automated and intelligent as possible, spanning network deployment, configuration management, and troubleshooting.

**Focus on computing-network collaboration to ensure differentiated application experience**

AI network is an end-to-end network that covers all scenarios spanning clouds, networks, edges, and devices. It includes the DCN, WAN, and networks that cover edges and devices. Computing-network collaboration is widely considered as a key technology that supports foundation model evolution from training to inference and from special-purpose to general-purpose. Real-time application awareness ensures differentiated experience of key applications, accelerates inference, and enables real-time interaction.

# Contents

01 Trend 1: Multi-Cloud Has Become the New Normal, and the Innovation of Elastic, Reliable, and Visualized Networks Is Accelerating

02 Trend 2: Boosting Foundation Models, Propelling Fundamental Transformation of DCNs

**03 Trend 3: The Sweeping Digital Transformation Propels Campus Networks into an Experience-Centric Future**

04 Trend 4: AI Development's Watershed Moment from the Point to System Level

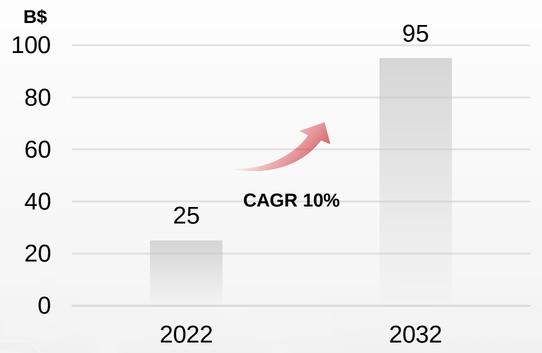
05 Trend 5: Integrated, XaaS, and Intelligent Become New Features of Network Security Construction

# Emerging Services Drive Campus Network Upgrade to Better Support Enterprise Digital Transformation

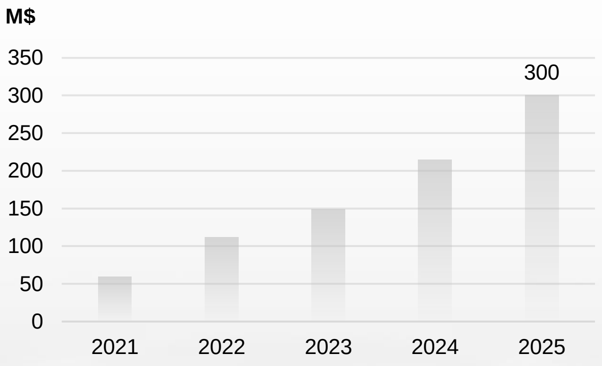
With the acceleration of enterprise digital transformation, new services and applications emerge, which not only improves enterprise office and production efficiency, but also poses new demands on campus networks. Currently, the main driving forces are as follows:

- **Widely-used video conferencing:** Video conference becomes an important tool for remote communication and hybrid office. The video conference market is expected to grow by 10% every year globally and reach US$95 billion by 2032. Take Huawei as an example. Video conferences connect nearly 400,000 users from employees to partners, covering more than 1000 office sites in 170 countries. The number of online users reaches 60,000 during peak hours, and more than 600,000 conferences are held every month. The quality of video transmission directly determines the communication efficiency.
- **Massive deployment of IoT applications:** In addition to traditional office terminals, enterprise digitalization will lead to massive IoT terminals, including devices for asset management, electronic shelf labels, and environment sensors. Take the retail industry as an example. A large number of supermarkets begin to replace traditional paper shelf labels (PSLs) with electronic shelf labels (ESLs) that support remote and real-time price update. The global ESL market will exceed US$3 billion by 2025. The rapid popularization of IoT applications and devices will further complicate enterprise IT systems.
- **Upcoming terminal upgrading:** Countries successively granted the 6 GHz frequency band, which will drive the upgrading of Wi-Fi terminals. By the first half of 2023, the number of terminals supporting 6 GHz has reached 2064, an increase of 260% compared with the first half of 2022. 67 devices support Wi-Fi 7, including 22 phones, 30 routers and gateways, 11 access points, and 4 laptops. In the process of upgrading terminals, enterprises will also expedite to upgrade WLAN for better accessing new terminals.
- **Emerging services (such as immersive experience):** As computing power and video display technologies mature, immersive applications are emerging, including holographic projection, naked-eye 3D, and metaverse office. If enterprises recently consider to innovate such technologies, campus networks will also be affected.

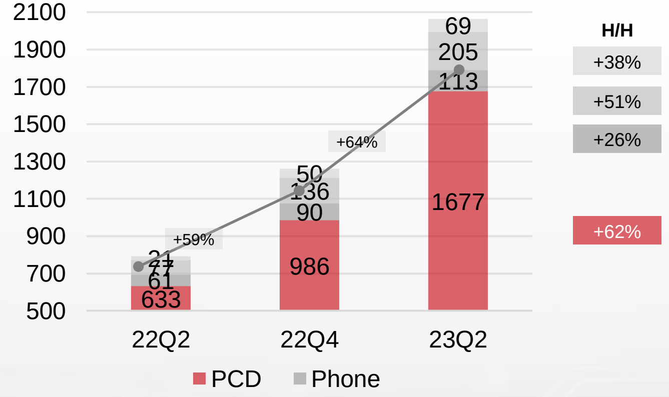
## Enterprise video conferencing market forecast



## Global ESL market forecast (2021 to 2025)



## Rapid popularization of 6 GHz terminals

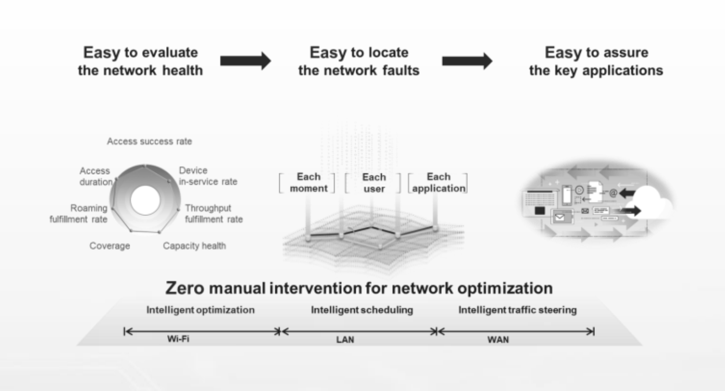


# Concept Upgrade of Enterprise Campus Network Construction, Building High-Quality Campus Networks

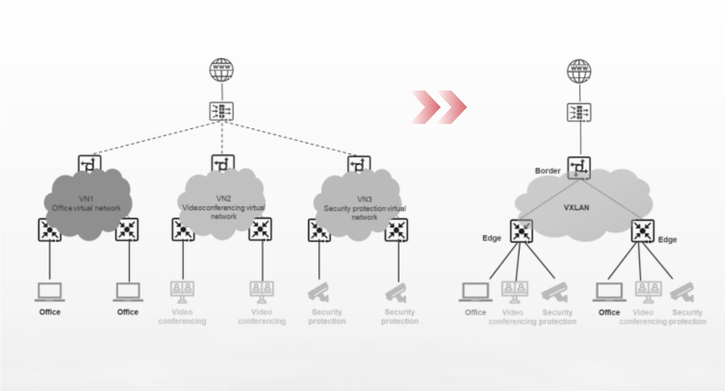
With the development of diversified applications, the number and types of services carried on enterprise networks increase rapidly. Also, diversified services have different demands on network bandwidth, latency, and security. Therefore, we need to change the traditional network construction concept and construct high-quality campus networks centered on user experience.

- **From bandwidth-driven to experience assurance:** Campus network services are evolving from traditional computer-based office services to HD video conferencing services. In addition, office applications are migrating from on-premises ones to cloud-based applications, posing higher requirements on campus network bandwidth and latency. Network bandwidth is costly, so we need to improve bandwidth efficiency, explore differentiated services based on application awareness to ensure user experience.
- **From one network for one purpose to one network for multiple purposes:** The rapid increase of service types on campus networks leads to high costs in dedicated network construction and O&M. Furthermore, network resource utilization is low and information silos hinders free transfer of enterprise data. The converged network for multiple services through network virtualization becomes indispensable for enterprises to reduce costs and improve efficiency.
- **From network self-construction to NaaS:** Traditional campus network planning, deployment, and O&M are usually performed locally and depend on the experience and professional skills of IT O&M personnel. This lead to inefficient network construction and O&M. As such, we need cloud-based automated network planning and deployment as well as intelligent O&M to improve IT O&M efficiency and enable enterprises to focus on business development.
- **From communication connection to converged communication and sensing**: Communication and sensing functions need to be implemented on enterprise wireless networks. They can complement each other for integration between communication and smart life, and promote green development across all industries.

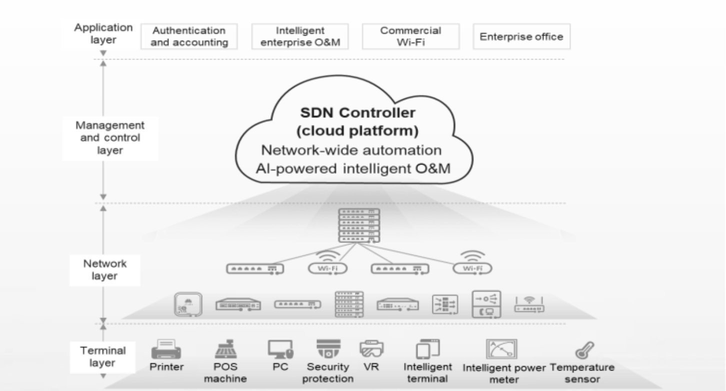
## Network status evaluation -> Experience assurance


Easy to evaluate the network health
Easy to locate the network faults
Easy to assure the key applications
Access success rate
Access duration
Device in-service rate
Roaming fulfillment rate
Throughput fulfillment rate
Coverage
Capacity health
Each moment
Each user
Each application
Zero manual intervention for network optimization
Intelligent optimization
Intelligent scheduling
Intelligent traffic steering
Wi-Fi
LAN
WAN

## Multiple networks -> One converged network



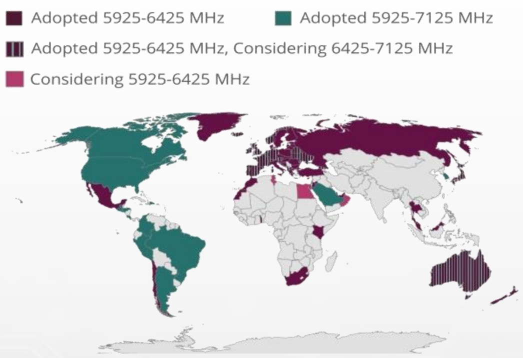
## Cloud-managed campus network



# WLAN Enters the Wi-Fi 7 Era, Accelerating Wireless Network Upgrade for Campuses

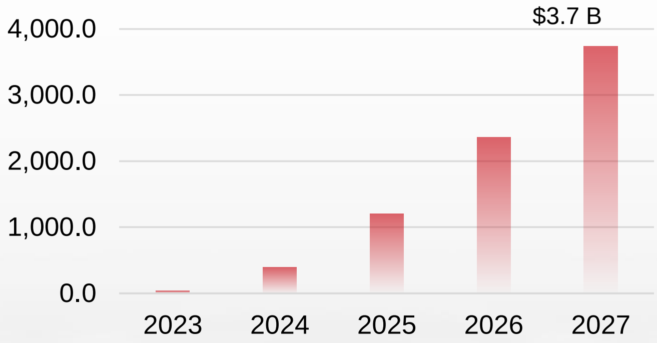
- **Wi-Fi 7 has entered the phase of commercial use.** Draft 4.0 of the Wi-Fi 7 standard (802.11be) was released in July 2023, and is expected to be finalized and officially released in Q1 of 2024. In terms of chips, vendors such as Qualcomm, Broadcom, MTK, and Intel have released Wi-Fi 7 chips since 2022. In terms of spectrum, 54 countries around the world have granted 6 GHz frequency band for Wi-Fi, covering Europe, Asia Pacific, Middle East, and Latin America. Huawei will officially launch the industry's first enterprise-level Wi-Fi 7 in September 2023. Most other vendors will launch Wi-Fi 7 products to the market in 2024.
- **The enterprise-level Wi-Fi 7 market is about to enter a period of rapid development.** According to Gartner's prediction, the shipment of enterprise-level Wi-Fi 7 APs will reach 12.4 million by 2027, equivalent to 27% of the total number of APs. In particular, more than 30% of organizations in the manufacturing market will upgrade to Wi-Fi 7, which will introduce more use cases to their business processes. At the same time, the combination of Wi-Fi 7 and time-sensitive networking (TSN) greatly improves network bandwidth and reliability, which can support key business processes in manufacturing and warehousing, and accelerates the adoption of next-generation Wi-Fi technologies.
- **Wi-Fi upgrade will also drive wired network upgrade.** The peak rate of Wi-Fi 7 exceeds 10 Gbps, and 2.5GE will become the minimum requirement for Wi-Fi 7 APs, which means that traditional GE switches cannot match the bandwidth demands of next-generation enterprise wireless networks. That is why access switches are being upgraded from GE to 2.5/5GE, which will drive the delivery of 25GE aggregation switches and 100GE core switches. According to the market share data released by IDC, the 2.5/5GE, 25GE, and 100GE port shipment grew 108%, 78%, and 62% in 2022, respectively. It is estimated that the shipment will continue to grow rapidly over the next few years, driving enterprise campus networks into the 10GE era.

## 6 GHz frequency band granted in 50+ countries

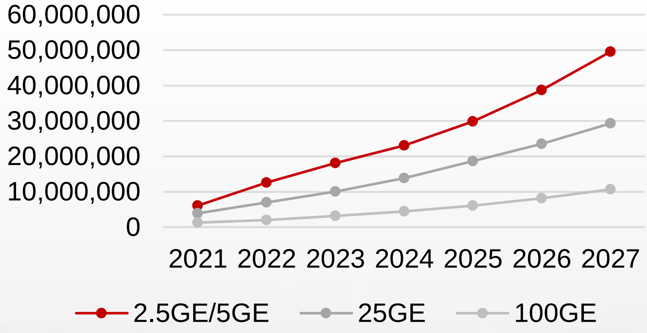


Adopted 5925-6425 MHz
Adopted 5925-7125 MHz
Adopted 5925-6425 MHz, Considering 6425-7125 MHz
Considering 5925-6425 MHz

## Global enterprise Wi-Fi 7 market revenue forecast



## Multi-GE/25GE/100GE port shipment forecast



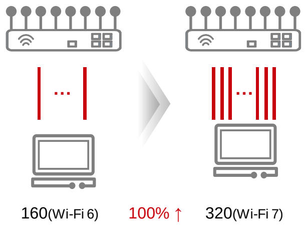
# Bandwidth & Reliability Improvement, Accelerating Scenario-Specific Application of Wi-Fi 7 Across Industries

- **Higher bandwidth:** Compared with Wi-Fi 6, Wi-Fi 7 supports the 6 GHz frequency band in addition to the 2.4 GHz and 5 GHz frequency bands. This reduces signal interference while providing wider spectrum resources, enabling 160 MHz continuous networking. Together with 4096-QAM, the bandwidth is increased by 2.4 times, meeting high-bandwidth needs of applications such as 4K video, AOI-based HD quality inspection, in-vehicle software installation, and AR/VR.
- **Lower latency:** Compared with Wi-Fi 6, Wi-Fi 7 uses multi-resource unit (RU) to flexibly combine RUs over the air interface and allocates multiple RUs to a single user, improving air interface resource utilization. In this way, the average latency is reduced by over 25%. Given this, Wi-Fi 7 is especially suitable for high-quality office scenarios, which can provide better assurance for delay-sensitive services, such as HD video conferencing, interactive office, and cloud multimedia rendering.
- **Higher reliability:** Wi-Fi 7 has been greatly improved in terms of link reliability and user experience assurance. With the multi-link operation feature, multiple data connections (2.4 GHz, 5 GHz, and 6 GHz) can be set up between terminals and APs. The three links can transmit and receive data at the same time to increase the link bandwidth. They can also transmit and receive the same data (multi-fed and selective receiving) to improve the link reliability. In addition, they support data link matching based on application identification to ensure differentiated experience. These features will provide better choices for AGV-based smart warehousing and flexible manufacturing.
- **Application scenarios:** Wi-Fi 7 can be widely used in wireless terminal reconstruction scenarios, such as smart production lines, smart warehousing, industrial terminal control, vehicle drive tests, and future metaverse.

## Higher wireless bandwidth

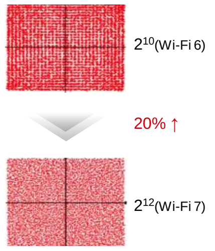
Peak rate of a single terminal: 5 Gbps

Spectrum bandwidth:
160 MHz -> 320 MHz



160(Wi-Fi 6) 100% ↑ 320(Wi-Fi 7)

QAM: 1024 -> 4096



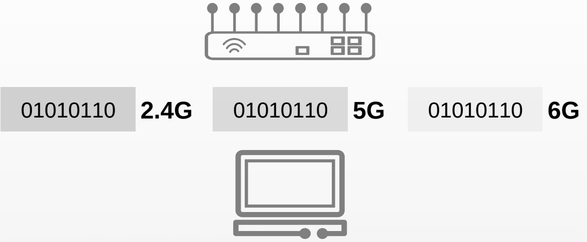
$2^{10}$(Wi-Fi 6)

20% ↑

$2^{12}$(Wi-Fi 7)

## Higher link reliability

Single Link -> Multiple links



01010110 2.4G

01010110 5G

01010110 6G

**Mode 1: Higher performance**
Load balancing among multiple links, improving link bandwidth

**Mode 2: Higher reliability**
Multi-fed and selective receiving, improving link reliability

**Mode 3: Better experience**
On-demand link selection for application, improving user experience

## More application scenarios

**Metaverse**

High bandwidth and low latency
10G+ bps, ms-level

**Manufacturing AOI**

High bandwidth
7- 8 Gbps (site)

**Industrial control**

Low latency
5 ms

**Telemedicine**

Low latency and high reliability
5 ms, multi-link

**Warehouse**

High reliability
Multi-link, 0 interruption

**AR/VR education**

High bandwidth and low latency
1+ Gbps, ＜5 ms

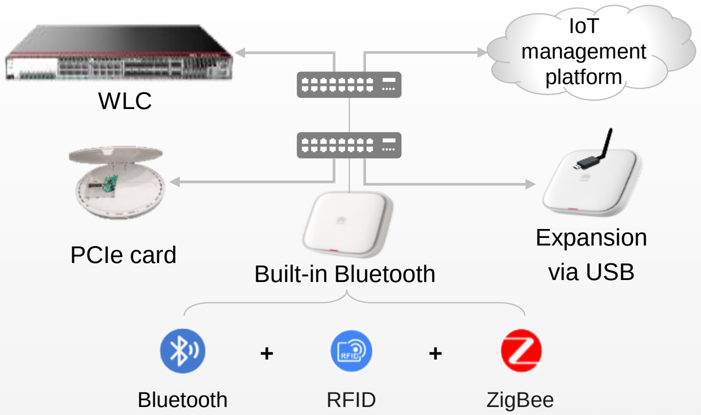
# Multiple Networks -> One Converged Network, Optimizing Enterprise Network Investment

- **Wi-Fi & IoT convergence:** The emergence of IoT applications in campuses increases the procurement, deployment, and O&M demands of enterprises for IoT base stations. Given IoT base stations and APs work in the similar way, plus, WLAN is the preferred access mode in the fully-wireless era, APs are widely used in enterprise office and production areas. Therefore, IoT application expansion based on WLAN APs becomes a better choice. Currently, IoT applications can be implemented through methods such as PCIe cards, USB dongle, and built-in Bluetooth. This solution greatly reduces the network construction and O&M costs.

- **Convergence of the production and office networks:** The campus network usually provides multiple services, such as office, video conferencing, security, production, and IoT. The dedicated network construction mode requires high costs, large equipment room space, and massive integrated cabling. However, the network utilization rate is generally lower than 5%. With the introduction of VXLAN into campus networks, converged bearing of multiple services becomes possible. In this way, one physical network can carry multiple services at the same time. In addition, technologies such as application identification and network slicing are used to provide differentiated priority scheduling for different services, ensuring service quality and user experience.

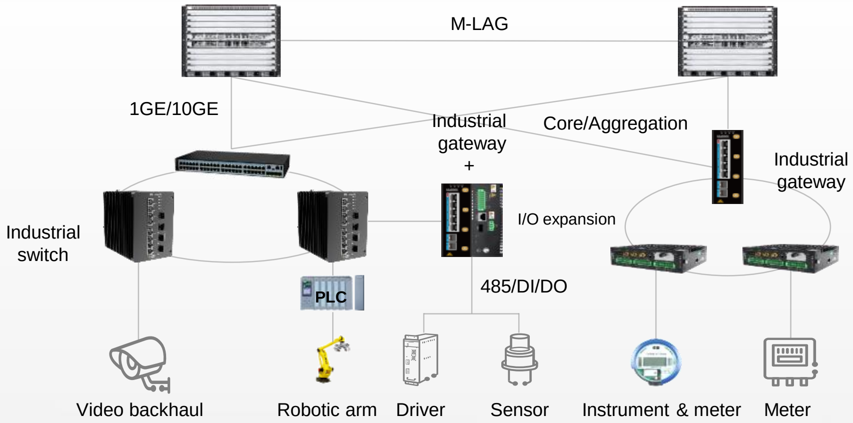
- **Application scenarios:** IoT application and multi-service converged bearing can be widely deployed and applied in industry scenarios such as education, healthcare, retail, and large enterprises.

## Converged IoT and Wi-Fi networking



**WLAN APs as access points to share wired backhaul resources, reducing network construction costs by over 30%**

## Converged production and office network

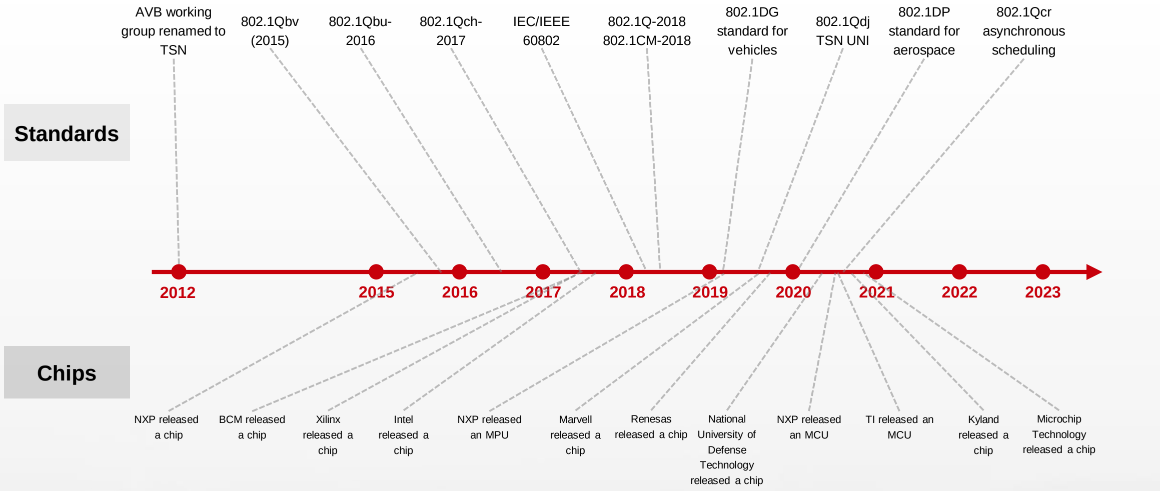


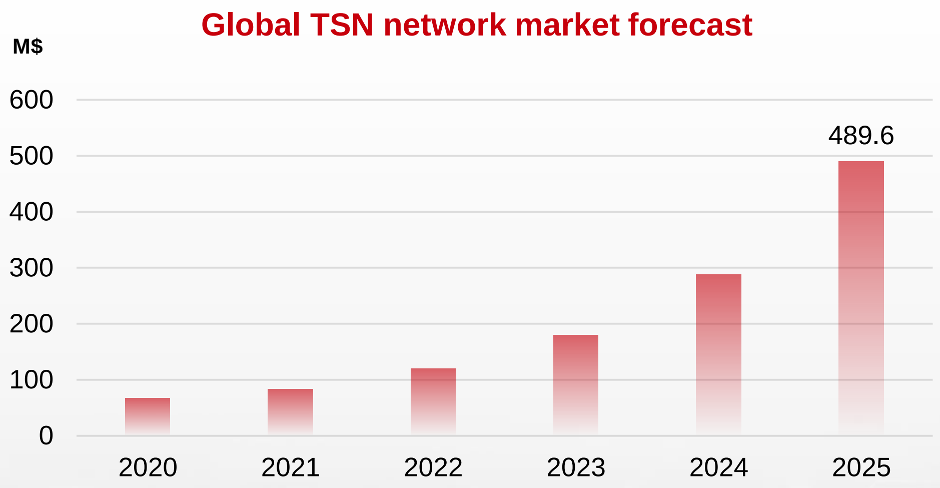
**One physical network for multiple services, securely isolating services and reducing network construction costs by over 50%**

# Non-Deterministic Latency -> Deterministic Latency, Achieving IT/OT Convergence via TSN

- **Rapid development of TSN:** Currently, Industrial bus and industrial Ethernet are the mainstream connection modes in the manufacturing industry. Although traditional IP/Ethernet networks have advantages such as good openness, good interoperability, mature industry, high bandwidth, and low costs, they can only provide best-effort services based on statistical multiplexing, which fail to ensure deterministic low latency required by the industry. The scenario-specific industrial Ethernet network uses specific methods to achieve bounded latency, but has poor interoperability and scalability. Worse yet, the usage of dedicated software and hardware leads to higher costs. TSN combines the advantages of both traditional and industrial Ethernets. It provides users with a network infrastructure featuring low costs, high bandwidth, and statistical multiplexing, solving the problem of difficult interworking between various buses and industrial Ethernet protocols. In addition, it boasts performance highlights such as bounded latency, ultra-low latency, automated network configuration, and high reliability.

- **Accelerated commercial use of TSN:** With the help of time synchronization and precise scheduling, TSN uses the periodic network transmission mechanism to ensure microsecond-level deterministic latency for services. This meets the transmission demands of time-sensitive services in production, manufacturing, and transportation scenarios. Currently, the technical standards (bounded latency, resource management, time synchronization, and high reliability) of TSN have been released. Alongside many chip vendors have launched chips that meet these standards, more than 10 device vendors, including Huawei, have released products and solutions, through which the interoperability between different vendors has been fully verified. In this context, TSN has been put into commercial use in North America and China. In addition, with obvious advantages in bandwidth, latency, and reliability, Wi-Fi 7 can provide flexibility and scalability for TSN. The combination of Wi-Fi 7 (wireless) and TSN (wired) provides more possibilities for applications such as industrial automation and robots, as well as accelerating their commercial use.

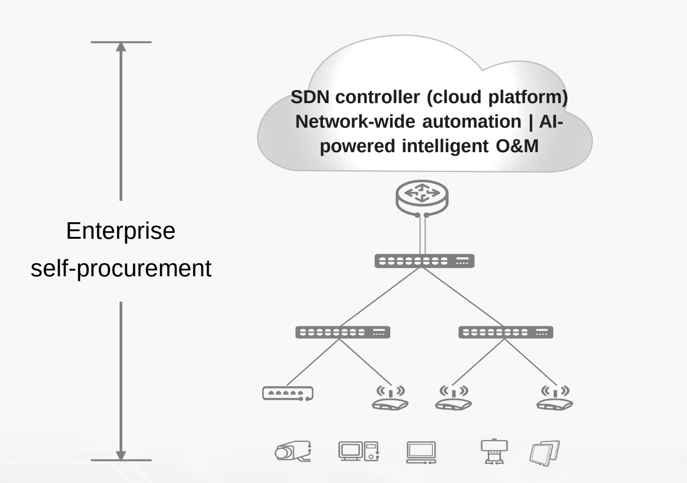



Global TSN network market forecast

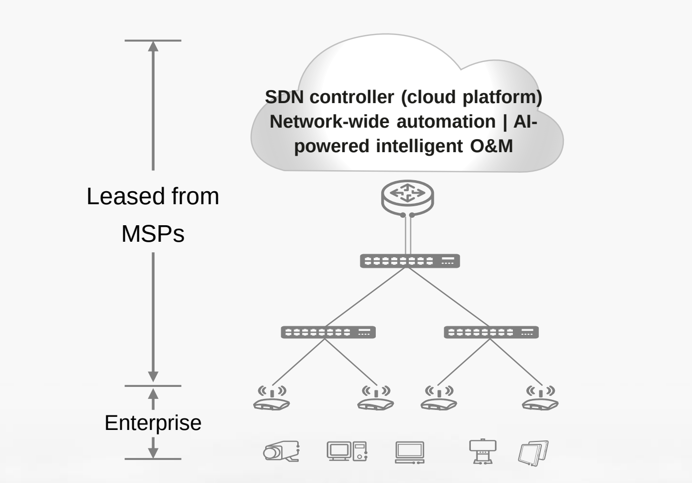
# Device Procurement -> NaaS, Turbocharging Enterprise Digital Innovation

- **More enterprises shift from device procurement to service procurement:** In the past, enterprises built campus networks through one-off procurement of hardware, software, licenses, and services, which are sometimes packaged together. Network operation management also required IT teams to have professional data communication planning, deployment, O&M, management, and optimization capabilities for stable and secure campus network running. Typically, traditional networks are less flexible due to the limitation of the purchased hardware and infrastructure. With the wide application of digital services, network complexity increases and enterprises have growing demands for flexible innovation. Based on the cloud management mode, highly-flexible network as a service (NaaS) comes into being, which allows users to customize network configurations and select specific services as required. It stands out for agile operations, service customization, and flexible charging modes to support complex networks and multi-cloud environments.
- **Benefits of NaaS:** First, enterprises can operate and control the network without purchasing, owning, or maintaining the network infrastructure. They can scale up or down the network as needed, quickly deploy services, and reduce or eliminate hardware-related costs. Second, enterprises can flexibly select service modes in the rapidly changing service environment through flexible business modes such as subscription by period and charging by usage. In addition, NaaS supports real-time software update and network security hardening, through which enterprises can speed up business innovation and reduce risks caused by security vulnerabilities.
- **Market trends**: The NaaS mode has been growing rapidly in recent years. According to Mordor Intelligence, the compound annual growth rate (CAGR) of the NaaS market will reach 34.5% from 2023 to 2027. Moreover, it is estimated that the revenue of the global campus NaaS market will exceed US$600 million by 2027.

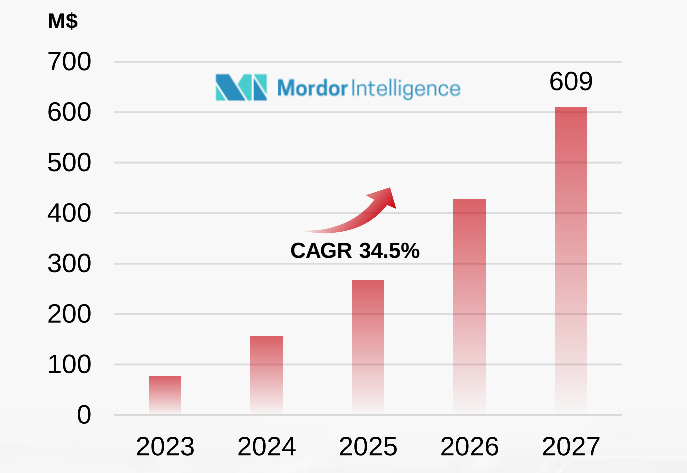
## Traditional mode: Enterprises purchase devices


Enterprise
self-procurement
SDN controller (cloud platform)
Network-wide automation | AI-powered intelligent O&M

## NaaS mode: Enterprises purchase services from MSPs


Leased from
MSPs
Enterprise
SDN controller (cloud platform)
Network-wide automation | AI-powered intelligent O&M

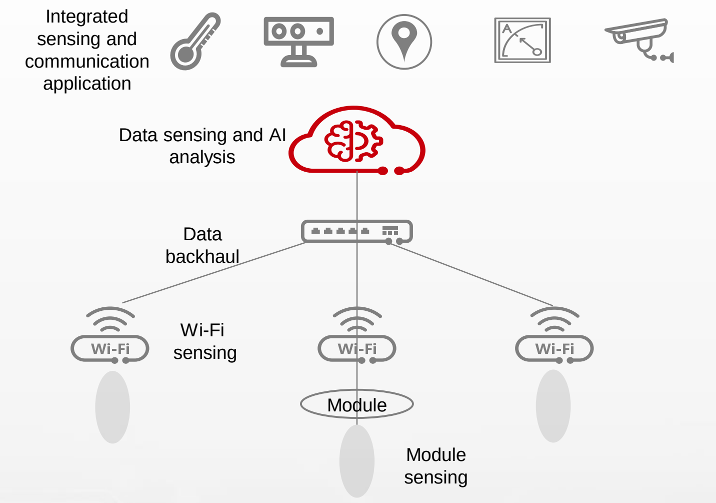
## Rapid growth of the campus NaaS market



# Network Communication -> Integrated Sensing and Communication, Building Networks Featuring All-Round Sensing

- **Wi-Fi devices develop towards intelligent sensing:** Wi-Fi can be used not only for communication, but also for sensing. Wi-Fi sensing uses Wi-Fi waves for motion and presence detection, and then applies machine learning algorithms to promote advanced applications. Wireless devices can be converted into sensors to perform high-precision body positioning and action recognition in a wireless manner, convert the recognition result into instructions, and backhaul them to the control system in real time. The control system then calculates the interference and reflection of signals in the physical space of people and objects, and collects data about them. Wi-Fi devices will participate in the network interaction for determining the locations of people and objects in a specific area.
- **Progress in Wi-Fi sensing standards:** In September 2020, Task Group IEEE 802.11bf was formed for standard research of integrated sensing and communication, which was used for sensing instead of data communication. Standards draft 1.0 and 2.0 were released in January and July of 2023, respectively. In addition, draft 3.0 is planned to be released in November 2023 and draft 4.0 as well as initial products and solutions will be released in January 2024. Besides, formal standards will be released in 2025.
- **Expanded application scenarios:** Integrated sensing and communication improves the performance and efficiency of wireless systems. This feature has the potential to be applied to more scenarios, such as high-precision physiological fall detection for health monitoring, and presence detection for energy saving and carbon reduction.

## Sensing-oriented Wi-Fi networks



Integrated sensing and communication application
Data sensing and AI analysis
Data backhaul
Wi-Fi
Wi-Fi sensing
Wi-Fi
Module
Wi-Fi
Module sensing

**Millimeter wave**

## Innovation in extensive application scenarios

**Low-altitude security**

**Smart transportation**

**Health monitoring**

**Navigation/Tracking**

**Intelligent electric power**

**Gesture sensing**

...

- Millimeter wave
- Home security
- Audio tracking
- Storage sensing
- Home control
- Gesture recognition
- Bio sensing
- Target recognition
- Remote situation detection
- Fall detection
- Remote diagnosis and treatment
- Sneeze sensing
- In-vehicle sensing

# Recommendations: Rapid Digital Application Promotion, Changing the Concept of Campus Network Construction

**Ultra-broadband access and IoT convergence**

Legacy Wi-Fi 5 APs of enterprises are facing the risk of aging and out-of-warranty. Therefore, device upgrade has become urgent. WLAN is also necessary in new office and production areas or wireless reconstruction scenarios, such as AOI for HD quality inspection, in-vehicle platform software installation and OTA upgrade, and AGV-based smart warehousing. Wi-Fi 7 is recommended in WLAN upgrade or construction scenarios as it features multi-fold bandwidth, low latency, and higher reliability. In IoT co-site scenarios, IoT converged APs are recommended to reduce comprehensive network construction costs for enterprises.

**Multi-purpose network and user experience assurance**

Multi-GE switches are recommended to be used as campus access switches to support high-performance Wi-Fi 6/7 bandwidth backhaul and provide GE or higher access services for wired terminals. High-density 25GE switches are recommended for campus aggregation and 100GE switches are recommended for campus core. In this way, a fully-wireless office network with 10GE access, 25GE aggregation, and 100GE core can be built to provide users with 10GE ultra-high-speed experience. The multi-service converged bearing solution is recommended for campuses with multiple services. This solution provides differentiated policies for multiple services on one physical network, which can ensure user experience, improve campus network resource utilization, and reduce network deployment costs.

**Cloud-based management and intelligent O&M**

The SDN controller is recommended to centrally manage and control wired and wireless networks, as well as automatically provision services. This enables more efficient network planning and deployment by IT O&M personnel. In addition, Telemetry can be used to gain real-time visibility into networks, devices, users, and applications. Specifically, this technology facilitates fast fault locating and demarcation, intelligent root cause analysis, and troubleshooting. In this way, it simplifies routine management, O&M, and troubleshooting of campus networks, as well as improving user satisfaction on campus networks.

- In large-sized network scenarios (with multiple branches) where enterprises have independent IT O&M teams, they are recommended to build their own cloud management platforms (controllers) for routine network O&M and management.
- In medium-sized network scenarios where the investment on enterprise networks is limited and there are no O&M capabilities, the NaaS mode is recommended for network construction and O&M hosting, reducing initial network construction costs and investment risks.

# Contents

01 Trend 1: Multi-Cloud Has Become the New Normal, and the Innovation of Elastic, Reliable, and Visualized Networks Is Accelerating

02 Trend 2: Boosting Foundation Models, Propelling Fundamental Transformation of DCNs

03 Trend 3: The Sweeping Digital Transformation Propels Campus Networks into an Experience-Centric Future

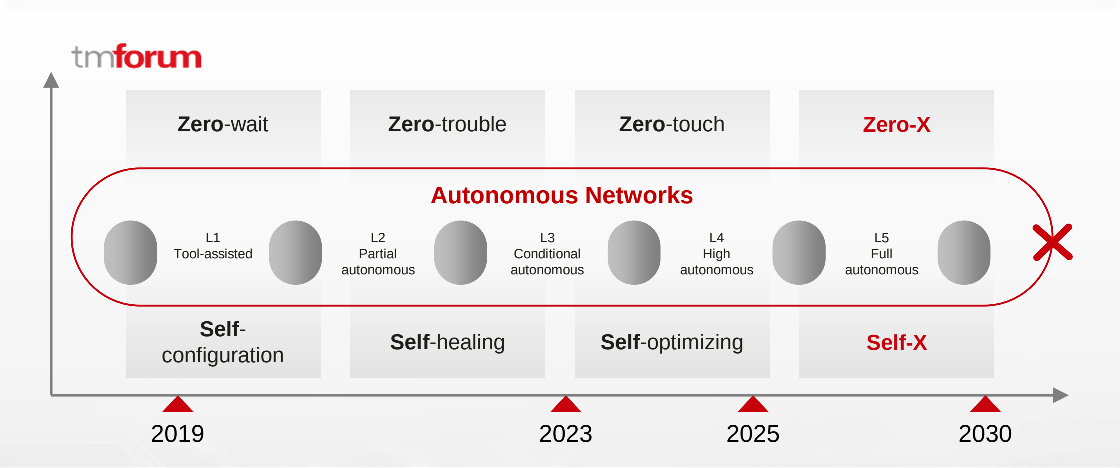
**04 Trend 4: AI Development's Watershed Moment from the Point to System Level**

05 Trend 5: Integrated, XaaS, and Intelligent Become New Features of Network Security Construction

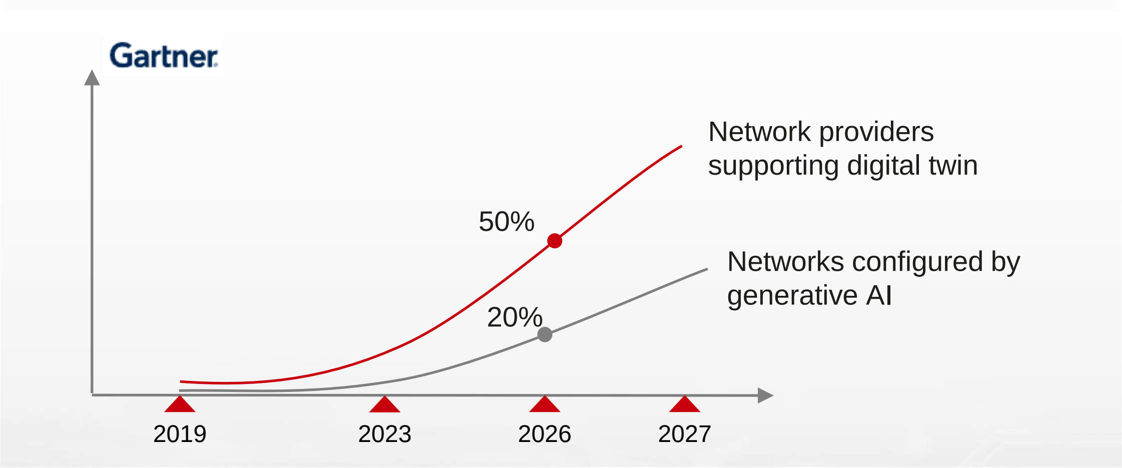
# The Rapid Development of AI Marks a Watershed Moment for the Large-Scale Deployment of Network Intelligence

- **Definition and development of network intelligence:** Network intelligence refers to quickly detecting and isolating problems through real-time data collection, prediction, and association, eliminating the need for network personnel to learn advanced network configuration and troubleshooting skills. AI plays an increasingly important role in coping with the increasing network complexity and demonstrates tremendous potential in transforming traditional network O&M, significantly improving productivity. In 2019, TM Forum proposed the concept of Autonomous Networks (AN), defined five AN levels (L1 to L5), and set the goal of achieving "full Autonomous Networks." Most communications networks are currently somewhere between L2 and L3.
- **Challenges to network intelligence:** Valid data is a crucial ingredient to network intelligence. Traditional NEs do not have advanced analysis capabilities, making it difficult for O&M engineers to identify valuable suggestions from a large number of logs and alarms. And even experienced network engineers could not provide accurate intents as the input. Even if reliable network suggestions are offered, O&M engineers have doubts about implementing them due to a lack of methods for visualizing the network in a comprehensive manner. Simply put, O&M engineers do not fully trust AI before they see the actual data and results. So far, most AI network applications are developed to solve specific problems and are difficult to deploy at the system level. The application rate of AI on the entire network is less than 10%.
- **Rapid development of network intelligence**: In recent years, a rapid increase in computing power has led to the birth of the digital twin and the large-scale application of various foundation models, significantly improving data validity and the effect of visualizing the impact on services. This will facilitate the rapid deployment of network intelligence. According to Gartner, by 2026, 50% of network providers will incorporate the digital twin into their solutions, and 20% of initial network configurations will be completed by generative AI; by 2027, the proportion of enterprises that implement AI automated network O&M will increase from 10% to 90%. The rapid development of AI will promote the large-scale deployment of network intelligence.

## Definition and timeline of L5 AN

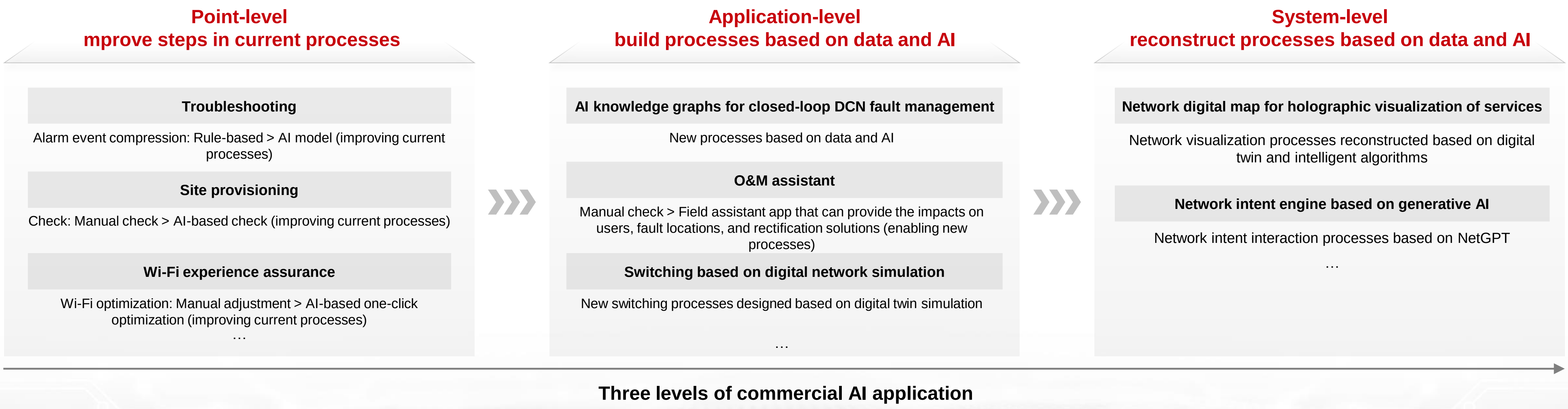


## AI network innovation trends



# System-Level Solutions for Datacom Network Intelligence Relies on Digital Twin and Generative AI

- **System-level network intelligence is key to large-scale deployment:** AI is unable to be deployed overnight, no matter which industry. Instead, it is usually deployed at three levels. The first level is point level, where AI is used to solve specific problems and improve current processes. This type of AI can be deployed independently without changing the system. Examples include alarm compression, Wi-Fi experience assurance, and automatic site provisioning. The second level is application level, where is used to solve many different problems and enable new processes that can be deployed independently without changing the system. Examples include automatic problem identification, locating, and solving through knowledge graphs on DCNs. This type of AI can adapt to some scenarios but cannot reconstruct the entire system. The third level is system level, where AI can improve multiple processes at the same time or enable new processes by transforming interdependent processes. In the datacom network domain, digital twin and generative AI are cutting-edge technologies that can reconstruct various processes. These technologies have been deployed in system-level solutions to support the large-scale deployment of network intelligence.
- **System-level solution 1:** digital twin. The network digital map based on digital twin technology was first deployed on campus DCNs in 2018 to visualize the network and rectify some of the network faults. Now it has been deployed on a large scale to leverage the network digital twin as a basic O&M platform, lowering the trial-and-error cost, accelerating innovation and iteration, and making network O&M more intelligent.
- **System-level solution 2:** generative AI. As ChatGPT made generative AI more popular at the end of 2022, network models started to develop rapidly. Generative AI will be further applied on networks to create detailed configurations and troubleshooting procedures based on manual inputs without any templates. The objective is to provide the key capability of converting service intents into network requirements and develop a service intent engine.



**Point-level**
**mprove steps in current processes**

**Troubleshooting**

Alarm event compression: Rule-based > AI model (improving current processes)

**Site provisioning**

Check: Manual check > AI-based check (improving current processes)

**Wi-Fi experience assurance**

Wi-Fi optimization: Manual adjustment > AI-based one-click optimization (improving current processes)

…

**Application-level**
**build processes based on data and AI**

**AI knowledge graphs for closed-loop DCN fault management**

New processes based on data and AI

**O&M assistant**

Manual check > Field assistant app that can provide the impacts on users, fault locations, and rectification solutions (enabling new processes)

**Switching based on digital network simulation**

New switching processes designed based on digital twin simulation

…

**System-level**
**reconstruct processes based on data and AI**

**Network digital map for holographic visualization of services**

Network visualization processes reconstructed based on digital twin and intelligent algorithms

**Network intent engine based on generative AI**

Network intent interaction processes based on NetGPT

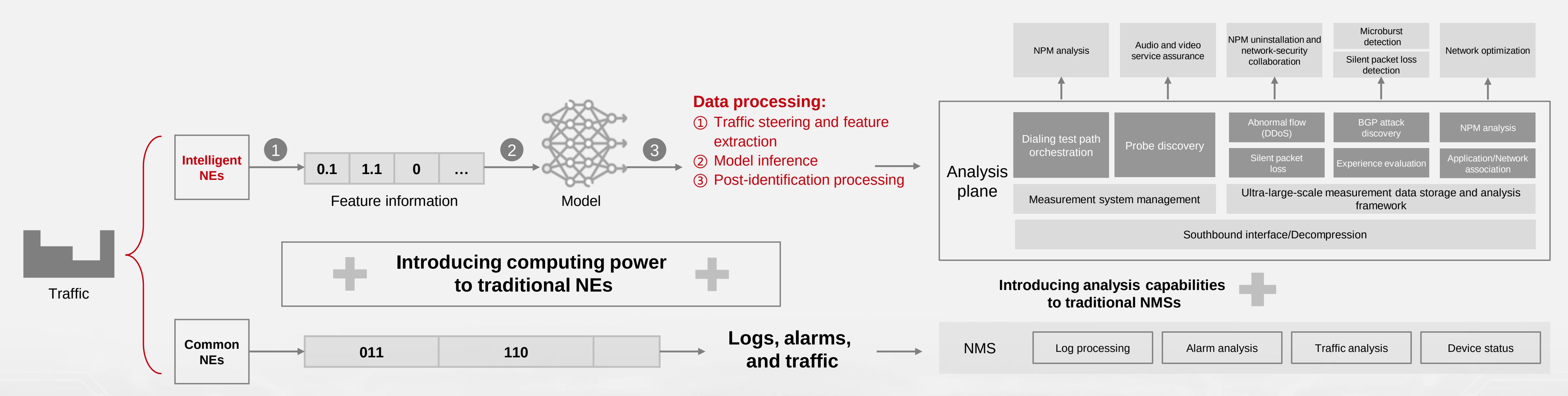
…

**Three levels of commercial AI application**

# From Receiving to Generating Intents, Intelligent NEs Lay a Solid Foundation for Network Intelligence

- **Intelligent NEs lays a solid foundation for network intelligence.** Network data, which is mainly generated by NEs, is the key to network intelligence. If network data contains only logs and alarms, it is difficult for O&M personnel to generate accurate network intents in most cases. The lack of accurate network intents results in passive network O&M. An intelligent network, which has become mainstream, is impossible without intelligent NEs. On an intelligent network, NEs will transform from receiving intents to generating intents, perform self-analysis, and collect, preprocess, and report multidimensional data. Based on traffic changes, the network generates intents that humans alone cannot provide, offering a solid data foundation for AI.

- **Crucial benefits of NE intelligence:** Intelligent NEs can use deep learning models to classify data flows based on packet traffic behavior features and take actions on the flows based on the inference result, improving key device capabilities, including service assurance and security detection. Take a financial backbone network as an example. Abnormal application traffic, which may affect high-value services and cause service faults, can be detected and reported by NEs alone. As such, network devices are upgraded to intelligent ones to identify abnormal traffic within seconds and perform closed-loop traffic limiting and steering, ensuring the SLA performance of high-value services. In terms of DDoS attack defense, the second-level DDoS attack detection mechanism of routers depends only on the change of the flow rate, and is unable to identify the types of attacks or accurately report them, including multi-to-one switching caused by link faults, whereas the device protocols are vulnerable to attacks. To address these issues, intelligent devices are used to identify the types of attacks from pass-by traffic, report attacks within seconds, and clean device traffic quickly for closed-loop management.



Traffic

Intelligent NEs

1

0.1 | 1.1 | 0 | ...

Feature information

2

Model

3

**Data processing:**

① Traffic steering and feature extraction

② Model inference

③ Post-identification processing

Introducing computing power to traditional NEs

Common NEs

011 | 110

Logs, alarms, and traffic

Analysis plane

NPM analysis

Audio and video service assurance

NPM uninstallation and network-security collaboration

Microburst detection

Silent packet loss detection

Network optimization

Dialing test path orchestration

Probe discovery

Abnormal flow (DDoS)

Silent packet loss

BGP attack discovery

Experience evaluation

NPM analysis

Application/Network association

Measurement system management

Ultra-large-scale measurement data storage and analysis framework

Southbound interface/Decompression

Introducing analysis capabilities to traditional NMSs

NMS

Log processing

Alarm analysis

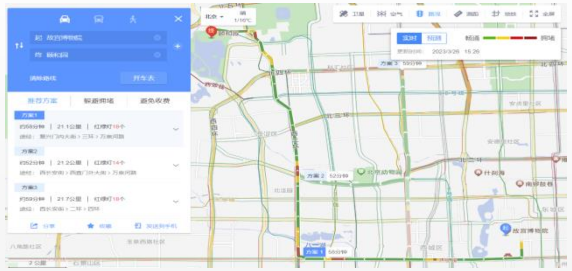
Traffic analysis

Device status

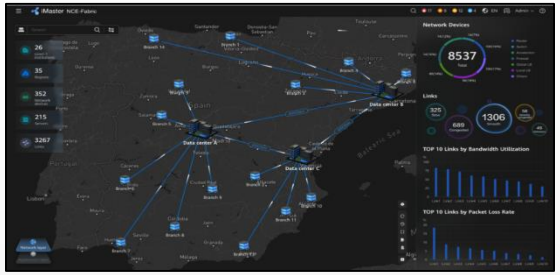
# Large-Scale Using of Network Digital Map, Extending from Multidimensional Visualization to Optimization Simulation

- **Full-scale application of the network digital map:** As the digital mirrors of physical network infrastructures, digital twins maintain almost the same topologies, services, and traffic data models as the physical network does. As refined full-lifecycle, multidimensional duplicates of the physical network, the digital twins provide a digital verification environment for network O&M. Unlike traditional simulation technologies, the network digital map is not a static snapshot of the physical network; rather, it is updated in real time based on the physical network status. With the help of AI technologies and self-learning, the network digital map will evolve by itself based on online pre-verification feedback and provide higher authenticity and reliability.
- **Technical trend 1: From one-dimensional to multidimensional visualization with effective service-network association.** Traditional network management systems (NMSs) visualize network quality from a single dimension only. However, in network O&M scenarios, it's quite often the case that network faults cannot be located although service faults did occur, or the network cannot quickly prove its own innocence. For example, a bank was struggling with low O&M efficiency because there was no visualized association between its hundreds of applications and tens of thousands of NEs. As such, achieving visualization at the network layer alone is far from enough — the network digital twin engine also needs to build a digital duplicate of the unified physical network model. In terms of O&M, multi-domain and multidimensional simulation and verification must be implemented to support network planning, construction, maintenance, and optimization activities, including displaying intra-/inter-application access relationships and detecting service exceptions in seconds.
- **Technical trend 2: From offline simulation to real-time simulation, reducing risks associated with network changes.** By scenario, network simulation technologies fall into two categories. The first category is offline simulation, which was prevalent in network planning in the past. However, as the network automation level continues to increase, the closure periods of network faults, self-service application, and other services are shortened (to minutes or even seconds as expected by customers). This drives the emergence and prevalence of the second category — real-time simulation. Based on NE configuration data, real-time simulation technology simulates the control plane and forwarding plane behaviors of devices' routing protocols and accurately generates NE protocol-specific routing tables and the global routing table. By analyzing entries in these routing tables, real-time simulation technology is able to learn and verify impact on the network.

## Digital Transit Map and Navigation



## Network Digital Map

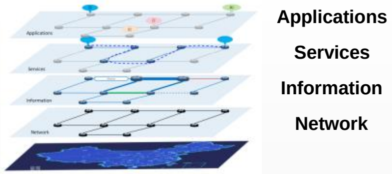


**Practice by Orange Spain**

- Holographic network visualization: Network-wide SLAs are visualized.
- Optimal network performance: Network latency is reduced by 30% (tested by P3).
- Autonomous network optimization: The path optimization time is shortened from 3 months to 3 minutes.

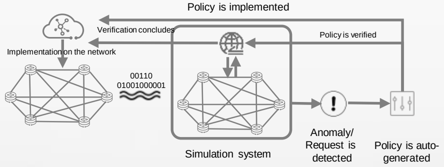
**Multidimensional Visualization**

- One map visualizing the entire network
- Real-time topology restoration
- Mutual visibility between applications and networks



**Real-Time Network Simulation**

- One-click path navigation
- Simulation of multi-cloud changes
- Holographic proactive awareness



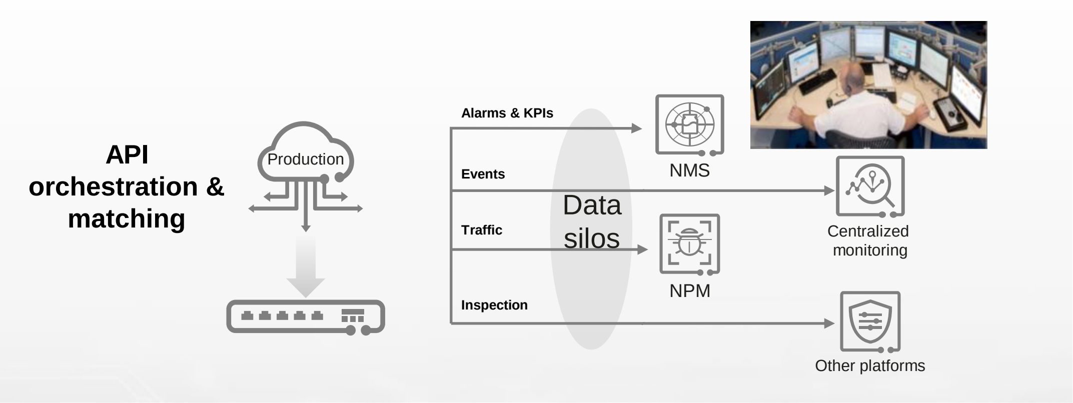
# Generative AI Creates Opportunities for NetGPT

- **Generative AI drives the development of NetGPT.** The personalized development of hundreds of models and thousands of modalities creates unparalleled opportunities for the emergence of large communications network models. Large models at the application layer must be connected to networks so that they can reach end users, but communications networks feature a "thin waist of the hourglass" and therefore expect one unified large communications network model (NetGPT) to process different network services at lower service complexity. This is the trend. Currently, NetGPT is still in its infancy, and its application scenarios are still under preliminary exploration — several known scenarios are knowledge management platform enablement and interactive network intent engine processing. Looking ahead, NetGPT may be deployed in a cloud-edge collaboration manner with the aim to effectively orchestrate heterogeneous distributed communications and computing resources, and this is a key step for NetGPT to leverage its full potential.

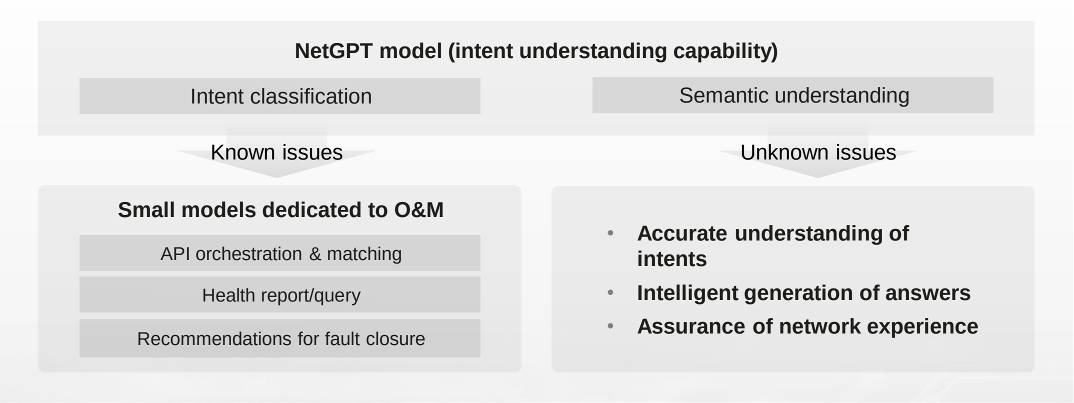
- **Scenario 1: Knowledge management platforms significantly increase the accuracy of interactive Q&A.** Traditionally, AI Q&A assistants are built on knowledge graph technology. If the keywords of a user's question are inaccurate, assistants tend to give inaccurate answers. To address this issue, the NetGPT model has been constructed using abundant datacom network corpus, and is retrained for knowledge expansion and adjusted based on tasks. This enables the model to accurately understand user intents, improve interaction efficiency, and effectively collaborate with dedicated small models in intelligent Q&A and network experience assurance scenarios. The accuracy of feature-related answers increases from 20% to 80%.

- **Scenario 2: The manual intent engine fully understands network intents.** Although we had NMSs in the past, there were too many network O&M tools and GUIs. When handling issues, we did not know where to start. Moreover, application/service flow data was not associated with network data, and only experienced experts could identify the association between them. Due to the ability to accurately identify, distinguish, and understand user intents, the NetGPT model can convert user intents into network requirements. Based on the input of maintenance personnel, it creates detailed configurations and troubleshooting procedures without any fixed templates, improving O&M efficiency and reducing professional dependency for O&M teams.

## Before: Complicated troubleshooting, difficult to get started



## After: Intelligently identifying intents and recommending closed-loop measures



NetGPT model (intent understanding capability)

Intent classification

Semantic understanding

Known issues

Unknown issues

**Small models dedicated to O&M**

API orchestration & matching

Health report/query

Recommendations for fault closure

- **Accurate understanding of intents**
- **Intelligent generation of answers**
- **Assurance of network experience**

# Recommendations: Combine AI Technologies with Human Intelligence to Accelerate Innovation and Improve Network Operations Efficiency

**Deploy intelligent network devices**

Prefer network devices with computing capabilities or even scalable computing capabilities. In terms of network intelligence, network controllers and analyzers alone are insufficient; rather, device intelligence is essential for global network visualization and network autonomy as it can shift from static inference and passive response to dynamic analysis and proactive recommendation.

**Adopt digital twin technology to achieve comprehensive network visualization**

For data communications networks, this is the first step to achieve network intelligence. With the digital twin network platform, network change operations such as adjustment, maintenance, and optimization can be fully tested and verified, and the operation solutions can be continuously evaluated, modified, and optimized based on feedback, with minimal impact on the physical network. In addition, the digital twin network records the states and behaviors of digital twins in real time to support history tracing and playback so that pre-verification can be completed without affecting network operations, significantly lowering the trial-and-error costs.

**Innovate generative AI applications at an accelerating pace**

Keep an eye on the research and application progress of NetGPT. Since generative AI will be widespread across industries at an accelerating pace, we can introduce innovations into intelligent Q&A, network configuration guidance, and O&M interaction by engaging far-sighted vendors and solutions and extending generative AI to local network domains.

**Embrace the power of AI**

Network intelligence has a palpable effect on network availability, performance, and operations efficiency. With the continuous progress of AI technologies, AI applications have demonstrated their value and realized a watershed moment for large-scale deployment. AI itself will not completely replace humans, but will better assist humans. AI networks will substantially improve network availability, optimization efficiency, and performance, enabling humans to do more with the same — if not fewer — resources.

# Contents

01 Trend 1: Multi-Cloud Has Become the New Normal, and the Innovation of Elastic, Reliable, and Visualized Networks Is Accelerating

02 Trend 2: Boosting Foundation Models, Propelling Fundamental Transformation of DCNs

03 Trend 3: The Sweeping Digital Transformation Propels Campus Networks into an Experience-Centric Future

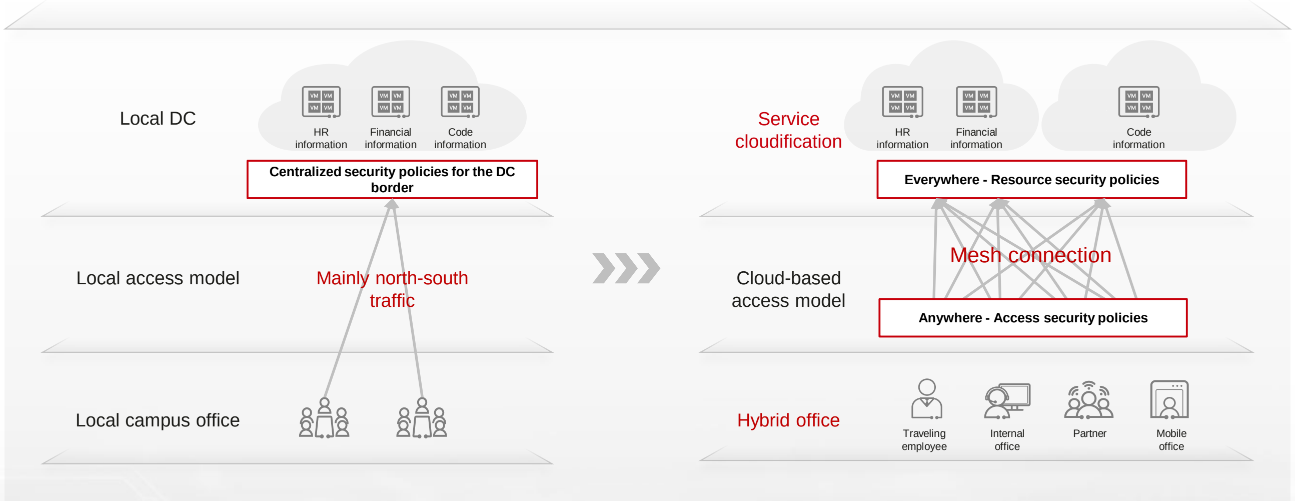
04 Trend 4: AI Development's Watershed Moment from the Point to System Level

**05 Trend 5: Integrated, XaaS, and Intelligent Become New Features of Network Security Construction**

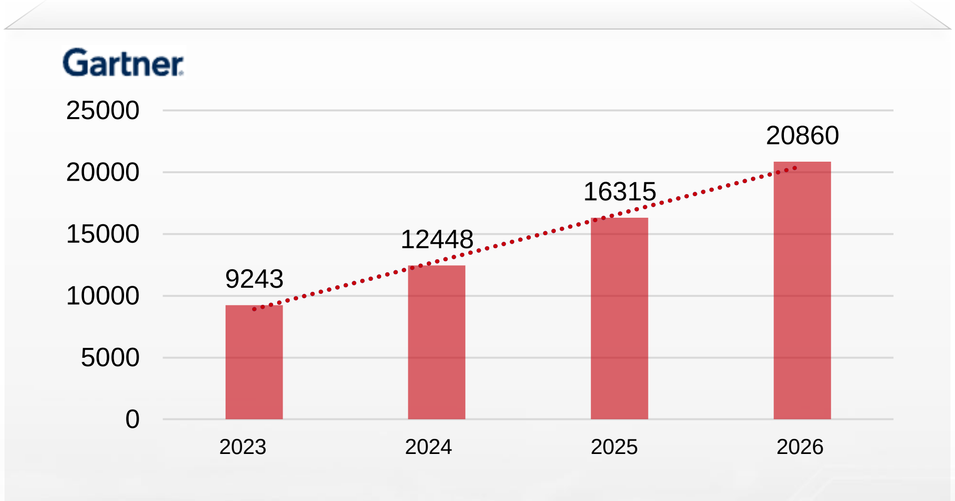
# Cloudification and Hybrid Office Break Security Boundaries, and Network-Security Collaborative Defense Becomes the Mainstream Choice

- **Enterprise cloudification breaks the defense boundary:** The traditional network security architecture focuses on the enterprise intranet and establishes a layered defense system at the enterprise boundary to safeguard data security. With more and more enterprises embracing cloud migration, the traditional closed architecture has evolved into a multi-cloud and multi-branch interconnected system. This has led to the breakdown of enterprise boundaries and increased exposure to network security risks. Moreover, there is a growing need for high reliability and security in production. The original centralized service access and security system have become ineffective and cumbersome. Traditional network security technologies are unable to address the increasingly sophisticated threats and vulnerabilities faced by the network periphery. As external access to the cloud accelerates, enterprises must implement advanced access control measures to ensure they have the capability to handle related network security needs and risks.
- **Hybrid office increases security risks:** Enterprise employees are not confined to fixed workplaces, and hybrid office has become a common practice. This means that employees may access the enterprise network through the insecure Internet anytime and anywhere. The original security architecture based on the enterprise LAN border is no longer effective, which brings new challenges to enterprise data security defense. Enterprises need to consider how to ensure secure access to the enterprise headquarters and multi-cloud platforms anytime and anywhere.
- **Network-security convergence represented by SASE has become a trend:** To address these changes, the adoption of zero-trust-based secure access service edge (SASE) has emerged as a popular trend, ushering in a new era of service evolution for branch networks and security convergence. SASE offers a range of converged network and security-as-a-service functions from a single cloud delivery platform, including zero-trust network access, cloud access security proxy, secure web gateway, firewall, and SD-WAN. This enables secure and seamless connections for any application across any network, location, or device. Gartner predicts that by 2026, 80% of enterprises will implement SASE solutions for architecture and networking reconstruction, with a market size of US$21 billion.

## Impact of enterprise cloudification and hybrid office on the network security architecture



## SASE end user expenditure forecast (US$)

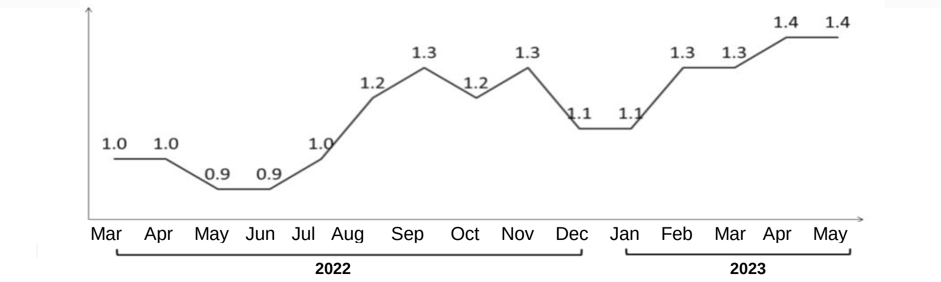


# High Network Security Construction Costs and Shortage of Professionals, Making the Shift Towards Cloud-Based Network Security an Inevitable Trend

- **High cost of network security construction:** Currently, the majority of enterprises invest their information security resources only after experiencing information security incidents. This reactive approach is driven by specific events and projects. When faced with related issues, enterprises tend to purchase corresponding devices without considering the overall system and design. Consequently, multiple types of security devices need to be deployed to address specific security incidents. This leads to excessive repeated construction and high costs in device procurement.
- **Lack of professional network security capabilities:** The rising complexity of network attacks presents significant technical challenges for enterprise IT personnel. They must possess a deep understanding of attack vectors, be proficient in defense strategies, and possess advanced skills in security data analysis and execution. However, for small and midsize enterprises or organizations, the cost of hiring professional security talent is often prohibitively high. Consequently, many of these organizations lack effective security management and are ill-equipped to handle security incidents. Furthermore, security threats can occur unpredictably, and even with security devices in place, O&M personnel may struggle to detect and respond to attacks due to a lack of professional security management processes and warning mechanisms for security incidents.
- **Network security moves towards cloud services:** For most enterprises, it is unrealistic to establish a comprehensive security service team that can operate under all circumstances if there is a shortage of time, capital, talent, and process. Building a modern security operation center to continuously defend against network threats is not feasible for every enterprise. However, the emergence of security cloud services has introduced a new approach to network security. By leveraging cloud technology, these services offer continuous and evolving security protection capabilities, along with convenient one-stop solutions for enterprises. The advantages of security cloud services include low technical requirements, cost-effectiveness, and the elimination of the need for specialized network security personnel. This addresses the security weaknesses commonly faced by small and midsize enterprises. Furthermore, through cloud service collaboration and continuous updates, these services can help provide global immunity upon any threat detected on one network node, significantly enhancing network security protection capabilities. A survey indicates that 85% of small enterprises are willing to adopt cloud management and services, particularly for building secure ICT infrastructure to support digital transformation efforts.

## China's network security talent demand will increase by 40% in 2023

**Monthly Change of Network Security Talent Demand Index (2022.3-2023.5)**



2023 Research Report on the Situation of the Network Security Talent Market

## From traditional onsite services to cloud-based network security service architecture

Security products + Personnel on-site

The L1 personnel on site lack the necessary skills, resulting in low efficiency when collaborating with the backend L2 personnel. Service delivery is currently being carried out manually.

- 5x8 security O&M and analysis of some logs
- Proactive discovery of some security issues
- 24/7 reactive emergency response, based on remote guidance

»»

Security products + Cloud services

Professional security experts and service personnel with strong skills and hands-on experience, implementing efficient service delivery with the help of the cloud platform

- 24/7 security O&M and full log analysis
- Proactive discovery of all security issues based on traffic
- 24/7 proactive emergency response and threat intelligence triggering

# Ransomware Attacks Become the Norm, Making It Crucial to Establish a Comprehensive Defense System to Create a Robust Line of Defense

- **Ransomware attacks become the norm:** In recent years, ransomware attacks have emerged one after another and have severely affected key sectors such as public service, finance, education, healthcare, manufacturing, and energy around the world. In some incidents, attackers hijack critical infrastructure to claim high ransom, which may even affect the normal operation of a country. So far, the average service interruption caused by ransomware has reached 16 days. An organization is attacked by ransomware every 11 seconds, and the largest ransom is up to US$70 million. Large companies with large and complex digital infrastructure have become one of the main targets of ransomware cyber criminals. According to an IDC report, 35% of global organizations have experienced three to four ransomware incidents. A successful ransomware attack requires an average ransom of about US$150,000, causing service interruption for five days on average.
- **The protection system moves towards defense-in-depth.** Ransomware attack tactics and variants are evolving. Traditional data backup, network border protection devices, and traditional antivirus software that relies on signature detection have become invalid. The number of ransomware variants increased exponentially, from 5400 in 2021 H2 to 10666 in 2022 H1, an increase of 98%. The encryption speed and permission theft speed of ransomware are very fast. The time window for administrators to handle ransomware is very short. The fastest time for ransomware to infiltrate the system to obtain permissions is 45 minutes, while the average encryption speed of 100,000 files is only 43 minutes. In addition, latest-generation ransomware attacks target backup systems, devices, and VMs. As a result, more than 46% of organizations that pay ransom after being attacked cannot fully restore data. New service changes and frequent new threats make security protection more professional and complex. Also, more intelligent security protection methods are required. Deploying multiple types of security products has been evolved to trusted infrastructure networks. The construction of an in-depth defense system has become a hot topic for enterprise investment.

## Continuously evolving of ransomware threats: rapid rise of variants, long service interruption, and frequent attacks

**57x**

In 2021, ransomware causes a loss of US$20 billion, which is 57 times the loss in 2015. The loss is predicted to be US$265 billion in 2031.

Source from: Cybersecurity Ventures

**16 days**

Ransomware attacks result in an average of 16 business days of system shutdowns.

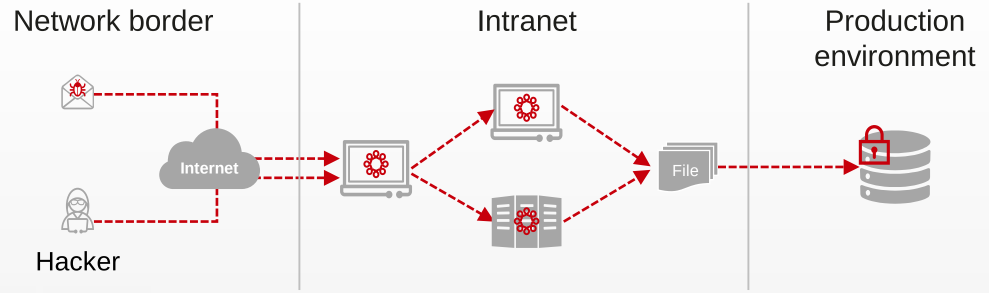
Source from: ZDNet

**11s**

In 2021, a ransomware attack took place every 11 seconds. By 2031, the interval is predicted to be just 2 seconds.

Source from: Cybersecurity Ventures

## Building an in-depth defense system for enterprises from the perspective of ransomware attacks



Network border

Intranet

Production environment

Internet

Hacker

File

**Line of defense 1**
**Network border intrusion prevention**

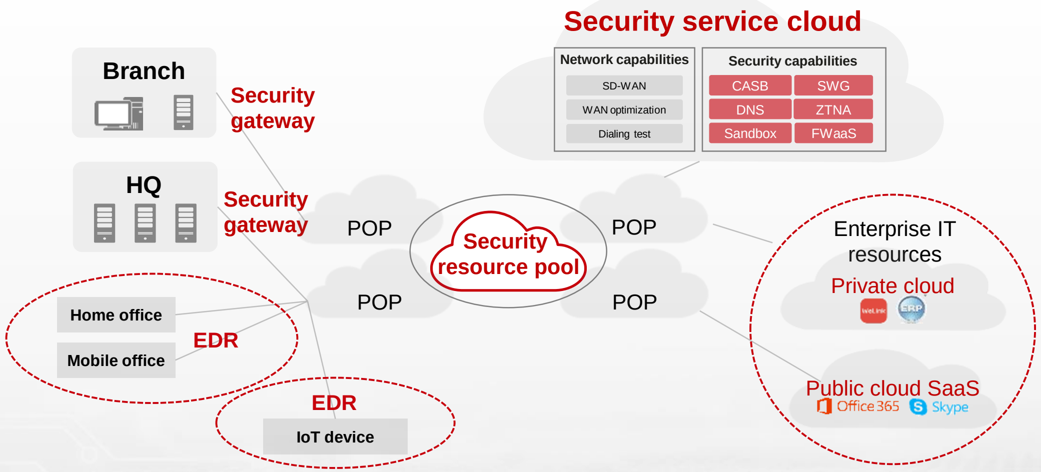
**Line of defense 2**
**Non-proliferation within the network**

**Line of defense 3**
**Anti-encryption in the production environment**

# From Network-Security Separation to Network-Security Integration, the Converged Architecture Comprehensively Improves the Overall Security Posture of Enterprises

- **Converged network-security architecture:** For enterprises, the primary network requirements include branch Internet access, interconnectivity between branches and the headquarters, and SaaS service access from branches. With the increasing cloud migration of enterprises and the growing use of mobile office by employees, a significant number of users, devices, applications, and data are now located outside enterprise DCs and networks. To address this, a converged network-security architecture deploys network and security protection capabilities on corresponding network nodes, implements flexible and distributed overlay networks through software definition, applies security protection capabilities to nearby entities, and provides unified policies and security posture awareness through collaboration with the operations brain. This architecture meets the network security interconnection requirements in various enterprise scenarios.
- **Advantages of network-security convergence technologies:** Compared to traditional network security architecture, network-security convergence represented by SASE has four key features: zero trust access, cloud-native architecture, support for all edges, and global distribution. These features enable SASE to better meet the increasing demands of enterprises for cloud application services and cloud network security products. Furthermore, these features reflect the integrated deployment of network and security. SASE offers comprehensive network and security services, establishing flexible overlay networking capabilities for branches to access the Internet, headquarters, and the cloud. It also establishes an end-to-end management mechanism and security protection measures based on overlay connections, eliminating physical network restrictions and simplifying complex networks. With its identity-centric approach, SASE provides ubiquitous defense capabilities. The centralized operation service simplifies policy management and security incident handling, providing customers with a simple, efficient, secure, and stable network access and service deployment experience.

## SASE-based converged network-security architecture and user benefits



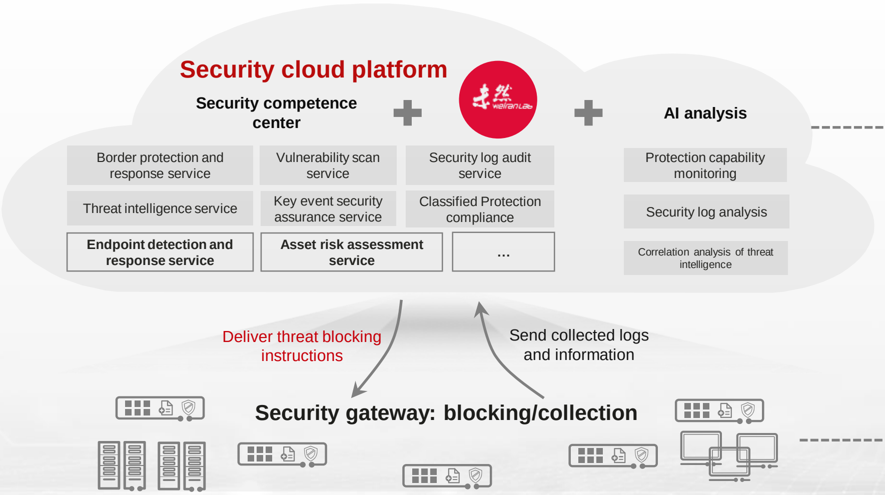
- **Reduced complexity and costs:** A single service provider reduces the number of physical or virtual devices at the branch border and the number of agents.
- **Enhanced performance/latency:** The SASE supplier provides POPs around the world to optimize the access latency and route selection.
- **Low OPEX:** Enterprises are no longer affected by hardware capacity expansion and EOL device updates. In addition, enterprises can quickly defend against new threats without paying attention to signature database updates.
- **Zero trust:** Multiple threat signals and context signals are used to ensure secure access to internal resources and the Internet.
- **Improved efficiency of network and network security personnel:** Build enterprise security strategies on a single platform.

# Upgrading Professional Security Capabilities: From Local Defense to Collaboration with the Cloud

**Cloud service-based network security system**: This architecture consists of the security cloud service platform and security gateway protection nodes on the customer's local network. It implements collaboration between cloud services and local devices to build a simple, efficient, and easy-to-use security cloud service solution. The division of responsibilities between the cloud, edge, and endpoint is clear. The security gateway leverages the advantages of local real-time protection and only migrates security log information and attack forensic data to the cloud. The security cloud platform leverages the advantages of computing power and threat intelligence to implement correlation analysis and comprehensive detection.

- **24/7 dynamic change protection:** Security detection devices are deployed at the Internet egress to detect traffic on-premises in real time and update threat intelligence on the cloud in real time. Expert models together with AI algorithms are used to implement intelligent aggregation and analysis of massive logs. Dynamic security protection capabilities are used to cope with dynamic security threats. Automatic threat analysis and handling capabilities are used to implement second-level threat identification.
- **Automatic real-time threat blocking:** After detecting external attack sources through cloud analysis, the cloud platform automatically delivers security policies and interworks with local hardware devices to block external attacks in minutes. Automatic threat analysis and handling capabilities help cope with professional and complex security analysis and manual handling. Enterprises can invest in analysis and utilize security capabilities at zero labor costs.
- **On-demand subscription of security services:** There are multiple types of cloud security services available, and additional services can be subscribed to in the future. New security service capabilities, such as cloud-based vulnerability scanning and log auditing, are regularly updated to continuously improve security protection capabilities. Enterprises have the flexibility to subscribe to and utilize security services based on their specific needs, thereby avoiding unnecessary investment in network security at the initial stage.

## Huawei Qiankun network security cloud service architecture



Security cloud platform
Security competence center
WeiRan Lab
AI analysis
Border protection and response service
Vulnerability scan service
Security log audit service
Threat intelligence service
Key event security assurance service
Classified Protection compliance
Endpoint detection and response service
Asset risk assessment service
...
Protection capability monitoring
Security log analysis
Correlation analysis of threat intelligence
Deliver threat blocking instructions
Send collected logs and information
Security gateway: blocking/collection

### Security cloud platform: Security competence center + AI-based analysis + Cloud expert service

- **Security competence center:** Provides platform capabilities and plans to support security capabilities such as endpoint security protection and asset risk assessment.
- **AI-based analysis:** Correlates and analyzes security logs and forensic files based on AI technologies to handle attacks rapidly and accurately.
- **Cloud expert service:** WeiRan Lab security experts experienced in attack-defense confrontation provide 24/7 online services.

### Security gateway: collection + blocking

- **Collection:** Collects and sends security logs to Huawei Qiankun cloud platform through encrypted channels to provide data for AI analysis.
- **Blocking:** Performs in-depth security detection on network traffic to discover and block attack traffic and malicious files, receives blacklist information from the cloud, and blocks IP addresses.

# Building an Intelligent Full-Process Ransomware Protection System: From Single-Point Defense to Multilayer Collaboration

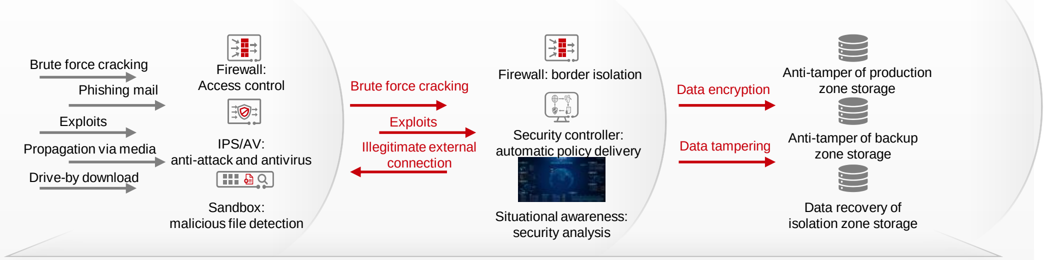
- **Full-process defense policy:** Based on the process of ransomware attacks, ransomware defense must evolve from traditional defense policies to full-process defense policies. During the pre-event network border intrusion prevention phase, implementing border isolation, anti-attack measures, antivirus software deployment, and malicious file detection can prevent up to 70% of attacks. In the in-event lateral movement phase, border isolation and automatic security analysis and handling can prevent an additional 20% of attacks. Finally, during the post-event ransomware encryption phase, the remaining 10% of attacks can be protected against through storage backup and isolation zone construction. However, the costs and effectiveness of protection during these three phases vary greatly. The network layer provides the most effective protection at the lowest cost. Implementing anti-intrusion and anti-proliferation measures at the network layer can guard against over 90% of ransomware attacks, which can effectively protect the network from hackers.
- **Multilayer network-storage collaboration:** The combination of storage and network infrastructure, along with multilayer and end-to-end effective protection, offers the strongest defense against ransomware. By implementing multilayer detection and collaborative data protection for network and storage devices, the defense against ransomware attacks shifts from reactive response to proactive defense. This approach includes effective prevention before attacks, accurate detection and response during attacks, and quick recovery after attacks. It enables users to promptly detect and intercept ransomware attacks, safeguard data from unauthorized encryption and theft, and efficiently restore data when necessary. Ultimately, it helps establish a comprehensive ransomware protection system that covers the entire process before, during, and after an event. This defense strategy is characterized by precise attack identification, comprehensive threat prevention, and rapid data recovery.

## Mainly border defense, internal protection as a supplement

**Pre-event: strengthen the border protection capability**

**In-event: network-wide anomaly monitoring and threat isolation**

**After-event: encrypted data storage and multi-copy backup**



**Investment cost: 1A**

**Prevents 70% attacks.**

**Investment cost: 2A**

**Prevents 20% attacks.**

**Investment cost: 3A**

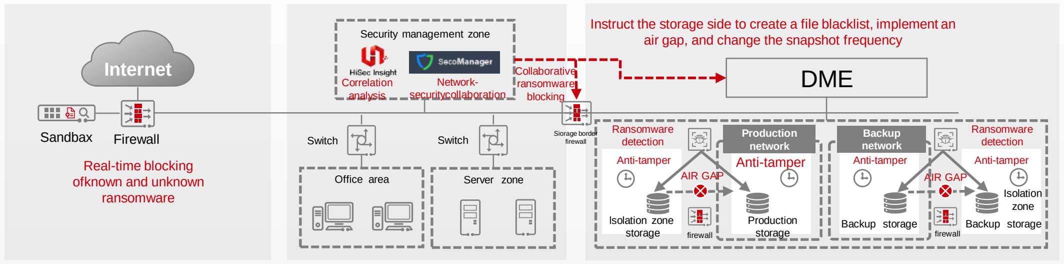
**Prevents 10% attacks.**

## Enterprise in-depth defense system construction

**Pre-event: network border intrusion prevention**
**Precise threat detection**

**In-event: non-proliferation within the network**
**Rapid threat handling**

**Post-event: anti-encryption in the production environment**
**Stable data recovery**



| | Industry | Huawei |
|---|---|---|
| **Ransomware detection rate** | 80% | 90% |
| **Ransomware source tracing and handling time** | Days | Minutes |
| **Data recovery speed** | 30TB/h | 170TB/h |

# Recommendations: Accelerate the Evolution of Security Protection Concepts and Technologies as Network Security Has Become a Core Element

**Put network security at the core of business**

The reactive defense approach of "considering network security after an incident occurs" is no longer effective. In order to achieve a stable digital transformation, we must view network security as a core element in digital business development and obtain strategic support from the enterprise. This entails developing network security plans not only to prevent network attacks, but also to enhance the enterprise's ability to effectively manage digital development risks.

**Use the cloud service-based network security mode to quickly improve security capabilities**

Network security affects the entire system. For enterprises that have no network security or IT professionals or have limited security investment budgets, it is recommended that security cloud services be used to safeguard their digital transformation.

**Embrace integrated network and security innovation and unifying network and security management strategies**

With the rise of connectivity and the widespread adoption of SaaS and cloud applications, the security attack surface of enterprises and organizations is expanding. As a result, they require broader visibility and unified policies to continuously monitor threats and risk exposure. To address this challenge, enterprises and organizations must establish an integrated protection system that can systematically detect, investigate, and respond to threats. This will enable security operation teams to gain a comprehensive understanding of risks and potential impacts.

# Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and organization for a fully connected, intelligent world.

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.

HUAWEI