NETWORK EVOLUTION FOR THE 5.5/6G ERA - NET5.5G ARCHITECTURE

WORLD BROADBAND ASSOCIATION

This white paper presents the technical leadership and recommendations of the World Broadband Association's working group (WBBA WG4) about network evolution for the 5.5G era (Net5.5G). WBBA WG4 suggests adoption of Net5.5G, which delves deeply into the development trends and blueprint architecture for network evolution and recommends the smooth modernization of the end-to-end data communication network to address the critical challenges associated with the demand for ubiquitous 10Gbps access for business, campus, and home; industry digitalization; and AI computing applications.

FOUNDING MEMBERS

MEMBERS

OBSERVERS

CONTENTS

AUTHORS

Sameer Malik
Senior Principal Analyst, Service Provider Routing and Switching Networks, Omdia

ZhongHua Chen
Senior Engineer and Project Manager (China Telecom), WBBA Working Group Chair

Tayeb Ben Meriem
Vice Chair (IPv6 Forum), WBBA Working Group Co-chair

EXECUTIVE SUMMARY AND VISION

Booming developments in ubiquitous artificial intelligence (AI) computing applications and the metaverse raise expectations of service agility for both business and consumer markets. While technological development speeding up, the critical question is whether service providers' existing underlying IP data communications networks can fulfill these expectations and deliver demanding services intelligently and with low latency, ensuring a deterministic experience over geographically distributed locations. The existing fragmented and silo-based traditional IP data communication network is evolving for the era of 5.5G and 6G to a single, end-to-end view that better reflects the trends of cloud-network synergy and convergence of information technology (IT) and operational technology (OT) networks.

IPv6-enhanced technology is considered to be the key enabler and fundamental innovation in network evolution for the 5.5G and 6G era for realizing intelligent network architecture. It offers flexibility in deploying emerging use cases and technologies such as 10G to the site, Wi-Fi 7, end-to-end 400GE, deterministic networking, and an application/computing-aware network with SRv6 (Segment Routing v6) for extensible capabilities. Therefore, cloud-network convergence with 10Gbps access technologies, SRv6 IP transport network, and data center network (DCN) overall single-view management is the key to enabling true business success.

One major shift associated with the advent of network evolution for the 5.5G and 6G era is the absolute necessity of including societal requirements and their related key value indicators, along with new performance requirements and associated KPIs, in the design blueprint for network evolution. These societal requirements are sustainability, inclusiveness, security, well-being, prosperity, affordability, and trustworthiness, and they are mainly associated with the UN's Sustainable Development Goals.

Moreover, regulatory frameworks such as AI acts (responsible AI, accountability, explainability, safety) and data acts (e.g., GDPR, data provenance, data sovereignty) as enacted in different regions, must be captured in this design blueprint for network evolution for the 5.5G and 6G era, to ensure compliance.

This trend in network evolution for the 5.5G and 6G era is a turning point, when standards development organizations/forums should expand their scope beyond their original responsibility demarcation to embrace a holistic, end-to-end view of the entire new ecosystems.

The aim of the World Broadband Assocation's working group (WBBA WG4) is to leverage its technical leadership and present recommendations on network evolution for the 5.5G era (Net5.5G).

VISION

AI and supercomputing power is one of the major driving forces in promoting digital transformation across numerous industries, boosting economic development. In parallel, it is placing new requirements on the underlay IP (datacom) network infrastructures in many verticals and domains, including campus networks and data center networks, to support larger AI training models. Net5.5G recommends the smooth modernization of the end-to-end IP data communication network.

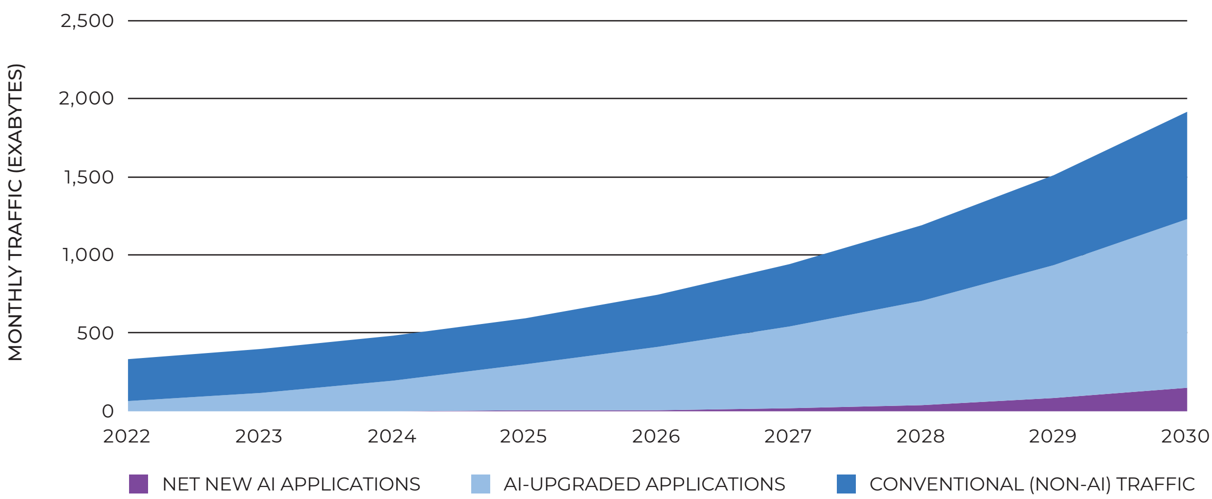
WBBA WG4 suggests adoption of Net5.5G, which delves deeply into the developmental trends and blueprint architecture for network evolution in addressing the critical challenges associated with the demand for ubiquitous 10Gbps access for business, campus, and home; industry digitalization; and AI computing applications. This aims to provide a good customer experience for services in need of high-speed (bandwidth), high-quality (reliability and low latency) communication.

Finally, it supports and guides key stakeholders and decision makers in implementing their enacted digital transformation strategies and related broadband policies and regulations in various regions. This process is already ongoing, and it is expected to continue beyond this decade.

KEY CHALLENGES AND SCENARIOS DRIVING EVOLUTION OF DATA COMMUNICATION NETWORKS

In the AI-enabled digital transformation era, secure and closer collaboration is needed between enterprises, networks, cloud, and applications. Omdia research on AI network traffic indicates that the traffic share of AI-related applications will be 56%, net new applications will account for 8%, and conventional (non-AI) traffic will only be 36% of network traffic in 2030. As shown in **Figure 1**, global traffic growth for AI-upgraded applications will increase yearly from 2022 to 2030 in comparison with non-AI application traffic.

FIGURE 1: PROJECTED GLOBAL NETWORK TRAFFIC GROWTH, 2022–30



SOURCE: OMDIA

The experience expectations of clients at home or in business are rising with the convergence of OT and IT in verticals, industry digitalization, and enterprise campus environments. As a result of these experience expectations, service providers are looking to maximize their revenue streams and manage their profit and loss while modernizing their IP transport networks. Current IP networks are inappropriate to meet the demand of bandwidth-hungry applications. In addition, the costs of operating aging brownfield datacom networks in terms of power, space, and processing requirements for AI training models are expected to rise:

- Many routing hops increase latency.
- Fast and flexible clouds are not matched by slow and rigid networks.
- Isolated clouds and networks slow down provisioning and the whole service lifecycle.
- Multicloud access is difficult because the existing network infrastructure foresees connection to only one cloud. The negative impact of IP address conflicts and network address translation mechanisms can be overcome by evolving and migrating current IP networks to IPv6.
- Traditional routing switches have low network throughput that cannot meet the requirements of AI training for high network throughput.
- There is no digital network for multi-tenant operation capabilities and end-to-end service management and orchestration.

More broadly, bandwidth-driven IP datacom networks are facing vital challenges and modernization requirements for different scenarios. Key modernization challenges to be addressed in each of the four identified domains are:

- Home consumer, business, and mobile networks requirements for a 10Gbps IP access network
- Larger-scale and higher-performance DCN (data center networking) in the AI era

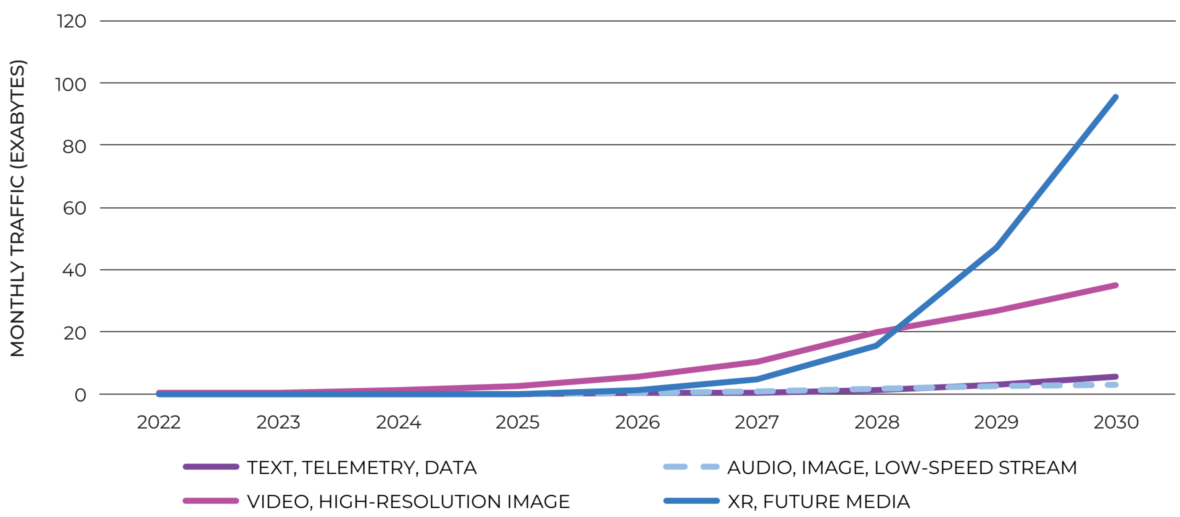
- Virtual/extended reality (VR/XR) experience and multicloud access for campus networks
- Automatic intelligent network operations and management (O&M) in an era when the rapid growth of general computing drives the speed of interconnection ports from 100G to 400G

IP TRANSPORT NETWORK: END-TO-END 400GE TO SUPPORT AT-HOME AND ENTERPRISE APPLICATIONS

The IP transport network must be upgraded to support the increased requirements from home, business enterprise, and mobile sites for broadband access to achieve 10GE home broadband, 10GE mobile access, and 10GE enterprise leased lines:

- **The upgrade of the at-home bearer network bandwidth and experience will be driven by emerging applications.** In the next 10 years, communications networks will need to support new human-computer interaction experiences. As shown in **Figure 2**, which illustrates new AI global network traffic growth by content type through 2030, Omdia research found that AI-enabled applications (such as XR, naked-eye 3D, digital touch, and digital smell) will become a sizable portion of the global traffic.

FIGURE 2: PROJECTED NET NEW AI GLOBAL NETWORK TRAFFIC GROWTH BY CONTENT TYPE, 2022–30



SOURCE: OMDIA

These future video applications make higher demands of data communications IP networks. In addition, new application scenarios such as digital tactile (a multidimensional somatosensory interaction) and digital olfactory (deep sensory interaction) require higher bandwidth and reliability from the network. With the gradual promotion of these new and high-bandwidth, high-definition (HD) video applications, the bandwidth requirements are about 10Gbps, the latency requirements are 1–5ms, and network availability must meet businesses' requirements of 99.999%. IP transport networks and data center backbone networks will be upgraded to 400GE, enabling sufficient network performance for immersive applications such as the metaverse and naked-eye 3D.

- **Enterprise cloudification and industry digitalization create new demands.** With the acceleration of the digital transformation process in thousands of industries, the focus of internet innovation will extend from "consumer internet" to "industrywide all-factor industry internet." It will eventually include new scenarios such as smart manufacturing and industrial internet. The existing IP data communication network does not have end-to-end networking and awareness capabilities in the access and cloud. It cannot meet the requirements for flexible scheduling of heterogeneous computing resources in different places. First, the network needs to evolve to the cloud-network synergy capability. Second, industry digital transformation accelerates the development of enterprise multicloud deployment services and drives networks to provide one-hop access to multiple clouds and intelligent networks, enabling autonomy and intelligent network O&M.

- **Development of 5.5G is far faster than development of 5G and 4G.** The 5.5G/6G era is set to build the foundation of a digital economy and enable an intelligent industry. Large-scale industry projects and 5G private networks are under construction in many large enterprises and verticals. Significant changes are seen in verticals, production processes, and supply chains. Global telcos are accelerating their “build process” efforts to deploy 5.5G wireless technology that shapes the required robust IP bearer low-latency backhaul networks with 10Gbps to the cell sites.

DATA CENTER: AI EXPLOSION DRIVES DCN INTO THE AGE OF MULTICOMPUTER POWER

AI represents a typical data-intensive computing scenario. It requires a large amount of sample data and generates a large amount of gradient data during computing. The explosion in the development of AI is introducing the intelligence era, when AI drives overall data center networking in the age of multicomputer power. Big AI models have been rapidly developed, especially since the launch of the GPT-3 as a heavyweight model that has dramatically sparked the development of the AI industry. Accelerated implementation of generative AI, model quantification, and AI leasing drive GPU clusters from 1,000 CPUs to 10,000 CPUs and GPU port rates to 400Gbps and then to 800Gbps in the future.

Take OpenAI as an example. Currently, GPT-4 has been trained and supports images and videos. The parameter scale and token scale of GPT-4 are respectively 10× and 43× those of GPT-3; the GPT-5 to be launched in the future may be 10× the current GPT-4. The GPT-5 cluster scale is estimated to reach more than 100,000 cards. Traditional TCP/IP protocol stacks involve multiple kernel context switchovers when sending and receiving packets, which will no longer be suitable for large-scale data interactions. Remote direct memory access (RDMA) has been introduced to address this scenario. The kernel bypass logic is used to achieve high throughput, low latency, and low CPU usage of computing units. Unlike TCP/IP, RDMA uses the Go-back-N retransmission mode and is sensitive to packet loss. If the packet loss rate is 1%, the RDMA throughput decreases to zero. In summary, the main requirements for DCN in the AI intelligent computing scenario are larger-scale networking, higher performance, and more intelligent O&M.

In conclusion, the network must provide high throughput, reliability, and automatic O&M capability to enable multi-DC comaintenance and management.

CAMPUS SCENARIO: 10GBPS HIGH-QUALITY SOLUTION

Digital transformation has brought about three significant changes in enterprise campus applications and terminals: wireless and Internet of Things (IoT), video-based office collaboration, and cloud-based services. Traditional enterprise campus networks are still in the era of simple data and are far from meeting the requirements of enterprise digital development. An intelligent campus network needs to be constructed with comprehensive awareness. In this way, the campus network is no longer isolated, and the different elements interact with and influence each other. In the next five years, the enterprise office campus will undergo the following changes:

- Massive wireless terminal access
- Video-based office collaboration
- Intelligent interaction brought about by AI and big models, promoting the cloudification of enterprise office services

In summary, massive wireless terminals and new immersive office collaboration technologies such as XR, intelligent interaction, and edge computing will bring new network challenges to campus office services and demand higher access bandwidth and lower latency from campus networks. In addition, network construction has shifted from being connection oriented to become experience oriented.

AUTOMATIC INTELLIGENT NETWORK O&M AMID THE RAPID GROWTH OF SUPERCOMPUTING

The demand for high-speed network connections will continue to grow with the rapid development of cloud computing, big data, and the explosion of IoT, and this places new requirements on existing O&M processes (fulfillment and assurance) that must be dynamic and intelligent. Service providers always complain that the extensive network is not scalable and cannot be monitored in real time, faults cannot be quickly solved, and the average training duration is short. To allow real-time provisioning and monitoring, servers' network interface controllers (NICs) are evolving from 10G/25G to 100G/200G, and interconnection ports are consequently evolving from 100G to 400G. At the same time, AI-based intelligence capabilities need to be included to allow programmable monitoring and preventive and proactive maintenance. From the perspective of vertical industries such as finance, shipping, and government, the network has to close some requirements gaps such as improving O&M efficiency and network reliability.

NET5.5G NETWORK ARCHITECTURE AND KEY TECHNICAL FEATURES

NET5.5G ARCHITECTURE

Net5.5G defines the sustainable evolution of data communication network infrastructure in the era of 5.5G and ubiquitous AI computing, underpinning the industry's digital and new application trends. It builds the digital IP transport network foundation, ensuring secure network operations, reduced network construction costs, and improved efficiency. The main capability advancement ensures intelligent internet and intersensing of everything by cloud-network convergence at the IP level with end-to-end IPv6/SRv6 technology as the fundamental innovation.

Net5.5G realizes the deployment of Wi-Fi 7, end-to-end 400GE, deterministic networking, and an application/computing-aware network, building an intelligent network infrastructure that connects the physical and digital spaces that are expected to be ready by 2030. Net5.5G is not expected to support simple traffic flows but ultra-reliable and low-latency use cases in fixed-mobile convergence.

Net5.5G defines three characteristics of the next-generation IP target network for network evolution in the 5.5G/6G era:

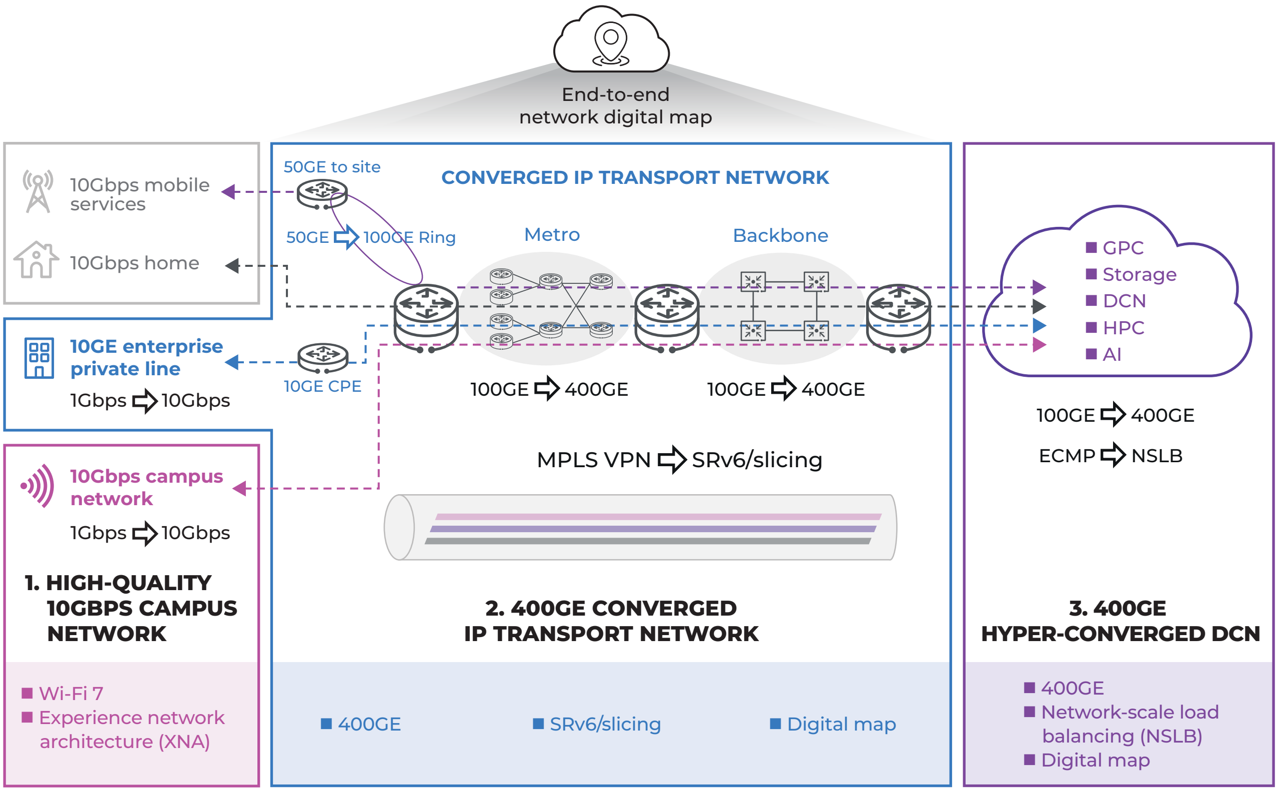
- One network for all services
- One network to multiple clouds for cloud-network synergy
- Network-as-a-service intelligent O&M

Net5.5G is the foundation that underpins the industry's digital and new applications trends with consideration of four main areas:

- IP transport network
- Data center networks
- Campus networks
- Intelligent O&M

From now through 2030 and beyond, remote-sensing technologies, extended/virtual/augmented (XR/VR/AR), remote AI applications, and supercomputing will be the main drivers for ubiquitous 10Gbps ultra-broadband deployments. To support these scenarios, networks must be implemented to fulfill requirements including high bandwidth and low latency and must be managed by intelligent end-to-end network digital maps.

FIGURE 3: NET5.5G ARCHITECTURE



SOURCE: WBBA

As shown in **Figure 3**, the end-to-end network architecture for next-generation networks consists of four parts:

- **Access networks** are upgraded to 10Gbps, including mobile broadband, home broadband, campus network, and enterprise private lines. Wi-Fi 7 and experience-centric network architecture are needed to guarantee service quality in campus networks.
- **Metro and core networks** need to be upgraded to 400GE, with SRv6 and slicing.
- **Data center networks** need to be upgraded to 400GE, with network-scale load balancing.
- **End-to-end network management** realizes an efficient network digital map including DCN, IP transport network, and access network.

NET5.5G KEY TECHNICAL FEATURES

Net5.5G proposes many innovations based on the standards, extending the network infrastructure from best efforts to deterministic and autonomous. Evolving technological capabilities underpinning the network will open the door to the predicted smart world for 2025–30. This will enable realization of a massive industrial digitalization meta-universe and help enterprises be flexible in using AI computing power from multiple clouds. The network will be capable of supporting diversified application scenarios by 2030:

- **IP transport network:** The base station access is upgraded from 10GE to 50GE, driving the transmission speed of the MAN (metropolitan area network) aggregation network and backbone network to 400GE.
- **Data center network:** The server network scale increases from 100,000 nodes to millions of nodes. The maximum interface rate of data center switches is upgraded to 400GE, reducing latency from microseconds to nanoseconds.
- **Campus network:** Wi-Fi is upgraded from Wi-Fi 6 to Wi-Fi 7, giving users a peak access capability of up to 23Gbps.
- **Intelligent O&M:** The network is upgraded from L3 (conditionally autonomous) to L4 (highly autonomous) or later to L4.5 to achieve an almost fully intelligent autonomous network.

Table 1 lists and summarizes the key technical features of Net5.5G.

TABLE 1: KEY TECHNICAL FEATURES OF NET5.5G

APPLICATION SCENARIO	IMPORTANCE
IP TRANSPORT NETWORK	Ultra-high bandwidth: the MAN aggregation and backbone network's transmission speed is upgraded from 100GE to 400GE.
	Efficient cloud migration: from multiple hops to the cloud to one hop to multiple clouds.
	Ultimate experience: tenant-level network slices are upgraded to the 100,000 level.
DCN	Ultra-high bandwidth: The maximum interface rate of data center switches is upgraded from 100GE to 400GE.
	High throughput: The network bandwidth utilization needs to be increased from 50% to 95%, to improve training efficiency.
CAMPUS NETWORK	Large bandwidth: Wi-Fi 7 and other technologies are supported with a maximum rate of 23Gbps or higher.
	Ultimate experience: differentiated granted service-level agreements (SLAs).
INTELLIGENT O&M	Upgradable autonomous network level: upgrade from L3 (conditionally autonomous) to L4 (highly autonomous) networks and later to L4.5 to achieve almost fully autonomous networks.

SOURCE: WBBA

NET5.5G DOMAINS: KEY TECHNICAL FEATURES AND ARCHITECTURE

IP TRANSPORT NETWORK

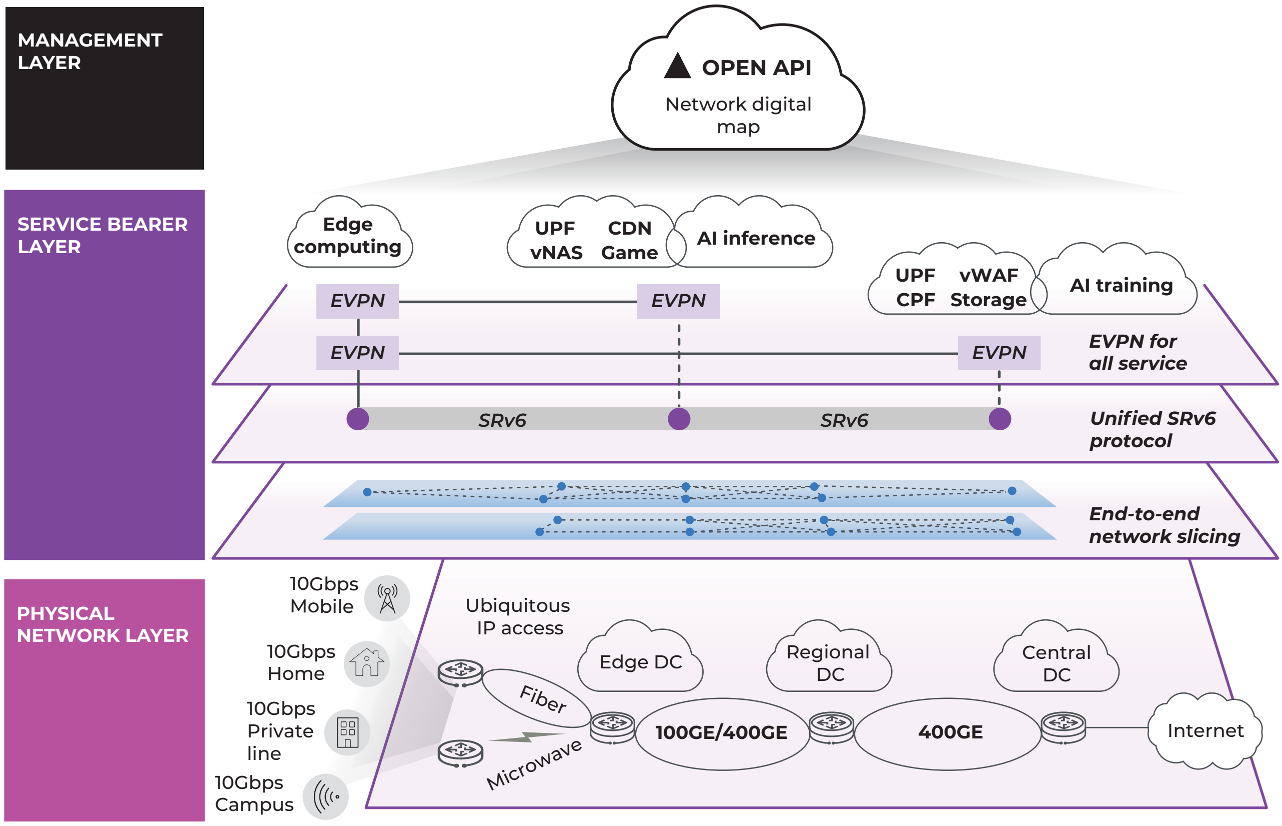
Upgrading the service bandwidth on the user access side to 10GE drives the carrier's IP transport network toward becoming a 400GE converged MAN and backbone network.

Network evolution must leverage IPv6-based IETF protocol innovations to increase the availability of IP addresses for new applications and to expand existing ones and also to enable automation and assure service levels and user experience. In this regard, SRv6 simplifies and unifies end-to-end IP addressing and tunnel signaling; connects data centers, end systems, and campus networks; and reduces cross-domain translation within the network.

Based on IPv6, SRv6 (with AI capabilities) allows new service innovation and enables the connection of intelligent networks. An IPv6 network that supports SRv6 and AI ensures fast computing at the edge. End-to-end SRv6 supports one-stop access to multicloud environments. Complex slicing based on FlexE technology enables a deterministic user experience. IPv6/SRv6 networks allow assurance of differentiated quality of service (QoS) through network slicing and can shorten service-provisioning time through automation and software-defined networking (SDN) control to meet cloud application requirements.

For the Net5.5G era, the IP transport network focuses on lower total cost of ownership of network construction, provides assurance of the ultimate experience at application and tenant level, and facilitates network service monetization. The network architecture is divided into three layers: the physical network layer, the service bearer layer, and the management and control layer. **Figure 4** provides an overview of the IP transport network architecture.

FIGURE 4: IP TRANSPORT NETWORK ARCHITECTURE



SOURCE: WBBA

LAYERING-BASED NETWORK ARCHITECTURE MODEL DESCRIPTION

- **Physical network layer:** Provide ultra-broadband connections, converged architecture, and high reliability using proactive methodologies:
    - **Ultra-broadband:** 50GE ports at the access, 100GE/400GE in access rings, and 400GE ports are deployed at the aggregation and core layers.
    - **Converged architecture:** The network uses a simplified and flexible architecture; any node on the network must be capable of carrying multiple services. The network is ready to cope with flexible future service changes and adjustments.
    - **High reliability:** Dual plane end to end ensures high network reliability and scalability.

- **Service bearer layer:** This is decoupled from the physical layer and focuses on agility. It includes the following:
    - **Control plane:** It provides interconnection and interworking for customer services, including control protocols and path programming.
    - **Forwarding plane:** It provides differentiated service level assurance at tenant level, including bandwidth resource isolation, high reliability, stable latency, low jitter, and zero congestion.

- **Management layer:** Network digital twin manages, controls, analyzes, and visualizes networks. It is the “brain” of the intelligent network.
    - **Visualization:** visualization of the perception of application-level and tenant-level experience and network KPIs associated with experience key quality indicators
    - **Automation:** automatic inspection, optimization, and online configuration simulation
    - **Intelligent:** AI capabilities, such as interactive intelligent diagnosis, predictive traffic grooming, and O&M efficiency improvement, in key scenarios
    - **Open northbound interface:** Tenant-level end-to-end operation capabilities

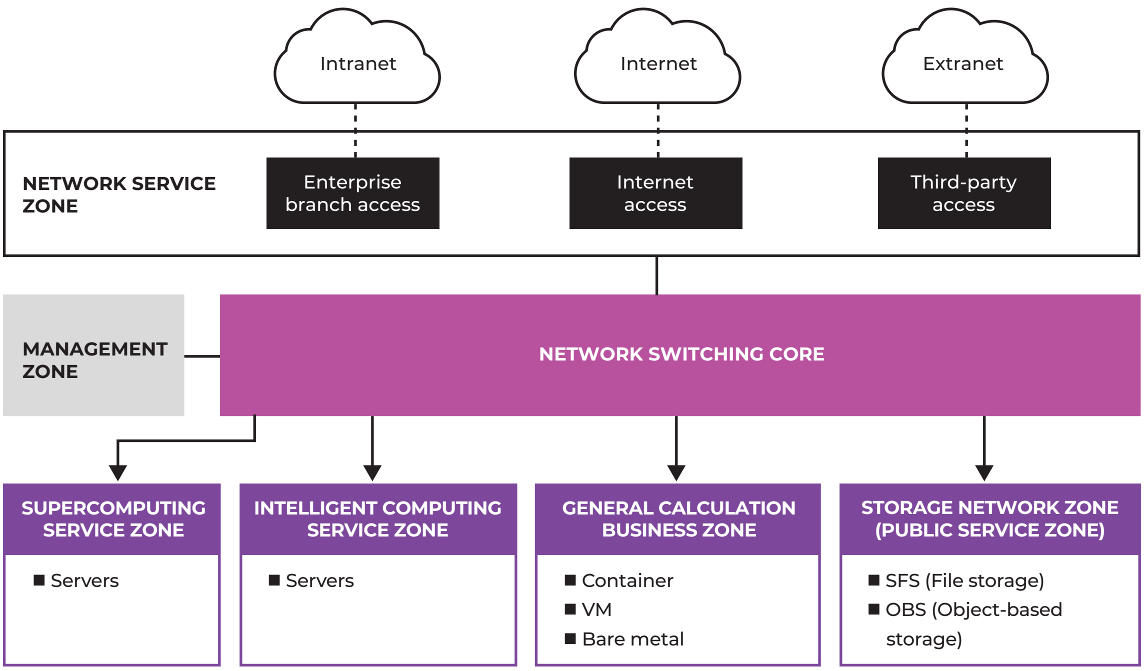
DATA CENTER NETWORK

The increasing utilization of generative AI will lead to increased requirements in data center design and efficient networking. Reducing the communication time for data synchronization across the data center to improve the effective GPU computing time ratio (GPU computing time / overall training time) will be critical to improving the efficiency of AI distributed training clusters:

- **Ultra-high bandwidth:** A large amount of communication data will be generated in the large-scale artificial intelligence model training scenario. Amdahl's lesser-known law indicates that in parallel computing, every 1MHz CPU can generate a maximum of 1Mbps I/O. Therefore, a server with a 32-core 2.5GHz CPU needs to be configured with a 100Gbps network adapter to maximize computing performance. In addition, the explosive growth of data drives the continuous growth of high-speed network connections. Therefore, 200GE ports will become the mainstream choice for data center server NICs, and 400GE interconnection and forwarding will become the future network architecture.
- **Ultra-high throughput:** The computing number of AI applications increases geometrically, and algorithm models are developing to a huge number. In AI training scenarios, the ultra-large data flows generated by large models require the network to solve the load-balancing problem. The network bandwidth utilization needs to be increased from 50% to 95%, to improve training efficiency.

The DCN is classified into general computing data center services, supercomputing data center services, and intelligent computing data center services based on large service scenarios. In addition, necessary storage network devices are required. **Figure 5** provides an overview of the Net5.5G DCN architecture.

FIGURE 5: DATA CENTER NETWORK ARCHITECTURE



SOURCE: WBBA

- **General computing service zone:** The general computing service area is the largest in the infrastructure of an enterprise data center. Generally, the area is divided into multiple physical points of delivery (PODs) based on enterprise service types. Different enterprise applications are deployed in different physical PODs. Containers, VMs, and bare-metal servers carry computing power based on application characteristics. Diversified services and computing capabilities make network deployment complex. On the one hand, the access side of servers will evolve from 10G/25G to 200G, and aggregation and core devices will evolve to 400G.

- **Intelligent computing service zone:** AI intelligent computing service consists of the management plane network, parameter plane (training) network, sample plane (storage) network, and service plane network. The parameter plane is the key network that affects the efficiency of AI model training. The key task of the network is to efficiently carry computing power and data between AI model training cards. First, high transport capacity requires high bandwidth. Currently, 200G is the mainstream technology for AI training access. With the boom in AI models, 400G access is rapidly maturing, and it is estimated that it will become mainstream in 2025. Second, AI training scenarios mainly involve a small number of high-density flows, which are few in number and high in flow density. Traditional equal-cost multipath (ECMP) routing uses the 5-tuple-based per flow hash algorithm. When the number of flows is low, hash unevenness is likely to occur, causing traffic conflict and back pressure, reducing traffic throughput, and affecting training efficiency. Traffic paths between AI training cards can be planned to avoid hash conflicts between traffic. Uplink conflicts on leaf switches and downlink conflicts on spine switches must be considered during path planning. Network-level load balancing is recommended for load balancing on the entire network.
- **Supercomputing service zone:** Supercomputing is a high-performance computing (HPC) service. More than 80% of the packets have payloads smaller than 16 bytes. End-to-end network latency is a key factor affecting supercomputing performance. Calculation in the traditional aggregate communication process is completed on the server side, and the network is only responsible for forwarding. Using an INC (in-network computing) function, part of the aggregate communication calculation process is offloaded to greatly improve efficiency. With the continuous development of HPC and AI applications, the supercomputing industry will evolve to the HPC+AI converged architecture.
- **Management zone:** Using the network digital map to perform visual management enables the use of many features such as multidimensional network visualization, service path navigation, and application experience assurance. The multilevel path visualization is implemented thanks to the unified management of all network devices in the data center, including devices of multiple vendors and resource types, such as computing, network, storage, containers, VMs, and bare metal.

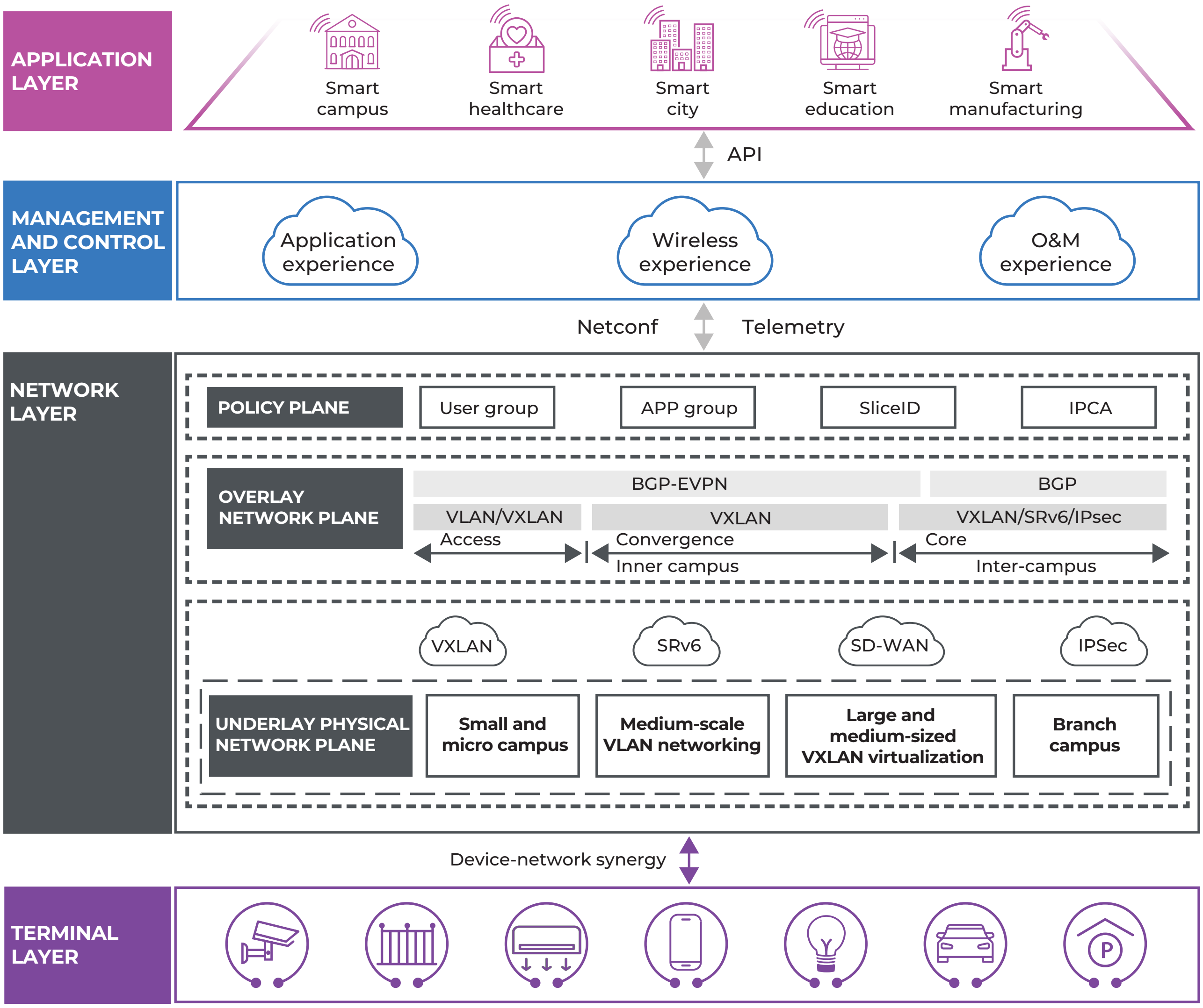
CAMPUS NETWORK

The digital transformation of enterprises drives the upgrading of campus network quality. HD videoconferencing, XR, and holography are widely used, requiring larger network bandwidth. As a result, the number and types of terminals increase by 20×, and the bandwidth requirement increases by 10×. The campus network architecture evolves toward ultra-broadband connection and high-speed access, improving customer experience in application, wireless, and O&M:

- **High bandwidth:** In a large-scale campus network, the number of terminals per user increases from one to four or more. The number of terminals connected to each access point (AP) increases up to 60. In peak hours, the download rate is sometimes less than 1Mbps, which cannot meet the requirements of rapid service development. Wi-Fi 7 further improves the channel width and transmission rate, enabling the maximum transmission rate of the campus network to be higher than 23Gbps.
- **Expected experience:** The network can identify applications for audio- and videoconference scenarios and ensures zero frame freezing in audio- and videoconferences when the network is congested. To achieve VIP/user experience assurance, the network can identify VIP users and provide them with a bandwidth guarantee so that no packet loss occurs when the network is congested. This feature ensures that VIP users can enjoy consistent network access rights anytime, anywhere, on any terminal and in any access mode.

The campus network architecture can be divided into the terminal, network, management, control, analysis, and application layers. Each layer has clear functional boundaries and provides different functions. The application and management, control, and analysis layers manage users, networks, and applications. The network layer implements high-quality bearer, and the terminal layer has to support multiple devices, including IoT. **Figure 6** illustrates an overview of the Net5.5G campus network architecture.

FIGURE 6: CAMPUS NETWORK ARCHITECTURE



SOURCE: WBBA

- **Terminal layer:** Terminals are various terminal devices that access the campus network, such as fixed, mobile, and IoT terminals. Terminals inform the network of their identity, and the network accurately identifies terminals to implement plug-and-play and precise policy control. Finally, the terminal and network layers collaborate with IoT applications to form an efficient cloud-device collaboration solution.
- **Network layer:** The network layer includes the physical layer (underlay), logical network layer (overlay), and service policy layer. Physical and logical networks are completely decoupled.
    - Physical network plane: It is recommended that APs supporting higher-bandwidth Wi-Fi 7 be used on the terminal access side to support mobile office scenarios and provide a better wireless experience than Wi-Fi 6. Core switches should use M-LAG (Multichassis Link Aggregation Group) to improve reliability.
    - Logical network plane: It provides a campus virtual switching network that is reachable by any device, implementing multipurpose underlay networks.
    - Service policy plane: Based on the logical network layer, the service policy layer implements flexible policy control between users and applications and ensures user experience. Services on multiple underlay networks in a traditional campus network can be integrated into one overlay network. Services on the overlay networks are completely isolated and share the same networking reliability, reducing costs and maintenance effort. The overlay network supports quick provisioning of campus network services.

- **Management and control layer:** Acting as a “smart brain,” the management and control layer provides network-level management capabilities such as configuration management, service management, network maintenance, fault detection, and security threat analysis.
- **Application layer:** The application layer of a campus network consists of independent servers that provide value-added services, such as AI application servers, email servers, instant messaging servers, and videoconference servers. Some servers with high network quality requirements, such as videoconference servers, need to grant proper service assurance. More service applications can be developed through the northbound and southbound interfaces provided by the management, control, and analysis platform to build a service application platform based on the campus network.

INTELLIGENT O&M

In the digital age, IP transport networks, data centers, and campuses face more diverse and complex service requirements. As a solution, integrating AI into network management, control, and analytics functions helps drive intelligent network automation. As defined by TM Forum, intelligent network automation is divided into six levels (L0 through L5) to measure network automation’s maturity and service experience.

Network evolution aims to upgrade from L3 (conditionally autonomous) to L4 (highly autonomous) networks and later to L5 to achieve fully intelligent autonomous networks. These three levels apply from single devices up to the overall network: devices, IP network, and service.

KEY ENABLING TECHNOLOGIES FOR NET5.5G NETWORK ARCHITECTURE

This section highlights the key enabling technologies targeted by the Net5.5G network architecture.

WI-FI 7

With the development of wireless LAN (WLAN) technologies, homes and enterprises increasingly rely on Wi-Fi to access their network. Wi-Fi 7 improves the data transmission rate and ensures low latency and high reliability. Therefore, Wi-Fi 7 better matches the requirements for robustness and delay performance for data transmission in scenarios such as voice conferencing, real-time operation, Industrial Internet of Things (IIoT), interactive telemedicine, and similar; for example, industrial automated guided vehicles (AGVs) require 100ms latency and 99.999% available services.

EXPERIENCE-CENTRIC NETWORK ARCHITECTURE

By improving network capabilities and performance, users can upgrade their network experience. In addition to connections, the experience-centric network architecture also senses service requirements and provides a better-personalized service experience.

The experience-centric network architecture includes three upgrades:

- **Wireless experience improvement:** Intelligent multimedia scheduling and multi-AP collaboration technologies forward high-priority wireless services preferentially, ensuring a smooth WLAN service experience without frame freezing.
- **Application experience improvement:** To increase the access experience of important applications, the intelligent audio and video application identification capability is used to identify service flows that need to be guaranteed. QoS, multimedia scheduling, and application-based intelligent traffic steering can be used to assure the user experience of important applications.
- **O&M experience improvement:** When users’ audio and video experience deteriorates, O&M personnel must quickly identify whether it is the network or the application causing the problem. The inline experience measurement technology can be used to color real service packets and accurately collect statistics on packet loss. This improves the O&M experience without fault reproduction.

NETWORK DIGITAL MAP

The network digital map is a solution applying the digital twin concept in communication networks. The network digital map provides a navigational traffic map experience. It supports agile service provisioning, automatic traffic optimization, intelligent fault analysis, and potential risk prediction. It is a virtual twin that digitally maps physical network entities and performs real-time interactive alignment with physical networks. It provides multidimensional dynamic topology, localization, network configuration verification, and service experience assurance capabilities through a map-based experience. It continuously guarantees service levels and improves service experience. The network digital map aims to build automatic, self-optimized, and self-healing intelligent O&M. It is a key foundation for the evolution of intelligent network O&M management.

Net5.5G recommends a digital map solution for the following supported capabilities:

- **Network resource visualization:** Network operations center staff expect to display the resources during network running in real time and to dynamically monitor network performance, improve network capacity, and locate and rectify faults visually. This includes mapping physical network connections, device physical components, and logical resources (slice, tunnel, IGP route, BGP route, and VPN services). The visual map includes real-time data such as application status, availability, and energy consumption.
- **Real-time network status awareness and proactive service experience assurance:** Based on the network digital map's basic resources and status information, proactive prediction, automatic recovery, network performance self-optimization, and automatic power consumption control can be further evolved to guarantee service experience efficiently. The network digital map must quickly detect network topology changes, node or link faults, and link bandwidth or delay changes; accurately detect application traffic packet loss information; and display correlation analysis KPIs of network link quality and service path quality KPIs. When network congestion occurs, network management can compute the optimal path based on service intent and perform fast automatic optimization and adjustment to guarantee services with high service level requirements preferentially. Ideally, network optimization can be completed before service experience deteriorates, greatly improving network O&M efficiency.
- **Network status correlation analysis and proactive fault management:** Fault management is one of the most critical processes during network operation and maintenance; it includes analyzing the IP network status, identifying network faults, and determining root causes and recommended recovery procedures. On an IP network, a single fault may cause many alarm events on multiple network elements and network layers, increasing the difficulty of locating the root cause. The network knowledge graph and the large language model are two key technologies that enable to address this issue.
- **Network digital twin emulation:** This supports intent-oriented configuration change and verification. The IP network has a large geographical coverage and carries many cross-city, cross-region, and even cross-country data services. Changing network configurations is a risky operation. The network digital map maintains the change configuration, interconnection routes, and traffic status of network devices. It can simulate the status and behavior of network protocols and traffic and emulate the routing table and forwarding entry of network devices. It provides real and objective basic data for network change risk assessment.

NETWORK PROGRAMMING (SRV6)

Segment Routing IPv6 (SRv6) is the last evolution of the source-routing technology. With the development of the industry ecosystem and related standards, SRv6 and SRv6 compression technologies are increasingly deployed on global IP networks, helping telecom carriers, industry customers, and enterprise customers deploy more cost-effective and intelligent networks and provide convenient and high-quality service experiences.

Net5.5G strongly recommends SRv6 routing because of the following key benefits:

- **Simplifying existing network protocols:** SRv6 does not use the LDP/RSVP-TE protocol and MPLS labels, reducing network protocols stack and reducing network management complexity. SRv6 can better cope with the challenges of 5G and cloud network development.

- **Specifying explicit forwarding paths:** SRv6 uses explicit forwarding paths to meet different service requirements, such as network optimization and backup path specification. Explicit forwarding path specification is essentially based on the IPv6 SRH extensible header. The explicit forwarding path of data packets is specified on ingress to implement network forwarding path programming.
- **Compatible with IPv6:** SRv6 can be deployed incrementally to support smooth evolution from IPv6 to SRv6. This eliminates the need for a one-time networkwide upgrade and supports fast service rollout. Incremental SRv6 deployment can protect network carriers' assets and investments to the maximum extent.
- **Cross-domain network path programming:** SRv6 is compatible with IPv6, enabling SRv6 to be better used in cross-domain scenarios and supporting end-to-end 5G and cloud network service development. With the network programming capability, SRv6 can perform path programming, meet service SLA requirements, and connect networks and applications to build intelligent cloud networks.
- **Header compression technology:** SRv6 header compression effectively resolves the problem of excessive SRv6 header overhead. It is a general mechanism compatible with SRv6 and supports multiple types of security identifiers (SIDs) with different lengths. These SIDs are referred to as C-SIDs (compressed SIDs). C-SID uses 32-bit C-SIDs to reduce the overhead of segment lists (SID lists) by up to 75% and reduce the overhead of SRv6 packet headers. In addition, SRv6 header compression supports hybrid programming of common SRv6 SIDs and compressed SIDs. SRv6 header compression can be deployed by upgrading some nodes on a network as required, implementing smooth upgrade and inventory evolution from SRv6 to SRv6 header compression.

NETWORK SLICING

Regarding standards, the drafts related to network-slice architecture have entered the RFC release phase, and several drafts related to the data plane, control plane, and management plane protocol extensions have also become workgroup drafts. It is expected that a batch of network-slice drafts will be released as formal standards in 2024 in IETF. As the network-slicing industry ecosystem and related standards develop and mature, IPv6-enhanced-based network slices are increasingly deployed on global IP networks to improve service experience for various industries and enterprise customers and enable telecom operators to monetize "to-business" business models.

The key benefits of network slicing are as follows:

- **Resource and security isolation:** Network slicing controls and prevents a service in a slice from affecting another slice in the same network and prevents data access between slices, thereby improving service security. Network slicing is of two types: hard and soft slicing. Hard slicing allocates exclusive network resources to different network slices.
- **Differentiated service quality assurance:** Network slicing uses shared network infrastructure to provide differentiated service-level assurance for different industries, services, or users. Network slicing enables operators to provide differentiated services for business to home (B2H), business to business (B2B), and business to consumer (B2C) instead of selling best-efforts traffic only. Providing differentiated services for tenants in the form of slice offerings will be the main model for carriers to provide services in the future. It will also generate a new value stream for carriers.
- **High reliability:** Network slicing provides local protection technologies for any fault point on the IP network. In addition, fault switchover can be performed within the scope of a network slice without affecting services in other network slices. High-value services of vertical industries and URLLC (ultra-reliable low-latency communications) services of mobile networks require high availability and millisecond-level fault recovery of IP networks.
- **Flexible and customizable topology connections:** Network slicing can support on-demand customization of logical network topology connections to meet the differentiated network connection requirements of different industries, services, or users. Network slicing does not need to detect the full network topology of the basic network, which simplifies network maintenance for network-slicing users, prevents excessive information exposure on the carrier network, and improves network security.

ULTRA-BROADBAND 400GE AND BEYOND

The 400GE Ethernet has a large transmission capacity, high transmission efficiency, and good interoperability, helping to meet the requirements of network evolution and existing services. It is the best solution for cloud service providers with requirements for high-density data centers and for telecom carriers urgently needing high-speed and high-bandwidth traffic growth solutions.

The key points and values of 400GE Ethernet are as follows:

- **Higher data rate:** Compared with previous generations of Ethernet, 400GE provides a higher media access control rate and supports higher data transmission speed per port for data centers, enabling cloud computing and high-performance computing scenarios that require higher network bandwidth.
- **Multichannel and wavelength division multiplexing (WDM):** Multiple channels transmit data simultaneously, increasing the overall data transmission capacity per port and improving network flexibility. WDM allows simultaneously transmitting “N” wavelengths (each carrying data) as optical channels in the same fiber, increasing the overall data transmission capacity. “N” can be programmable and reconfigurable, which improves network flexibility and scalability.
- **Higher efficiency:** The main advantage of 400GE is in reducing network construction costs. Higher capacity per switch and a higher rate per channel will significantly reduce the overall racks and fiber counts for the same data center capacity, lowering capex requirements. Lower power per bit using higher-speed SerDes (Serializer/Deserializer) and optical devices will reduce the long-term power bill.
- **Four-level amplitude modulation (PAM4):** The 400GE high-speed optical modules use the PAM4 modulation technology. Each symbol consists of two bits of information and is represented by four different power levels during transmission. The 400GE high-speed optical modules have higher spectral efficiency and reduce the power per bit by more than 15% compared with the 100GE optical module.
- **Compatibility and interoperability:** Because 400GE devices can be used to upgrade the end-user network in existing network architecture, large-scale infrastructure replacement can be avoided. Interoperability is guaranteed between devices from different vendors, which could promote market competition and technological innovation in the field.

NETWORK-SCALE LOAD BALANCING

Network-scale load balancing is a technology tailored for AI training scenarios. It calculates the optimal traffic distribution by drawing a global traffic matrix and automatically diverts traffic through network devices. The network-scale load-balancing technology takes advantage of the global perspective of both network and AI training tasks, achieving 100% networkwide traffic balancing and improving performance in AI training scenarios.

During AI large model training, parameters are synchronized and exchanged between servers through the high-speed network. The communication traffic has the following characteristics: periodic traffic, a small number of flows, persistently connected traffic, and strong real-time synchronization between parallel tasks. Communication efficiency depends heavily on the slowest node. In addition, a large amount of data is transmitted in the AI cluster training scenario. Given the current per flow ECMP mechanism, load balancing is relatively uneven, and congestion happens frequently on the network, resulting in a sharp performance decrease in AI cluster training. Therefore, improving network throughput by adopting networkwide scheduling is important for effective load balancing.

The key points and benefits of network scale load balancing are as follows:

- **Effective bandwidth improvement:** Obtain global topology and task information, apply network-scale load balancing to compute paths, and dynamically deliver device policies quickly. This prevents traffic conflicts caused by random ECMP hashing and improves effective bandwidth during training. In the AI single-task scenario, the effective bandwidth utilization is increased to more than 90%, and the efficiency of the AI training task is improved by more than 20%, greatly shortening the AI task completion time and saving customers’ costs and expenses.

- **AI training performance improvement in multitask parallelism:** Centrally manage and analyze multitask information and use global path computation to relieve multitask communication conflict, efficiently carry multitask traffic, and thus improve AI cluster utilization. AI training performance is improved by more than 50% in multitask scenarios.
- **AI technologies bring new requirements and challenges to networks.** Networks need continuous technological innovation to better adapt to the future. Improving AI training efficiency by achieving overall load balancing is the beginning of network technology innovation.

CONCLUSION: A WAY FORWARD

NET5.5G ROADMAP AND NET6G PROSPECTS

The success of the digital economy depends on new requirements being placed on network connectivity and, more specifically, robustness and intelligence. But this robust intelligent network connectivity demands cloud-network convergence, including supercomputing power, extensive data integration, native AI-enabled network intelligence, and network security automation. The growth of the digital economy is catalyzing IP datacom bearer network access, metro, and backbone network bandwidth roadmaps. Accelerating end-to-end IPv6 adoption, encouraging SRv6 deployments, FlexE for bearer network, network slicing for physical isolation, and network digital mapping for intelligent O&M in multiservice, converged IP bearer networks will make a big difference for service providers in winning connectivity business.

Service providers in the 5.5G/6G era must meet requirements from enterprises and industries for an enormous flow of data and AI computing capabilities and optimize the capability of data center interconnection for the AI infrastructure. In 2025 and beyond, diversified at-home consumers, enterprises, and vertical industries (such as healthcare, education, transportation, manufacturing, logistics, and mining) will aggressively demand a strong end-to-end transport bearer network foundation with integrated secure devices to the cloud, automated network openness with on-demand multicloud interconnections, and terminal capability enhancement.

Access at 10Gbps in the home, campus, and enterprise is the key enabler to grant proper quality of experience to end users, enabling AR, unmanned industries, and proper interaction with AI-based services. Wi-Fi 7 will play a fundamental role in this context.

In addition, 400GE DCN is needed to support complex AI-based applications in data center and cloud in order to provide nonblocking computational and latency reduction.

End-to-end digital mapping will have to span all the network domains, providing flexible service and infrastructure lifecycle.

The Net5.5G framework and proposed recommendations empower all industry stakeholders by envisioning a modernization of IP bearer network architecture to offer URLLC networks for AI-enabled digital hubs and industrial intranets.

Net5.5G uses the end-to-end IPv6/SRv6 technology as the core. Upgrade key network capabilities—such as Wi-Fi 7 access, end-to-end 400GE, network digital maps, network-scale load balancing, experience-centric network architecture—and build an intelligent network infrastructure that connects physical and digital spaces.

WBBA STRATEGIC RECOMMENDATIONS TO REGULATORY VITAL STAKEHOLDERS AND POLICYMAKERS FOR BEARER AND DCN CONSTRUCTION INCLUSIONS

The aim of WBBA WG4 is to demonstrate technical leadership and present recommendations about network evolution for the 5.5 era (Net5.5G).

All industry ecosystem stakeholders, including service providers, government, regulators, and vendors, should collaborate to tap into the enduring value of Net5.5G-enabled cloud-network convergence. Aiming at the evolution of future services and new technologies, it will open up the opportunity to discuss the idea of Net6G.

WG4 members are encouraged to submit proof-of-concept (POC) project proposals within the POC program governance and form consortiums under the leadership of key industry ecosystem stakeholders, including regulators and enterprises, as instantiation and operationalization of the suggested recommendations.

Below are WBBA's key strategic imperatives and suggestions regarding IP transport and campus and DCN network construction for policy planning and inclusion:

- The end-to-end IP bearer network must be gradually upgraded to 400GE technologies and IPv6 protocol stack. They represent a sustainable path and logical evolution for internet development. This will help to facilitate the applications of segment routing, network slicing, and network digital maps in the end-to-end networks for 10Gbps access for enterprises and homes.
- Evolve campus network architecture toward ultra-broadband connection and high-speed access, improving customer experience in application, wireless, and O&M.
- AI will become a dominant factor from the perspective of the DC computing workforce. To meet the requirements of these AI computing cases, the following key features are necessary: ultra-large bandwidth supporting 400GE access; high throughput supporting 98% utilization of GPU computing power; and lossless Ethernet network with zero packet loss.

WORLD BROADBAND ASSOCIATION

Join the World Broadband Association

We encourage your feedback and would welcome the chance to discuss with you how you can benefit from, and contribute to, the success of the WBBA. Please submit enquiries for free membership via *https://worldbroadbandassociation.com/*