

Issue 096

Volume 01 (2024)

BUSINESS SUCCESS

Embracing 5.5G to
Unleash Industrial Dividends

INTELLIGENT TRANSFORMATION

EM 2.0: A Road for the Digital Intelligent
Transformation of African Carriers

INNOVATION FOR THE FUTURE

Jointly Defining ICT Architecture for 5.5G:
**The Key to Unlocking
New Opportunities**

HuaweiTech

EXPLORE · INNOVATE · INSPIRE

5.5G

Copilot

Intelligence

Agent



Embracing 5.5G
To Advance the
Intelligent World

SCAN TO VIEW ON YOUR PHONE

Restricted Publication

Free Distribution

PUBLICATION REGISTRATION NO.:
YUE B NO. L0230032

5.5G Use Cases

5.5G Era: 4 Business Trends and 10 Application Scenarios



- **People**

3D Content

Glasses - free 3D Video, 3D Live, XR...
- **Home**

Large-screen smart connection

Large-screen/multi-screen entertainment, Car road synergy...
- **Car**

AI-based content generation

ChatGPT, Luma AI...
- **Thing**

Full Connection and intellectualization

Passive IoT, flexible manufacturing, ...
- **Government and enterprise**

Building a Fully Connected, Intelligent World

Embracing 5.5G to Advance the Intelligent World



Corporate Senior Vice President and President of ICT Sales & Service, Huawei
Li Peng

An opportunity may come only once, so we must seize it before it slips away.

Over the past 20 years, informatization and digitalization have brought opportunities worth trillions of dollars to the ICT industry, and today, we're rapidly approaching a fully intelligent world. According to the *Intelligent Economy Enabling Sustainable Growth* report, the Intelligent Economy could be worth US\$18.8 trillion by 2030. As intelligence continues to redefine information and value streams and unleash the power of ubiquitous connectivity, the ICT industry will embrace new development opportunities.

Intelligence will make the acquisition, presentation, transfer, storage, and processing of information more efficient, redefining value streams. Huawei predicts that, by 2026, AI will create more than 250 billion high-quality images and over 70 million short videos, representing exponential growth in AI-generated content. Additionally, new types of connected objects and scenarios will emerge, including AI phones, digital humans, and intelligent vehicles, while new models for collaborative storage and computing between cloud, edge, and devices will be created. These changes will stimulate the generation of over one trillion gigabytes of data traffic.

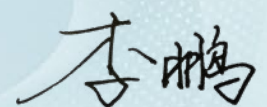
Innovative applications are driving users to focus on more dimensions, including uplink, QoS, and latency, creating new value. New business models centering on metrics like bundles and SLAs are leading to new monetization methods, opening up new value streams worth hundreds of billions of dollars to the ICT industry. For example, one carrier launched a 5G livestream package to provide guaranteed uplink speeds for seamless live streaming. The package's monetization model based on multiple metrics, like uplink speeds and guaranteed QoS, helped increase the carrier's ARPU by over 70%. Another carrier launched 5G New Calling services that provide capabilities like positioning and real-time interaction. These capabilities have made mobility services and car insurance claims more convenient, and have helped the carrier successfully monetize the B2B2C market.

Ubiquitous connectivity is pushing the boundaries of both time and space, allowing information and value streams to flow faster and more broadly. Networks are delivering stronger capabilities, including large bandwidth and low latency across different scenarios. Intelligent autonomous vehicles can operate around the clock; production systems can be controlled in real time; and cloud phones enable consumers to access online games and workplace applications far beyond the limitations of local device storage and processing power. In wide-area IoT, connectivity density and precision accuracy must increase by over 10 times to support the needs of any scenario, from smart grids and manufacturing, to storage and logistics.

The redefined information and value streams, alongside ubiquitous connectivity, will raise the bar for networks, and 5.5G will be our path forward in an intelligent world. In China, the Middle East, and Europe, leading carriers have made nonstop innovations and verified advanced 5.5G capabilities on commercial networks. Their tests cover a wide range of scenarios, including smart connections for people, homes, vehicles, and stadiums. Furthermore, an increasing number of 10-gigabit smart cities powered by 5.5G are popping up across the globe.

Huawei is fully prepared for the first year of commercial 5.5G. At MWC Barcelona 2024, we released a full series of 5.5G products and solutions for a wide range of scenarios, along with the industry's first Telecom Foundation Model. They will help carriers provide better user experience and improve O&M efficiency.

The poet Kahlil Gibran once said, "Progress lies not in enhancing what is, but in advancing toward what will be." Huawei will continue working with global carriers and partners to innovate and build today's networks for tomorrow's applications. Together, we can advance the intelligent world!



Publisher ICT Strategy & Marketing Dept.
Huawei Technologies Co., Ltd.

Presented By Zhou Jun

Consultant Song Xiaodi

Editor-in-Chief Xing Jingfan

Editor-at-Large Tang Xiaoqiang

Editors Fu Dongwei
Gary Marcus Maidment

Art Editors Zhou Shumin
Xu Chuangliang

Contributors Wang Xiangtian, Wang Yuhua,
Zhu Wei, Gong Tao, Liu Haoifei,
Qi Fushui, He Hui, Tang Jiang,
Zhang Liqiang, Zeng Xiaofei,
Wei Yongfu, Zhao Yingnan,
Du Yanchen



HuaweiTech:
Trends.
Insights.
Innovation.
Success.

Please contact the
Huawei Tech editorial office to
request a copy of, contribute to,
or comment on Huawei Tech.

E-mail HWtech@huawei.com
Tel +86 755 89241326
Address G1, Huawei Industrial Base,
Bantian, Longgang,
Shenzhen 518129, China

Publication Registration No.:
Yue B No. L0230032

Printed By Artron Art (Group) Co., Ltd.
Address 19 Shenyun Road, Nanshan
District, Shenzhen, China

Printed On May 10, 2024
Copies Printed 3000

Copyright © Huawei Technologies Co., Ltd.
2024. All rights reserved.

No part of this document may be
reproduced or transmitted in any form or
by any means without prior written consent
from Huawei Technologies Co., Ltd.

NO WARRANTY
The contents of this document are for
information purpose only, and provided "as
is". Except as required by applicable laws,
no warranties of any kind, either express
or implied, including but not limited to, the
implied warranties of merchantability and
fitness for a particular purpose, are made in
relation to the contents of this document. To
the maximum extent permitted by applicable
law, in no case shall Huawei Technologies
Co., Ltd. be liable for any special, incidental,
indirect, or consequential damages, or lost
profits, business, revenue, data, goodwill
or anticipated savings arising from or in
connection with any use of this document.

This publication is printed using eco-friendly
paper and ink.

Business Success

Embracing 5.5G to Unleash Industrial Dividends 08

Liu Kang
President of the ICT Marketing
& Solution Sales Dept, Huawei

Ubiquitous Fiber Networks with Huawei ODN 3.0 14

Zhao Maiqing
CTO, Home Broadband Solutions, Huawei

5G & Rail: The Road to the National Championship 30

He Tao
Senior Consulting Expert and Marketing
Expert, ICT Sales & Service Dept, Huawei

Network Evolution

Middle East: Leading the 5.5G Era and Striding Towards an Intelligent World 42

Tang Zhentian
Director, Middle East & Central Asia ICT
Marketing & Solution Sales Dept, Huawei

Future All-optical Network Architecture and Key Technologies 50

Tang Xiaojun
Chief Technology Planner, Optical
Business Product Line, Huawei

APN6 Enables Innovation in Cloud-Network-Edge-Device Collaboration 58

Li Zhenbin
Chief Protocol Expert of Huawei and Former
Member of the Internet Architecture Board (IAB)
of the Internet Engineering Task Force (IETF)

AI Data Lake: Breaking Silos and Accelerating Intelligence 68

Michael Fan
Director, Data Storage Marketing
Execution Dept, Huawei

Intelligent Transformation

Accelerating Intelligent ICT with a Four-layer Framework and Transformation in Three Areas 78

Wang Su
Director, Integrated Solution Marketing Dept, Huawei

EM 2.0: A Road for the Digital Intelligent Transformation of African Carriers 86

Chris Meng
Director, Northern Africa ICT Marketing & Solution Sales Dept, Huawei
Wang Jie
Director, Northern Africa Marketing Dept, Huawei
Liu Jifan
CEO, Huawei Ethiopia

Accelerating Intelligent Transformation with Application-driven Industry Collaboration 96

Li Changwei
Chief Strategic Marketing Expert, Huawei

AI-ready Cloud Drives Intelligent Digital Transformation for Carriers 106

Wang Xiaobin
Chief Architect, ICT Computing Products and
Solutions (Carrier Domain), Huawei

Innovation for the Future

Jointly Defining ICT Architecture for 5.5G: The Key to Unlocking New Opportunities 116

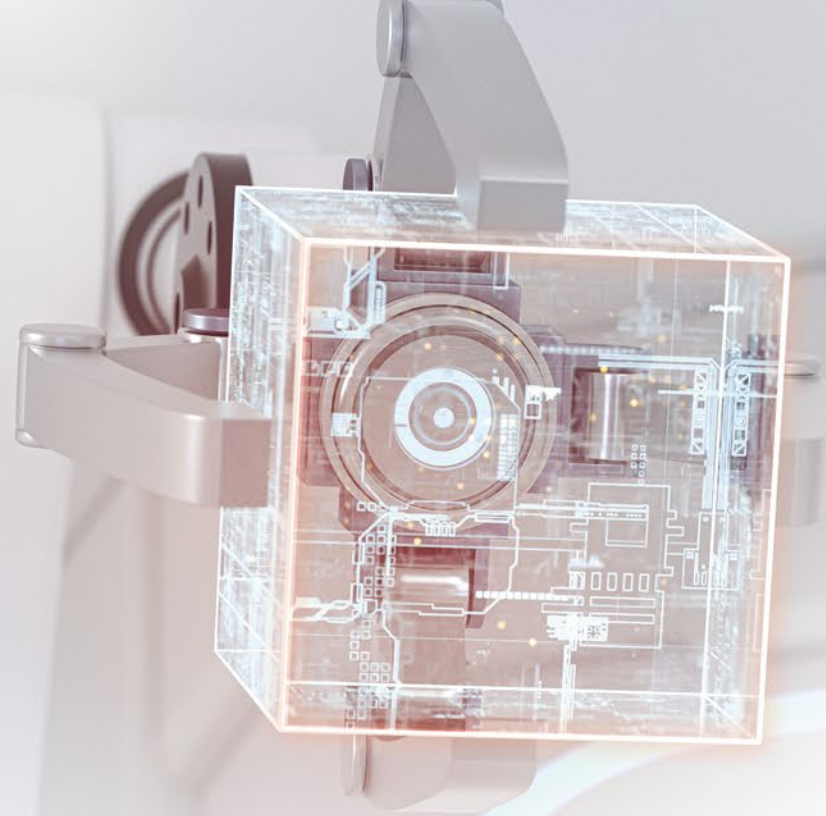
Dang Wenshuan
Chief Strategy Architect, Huawei

AI: The Bridge to 6G 124

Tong Wen
Huawei Fellow, CTO of Huawei Wireless
Zhu Peiyong
Huawei Fellow, Senior Vice President of Wireless Research, Huawei
Ma Jianglei
Technical Vice President of Wireless Research, Huawei
Chen Yan
Senior Expert of Wireless Research, Huawei



01.
Business
Success



Embracing 5.5G to Unleash Industrial Dividends



Liu Kang

President of the ICT Marketing & Solution Sales Dept, Huawei

5.5G is the natural evolution of 5G. With a tenfold increase in network capabilities, 5.5G is becoming hugely popular among industries. At the 5G Advanced: Completing the Enterprise Opportunity forum during MWC Barcelona 2024, Huawei proposed that 5.5G will enable industries to accelerate both digital and intelligent transformation.

5G has achieved significant success since its commercial use. By the end of 2023, 294 commercial 5G networks had been deployed worldwide, serving more than 1.4 billion 5G subscribers and establishing 5G as a major engine for carriers' revenue growth. In 2023, 90% of carriers that invested in 5G saw revenue growth and achieved positive business cycles. 5G has helped industries go digital, and has been used at scale in more than 10 high-value scenarios, including 5G-enabled machine vision and 5G-enabled remote control. 5G has also become a new engine driving the development of the digital economy.

The industry is undergoing tremendous changes, making it essential to upgrade connectivity technologies. The rapid development of generative AI presents huge opportunities, driving the exponential growth of traffic.

In the B2B domain, 5G has been integrated with core production activities and glasses-free 3D technology has reached an inflection point in terms of experience. The emergence of new services, connections, and experience is bringing unprecedented opportunities, driving leading carriers to shift from traffic monetization to the monetization of deterministic experience and convergent new services. These changes place higher requirements on the capabilities of existing networks.

5.5G is the natural evolution of 5G. 5.5G delivers higher speeds, lower latency, higher network reliability, wider connectivity, and native intelligence, allowing it to meet the higher requirements of communications networks in the future.

5.5G supports multi-carrier aggregation, delivering a 10-Gbps downlink rate,

1-Gbps uplink rate, and a 10-fold improvement in network capabilities

Evolution towards 5.5G is a continuous process. In the first phase, TDD 3CC aggregation above 260 MHz is able to achieve a peak rate of at least 5 Gbps. This has enabled wireless networks to deliver a deterministic experience for people-connected services for the first time. In the second phase, a downlink peak rate of 10 Gbps can be achieved.

5.5G networks that deliver a deterministic experience will significantly improve consumer perception. Voice and 2D videos used to be the main methods of content interaction, but immersive 3D videos, XR, and holography are developing into the major forms of future content interaction. This upgraded experience will raise the bar for connectivity. Previously, services have generally required a latency of 100 ms, but the latency required for future services will be as low as 10 to 20 ms. This tenfold decrease in latency means that 5.5G networks enabled by 3CC aggregation will be integral to network deployment in order to provide a better experience. The high bandwidth delivered by 3CC aggregation will give full play to its strengths in transportation hubs, such as high-speed railway stations, subway stations, and airports, and venues like commercial offices, stadiums, tourist attractions, and urban villages.

In the home market, FWA will continue evolving towards FWA2, leading to the emergence of three new application scenarios. First, FWA Pro can deliver a bandwidth of 1-2 Gbps, meeting the demand for high-speed services while providing a deterministic experience. This has allowed home users to enjoy better services, including ultra-HD videos and VR applications. Second, the cost-effective FWA Lite is primarily intended for

markets evolving from 4G to 5G and new markets without connectivity. This solution will help carriers obtain new users, but help release 4G spectrum resources to 5G as home users migrate from 4G to 5G, thus improving network efficiency. Third, FWA Biz is a solution for enterprises across different industries. Built on FWA, the solution can achieve 99.99% network stability and 20-ms low latency. This has made it easier for carriers to replace microwave private lines and low-speed copper private lines, accelerating the interconnectivity of small- and medium-sized enterprises.

With 5.5G, carriers can monetize network capabilities across more dimensions and provide differentiated packages based on deterministic experience. Carriers are now able to meet diversified customer demands and increase their ARPU.

5.5G enables native intelligence, meaning networks will be capable of self-optimization and self-management to better adapt to different scenarios and service requirements

As wireless networks continue to evolve, they will be able to deliver a greater range of services. Networks will also inevitably become more complex, bringing challenges for carriers in terms of offering simplified O&M and lower energy consumption while delivering diversified services and guaranteed service experience. The 5.5G intelligent solution, IntelligentRAN, will help carriers build autonomous driving networks that feature intelligent and simplified O&M, intelligent network optimization, and intelligent service operations.

- **5.5G-enabled intelligent and simplified O&M:** The solution provides capabilities like wireless FME copilots and key service assurance agents

based on the telecom foundation model. These facilitate wireless intelligent alarm management, accurate alarm identification, quick fault location, and fault prediction and prevention. With this solution, carriers can shift from responsive O&M to predictive and preventive O&M, achieving "zero" network faults.

- **5.5G-enabled intelligent network optimization:** Intelligent NEs have been introduced to implement intelligent resource scheduling, optimize the experience and capacity of multi-band and multi-site heterogeneous networks, and maximize spectrum efficiency. Features such as topology, time sequence, and grids have also been integrated to create a unified and general performance prediction model, resulting in a lower number of accumulated errors of multi-objective serial prediction, and the accurate and fast prediction of multiple KPIs. By employing the solution, carriers can realize optimal network performance and energy efficiency, meaning carriers can maximize energy saving results while guaranteeing stable network performance.
- **5.5G-enabled intelligent service operations:** User-level dynamic simulations help implement precise network planning based on coverage, speed, and latency, and enable fast service provisioning, which can meet differentiated service SLA requirements. Real-time, dynamic resource scheduling can be achieved based on prediction capabilities guaranteeing smooth service experience and service-based network adaptation when faced with challenging scenarios such as service fluctuations, burst traffic, and large events.

5.5G marks the first time that communications capabilities have been expanded from connectivity to integrated sensing and communications, opening up new market opportunities

- 5.5G marks the first time that integrated sensing and communications have been possible, meaning that 5.5G can deliver both communications and sensing capabilities. The large-scale deployment of commercial 5.5G networks can deliver full-area, grid-based, and low-cost detection capabilities. Compared with traditional radars, 5.5G networks have the following advantages:
 - **Spectrum sharing:** The allocation of 10% of spectrum resources for sensing and 90% for communications maximizes spectrum resources.
 - **Base station sharing:** A wide-area native sensing network can be built based on millions of base stations.
 - **Low cost:** Communications can share most deployment costs.
 - **Network collaboration:** The blind spots of traditional radars can be overcome by multi-site collaboration.
- 5.5G can be applied in a variety of sensing scenarios, including low-altitude unauthorized drone flight warnings and route supervision; offshore sea-surface monitoring; inland waterway management; vehicle and pedestrian sensing; the sensing of high-speed railways and related perimeter intrusions; border intrusions; thrown objects; meteorological monitoring and predictions; the monitoring of micro-deformation of buildings, bridges, and mountains; and crop monitoring.
- Based on the native sensing and communications capabilities of 5.5G networks, carriers can fully utilize the strengths of cloud, computing, and connectivity to fill in the gaps of industry capabilities. They can also provide new and better key solutions to avoid homogeneous competition, greatly expanding carriers' business boundaries and creating new business opportunities.

5.5G improves 5G IoT capabilities and enables IoT across various scenarios, making all things connected possible

5.5G drives IoT development, as it is capable of breaking the limits of 5G in some scenarios. 5.5G also improves network performance and connectivity and reduces costs, meaning that IoT-related goals that cannot be achieved with 5G become achievable. This makes all things connected possible, enabling carriers to maximize the commercial value of 5.5G.

- 5G RedCap and 5.5G RedCap intrinsically support 5G capabilities like large uplink rates, positioning, and network slicing. These lightweight solutions provide medium- and high-speed connections that meet the high SLA requirements of industry unified communications, such as smart power distribution, smart city monitoring, and industrial sensing, and cut deployment and usage costs.
- The RedCap industry ecosystem is rapidly maturing. In 2023, eight RedCap modules and more than 30 types of RedCap devices (DTUs, cameras, and CPE) were launched worldwide. It is estimated that over 100 types of RedCap devices will be launched for commercial use in 2024, and the average cost of a RedCap module is expected to drop from US\$30 to US\$20 this year. Furthermore, the average cost of RedCap-based data collection devices has decreased to just US\$150. These costs will continue to fall as millions of RedCap devices start to be shipped from China.
- 5.5G P-IoT features large bandwidth, continuous networking, long-distance coverage, and low power consumption, allowing it to support IoT service scenarios that require speeds of

10 Kbps or lower. Such scenarios include warehousing, production and manufacturing, logistics, and remote monitoring.

5.5G further unleashes industry productivity and supports more industry scenarios, creating greater economic value

5.5G is expanding private network services from industry private lines only to private lines and ICT services, thus accelerating industry digital transformation. The 5G private network and private line market is developing fast and is set to reach US\$20 billion by 2025. 5G private lines are extensively used across industries thanks to fast, low-cost deployment. 5.5G can further improve the network performance of private networks, including guaranteed 300 Mbps and millisecond-level latency. 5.5G can also be integrated with networking, computing, cloud, IoT, and other industry requirements to open up more possibilities for industry digital transformation. 5.5G supports an uplink transmission rate of 1 Gbps for a single user. This can meet the requirements of interactive, immersive services such as AI training data cloudification, cloud photos, and cloud conferencing. In industrial production, 5.5G can enable services such as AI quality inspections, security monitoring, and remote control. And in healthcare, the low latency and high reliability of 5.5G can make telemedicine applications possible.

With the enhancement of 5.5G capabilities, related technical applications are expanding from enterprises' auxiliary systems to core production systems. 5G serves as an enabler and accelerator in the digital transformation of various industries. For example, 5G has been applied in more than 400 mines in countries around the world, including China, Thailand, Brazil, and South

Africa, with more than 20 carriers now providing 5G services for mines. 5.5G technologies bring new momentum to the digital upgrade of industries, with 5.5G supporting ultra-large uplink rate of 1 Gbps and latency as low as 4 ms. These features can support enterprises' core production processes such as high-density quality inspection and flexible production, greatly enhancing the productivity of enterprises. In the mining industry, for example, the length of a fully-mechanized mining surface in a coal mine ranges from 100 to 300 meters. Previously, it was difficult to clearly see the entire surface. However, 5.5G wireless access technology can be employed to provide network coverage for underground mining surfaces and areas in which explosions are set off, solving issues related to fiber disconnection and data collection failures. Low-frequency and large-bandwidth resources are used to backhaul videos from more than 100 cameras, and AI is used to stitch these videos into a panoramic video of the mining surface, providing a comprehensive and clear picture of underground operations.

5.5G expands IoV applications from in-vehicle entertainment to vehicle-road synergy, accelerating the arrival of smart transportation

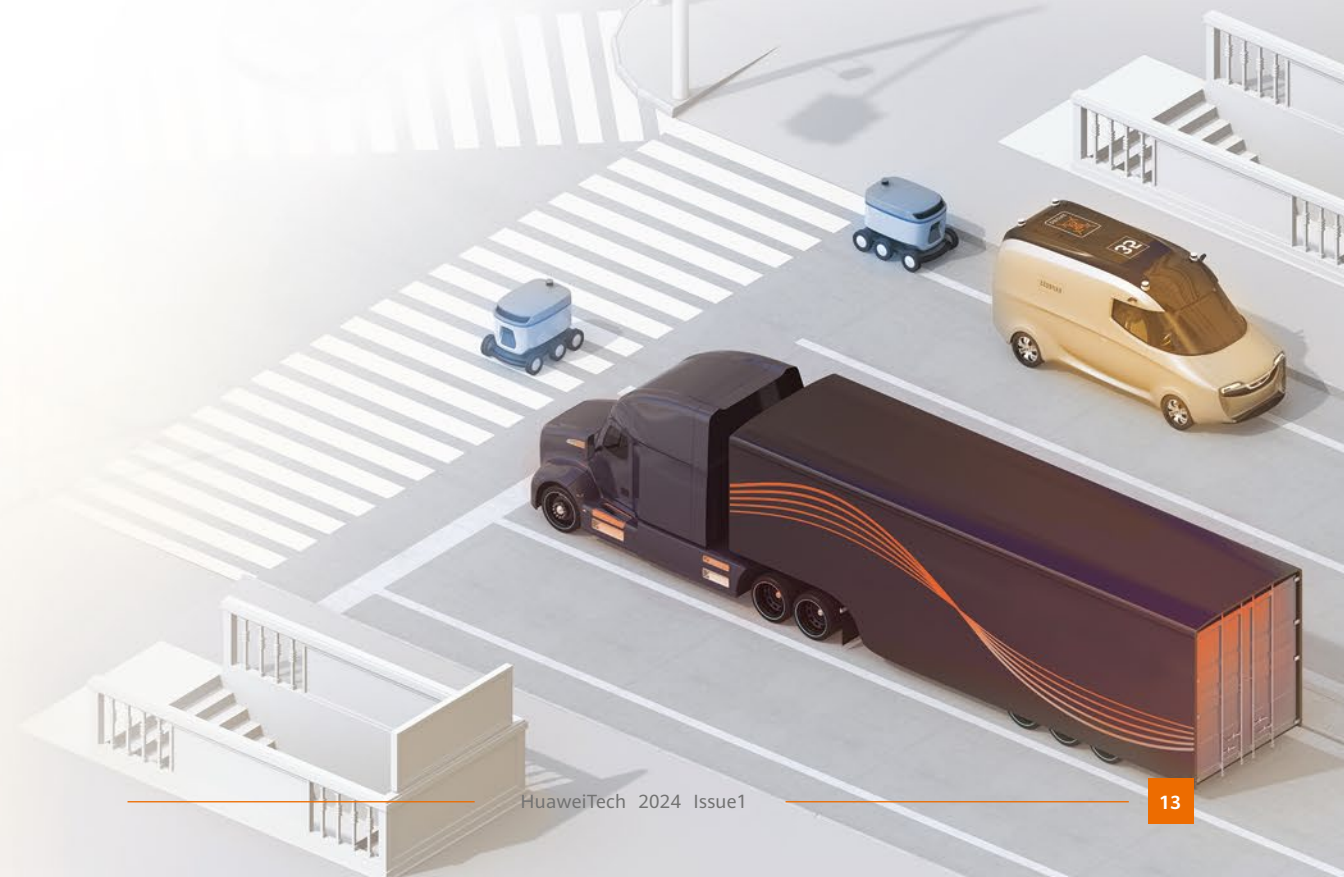
As the automotive industry becomes more autonomous and electric, the number of connected vehicles worldwide is soaring. By the end of 2023, the number stood at 350 million, and 45% of all vehicles on sale today support autonomous driving at L2 or higher. These trends will see the traffic of connected vehicles soar by 100 times, and the data of usage (DOU) in vehicles to exceed 100 GB. These developments will require the support of high-speed 5G networks. To ensure transportation safety and

improve efficiency, vehicle-road synergy requires networks with 99% reliability and 20-ms latency. This is where 5.5G comes in. New E2E technologies have already been verified in Shanghai, where a 1-km demonstration route boasts reliable 5.5G networks with stable latency. With vehicle-road synergy supported by 5.5G's sensing capabilities, drivers no longer need to worry about complex traffic situations, while accidents caused by blind spots and other unexpected factors can be prevented. It is forecast that traffic management efficiency will improve by 30% and average commute times will be 20% faster.

Global carriers have reached an industry-wide consensus and are actively deploying 5.5G. During the Global Mobile Broadband Forum in October 2023, 13 world-leading carriers jointly announced the launch of the Global 5.5G Network Pioneers, marking 5.5G's transition from technical verification to commercial deployment. At the 17th Telecom Review

Leaders Summit in December 2023, experts and representatives from authoritative organizations and enterprises jointly announced the first year of 5.5G deployment. These organizations include ITU, du and e& (UAE), Ooredoo (Qatar), Vodafone Oman, Huawei, Ericsson, and Nokia. Over the past year, leading carriers from the Middle East, Europe, and Asia Pacific (including China), have fully verified TDD 3CC on live networks and started commercial deployment.

Close industry collaboration and ecosystem collaboration are also key to unleashing the value of 5.5G. Huawei stands ready to work with industry and ecosystem partners to explore its extensive applications and provide end-to-end intelligent 5.5G solutions. Together, let's usher in a new era of 5.5G-enabled intelligent connectivity and applications so that all participants will benefit from the development of 5.5G and achieve shared success. T



Ubiquitous Fiber Networks with Huawei ODN 3.0



Zhao Maiqing

CTO, Home Broadband Solutions, Huawei

With Huawei's core concept for ODN construction centering on full and dense coverage coupled with short and easy access, Huawei's ODN 3.0 solution uses two transformative technologies to support five typical network scenarios.

Two key transformative technologies in ODN 3.0

In the earliest FTTH solution, ODN 1.0 optical splitting was used for optical splitters, while fusion splicing or mechanical splicing was reserved for fiber connections. In 2015, some vendors implemented drop cable pre-connection by connecting fiber drop cables to fiber access terminals (FATs), forming the ODN 2.0 solution. Since 2018, based on ODN 2.0, Huawei has gradually realized pre-connection between distribution optical cables and level-2 optical splitters, uneven optical splitting of level-2 optical splitter FATs, and pre-connection between fiber feeder cables and level-1 optical splitters. This has resulted in a comprehensive solution that implements full pre-connection, cascading, and uneven optical splitting technologies, culminating in the ODN 3.0 solution.

1. Uneven optical splitters: Efficient connections and reduced costs

An uneven optical splitter (as shown in Figure 1) unevenly splits 100% of optical power signals from COs, with 70% of output allocated to remote ends and 30% retained for local users, and then evenly splits the portion of optical power between local users. Although only 30% of optical power is retained for local users, it is sufficient to meet local users' needs.

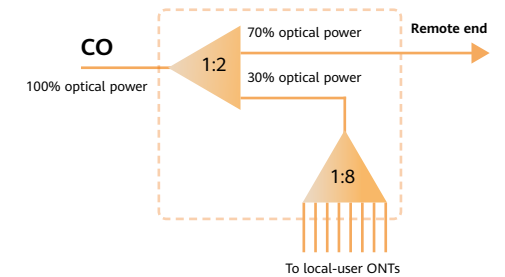


Figure 1: Uneven optical splitter

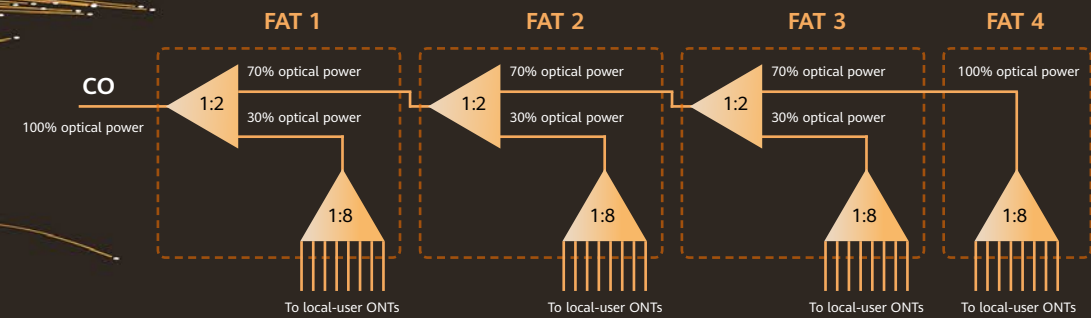
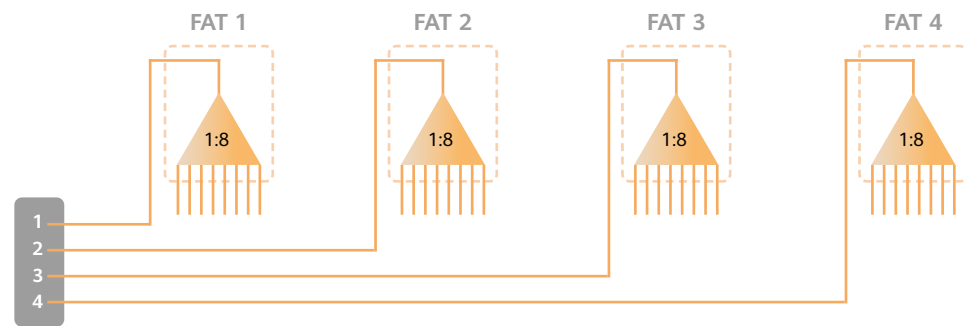


Figure 2: Four-level cascade with uneven optical splitting

Several uneven optical splitters are usually cascaded together. For example, for four-level cascading, as shown in Figure 2, the first three splitters perform uneven optical splitting, and the fourth performs even optical splitting. This means 100% of the optical power from the CO is retained for FAT 4 users. Although the third and fourth splitters receive only a relatively small amount of CO optical power, it is still sufficient to meet the needs of the local users.

Uneven optical splitters are advantageous in that they significantly reduce the number of optical cables required. Only a single-core optical cable is needed between the four optical splitters and between the optical splitters and the CO. By contrast, an even optical splitting solution would require four optical cables, as shown in Figure 3.



Generally, in an ODN project, costs related to optical splitter material and installation account for less than 15% of total investment, while 85% of investment relates to optical fibers.

Therefore, uneven optical splitting is an advanced and revolutionary development for FTTH network solutions.

2. Pre-connection: Reduced faults for ODNs

An FTTH ODN is a network of optical fibers connected to different devices, such as optical splitters, FATs, and optical cable junction boxes. One way to connect these fibers is using fiber fusion splicers onsite, but fiber splicing requires highly skilled technicians and can be very difficult if operation sites are unfavorable.

Another way is mechanical splicing, which is simpler than fiber splicing. However, poor connection quality can result in faults or excessive signal attenuation within just three years. Furthermore, both methods require device cover plates to be removed from boxes, meaning exposure to the weather and compromised protection of components over time. Most ODN faults actually occur at these junctions. A typical example is tail-end optical splitters inside FATs. Frequent removal of FAT cover plates during onsite installation and repairs often leads to faults over time from compromised protection.

Pre-connection (Figure 4) means fiber sockets are reserved outside an FAT, and a fiber plug is reserved in the factory, so that the plug can easily be fixed to the socket at the construction site. Pre-connection ensures that FATs never need to be uncovered. Of course, reliable pre-connection and adequate optical attenuation requires firm conjunctions, meaning both the sockets on devices like FATs and the pre-connection plugs of optical fibers must strictly be manufactured to a high standard.

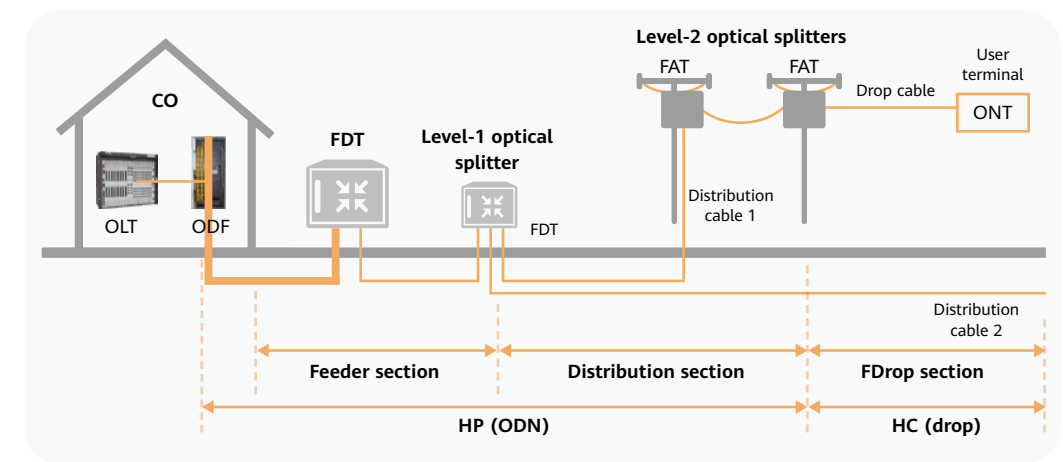


Figure 5: FTTH network

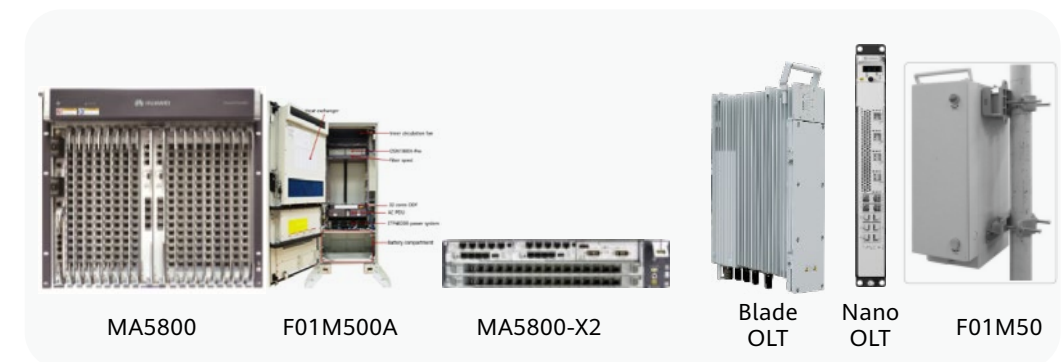


Figure 6: Different types of OLTs



Figure 4: Pre-connection devices

OLT options in FTTH networks

An FTTH network (Figure 5) consists of three parts: OLT, ODN, and drop cable segment. Generally, the ODN part is referred to as home pass (HP), and the drop part is referred to as home connection (HC). HP consists of the feeder section and distribution cable section. Several types of OLT can be used depending on the actual scenario (Figure 6).

1. Indoor OLT

Applicable to COs or indoor base stations, each OLT can typically carry 4,000 to 6,000 users and each OLT rack can house two or three OLTs. Generally, the coverage radius, that is, the distance between an OLT and the farthest

optical network terminal (ONT), should be a maximum of 6 kilometers.

2. Outdoor OLT

When no indoor equipment room is available, an outdoor OLT cabinet is typically installed in an open outdoor space, preferably where base stations are located. A dedicated concrete base will be built in advance for such OLTs to keep water out. Each outdoor cabinet houses only one OLT, typically covering 4,000 to 6,000 users. WDM and IP equipment can also be installed in outdoor cabinets as required. Power supply and storage batteries are required in the cabinets, while natural air cooling, fan-based air cooling, or dedicated air-conditioning can be selected based on local temperatures.

3. Compact OLT

- **Blade OLT:** OLTs of this kind are compact and weigh about 13 kilograms. These OLTs can be installed on a pole or in an open space within a base station. They can carry 1,000 FTTH users each, or 2,000 FTTH users when two units are installed back to back and share two uplink optical fibers to the CO.
- **MA5800-X2:** This OLT model can be installed inside a mini outdoor cabinet which is then fixed at a base station or street cabinet to support up to 2,000 users per unit.
- **Nano OLT:** This type of OLT provides only four GPON ports and can carry 256 broadband users. It is smaller and lighter than other OLTs, and can be installed inside a mini outdoor cabinet which is then fixed outdoors. Nano OLTs are particularly suitable for rural areas, as they are an extraordinarily low-cost solution for areas with small populations and low household density.

mean sparser coverage and longer access lines, leading to repeated construction that wastes both time and investment. Generally, a 30% take-up rate within three years after an FTTH project is completed is already considered fair performance. For a large-scale, multi-area FTTH network construction project, a total take-up rate of over 50% within five years of project completion would be considered impressive performance. Some carriers require their local business units to achieve a take-up rate of more than 70%. This is incredibly difficult to achieve when they are developing the business at scale.

$$\text{Household Penetration Rate of Ports} = \frac{\text{Number of FAT Ports (in an area)}}{\text{Number of Households (in an area)}}$$

Household penetration rate of ports means the number of optical fiber ports to be deployed for every 100 households in an area. Carriers that emphasize lower CAPEX will reduce the number of ports planned for the first phase of construction projects, resulting in low household port penetration rate.

Core concepts in ODN design

1. Another look at two carrier KPIs

$$\text{Port Take-up Rate} = \frac{\text{Number of Subscribers (in an area)}}{\text{Number of FAT Ports (in an area)}}$$

Traditionally, carrier performance is evaluated according to take-up rate. This is problematic because carriers tend to require high port-take-up rate soon after project acceptance. Setting a high port-take-up rate as a KPI for the first two years after project completion will inevitably result in fewer ports being planned for the initial phase of the project. This will

The CAPEX of an FTTH project typically comprises construction labor (60%), optical cable (25%), devices like optical splitters (15%), and minor costs such as design and supervision. Labor cost clearly accounts for the majority of the total cost. A low household penetration rate of ports will mean two, three, or even more phases of construction are required over just one or two years, with each phase resulting in basically the same labor cost. Three more phases of construction mean three times the required investment compared with everything being done at once.

Uneven optical splitting is an advanced and revolutionary development for FTTH network solutions.



Therefore, a 70% or higher household penetration rate of ports is recommended to be achieved following the very first phase of FTTH construction.

2. Full and dense coverage, short and easy access

- **Full coverage: Key to one-time holistic planning**

One-time holistic planning should be performed for small- and medium-sized cities to achieve full coverage in one deployment, without requiring secondary design and construction. The FTTH construction project of a city should have as many areas as possible contiguously and fully covered, again in one deployment. This is especially true for traditional fixed network carriers, because they need to migrate existing xDSL users across all areas to FTTH.

Full coverage is required for user volume to ramp up across an area or city and for carriers to build their brand influence, because existing users will pass on their experience to potential new users through word of mouth.

By contrast, interspersed coverage may result in consumer complaints and negative market effects in non-covered areas. This may encourage competitors to fill coverage vacancies through repeated construction and compete for users.

- **Dense coverage: Improving return on investment and market confidence**

Dense coverage means a 60% or higher household penetration rate of ports. This reflects carriers' expectations for and confidence in the market.

Dense coverage can improve ROI, shorten project duration, and ensure high network operations quality, while enhancing confidence in market development, reducing management costs, and reducing competitors' market expectations.

A 60%–100% household penetration rate of ports is recommended, as the number of FAT ports in an FTTH network is a rational standard for measuring a carrier's network capabilities, which represents medium- and long-term market development potential. This is intended to support development over at least the next five years, and carriers should not expect a high take-up rate in the first year of project completion.

- **Short access: Improving FTTH installation efficiency**

For areas with average resident density, 8-port FATs are recommended. For areas with low resident density, 4-port or 2-port FATs can be used. The average access distance between an FAT and a household ranges from 30 meters to 80 meters.

Shorter access ensures efficient installation. Broadband installations per capita per day is a major factor affecting the scale of carriers' user development and total HC installation costs. FTTH networks are a key foundation for developing broadband

users at scale. High-quality network construction will mean efficient device installation, low fault rate, good user experience, and a good service brand.

- **Easy access: Quick installation and troubleshooting**

The end-to-end pre-connected FAT solution eliminates the need for high-altitude fiber splicing while decreasing the failure rate caused by fiber clamping. When a FAT is faulty, it can simply be replaced. This ensures quick installation and troubleshooting for FATs.

3. Drawbacks of sparse coverage and long access

Sparse coverage and long access are the main reasons for poor ROI for global carriers in FTTH deployment over the past decade or so. Repeated construction in the same area is the primary reason why some carriers fail to succeed in FTTH, as this wastes investment, time, and market opportunities.

- **Sparse coverage means wasted investment for secondary construction**

Sparse coverage in the initial phase

Huawei's core concept for ODN construction centers on full and dense coverage coupled with short and easy access.



of construction results in a huge waste of investment for a second phase of construction.

Figure 7 shows a typical case of secondary construction. The total investment required to achieve a 70%–90% household penetration rate of ports over two phases of construction is twice that required to achieve this over a single phase. Some carriers encounter a shortage of available ports and a budget crunch despite constant construction projects in virtually the same area year after year. Typically, this situation is primarily caused by the dogmatic practice of evaluating performance according to take-up rate, resulting in a shortage of upfront investment and subsequent repeated construction. Generally, during the second and third phases of construction, identifying idle fibers in fiber distribution terminals (FDTs) is more time-consuming than laying new optical cables. Some carriers end up having as many as three to five phases of construction in the same areas, wasting huge investment.

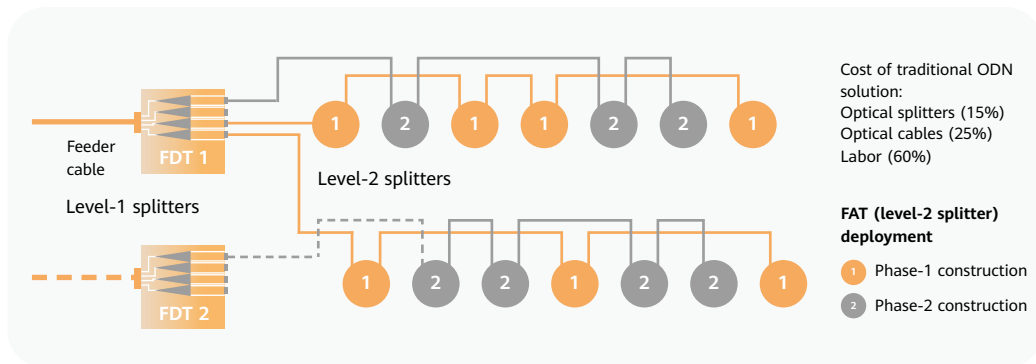


Figure 7: Investment wasted due to secondary construction

- **Long access leads to high installation costs**
 Figure 8 shows a typical case of long access. Due to a sparse coverage approach, users in the area are covered by only FAT 1, deployed during the first phase of construction. When users 4, 10, and 11 subscribe to the broadband service, their access to an FAT can only be delivered over a long distance.

As FAT 1 ports become almost fully loaded, FAT 2 is installed through second-phase construction to address additional user needs. However, users 4, 10, and 11 who were connected to FAT 1 cannot be migrated to FAT 2.

When FAT 1 is fully loaded and no idle port is available, new users near FAT 1 (e.g., user 7) can only be connected to FAT 2. In this case, the HC installation distance is very long, significantly increasing installation costs.

Deploying the two FATs at the same time can prevent the long-access problem caused by sparse coverage and address the high

total installation cost resulting from high labor costs of HC installation. The criteria for evaluating FTTH network construction involves how many broadband lines an engineer can install per day. Therefore, it is recommended that the average distance between an FAT and a household be a maximum of 100 meters.

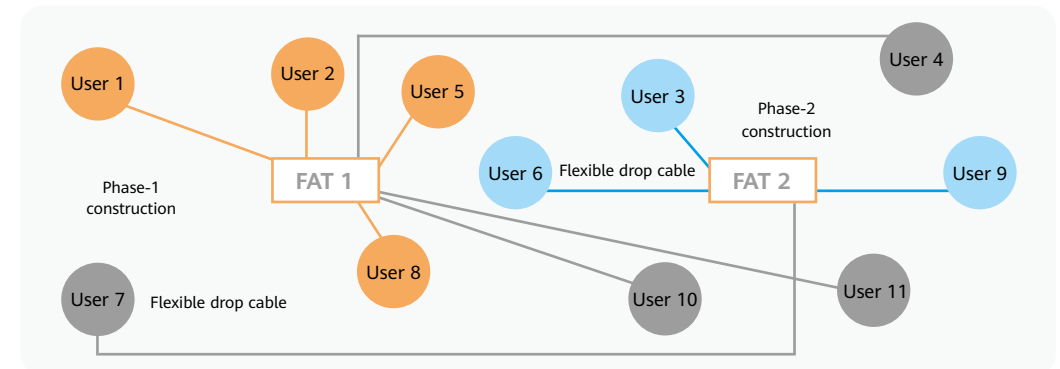
FATs should be located as close to consumers as possible. For example, an 8-port FAT, which can connect to eight households, should be located with the shortest average distance to the eight households it connects to.

The capacity of a hub box should first be identified to determine the maximum number of FATs that can be connected to it. For example, a hub box that can cover eight groups of FATs, with four FATs in each group, should be located with the shortest average distance to the eight groups of FATs it connects to.

Three keys to ODN design

1. Locations of level-1 splitters (hub boxes) and level-2 splitters (FATs)

✗ FAT 1 and FAT 2 separately deployed over two construction phases (sparse coverage, long access)



✓ Two FATs deployed at one time (dense coverage, short access)

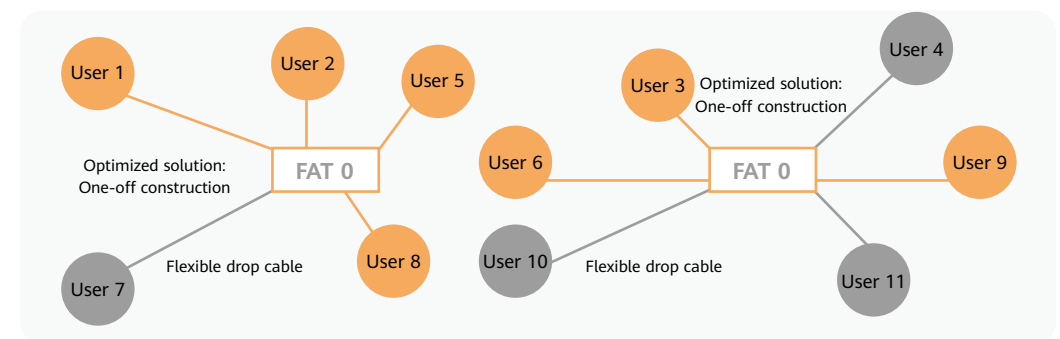


Figure 8: Long access increases the HC installation cost

Of course, the exact locations of FATs and hub boxes are the result of a comprehensive analysis of onsite conditions.

This is the only way to ensure the minimum number of optical cables used between hub boxes and a CO, as well as the shortest distances between hub boxes and FATs and between FATs and consumer homes. Ensuring this minimizes the optical cable investment

required in the feeder section, distribution cable section, and HC section.

Neither putting level-1 splitters inside COs nor installing multiple level-2 splitters in a street cabinet is advisable, as both approaches would result in unsuitably long distances between the FATs and consumers' homes — far more than 100 meters. Laying a large number of drop cables inside street cabinets and underground

pipes requires enormous investment and makes future maintenance more difficult.

The feeder cables between COs and level-1 splitters are considered scarce resources, so using the minimal number is recommended.

Given the fixed number of optical cables between level-1 and level-2 splitters and between level-2 splitters and ONTs, their lengths should be minimized.

2. Leveraging existing network resources

COs, wireless base stations, and existing FDTs already have the power supply and optical cable resources required for FTTH networks, such as outgoing optical cables from COs and redundant optical cables between base stations or optical nodes and COs. Deploying OLTs where these existing resources can be leveraged will help reduce the need to deploy new optical cables in the feeder section, which is typically long and comes with a large number of fiber cores, and will help reduce additional investment in deploying the power supply.

3. Replacing traditional ODNs with ODN 3.0

The technical design and construction of ODN 3.0 are simple and efficient, and the ODNs can run stably with low fault rates, reducing ODN maintenance costs. If the fault rate of carrier ODNs that are built earlier is too high to keep maintenance costs down, then carriers can consider replacing legacy ODNs through maintenance investment.

Overview of the ODN 3.0 solution

1. Solution for high-density, low-rise, overhead scenarios

As shown in Figure 9, this solution consists of three types of boxes (hub boxes, sub-boxes, and end boxes) and three types of optical cables (MPO cables, single-core distribution cables, and drop cables). Each hub box performs 1:2 optical splitting on the fiber cores of the four GPON ports and outputs eight links. Sub-boxes use uneven optical splitting and can consist of four cascaded FATs, with each connected to eight households. For areas that are far away from an OLT, a regular optical cable and an MPO X-box can be used. X-boxes can be directly connected to both 12-core pre-connected optical cables and regular optical cables for output. This solution uses end-to-end pre-connection, and is suitable for overhead-deployment scenarios, with high installation efficiency and low fault rates.

2. Solution for overhead deployment in low-density areas

In this solution, as shown in Figure 10, each hub box has a 1:2 optical splitter built in. FAT 1–FAT 3 use uneven optical splitting with eight ports, while FAT 4 uses even optical splitting. The three types of optical cables involved are 4-core pre-connected cables, single-core distribution cables, and single-core pre-connected drop cables. If the covered area is close to the CO, a 4-core pre-connected cable can connect the four GPON ports of the OLT with four hub boxes, thus supporting up to 256 households in total.

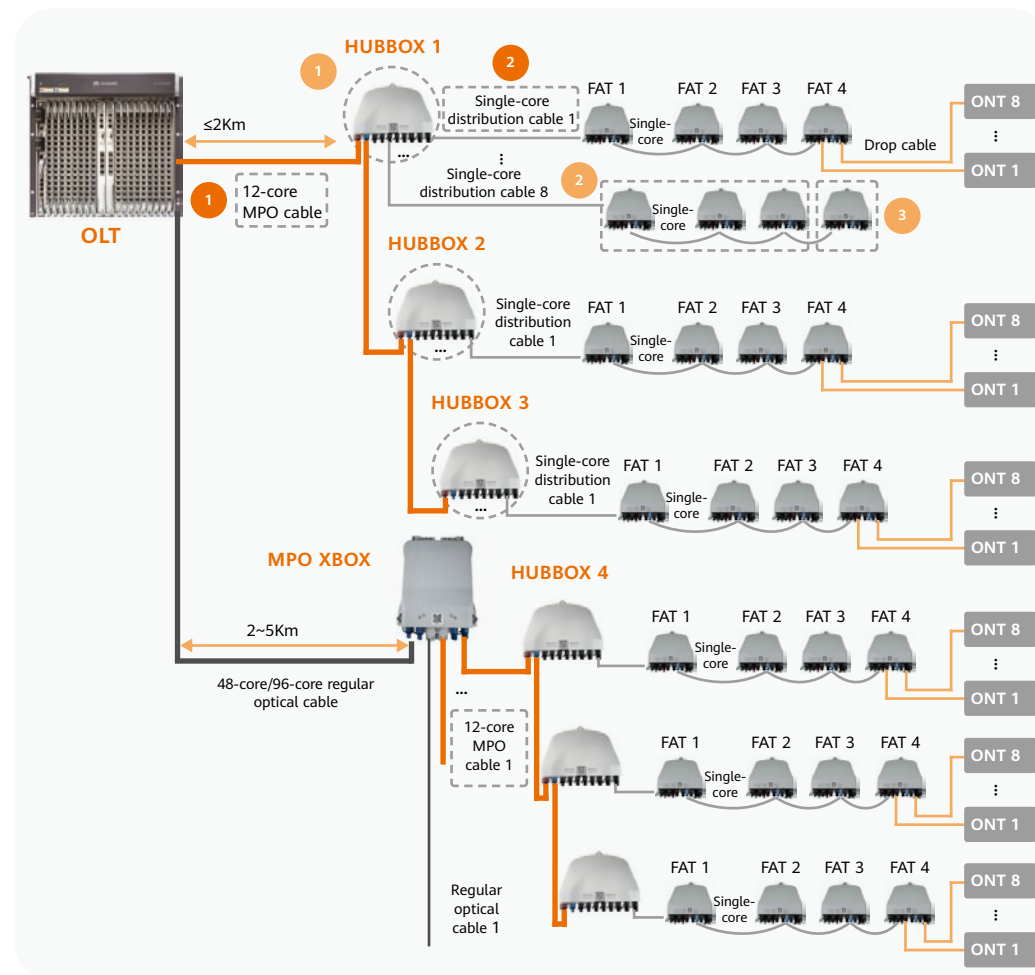


Figure 9: Solution for overhead deployment in high-density, low-rise facilities

3. Solution for multistory buildings

This solution, as shown in Figure 11, uses a traditional feeder cable connected to a hub box in the building. Each hub box has four built-in 1:2 optical splitters to output eight links, each

with with four cascaded FATs, covering up to 32 households in a building unit. With eight building units covered, where each unit has four stories and each story has eight households, up to 256 households can be supported in total.

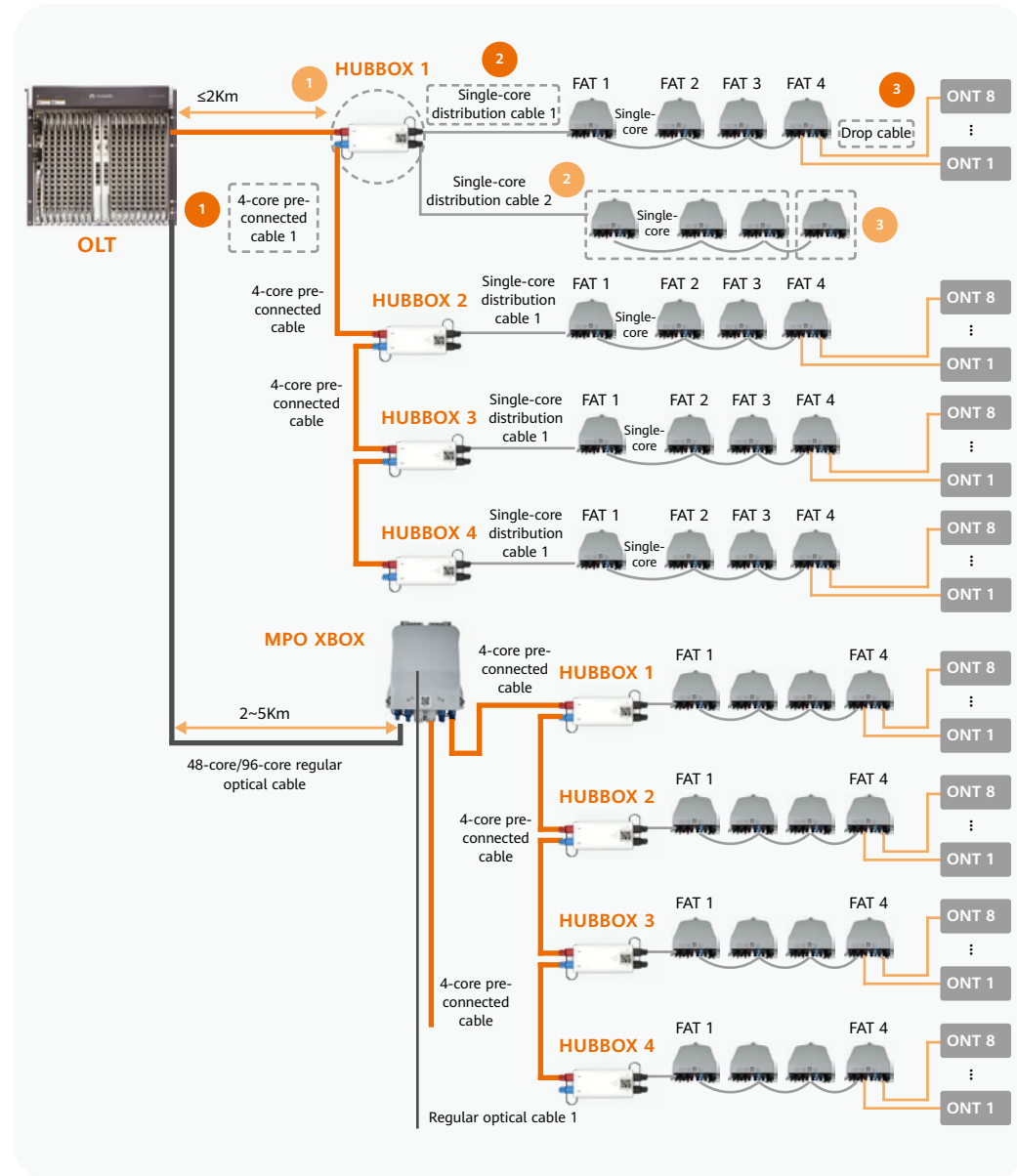


Figure 10: Solution for overhead deployment in low-density areas

4. Solution for underground deployment

In this scenario, it is recommended that level-1 optical splitters (hub boxes) are installed in street cabinets. If possible, the coverage area of level-1 splitters should be narrowed so that hub boxes are as close to the covered FATs as possible to shorten the distance between them.

The FATs shown in Figure 12 are level-2 splitters. It is recommended that these FATs are installed on walls inside manholes. 8-port or 4-port

FATs can be used depending on the household density in the area. Pre-connection is used between FATs and households, and OneTB's (underground) are deployed outside household doorways. Pre-connected cables used between FATs and OneTB's can be vertically deployed in the same trenches as distribution cables. In addition, protection pipes that lead all the way to household doorways can be added for these cables, where a trench is dug horizontally, and the cables are connected to OneTB's.

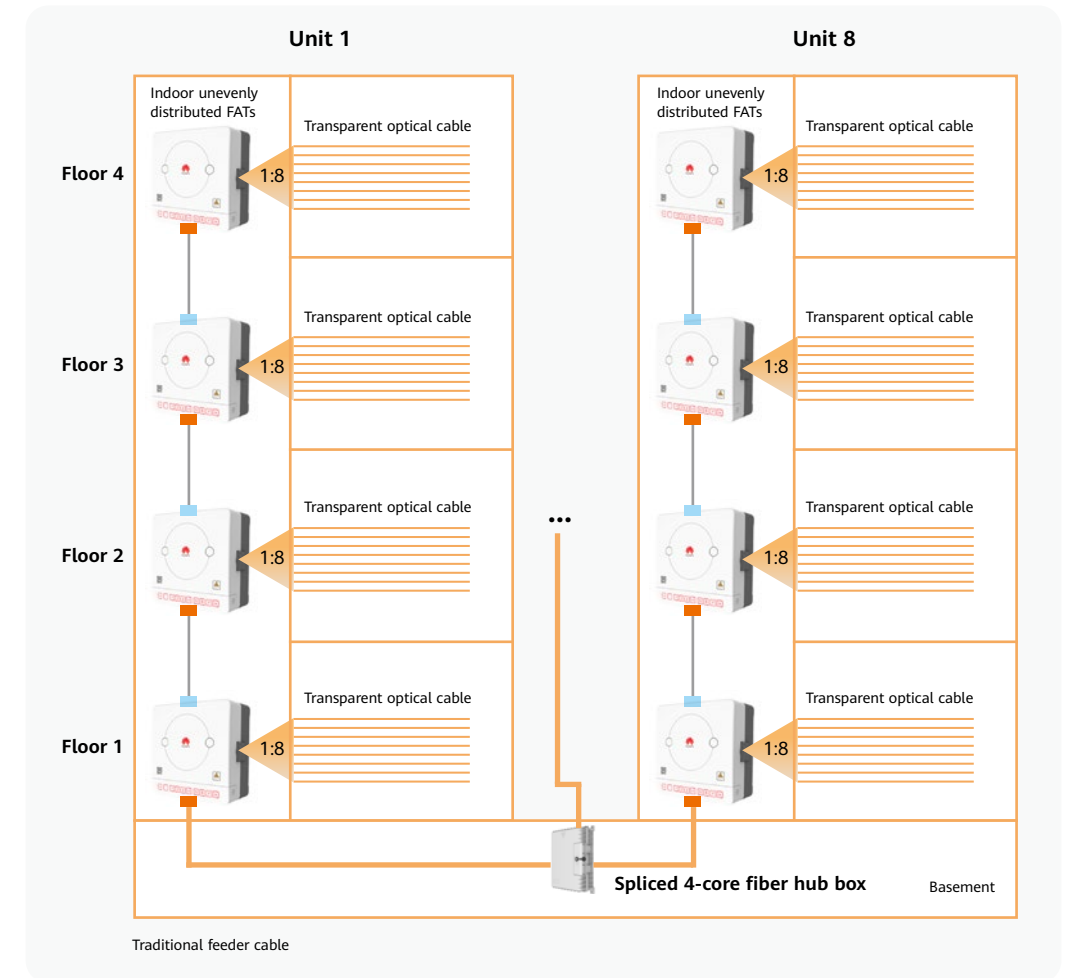


Figure 11: Solution for multistory buildings

Optical cables between hub boxes and FATs and between different FATs can be pulled through conduits or blown through microducts. The greatest benefit of this solution is that it does not require secondary ground excavation at household doorways. Of course, it also has other significant benefits, such as shortened HC length and distribution optical cable length, as mentioned previously.

5. Leveraging existing backbone cable resources

One solution is to reuse existing FDTs to expand the network (Figure 13).

The benefit of this solution is that it allows the utilization of existing feeder optical cables and FDTs, with pre-connection used on sections beyond the FDT, removing the need for fiber splicing. Fiber splicing needs to be performed only once at the FDT to connect the traditional optical cable with the 4-core pre-connected optical cable that is connected to FATs. Each FAT performs 1:8 splitting on an optical fiber to connect to eight households.

Another solution is to reuse existing backbone optical cables to develop B2B services and fiber to the site services (Figure 14).

The biggest problem with traditional ODNs is low optical cable utilization due to multi-core cables laid over long distances. Along the routes from the CO to optical splitters, there are multiple optical cable splicing points such as FDTs and optical closures. These cause high fault rates and optical attenuation, difficulties with maintenance and troubleshooting, and complex resource management.

The P2P box used in the ODN 3.0 solution can directly output dual-core optical cables to P2P private line users and mobile base stations. Therefore, this solution is particularly suited to supporting integrated carriers with deploying networks and fully leveraging their backbone optical cable resources.

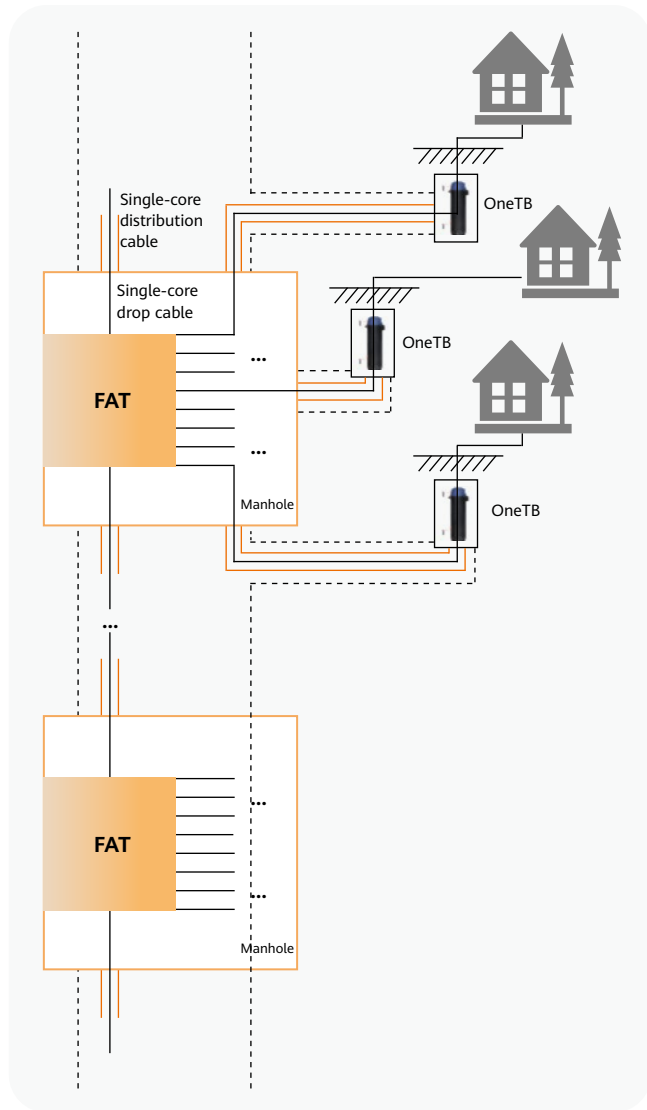


Figure 12: Solution for underground deployment

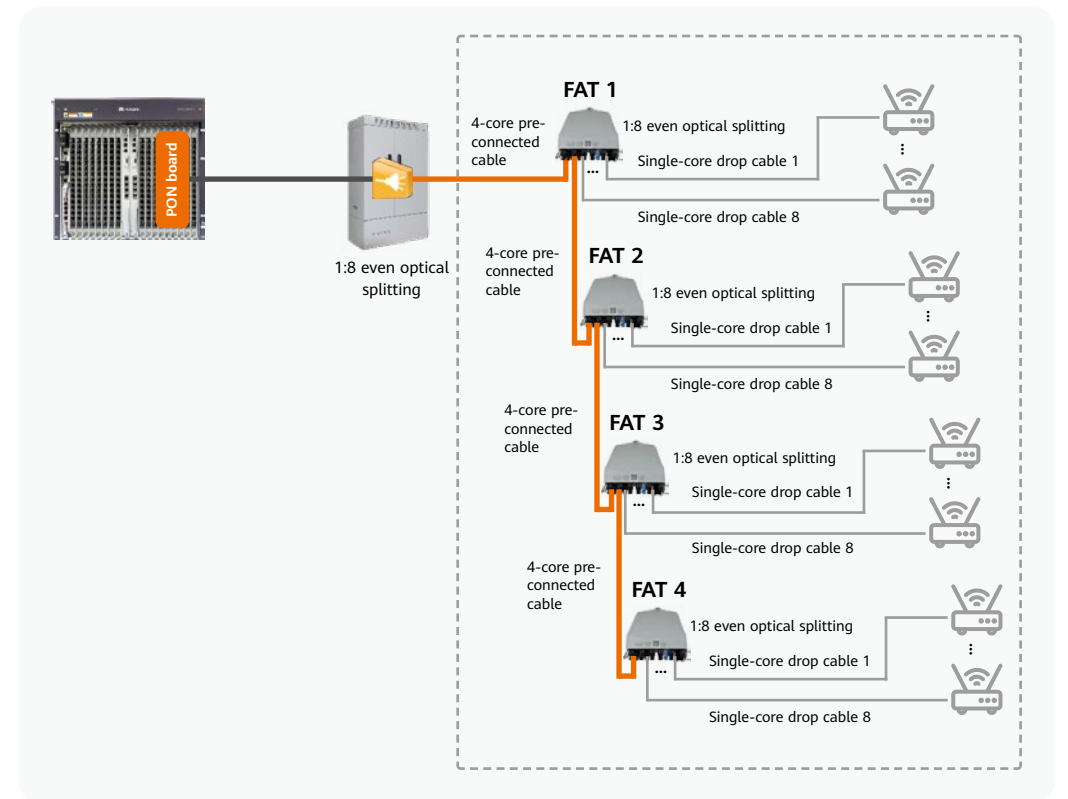


Figure 13: Reusing existing FDTs to expand the network

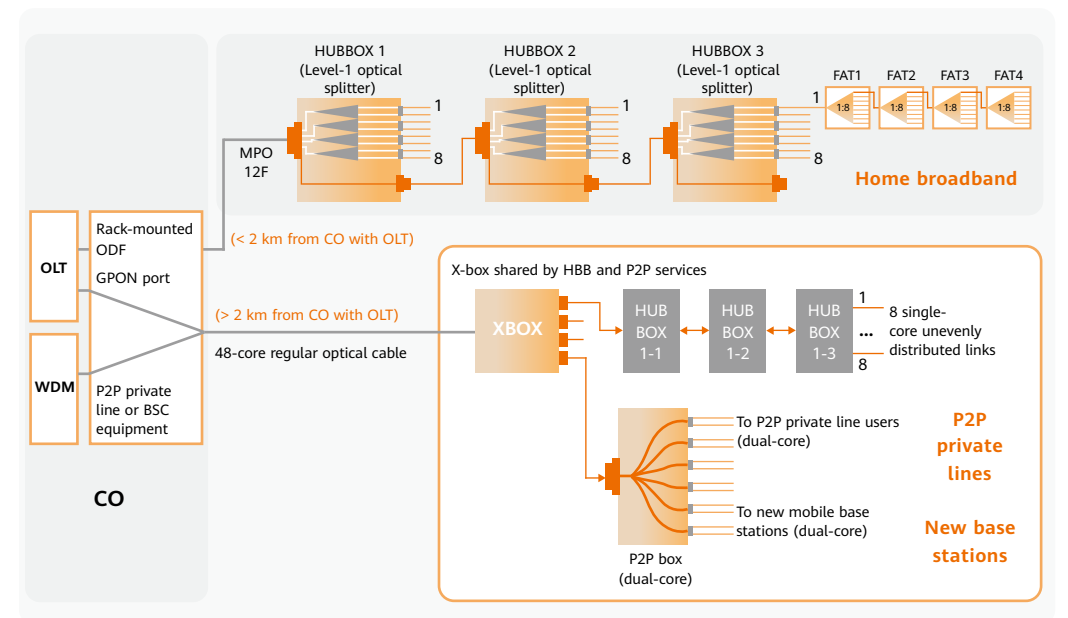


Figure 14: Reusing existing backbone optical cables



5G & Rail: The Road to the National Championship



He Tao

Senior Consulting Expert and Marketing Expert,
ICT Sales & Service Dept, Huawei

Wuhan Metro's 5G-powered Smart Urban Rail project recently won a national award for being one of China's most exciting 5G applications in 2023. The project now serves as a benchmark for helping industry customers to implement dedicated public 5G network projects.

On October 17, 2023, China's Ministry of Industry and Information Technology (MIIT) kicked off the final of their annual 5G application competition in Shanghai. This year's 6th "Blooming Cup" Competition challenged participants to think about how "5G Drives Digital and Real Economy Development" and find new solutions that would deeply integrate 5G into industrial digital transformation.

Of the 45,728 projects submitted from across China this year, Wuhan Metro's 5G-powered Smart Urban Rail project stood out and took first prize. The project's application of public 5G networks in rail transit attracted significant praise from state agencies, China's carriers, and enterprises in related fields, as it presented a novel, and more importantly, feasible, business model for public 5G networks.

This 5G-powered Smart Urban Rail project, while a success, had not progressed without its own challenges and setbacks. And so, the summary presented at the Blooming Cup interested many as it answered two important questions: How did China Mobile persuade a customer from the rail transit industry to put its services on a 5G network? And, what implications did this project's success have for future 5GtoB services?

The road to a national championship

China Mobile initially kicked off this project to figure out how to market connectivity to the rail transit industry. As they dug deeper into how the industry's complex production systems worked, however, it became clear the

It became clear the project needed to refocus on broader industry digitalization, as true success could not be achieved by targeting an individual metro company – they needed to target the industry as a whole.

project needed to refocus on broader industry digitalization, as true success could not be achieved by targeting an individual metro company – they needed to target the industry as a whole.

China Mobile had previous experience in 5GtoB services. They knew that industry customers often need patient persuasion before they decide to adopt 5G. They needed to take a chain-like approach to bridge gaps in the industry, going slowly and steadily in the early stages to pave the way for a boom in adoption at the later stages. They broke this strategy down into three parts:

First, they had to identify a correct overall direction at the beginning of the project. Once that direction was defined, all further efforts had to strictly stay the course. China Mobile chose the rail transit industry based on the logic presented in Figure 1. Many carriers are installing 5G networks in metros, but those network resources are often not being fully utilized due to the tidal effect in the number of passengers at different time periods. Traffic analysis showed that the physical resource

block (PRB) usage of carrier 5G networks along most metro lines is only at about 20%.

At the same time, metro operators traditionally build their own private networks, which can cost over 1 million Chinese yuan per kilometer of track. Metro wireless network bandwidth is also typically not large enough to carry more data services, and the network can only run in a best-effort manner. Service data that cannot be transmitted needs to be manually copied, which increases OPEX. Idle carrier network resources can be leveraged to address this private-network bandwidth problem—one of the biggest headaches of metro operators. Fully transitioning to carrier networks, which could simultaneously serve multiple purposes, presented a win-win situation.

A smooth and workable business model required thorough analysis to determine the base logic. Once the general model was decided, China Mobile was able to call upon resources from across the industry to address the specific challenges that arose during implementation.

The second step was to establish a defined industry ecosystem to better aggregate industry resources. 5G application generally needs significant support from not only the carrier industry, but also the target industry. The more robust the industry ecosystem, the more collaborative resources they had to achieve shared success.

Industry adoption of a new technology is never just a technical matter. It has to be tackled from both the technological and the commercial ends. Technologically, carriers must ensure that industry customers are confident about 5G. 5G network slicing had already been piloted with Nanjing Metro, using dedicated resources to carry rail transit services. That pilot project set a benchmark

that improved the industry's confidence. Carriers' existing expertise in 5G cybersecurity from other industries has also improved the rail transit industry's trust in the reliability of these services. Commercially though, many of China's carriers had years of experience in leasing, instead of building, networks that can best serve another industry. Under this model, customers pay carriers annually to use carrier 5G networks for around the same total cost as building their own traditional network.

By combining the technological and commercial elements, China Mobile was able to resolve 14 of the key problems facing the rail transit industry within just 3 years. This finally paid off in the Wuhan Metro project.

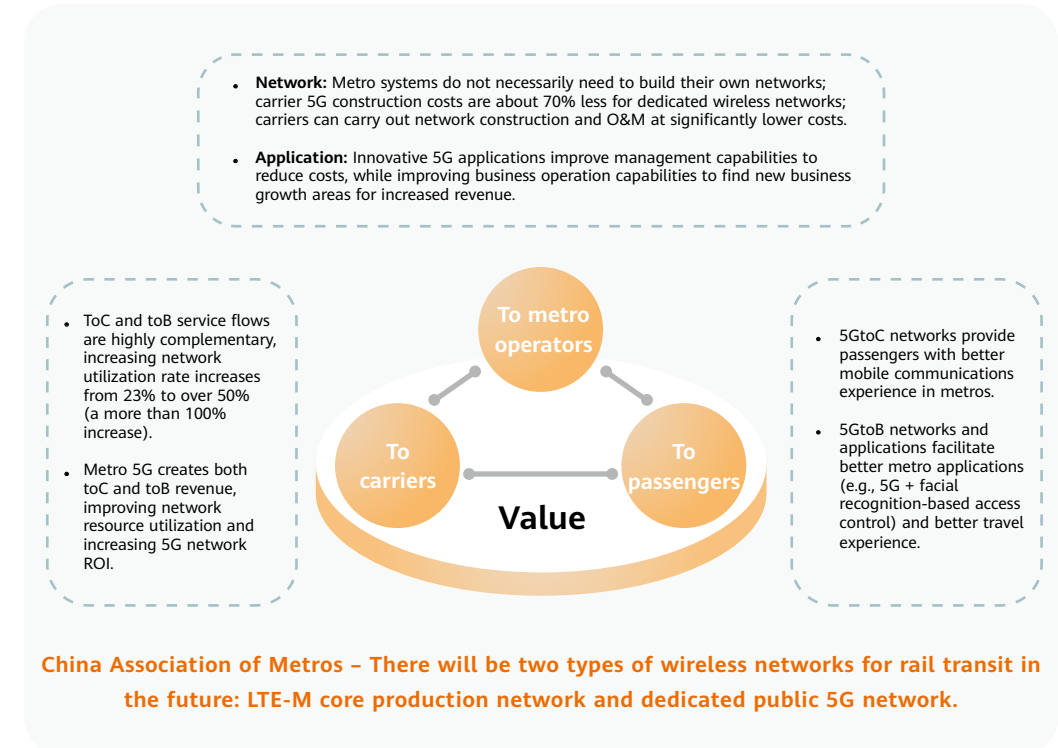


Figure 1: The value of dedicated public 5G networks for rail transit

The third step was to develop scenario-based solutions to resolve key implementation issues. 5G networks are designed to essentially amplify connectivity for terminal devices, but data connections alone do not create real value. To effectively support metro operation services, technologies like AI, big data, and integrated communications were also needed

to make the collected data a key part of production. Industry customers who actually use these solutions are most concerned about the practical value that they can create. The development of dedicated public 5G networks in Wuhan Metro therefore underwent three phases: reshaping connectivity, reshaping the platform, and building intelligence (Figure 2).

The 6th "Blooming Cup" Competition has helped promote 5G applications and explore new business models. Dedicated public 5G networks have been recognized across the rail transit industry as a feasible solution.

Planning is key

The bidding for the 5G 800 MHz trunking communications on Wuhan Metro's Line 19 was completed in early 2023. Since then, many other metro operators in China initiated similar projects. For example, Shanghai Metro plans to have all of its 23 lines covered by a dedicated public 5G network by the end of 2024. Guangzhou Metro also plans to expand the scope of its existing coverage from a single line to their entire metro system. Other cities have also produced blueprints for future rollout, including Nanning, Wuxi, Tianjin, and Suzhou. The clear industry demand presents large opportunities for dedicated public 5G network projects, but carriers still struggle with implementation speed and customer confidence.

Notably, this project has provided carriers with clear principles for future planning.

First, business models: The business model design must be based on actual service scenarios to keep the price of 5G reasonable for the target customer industry. In metro rail, a complete dedicated public 5G network project covers two parts: main lines and rail yards. For carriers, the cost of providing private network services mainly includes the following items:

1. **C (Construction cost):** The full cost of ensuring full coverage of new unmanned sections like train depots and turn-back tracks
2. **B (Basic sharing):** The cost of deploying public 5G networks that are under construction
3. **X:** The percentage of the public 5G network's uplink resource blocks used by metro operations within a security baseline
4. **A (Annual fees):** The fees charged for assurance services, such as slicing service, network optimization, and network maintenance over a 10- to 15-year period
5. **T:** Off-peak usage duration

The cost of using dedicated public 5G networks can then be calculated as:

$$C + B \times X + A \text{ (by bandwidth)}$$

$$\text{or } C + B \times T \div 24 + A \text{ (by time)}$$

Given this, the cost of using wireless 5G network resources on demand roughly equals that of building the metro operator's own wireless private network. Leaving resources unused outside metro service hours wastes 5G coverage along the line. Metro operators can also apply unmanned inspection robots to tunnel inspection, which use 5G networks for data backhaul, achieving unmanned and remote operations.

For metro ground services, 5G can provide the following alternative private network capabilities (Table 1).

The second principle relates to the network solution. To support train-ground backhaul

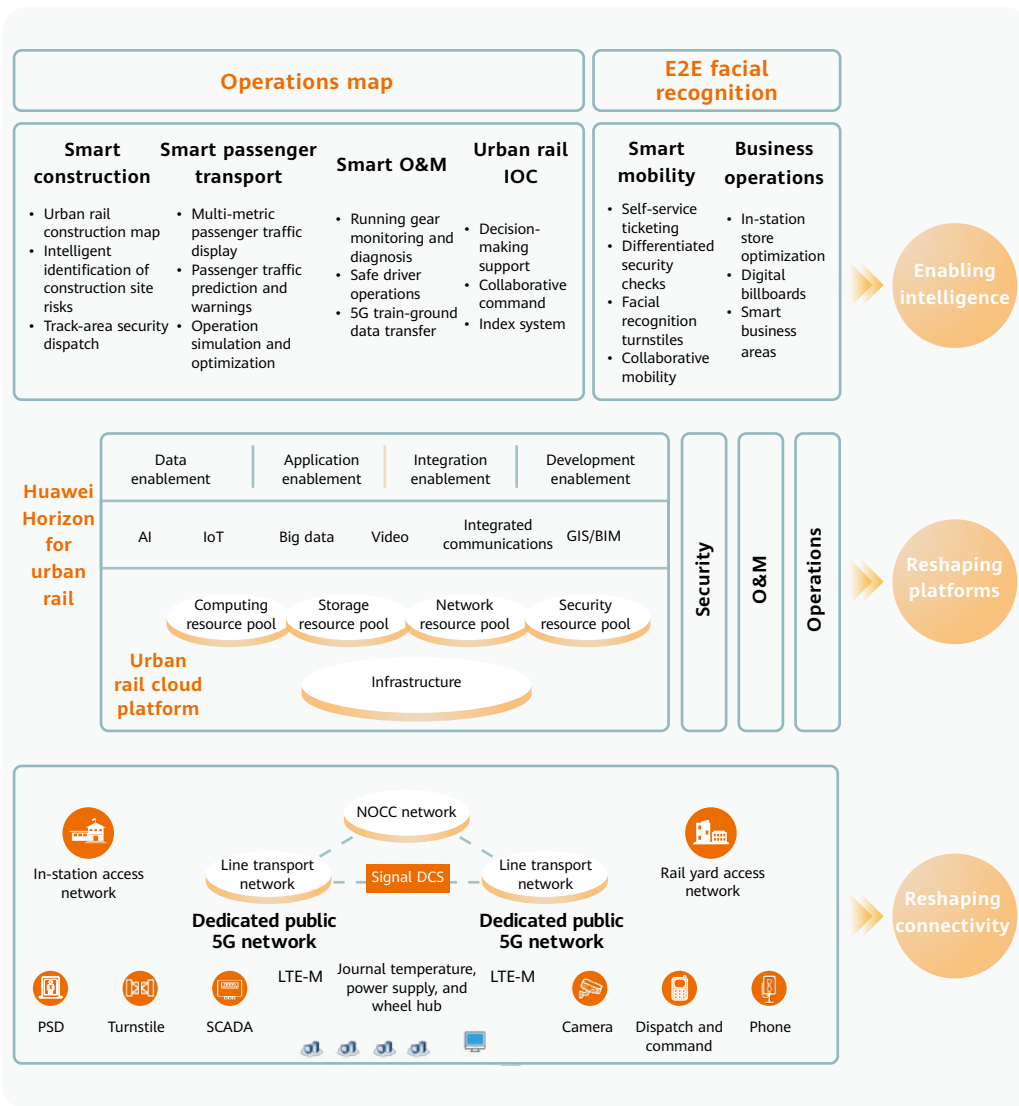


Figure 2: Blueprint architecture for smart urban rail

Network Built by Metro Operator	Metro Service	5G Network
Wi-Fi (target network migration)	CCTV service (partial)	Guaranteed 12-Mbps uplink
	PIS service	Guaranteed 16-Mbps downlink for service migration
TETRA 800 (target network migration)	Narrowband trunking	Guaranteed 4-Mbps uplink & downlink for broadband video communications
LTE1800 reuse (adding links)	TCMS 500 Kbps	Guaranteed 5 Mbps for TCMS
	CBTC redundant networks	
	Train trunking communications	Integrated 5G communications

Basic 5G network packages provide private uplink lines with 26 Mbps bandwidth for real-time services, 80 Mbps bandwidth for quasi-real-time services hours, and 200 Mbps bandwidth for off-peak services, with a 99.9% network availability over a 10-year contract period. 5G inherits the services previously carried on Wi-Fi and TETRA networks, while improving TCMS services and providing backup for LTE-M network services.

Table 1: 5G private network capabilities for metro services

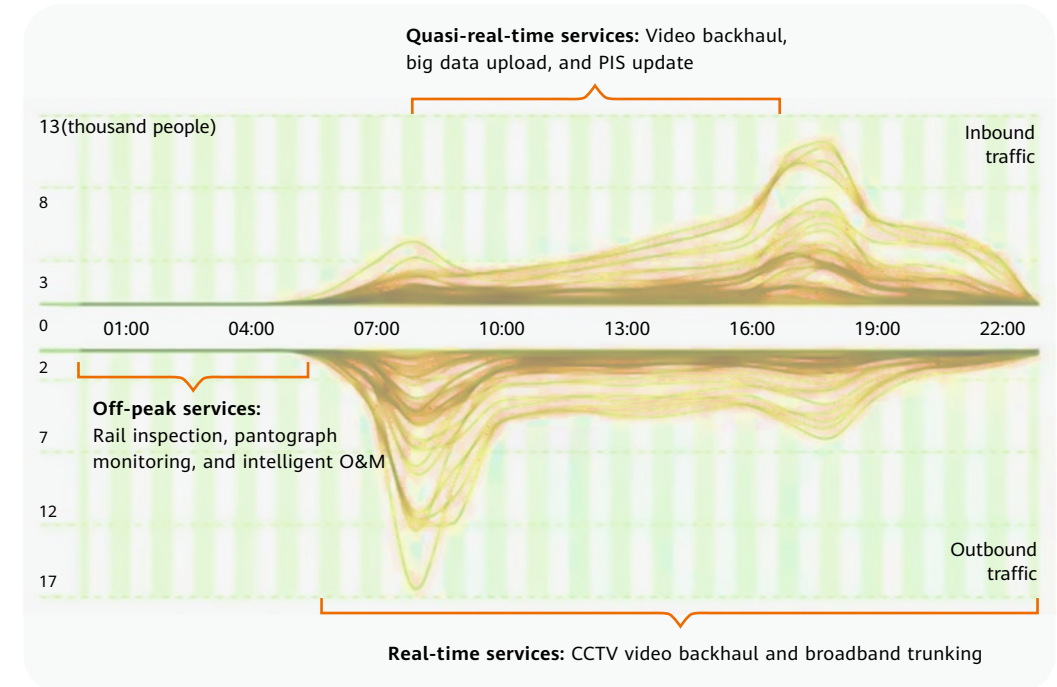


Figure 3: Traffic peaks and troughs of the three types of metro backhaul services

services, 5G network transport must meet SLA requirements. For individual users, disconnections and dropped calls are not likely to cause serious loss. However, carrier 5G networks must ensure the services they carry are always available. Metro operators need these high-quality network capabilities to ensure continuous 5G access. They are also key for further application development. During the Wuhan Metro project's bidding, the key factor that drove the customer to choose China Mobile was not price, but network reliability, availability, system security, and data redundancy design.

China Mobile uses a customizable five-domain

the tidal effect in metro services. This solution consists of three steps: service analysis, existing network assessment and optimization, and slicing customization and provisioning.

During service analysis, plans are based on the actual needs of metro backhaul services. Depending on how services are used, services can be classified into three types: real-time services, quasi-real-time services, and off-peak services (Figure 3). Real-time services are identified and basic bandwidth is provided to guarantee these services.

During existing network assessment, they analyze the 5G networks already deployed in metro facilities. Due to the tidal effect in

passenger traffic, network resource usage also sees peaks and troughs. By accounting for these traffic flows, they can use carrier networks for quasi-real-time metro services during off-peak hours. At night, when the trains are not in service and tunnel networks are completely idle, carrier networks can be utilized for the backhaul of off-peak services. This model maximizes the utilization of 5G network resources and reduces costs, and can be quickly replicated by the rest of the industry.

Finally, during slicing customization and provisioning, they determine time- and space-based 5GtoB service policies through network analysis (Figure 4). By setting

parameters for each of the three domains – wireless network, transport network, and core network – they can guarantee SLAs for industry applications. The dedicated public 5G network slicing solution is customized for metro service scenarios to maximize resource utilization. This solution monitors the resource availability of each cell at different periods, and provides time- and space-domain service policies. It also sets up automatic parameter configuration for the core and wireless networks to implement a site-specific service design.

The third principle from this project relates to value-added innovation. An example of such innovation can be seen in the figures

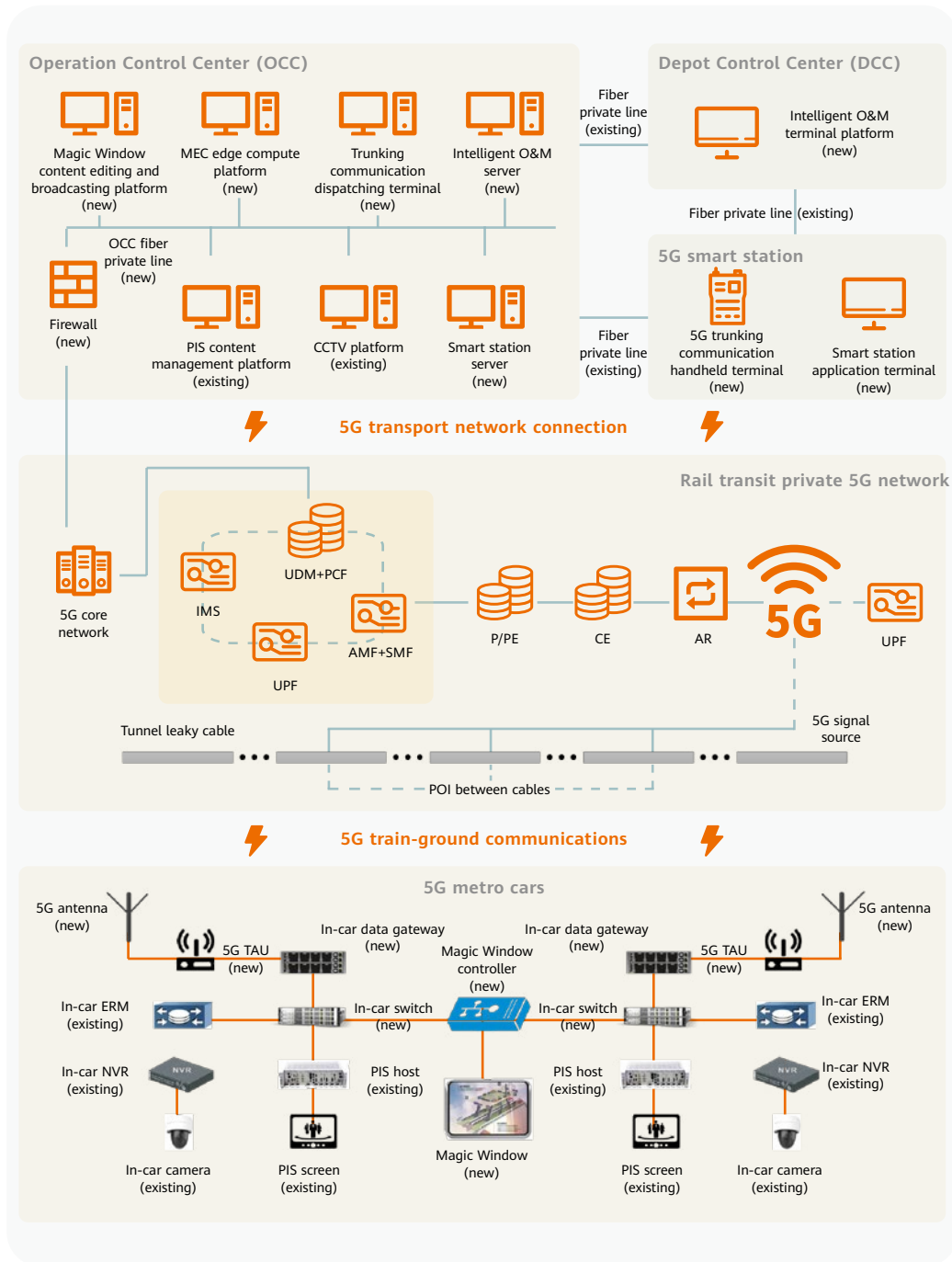


Figure 4: Networking architecture of the dedicated public 5G network solution for rail transit

below, which show a 5G-based leaky cable positioning solution that aims to support positioning in rail transit (Figure 5). Customized leaky cable and wireless UTDOA features can help position people and vehicles in rail areas. People positioning is mainly used by maintenance personnel to accurately and quickly locate faults during track inspections. Vehicle positioning is used by the pantograph monitoring and O&M system to locate faults during train operations. Many metro systems currently use axle counters and UWB base stations for positioning, but this is costly, requiring about 2 million Chinese yuan in investment per line (for a standard 30-kilometer line). In contrast, utilizing the 5G networks already deployed in tunnels for

positioning allows metro operators to save the cost of building their own UWB networks.

The success carriers are finding in rail transit is an important step in realizing the full value of 5G. The dedicated public 5G network model is rapidly becoming the new standard for the rail transit industry, and one that can potentially be applied to more scenarios, including civil airports, high-speed railway stations, and healthcare campuses.

More technological advancements and the promotion of other successful applications will help 5G continue creating value in more fields and help industries go digital and thrive. **T**







Tunnel positioning requirements	Current worker positioning solution	Current rail transit positioning solution
<p>Worker positioning</p> <p>Inspector positioning and real-time fault location reporting</p> 	<p>Positioning with wall tags: high error rate; real-time monitoring unfeasible</p> 	
<p>Vehicle pantograph monitoring, locating, and data backhaul</p> 	<p>High UWB construction (CNY60,000-80,000/km) and maintenance costs</p> 	
		<p>Axle-counter-based positioning error: 20m/km</p>

Figure 5: Rail transit positioning requirements and solutions



02. Network Evolution

Middle East: Leading the 5.5G Era and Striding Towards an Intelligent World



Tang Zhentian

Director, Middle East & Central Asia ICT Marketing & Solution Sales Dept, Huawei

The advancements in 5G and 5.5G that carriers have made in the Middle East's Gulf Cooperation Council (GCC) countries have set new benchmarks in scenario-based technology verification and growth opportunities, setting the stage for region-wide digital and intelligent upgrades.

The Middle East has always played a unique role in history as a place of connections, bridging great geographic, economic, and cultural divides.

During the Middle Ages, Arab merchants used camels and ships to traverse the Silk Road, bringing silk and porcelain from China to India and the Malay Islands, where they would pick up spices, minerals, and cane sugar to trade for gemstones in Central Asia and ivory and gold sand from East Africa. Their powerful networks then went on to connect Northern Europe with honey, fur, and wood. This mercantile influence spanned the world and they played an important role in the Arab Empire's prosperity, in turn influencing the world's economy, science, technology, culture, and art.

Today, the six nations of the Gulf Cooperation Council (GCC) are returning to this model through economic restructuring. Digital and intelligent transformation are presenting the

perfect opportunity to reduce their economic reliance on oil. The United Arab Emirates and Saudi Arabia, for example, have set in motion plans to become digital hubs for their surrounding regions. Multinational carriers in the region, like stc, e&, and Zain have jumped at the chance to become the new merchants of the digital era by supporting digital upgrades across the Middle East and Africa, and developing new strategies for cloud, DCs, and toB integration.

All have turned their attention towards building world-class ICT infrastructure.

The first year of commercial 5.5G in the Middle East

During the 2023 Global Mobile Broadband Forum (MBBF), six Middle Eastern carriers kicked off the region's commercial launch of 5.5G. Shortly after, those same carriers successfully completed 10-Gbps testing and more than 20 scenario-specific

“ Qatar has already put a roadmap in place to transition to 5.5G in partnership with key stakeholders, like vendors, regulators, and strategic partners. By planning for upcoming events and digital transformation, we've positioned ourselves at the forefront of this technological revolution.^[1] ”

— Ramy Bactor, Chief Technology Officer, Vodafone Qatar



pilots of 5.5G technologies like RedCap, Passive IoT, and glasses-free 3D.

Exploring new 5.5G services for sustained 5G business success

So far, these carriers have been pragmatic in their approach to 5.5G. Their existing success in 5G has been the core driver of their continued investment.

By the end of 2023, 5G had a user penetration rate of over 40%, and over 40% of mobile data traffic came from 5G connections. Around 2018, many carriers had begun experiencing a decline in revenue as 4G reached maturity, but 5G has reversed this trend. Leading carriers have realized 20% to 45% revenue growth thanks to 5G, and FWA services are rapidly transforming the industry landscape in the Middle East.

Middle Eastern carriers are therefore eager to capitalize on 5.5G. Many have already identified a range of high-value 5.5G use cases that will help improve user experience and foster innovative new applications for individuals, homes, and industries.

Individual services: Immersive glasses-free 3D experience stimulates 10-fold traffic

Glasses-free 3D started making waves back in 2009 with the release of the sci-fi movie *Avatar*. Now, more than 10 years later, the core technologies behind glasses-free 3D are beginning to see large-scale commercialization as related technologies become increasingly reliable, with better experiences delivered, 3D content easier to produce, and more affordable devices available.

Shopping malls in the UAE and Oman are already using glasses-free 3D screens for advertising,

searches for "glasses-free 3D" have surged over the past half year according to Google Trends, with over half of those searches being made in countries like Saudi Arabia, the UAE, and Oman. This is a prime example of how carriers are exploring new approaches to 5G monetization. One carrier found more success with 3D tablets than with glasses-free 3D phone screen protectors thanks to the better visual experience they deliver, deciding to bundle 3D tablets with FWA and phone screen protectors with toC services. These sales plans are set to launch in the near future. A second carrier plans to introduce glasses-free 3D use cases to their 5.5G Innovation Center and work with OSN (a local OTT service provider) and YouTube to build a 3D content ecosystem. A third carrier is planning to launch its own 3D livestreaming services and working with OTT service providers to establish a dedicated 3D service section and also plans to bundle 3D services into 5G subscription packages.



“5G is going to deliver an experience that will be close to or even match the fiber experience. We believe that by moving from 5G to 5G SA, and to 5G-Advanced, we are halfway to 6G.”^[3]

— Karim Benkirane, Chief Commercial Officer, du

And a fourth carrier has decided to first explore 3D screen protector services before turning to 3D content.

Many carriers are already designing premium service packages that will increase monetization and are directly working on developing a 3D content ecosystem. Many of these glasses-free 3D services are expected to launch in the first half of 2024.

Home services: Fiber-like to air-fiber FWA experiences

Another top area of focus for carriers is sustainable FWA growth.

The household penetration of FWA services in the Middle East already exceeds 21%, and the service is growing rapidly, even in markets with extensive optical deployment, such as the UAE, Qatar, and Bahrain. However, this rapid growth has created two

pressing challenges for carriers. First, per-site user volume growth is increasing network congestion, which in turn harms user experience and increases complaints. This has made FWA user satisfaction a key area of concern. Second, fiber downlink rates are growing by over 40% every year. Mainstream fiber-based packages already deliver 300 Mbps to 500 Mbps rates, so, how can FWA maintain a competitive edge?

Over the past year, multiple carriers have worked with Huawei to explore these issues. What we've found is that there is still tremendous untapped potential in FWA if we upgrade offerings from unguaranteed 100-Mbps experience packages to reliable 300-Mbps experience packages. Our surveys have found that 70% of user packages can be upgraded, and 67% of users are willing to pay over 20% more for better experiences. This means that a reliable user experience is required throughout the

We are proud to launch our white paper, which outlines our commitment to technological excellence and leadership in the advancement of 5G technology in the UAE. It reflects our proactive approach to adopting advanced technologies and our dedication to pioneering innovation.^[7]

— Khalid Murshed, Chief Technology and Information Officer, etisalat by e&

user journey. Carriers will need to assess available network resources, coverage, and interference before high-speed FWA services can be provisioned. Provisioning will also often need higher resource scheduling priorities and higher performing CPEs. In some cases, onsite assisted installation will be needed to ensure the optimal installation position. Once services go live, carriers will need to be able to monitor user experience KPIs in real time for preventive optimization, and provide dedicated VIP service channels for consumers.

The UAE's du is an FWA pioneer. During MWC Barcelona 2024, du announced that it was ready to launch high-speed, low-latency packages for gamers and bundle more TV privileges with FWA packages^[2].

FWA's potential remains huge as optical technologies evolve.

Industry services: 10 Gigabit City infrastructure facilitates industry digitalization

In 2022, Huawei signed an MoU with Saudi Arabia's Ministry of Communications and Information Technology on building a 10-Gbps Society. The country's 10 gigabit vision is now

being implemented as the nation plans new urban areas. The enhanced network capabilities brought by 5G and 5.5G will enable carriers to provide ubiquitous basic connectivity while also helping them build global leadership in integrated national ICT solutions.

- **A low-carbon "cognitive city":** Saudi Arabia's NEOM aims to use 5G to connect people and things to build a "cognitive city". The tourism developer Red Sea Global then worked with Zain to build zero-carbon 5G networks for a massive future city project that would deploy 5G-powered IoT solutions to monitor important ecological sites, like coral reefs and sea turtle nesting sites^[4]. Carriers are now verifying multi-scenario IoT applications using 5G smart lamp poles. Some of these applications include intelligent security, electronic screens, EV charging piles, environmental monitoring, and autonomous driving.
- **Digital and intelligent industries:** Industries are now increasingly receptive of 5G. According to a survey by the UAE government, about 40% of enterprises plan to use some form of 5G application within the next year. For example, stc Group and the Saudi

Ports Authority launched the Smart Ports Initiative to automate port operations through different scenario-based applications including remote crane operation, 3D visualization of cranes, license plate recognition, and facial recognition. etisalat by e& also released a 5G-Advanced white paper on multiple 5G products, including 5G mobile edge, 5G campus, and 5G networks for cars to promote the application of 5G and 5.5G in the government, transportation, healthcare, aviation, and oil and gas sectors^[6].

- **Digital twins:** Carriers in the Middle East are also actively testing and applying 5.5G technologies such as passive IoT and RedCap. In the UAE, passive IoT's long standby times and low costs are enabling a number of smart retail applications, connecting a wide range of goods to digital twins. This has increased tracking and stock allocation efficiency. In Saudi Arabia, low-cost 5G RedCap modules are also being used to explore new CCTV and smart security applications.

Building 10G networks using leading technologies

Leading carriers in the region have already completed 5.5G target network planning.

User-perceived speeds are expected to double in the next three years, with the average 5G downlink rate expected to jump from 100 Mbps to 200 Mbps by 2026, and the user-perceived FWA rate spiking from 165 Mbps to 500 Mbps.

Once this is achieved, carriers are likely to shift their focus towards pursuing ubiquitous 10G connectivity. By 2026, peak user rates will already be at 5 Gbps with under 20 ms latency,

so the next milestone will be achieving 10-Gbps user-perceived downlink and sub-10-ms latency by 2030.

A consensus has already been reached by the region's carriers that TDD 3CC aggregation (300 MHz band) will be one of the fundamental differences between 5.5G and 5G. Because of this, they are working with Huawei on joint innovation in this area. For example, the region has already taken the lead in fostering dual-band META (2.6/2.3 GHz + C-band, dual-band 3CC) products that support simplified 5.5G deployment, as well as 3CC small cell products, to improve indoor user experience. The UAE's du is also already proceeding with large-scale TDD 3CC deployment. The extremely large antenna array (ELAA) technology has increased the number of dipoles from 192 to 384, enabling a peak rate of 5 Gbps, 30% higher user-perceived rates and 30% lower energy consumption^[8].

Saudi Arabia's stc is similarly deploying 4CC, which includes C-band and low and medium bands, to improve network performance and user experience, and speed up 5.5G evolution^[9].

At MWC Barcelona 2024, Huawei signed three MoUs with:

- du to transform the UAE into a 5.5G country
- Kuwait's Communication and Information Technology Regulatory Authority (CITRA) to build 5.5G smart cities
- Zain KSA on "All in 5.5G".

Many carriers are working with vendors like Huawei to seek smooth 5.5G network evolution, 5G super packages (which use new services like New Calling, 3D, and MR to build the 5.5G brand), and rate-based FWA pricing evolution.

These demands have become an important factor driving innovation at Huawei.

High-quality 5G construction is creating a digital oasis

More carriers and governments are pursuing digital, intelligent, and low-carbon transformation.

The GCC nations have all launched national visions that will significantly increase the prominence of their digital economies and help them become global leaders, such as Saudi Arabia's Vision 2030 and the UAE's "We the UAE 2031". These governments are encouraging 5G network deployment through policy and funding support, which is successfully driving improvements in mobile network quality. Currently, four out of the world's 10 leading countries according to Speedtest are from the Middle East. Bahrain, the UAE, Saudi Arabia, and Oman have made significant commitments to carbon neutrality. Digital and intelligent transformation are also being clearly seen in many cities and industries across the region.

At the carrier level, Middle Eastern carriers are continuing to build high-quality communications networks as they digitalize their own operations.

Success in 5G has helped many of these carriers improve brand value. From 2018 to 2023, stc's brand value increased by 85%, helping them move up 19 places in global rankings and making them the 11th top global carrier. e& also moved up to 14th place. Middle Eastern carriers are pursuing industry leadership with their 5.5G evolution. 2024 will be the first year of large-scale 5.5G construction in the Middle East, with leading countries like the UAE expected to achieve nationwide 5.5G deployment.

This has made the Gulf region significantly more economically dynamic. In 2023, Dubai Chambers saw a 43% increase in the number of registered new enterprises, with the growth rate of Abu Dhabi's startup ecosystem ranking sixth globally. The UAE plans to incubate 20 unicorn companies by 2031 and Saudi Arabia saw a 33% increase in its number of startups in 2023. Saudi Arabia also led the Middle East and North Africa with a record US\$1.38 billion venture capital invested last year. This has made the Gulf region one of the prime destinations for global startups^[11].

5.5G will likely continue to drive growth in the Middle East's economy and society. This will trigger a chain reaction, supporting national visions like Saudi Vision 2030 and We the

UAE 2031 and stimulating new digital vitality by attracting global talent and encouraging more innovation. This growing digital oasis will undoubtedly create a new benchmark for the world in digital transformation. **T**

References:

[1] <https://www.marhaba.qa/vodafone-qatar-hosts-groundbreaking-discussion-on-5-5g-technology/>

[2] *du, Huawei sign strategic cooperation to boost 5G-A in the UAE.* <https://www.rcrwireless.com/20240228/5g/du-huawei-sign-strategic-cooperation-boost-5g-uae> Feb 28, 2024

[3] *Du to bolster FWA 5G play to challenge rival.* <https://developingtelecoms.com/telecom-business/operator-news/16333-du-to-bolster-fwa-5g-play-to-challenge-rival.html> Feb 29, 2024

[4] *An Inspiring Journey: Zain KSA and Red Sea Global Pioneering World's First Zero-Emission 5G Network.* <https://www.telecomreview.com/articles/exclusive-interviews/7406-an-inspiring-journey-zain-ksa-and-red-sea-global-pioneering-world-s-first-zero-emission-5g-network> Oct 09, 2023

[5] *5G-Advanced: The Future of Wireless Communications in the UAE.* https://www.samenacouncil.org/samena_daily_news?news=94414 March 09, 2023

[6] https://www.eand.com/content/dam/eand/assets/docs/general/5G-advanced_the_future_of%20_wireless_communications_in_uae.pdf

[7] *etisalat by e& launches white paper on UAE's transition to 5G-advanced and commitment to global standards.* <https://wam.ae/article/b0ys2bv-etisalat-launches-white-paper-uaes-transition> Jan 03 2024

[8] *du Spearheads Commercial Deployment of 5G-Advanced.* <https://www.telecomreview.com/articles/telecom-operators/7818-du-spearheads-commercial-deployment-of-5g-advanced> Feb 07, 2024

[9] *stc Achieves Substantial Progress in Advancing 5G.* <https://www.telecomreview.com/articles/telecom-operators/7719-stc-achieves-substantial-progress-in-advancing-5g> Jan 02 2024

[10] *du Commits to Building UAE as 5G-Advanced Country.* <https://www.telecomreview.com/articles/telecom-operators/7882-mwc24-du-commits-to-building-uae-as-5g-advanced-country> FEB 27 2024

[11] *Accelerating Growth: the GCC's Start-up Ecosystem.* <https://communicateonline.me/category/industry-insights/post-details/accelerating-growth-the-gccs-start-up-ecosystemhttps://communicateonline.me/category/industry-insights/post-details/accelerating-growth-the-gccs-start-up-ecosystem> Jan 31 2024

This is a key milestone for du's commercial 5G-Advanced deployment. Huawei's leading technologies will support the realization of du's strategic vision, play an important role in 5G-Advanced service innovation and industry digitalization, and jointly accelerate the UAE's digital transformation.^[10]

— Fahad Al Hassawi, Chief Executive Officer, du

Future All-optical Network Architecture and Key Technologies



Tang Xiaojun

Chief Technology Planner, Optical Business Product Line, Huawei

Evolving towards the 2030 optical communications network system and architecture is a key issue facing the optical communications industry and requires viable technical options for building future-oriented and novel optical communications network systems.

Optical networks form infrastructure that deliver ultra-broadband, large-capacity, and low-latency connectivity for the digital world. From voice-based communications in the 1980s to modern broadband communications that typically support video applications, fixed networks have evolved through five generations. 2024 is the first year of F5G-Advanced (F5G-A), and AI-driven applications and the 3D display industry are now on the rise. Breakthroughs in AI foundation model technology, represented by OpenAI, are leading to the construction of numerous AI clusters and the adoption of many new applications. Furthermore, products based on immersive 3D display technology, represented by Apple Vision Pro, are entering the market on a large scale. The enormous computing and transport power that accompanies these developments urgently needs all-optical infrastructure. Optical communications networks have been presented with unprecedented opportunities and challenges, which can only be addressed through breakthroughs in the architecture and

key technologies in the following areas:

- **Backbone networks:** Improvements in single-channel rates and system capacity to support increasing traffic between and within data centers – such as that driven by China's "East Data, West Computing" project – remain a major force driving evolution.
- **Metro networks:** Changing service traffic directions are driving transition from the traditional north-south network architecture to a cloud-centric, low-latency, high-quality, and T-shaped network architecture.
- **Access networks:** As 3D applications make their way into homes and AI foundation models into devices, the focus of evolution will shift from bandwidth improvement to high-bandwidth and high-quality experiences.

backbone networks. The acceleration of AI adoption is further increasing requirements for the transmission capacity of backbone networks. Backbone connectivity capacity can be improved in three directions: improving single-channel rates, broadening frequency bands, and evolving towards space-division multiplexing (SDM) systems.

Direction 1: Improving single-channel rates

Improving single-channel rates is a major method of increasing backbone capacity. The evolution from coherent 100 Gbps with 50 GHz channel spacing to 400 Gbps with 150 GHz channel spacing has improved both spectral efficiency and capacity by 33%. Requirements for higher single-channel rates are often accompanied by expectations for lower per-bit costs and power consumption. To meet such requirements, single-channel rates will evolve towards coherent 1.6 Tbps with 500 GHz channel spacing. As we approach the Shannon limit, spectrum improvement is becoming increasingly difficult, requiring breakthroughs in both algorithm and component innovation. The current technical challenges are as follows:

Backbone network architecture: Evolution of single-channel rates and system capacity

Capacity improvement is a key driving force behind the generational evolution of

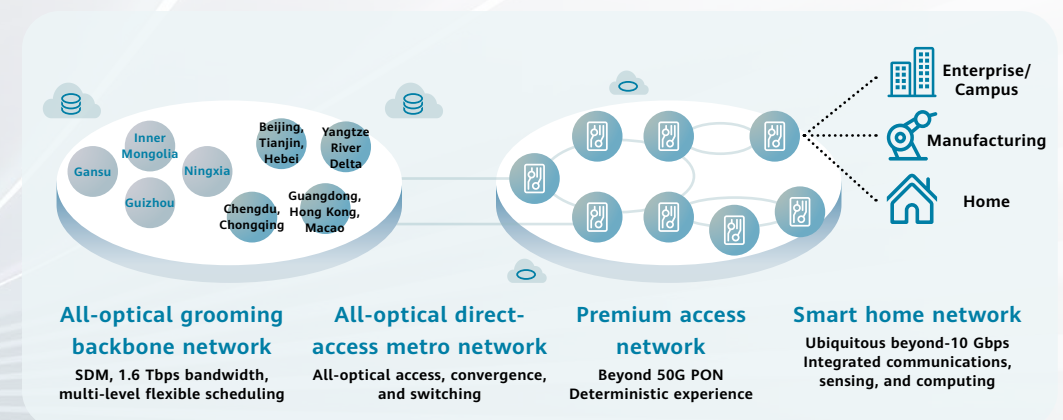


Figure 1: 2030 all-optical network architecture

- Improving spectral efficiency with higher-order modulation formats and optimized algorithms:** The spectral efficiency of 1.6 Tbps with 400 GHz channel spacing (4 bit/Hz) is 50% higher than that of 400 Gbps (2.67 bit/Hz). In addition, the modulation format is changed from QPSK to CS16QAM, and long-distance backbone transmission must be covered. These changes require better modulation format designs and optimized algorithms.
- Components with 400G baud rates:** Achieving 1.6 Tbps coherent transmission on a single transceiver requires components that support close-to-400G baud rates, such as AD/DA converters, modulators, and photodiodes. The latest industry developments can achieve modulation and demodulation at baud rates of about 250G, but there is still a significant gap to 400G baud rates.
- High-integration superchannel technology:** Limited component techniques mean baud rate increases are becoming increasingly difficult. An alternative way to achieve 1.6 Tbps rates per channel is multi-channel integration, which means integrating two 800-Gbps channels or four 400-Gbps channels. A multi-channel solution uses lower-bandwidth components, but requires better photonic integration technology that is capable of solving issues like relative frequency control, multi-channel component yield improvement, reduced raw-material usage, and the warping of boxes and substrates.

Direction 2: Broadening the spectrum

System capacity is equal to the single-channel rate multiplied by the number of channels in the system. Increasing the number of channels by broadening the spectrum is another key direction for the evolution of optical system capacity. A C120+L120 system

delivers two-times wider spectrum than a C80 system, as well as higher single-channel rates, increasing system capacity from 8 Tbps, which was delivered in the early days of coherent systems, to today's 32 Tbps. Beyond the C-band and L-band, the S-band and U-band have the potential to help broaden the spectrum. However, the optoelectronic components for the S-band and U-band, such as tunable lasers, photodiodes, and optical amplifiers, are not yet mature. Moreover, modulators and receive-

end mixers are wavelength-sensitive and must be re-designed. Adding these two bands will increase stimulated Raman scattering (SRS), a non-linear effect in the system that causes the transmission performance of the C-band and L-band to deteriorate, increasing system-design and O&M complexity. These challenges require further research and investment by both academia and the industry. Even if the challenges are successfully overcome, system capacity can still only be doubled, meaning a

low return on investment and difficulties for long-term evolution.

Direction 3: Evolution towards SDM systems

The growing complexity of technical challenges in relation to the two aforementioned solutions means that backbone network evolution towards SDM systems is becoming increasingly important and urgent. SDM was proposed 60 years ago, and there has been a large amount of academic research into key SDM technologies, with solid progress being made over the past dozen years. However, the following issues regarding SDM productization remain unresolved:

- Academic research has proposed multiple technical routes for SDM, including uncoupled multi-core fiber, coupled multi-core fiber, multi-fiber, few-mode fiber, multi-mode fiber, and orbital angular momentum (OAM) solutions. These routes each have their own advantages and disadvantages, and it is too early to say which of them will ultimately become dominant.
- SDM can be carried by different types of optical fibers, such as G.652, G.654, and hollow-core.
- It is not yet clear whether the system architecture would be WDM first followed by SDM or the other way around.
- SDM can have two, four, or more channels.

According to Huawei's analysis, the multi-fiber solution and uncoupled fiber solution are currently the two most mature. The former is a mature design based on parallel single-mode fibers, while the latter has lower crosstalk than other solutions and is made with ground-breaking drawing techniques. Huawei believes that breakthroughs and consensus are needed in the following four areas for SDM be commercialized across the industry:



The first is compatible system architecture. The WDM system was the key to the past success of optical backbone networks, meaning compatibility with the WDM system is necessary to commercialize the SDM system. SDM and WDM multiplex different resources, so the design through which these two systems are made compatible must be carefully considered. A solution with WDM followed by SDM is recommended because the compatibility in question can be realized through the use of simple-structure components, including WDM/SDM multiplexers, optical amplifiers, optical cross-connect (OXC) devices, and fan-in/fan-out (FI/FO) devices. This solution must support both multi-fiber and uncoupled multi-core fiber technologies. The two technologies are suitable for different types of networks, with the former designed for the upgrade of legacy fiber networks, and the latter designed for newly-deployed fiber networks. It is also possible to use the two technologies together by adding new multi-core fibers to existing single-mode optical fibers in order to form a hybrid SDM network. In terms of fiber types, SDM systems must support both single-mode fibers and hollow-core fibers. Hollow-core fibers have attracted a lot of attention in recent years due to their low nonlinearity, low dispersion, and low latency, and have the potential to become a disruptive fiber technology.

The second is simplified management system. Another key issue to consider when productizing

SDM is adding one more dimension of multiplexing without increasing the complexity of network management. Huawei believes that WDM-like network management on an SDM system, despite the added space dimension, is the main focus and area of breakthrough for management and control technologies. Traditional WDM systems, multi-core SDM systems, and multi-fiber SDM systems all use the same network management system (NMS) to centrally manage OTU ports, wavelengths, and fiber cores. This involves a range of technologies including network resource pooling (establishing an OTU port resource pool, a wavelength resource pool, and a fiber core resource pool to share resources among different systems) and network as a service (NaaS) application programming interfaces (APIs).

The third is highly-integrated system architecture. Limited equipment room spaces require more closely integrated network equipment, meaning that increasing system capacity without requiring larger equipment is critical to productizing SDM. From a system composition perspective, a more integrated SDM system requires optimized component sizes including optical modules, optical amplifiers, OXC, and FI/FO devices. Optical module improvement in baud rates will inevitably encounter a bottleneck. Superchannel optical modules supporting higher port rates may necessitate a change in technologies, which will present the same challenges as mentioned earlier,

including relative frequency control, multi-channel component yield improvement, reduced raw-material usage, and the warping of boxes and substrates. Challenges to improving optical amplifier integration include: multi-component rare earth doping used for integrated wide-spectrum amplification, high-power and low-cost pumps used to support multi-channel amplification, and multi-channel passive components with high integration and low loss. The main challenge in relation to OXC is that the additional space dimension requires high-degree OXC, posing serious challenges to 128-degree wavelength selective switches (WSS) that can only be addressed with OXC architecture innovation. In terms of optical connections for an SDM system, another technical challenge is direct connection between multi-core fibers and single-mode fibers, and also between multi-core fibers and optical modules.

The fourth is converged applications such as communications, sensing, and security. SDM can support more application scenarios than large-capacity communications. For example, technologies like distributed optical fiber sensing and quantum key distribution (QKD) that

have emerged in recent years are significantly expanding the application scenarios of optical fibers, enabling converged applications of integrated sensing, communications, and encryption. With multiple spatial channels, SDM can be used for applications that integrate communication capacity expansion, integrated communications and sensing, and security. For example, in a four-core system, one core is used for traditional optical communications, another two cores are used for QKD-based communication and key negotiation, and the fourth core is used for distributed optical fiber sensing.

Metro network architecture: Cloud-centric low-latency and high-quality networks

AI's massive requirements for computing power will transform metropolitan area network (MAN) traffic from a north-south model to a T-shaped model that goes both north-south and east-west, and cause traffic to shift from user-to-user to mainly user-to-cloud. In addition, services like AI, VR, and smart manufacturing require higher bandwidth and lower latency. It is essential that the metro

AI's massive requirements for computing power will transform metro network traffic from user-to-user to mainly user-to-cloud.

network architecture supports one-hop access to the cloud and ultra-low latency between AZs to meet user requirements for high-quality experiences. Legacy networks with an architecture that features hop-by-hop forwarding and multi-ring stacking can no longer meet the latest requirements. Key technologies like all-optical interconnection, fine-grain OTN (fgOTN), and optical-layer digitalization are required to ensure high bandwidth and low latency for the optical metro network architecture.

- **All-optical interconnection**

Centrally managing the wavelength resources of metro network access rings through OXC, with the resources in optical fibers shared among the rings, will enable the resources of each ring to be flexibly adjusted in order to effectively cope with traffic imbalance between rings. In addition, OXC eliminates the need for optical-to-electrical conversion, ensuring one-hop access to the cloud with ultra-low latency.

- **fgOTN**

OTN technology is deployed at the network edge and customer edge to provide hard pipes for services. The hard pipes feature physical isolation and high security, while supporting the continuous evolution of single-channel rates as well as hitless and fast bandwidth adjustment, thus meeting service requirements for increased and more flexible bandwidth. Furthermore, fgOTN hard pipes can provide deterministic latency, enabling deterministic experience for premium industry customers.

- **Optical-layer digitalization**

Optical performance visualization can be extended to the edge of optical networks.

Traffic flows are changed to be based on a deterministic cloud-centric model that facilitates optical network planning and O&M. Digital optical-layer technology can be used to accurately model the optical layer, thereby visualizing the optical network. This supports accurate service provisioning in the planning phase, intelligent fault location in the O&M phase, and quick service recovery on highly-reliable automatically switched optical networks (ASONS). When used alongside an operation app, this will help monetize network O&M.

Access network architecture: High bandwidth and high-quality experience

Optical access network technology has evolved along the following path: PON > GPON > 10G PON > 50G PON > Beyond 50G PON, providing high-bandwidth networks for access users. Emerging new services (e.g., AI, AR, VR, and holographic display) and applications (e.g., smart manufacturing) require access networks that deliver not only higher bandwidth, for example, 2 Gbps for 8K VR with high-quality experience, but lower latency, lower jitter, and security isolation. For example, manufacturing requires deterministic μ s-level latency and jitter. In addition, vertical industries require multiple service networks to be carried on a single physical network and be strictly and logically isolated to uphold the service level agreements (SLAs) of each service network. Therefore, future optical access networks will not only increase bandwidth towards beyond 50G PON, but transition from a best-effort model to one that ensures differentiated services with guaranteed quality. Deterministic

bandwidth, latency, and jitter, as well as high reliability, will be provided for diamond services (e.g., for high-end villas), while deterministic bandwidth and sub-ms-level latency and jitter will be provided for silver services (e.g., for apartments). On-demand bandwidth and ms-level latency and jitter will then be allocated for copper services (e.g., for remote rural areas) at affordable prices.

The high-bandwidth and low-latency requirements of beyond 50G PON can potentially be met by the following three technical solutions:

- **Direct modulation and detection**

This solution is based on the continued evolution of current-generation technologies. It provides a time division multiple access (TDMA) mechanism and supports low-cost, point-to-multipoint (P2MP) access, but cannot provide deterministic latency. The solution relies on higher-bandwidth optoelectronic components. Furthermore, direct detection means a 3 dB decrease in receiver sensitivity each time bandwidth doubles. Bandwidth improvements will also bring higher dispersion penalty. Reusing the optical distribution network (ODN) will require much higher transmit power than that provided by a 10G PON network, posing a great challenge to high-power lasers.


- **WDM/FDM & direct detection**

Frequency-division multiple access (FDMA) or wavelength-division multiple access (WDMA) is implemented through multiple wavelengths, and independent high-quality P2P access is implemented through a single

frequency or wavelength, providing users with μ s-level latency and jitter. This solution does not require a high power budget, but does require optical network units (ONUs) equipped with tunable lasers. Access networks are highly cost-sensitive, meaning low-cost tunable lasers are a key technical challenge for this solution.

- **Coherent solution**

Through the application of digital subcarrier technology, coherent networks can also deliver high-quality P2P access and ensure deterministic latency with physical isolation. Coherent systems can achieve high bandwidth with low-bandwidth components and higher-order modulation, which both realizes higher receiver sensitivity and meets the power budget requirements of ODNs. However, this requires ONUs, which are cost-sensitive, to be equipped with intrinsic lasers, creating a challenge regarding laser costs.

Looking ahead to 2030, ground-breaking transformations in fields like AI foundation models and immersive 3D interaction will require an all-optical network architecture with higher bandwidth and lower latency. Developing an all-optical network architecture system will require breakthroughs in key technologies related to backbone networks, metro networks, and access networks to support the connectivity required by the massive and constantly surging number of digital services. At Huawei, we believe that all-optical network architectures will help make people's lives more digitalized and intelligent and serve as the connectivity foundation for an intelligent world. 

APN6 Enables Innovation in Cloud-Network-Edge-Device Collaboration



Li Zhenbin

Chief Protocol Expert of Huawei and Former Member of the Internet Architecture Board (IAB) of the Internet Engineering Task Force (IETF)



With technological evolution, APN6 is reshaping the future of network services, and is likely to initiate a major transformation in Internet architecture. This article explores how APN6 technology is transforming the cloud-network-edge-device collaboration system, as well as its potential influence in emerging business domains.

In the digital age, every innovation in network technology has the potential to initiate a major industry transformation. Application-aware IPv6 Networking (APN6) uses the programmable space in IPv6 packets to bring application information (parameters such as identifiers and/or network performance requirements) into a network. This makes the network aware of applications and their requirements, and enables refined network services and accurate network O&M, facilitating innovation in cloud-network-edge-device convergence.

As differentiated application requirements and diverse network technologies and services emerge, a wide range of applications designed for these requirements are being developed. Specific application scenarios include:

- Mobile interconnection applications oriented towards enhanced Mobile Broadband (eMBB) such as HD video, virtual reality, cloud storage, high-speed mobile Internet access, and AI
- Device interconnection applications oriented towards massive Machine-Type Communications (mMTC) such as environmental monitoring, smart meter reading, and smart agriculture
- Special applications oriented towards Ultra-Reliable and Low Latency Communications

(URLLC) such as Internet of Vehicles (IoV), industrial control, smart manufacturing, and remote surgery

- As industries migrate to the cloud, cloud applications are used over the Internet in industries such as finance, manufacturing, education, and healthcare, as well as by individual users and households. They are reshaping industries and people's social activities, entertainment, and everyday lives. Major applications include smart cities, private networks for financial cloud, cloud-based healthcare, online education, remote offices, cloud private lines for e-commerce, and cloud gaming.

Diversified application scenarios present new challenges for network O&M. Effectively

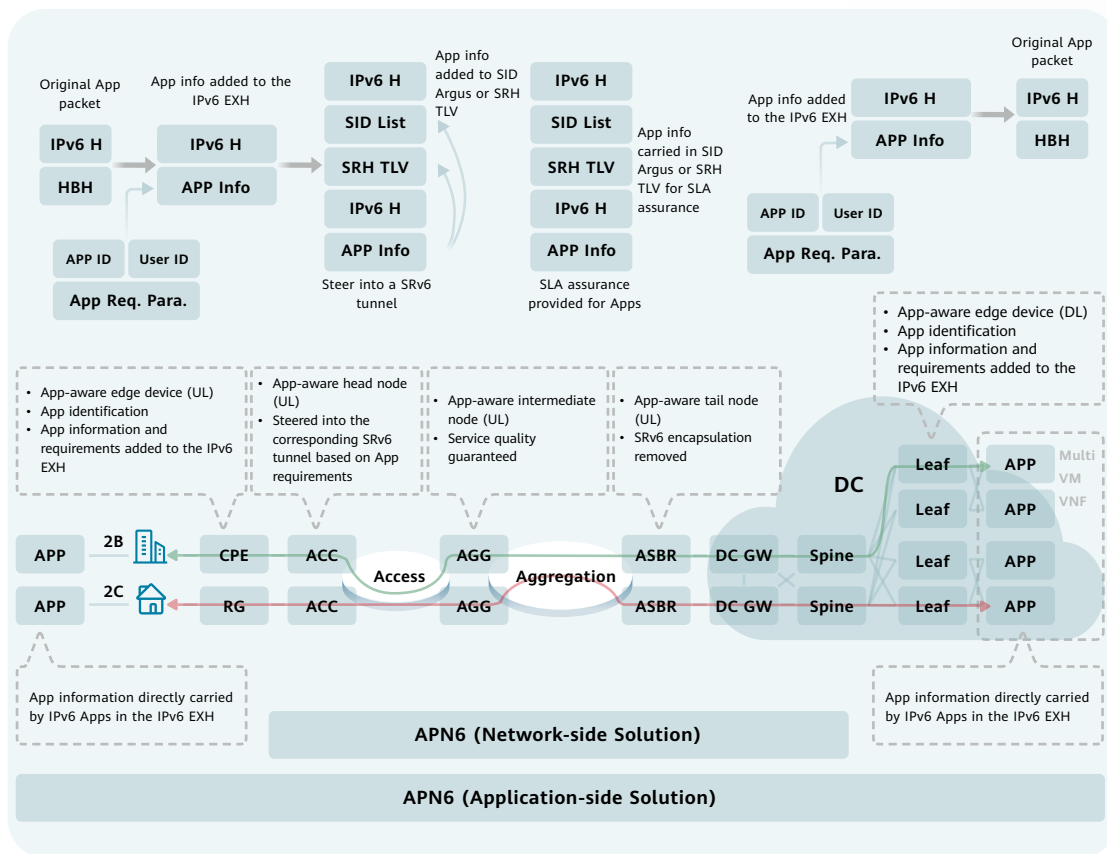


Figure 1: APN6 framework

realizing refined network services and accurate network O&M is key to meeting differentiated application requirements, ensuring SLAs and promoting sustainable network development and evolution. In-depth application-network integration is a feasible path to achieving this goal and has catalyzed the emergence of the APN6 technology.

APN6 technology: Two solutions

How does application-aware IPv6 networking realize intelligent conversations between networks and applications? Key components of the

APN6 framework include applications, network edge nodes, and head nodes, intermediate nodes, and tail nodes that provide network services based on APN information (Figure 1).

Depending on where APN information is added, there are two main types of APN6 solutions, application-side and network-side, each with its own advantages and disadvantages.

In the **network-side APN6 solution**, application and user information is added to packets by network border equipment. The advantage of this solution is that network border equipment

APN6 means more than just enhancement and innovation in IPv6. It provides a foundation for the next-generation Internet protocol.

and network equipment providing services based on APN information are managed and controlled by the same carrier or enterprise and belong to the same trusted domain. Therefore, there are no privacy or cybersecurity issues. Its disadvantage is that the application and user information added to packets may be inaccurate or incomplete. This is because the information is added by network border equipment, instead of applications, but the equipment cannot acquire the information of certain applications.

In the **application-side APN6 solution**, application and user information is directly added to packets by applications. Its advantage is that it ensures the application and user information added to packets is accurate and complete. The disadvantage is that it creates privacy and cybersecurity challenges, as information must be transmitted between different trusted domains such as terminal devices, network infrastructure, and cloud services.

The IETF draft (draft-li-apn-framework) defines the application information (APN attributes)

carried in APN6 packets, including application identifier information (APN IDs) and application requirement parameter information (APN parameters). An APN ID provides information that helps the network distinguish between different application flows and different users (user groups) of a certain application or certain type of applications. Such information may include Application Group IDs and User Group IDs. APN parameters optionally carry information that may include parameters related to application requirements for network performance, such as bandwidth, latency, jitter, and packet loss rate.

APN6 promotes innovation in the next-generation Internet protocol

APN6 generates significant changes to Internet architecture and promotes innovation in the next-generation Internet protocol.

Network architecture consists of three key aspects: identification, forwarding, and control. These aspects have different impacts

on the network architecture. Identification can cause fundamental architecture changes as, when it changes, both forwarding and control must also change, which alters the entire architecture. However, identification changes are also extremely difficult. For example, changes in address identification from IPv4 to IPv6 require upgrade in the entire Internet infrastructure.

The US launched five Internet architecture research projects through Internet2 around 2010. These projects covered Internet architecture transformation, with the more well-known ones including Mobility First and NDN. These projects provided a wealth of valuable ideas about Internet architecture, but were ultimately unsuccessful for three reasons:

1. The Internet's influence is too large, meaning it was incredibly difficult to push for transformation.
2. Transformation was incompatible with the evolution of existing Internet technologies, and fully upgrading network infrastructure was too costly and time-consuming.
3. Existing network software and hardware capabilities could not fully support new network technologies.

APN6 brings important changes to the network identification system. APN6 can be understood as introducing APN IDs, in addition to IPv6 addresses, so that packets contain both IP addresses and identity-like IDs. APN6 means more than just enhancement and innovation in IPv6. It provides a foundation for the next-generation Internet protocol, creating new

space for Internet development. As a result, the IP technology architecture will shift from network services based on IP addresses to network services based on APN IDs or on both IP addresses and APN IDs.

The network architecture upgrade enabled by APN6 differs from the upgrade from IPv4 to IPv6. A key issue encountered during the IPv4-to-IPv6 upgrade was compatibility. Incompatibility between IPv6 and IPv4 addresses means that Internet infrastructure must be upgraded to support IPv6. This is partly why IPv6 deployment has been slow. The APN6-enabled network architecture upgrade is based on an identification system that consists of IPv6 address and APN ID, thanks to two factors:

First, the IP address system and functions are already well defined, and the address space issue is resolved by the upgrade from IPv4 to IPv6. Therefore, we should shift our efforts from an IP-address-based identification system to an identity-like APN-ID-based identification system, which involves allocation, use, and management of APN IDs, and define various protocol extensions on that basis.

Second, by learning lessons from the IPv4-to-IPv6 upgrade, incremental deployment and upgrade can be performed based on IPv6's backward compatibility mechanism. With an APN ID carried in the IPv6 extension header, an IPv6 node on the network provides services based on the APN ID if the node identifies the APN information in the packet. If the IPv6 node cannot identify the APN information in the packet, it considers the packet to be a common

IPv6 packet, and forwards it according to the destination IPv6 address. This allows networks to evolve incrementally.

The combination of IPv6 addresses and APN IDs can be regarded as another expansion of the IP address space. The identifier, which was previously based on a 128-bit IPv6 address, has been lengthened to include more bits — there are now

three types of APN IDs: 32-bit, 64-bit, and 128-bit. This means an APN ID used in combination with a 128-bit IPv6 address is equivalent to a 160-bit, 192-bit, or 256-bit identifier.

The network architecture upgrade in question is made possible with a new identifier—APN ID—in addition to the IPv6 address.

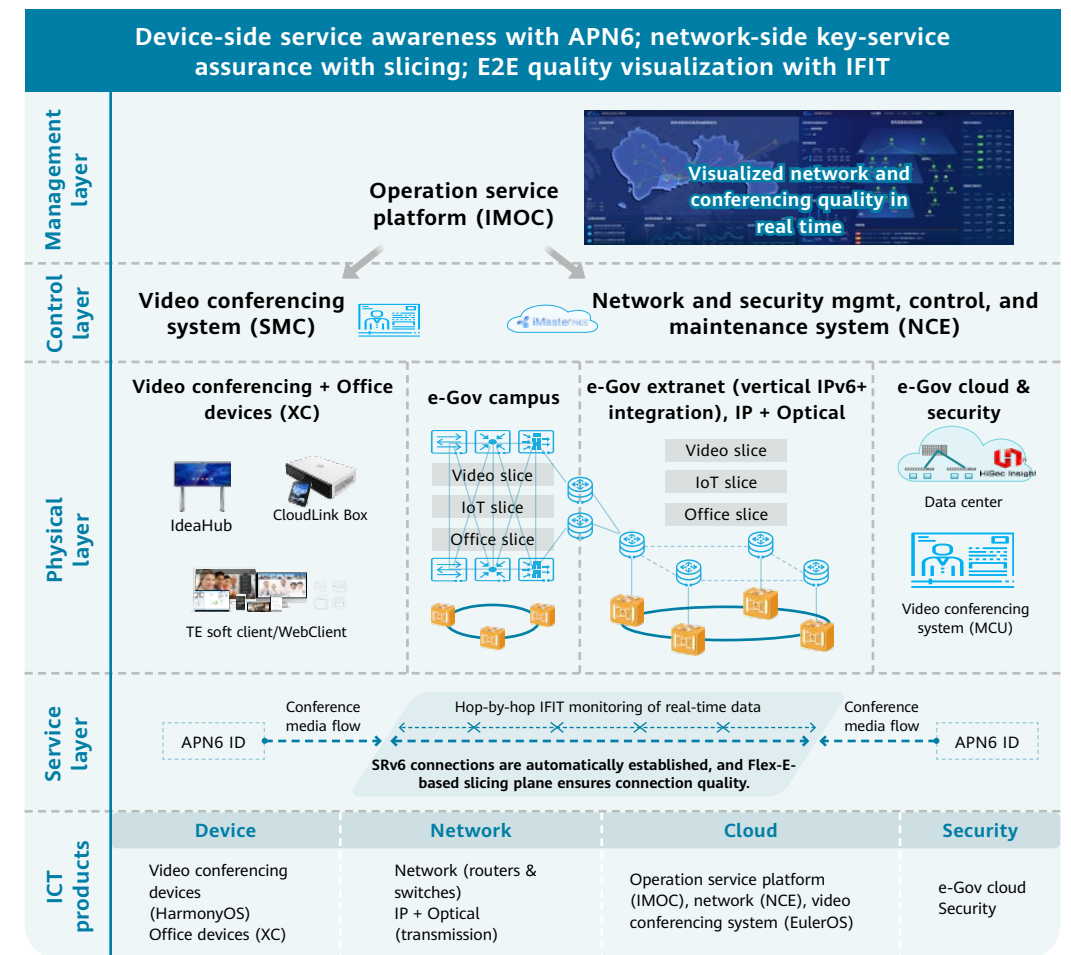


Figure 2: APN6-based video conferencing experience assurance solution

The combination of IPv6 addresses and APN IDs can be regarded as another expansion of the IP address space.

APN6 deployment use cases and technical value

An APN6-based video conferencing experience assurance solution has been deployed on an e-Government extranet (Figure 2). This means APN6-based solutions are already commercially viable for live networks.

Traditional video conferencing assurance is difficult as it is highly labor-intensive and networks cannot support targeted assurance services due to a lack of integration between the video system and the network, meaning faults during video conferencing are difficult to locate.

With the APN6-based video conferencing assurance solution, IPv6 packets sent by the video conferencing application carry the application ID corresponding to the video conference that needs assurance. Equipment

at the border of the IP transport network can then steer the traffic into a network slice dedicated to the video service in accordance with the application ID. With resource isolation supported by IP network slices, other services will not affect traffic in the video slice, thus guaranteeing good experience for video conferencing users. Furthermore, the IP transport network performs in-situ flow information telemetry (IFIT) for applications based on application IDs to provide application-level, service-quality visualization. This makes it possible to both quickly locate issues and optimize traffic when video conferencing experience is poor.

APN6 can also be applied to scenarios such as computing-network convergence (Figure 3). Latency is critical to good XR service experience. Latency assurance depends not only on network load, but on computing power load on the edge cloud. Therefore,

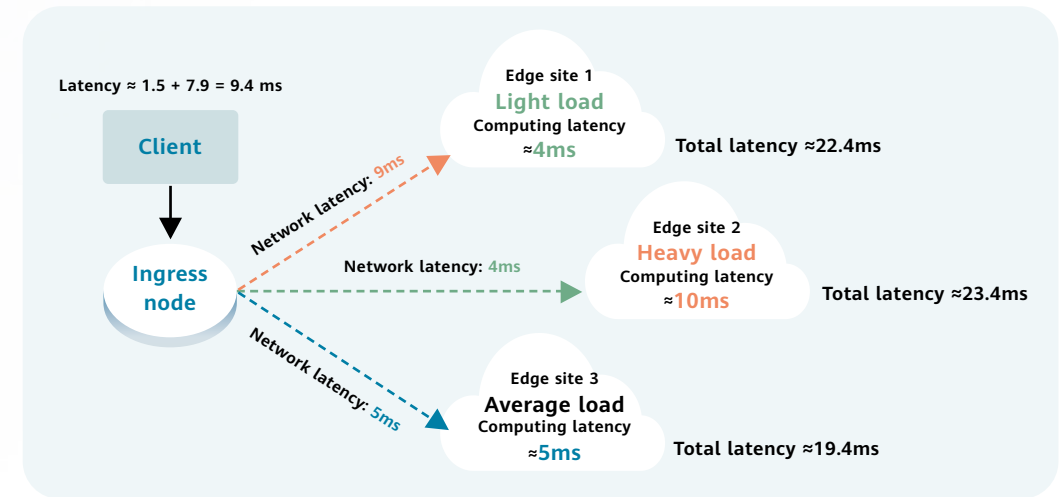


Figure 3: XR experience assurance through optimization based on both network and computing power loads

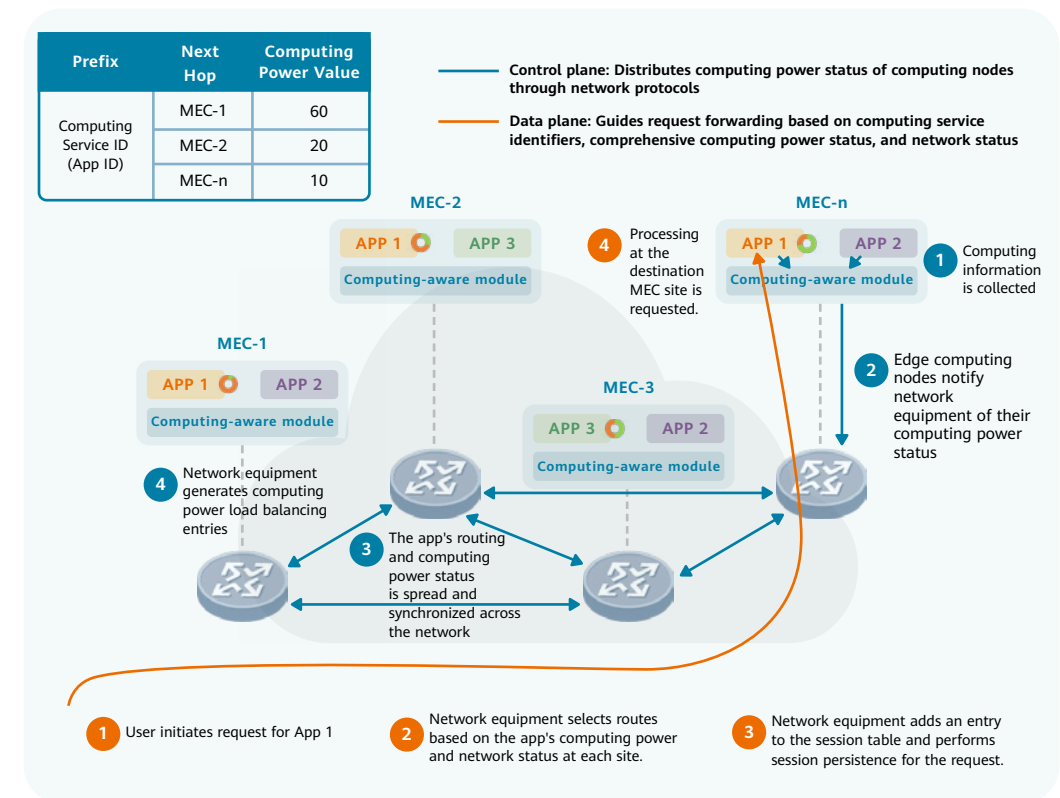


Figure 4: Computing-aware traffic steering solution

traffic optimization must be performed with consideration for both network load and computing power load.

A computing-aware traffic steering solution implements comprehensive scheduling based on the network and computing power loads (Figure 4). With this solution, the network must be notified of the location and application identifiers and load information of a computing service. The head node of the network forms multiple path-forwarding rules based on received information like different computing power locations and loads of the same application, points to different computing service locations, and identifies different comprehensive load information relating to network load and computing power load. When receiving new service

flow packets for accessing an application, the head node can select a path with a light load, and direct the flow to the corresponding computing service location, thus guaranteeing optimal application access experience.

The aforementioned two sample APN6 applications also show the benefit of separating locators from IDs. Traditional Internet packets have only an IP address as the identifier, meaning the IP address serves both as a locator and an ID. In mobile settings, application addresses constantly change, and an IP-address-based identifier system means the address and ID will both change, making it necessary to adjust the IP-address-


based traffic steering policy. APN6 introduces an identity-like ID (APN ID) beyond IPv6 addresses, so that the ID remains unchanged when the application address changes. This means an ID-based network policy can still work without adjustment. This mechanism greatly simplifies network O&M.

More agreements reached on standardization

Since early 2019, over a dozen IETF drafts on APN6 technology have been submitted, including the problem statement, framework, IPv6 encapsulation, YANG model, and BGP extensions. APN6 poses challenges to traditional

Internet design principles such as the end-to-end principle and the application-network separation principle.

Early on, APN6 caused many debates in the standardization community. But over time, a greater consensus has been reached. APN6 can be preferentially applied to limited domains where both the application and network are controlled, but further development is required for Internet applications in which more third parties are involved.

Regardless, APN6 has made huge progress in terms of both theory and practical applications. The advancement of emerging services like AI computing is leading to a clearer trend in cloud-network-edge-device convergence, which will pave the way for the further development of APN6. 

AI Data Lake: Breaking Silos and Accelerating Intelligence



Michael Fan

Director, Data Storage Marketing Execution Dept, Huawei

Huawei has launched a three-layer AI data lake solution to break down the data silos that are preventing many enterprises from reaping the benefits of AI models.

Since the release of ChatGPT in November 2022, large AI model technology has been on a rapid upward trajectory, with model training emerging as the driving force behind technological progress. However, the data volume and diversity of data types involved in model training are increasing exponentially, meaning that data silos pose a major barrier to progress.

Many organizations are struggling to see the way forward.

Huawei's AI data lake solution, however, does not just create bridges between data silos, it lights the way to a more intelligent world and a new era of AI innovation and development.

Trends and challenges of large AI model services

In 2023, Google released the Gemini multimodal model, which is able to understand, operate, and integrate different types of information, including text, code, audio, images, and video. In February 2024, OpenAI released a video model called Sora, which can create realistic scenes from text instructions. By combining a diffusion model with the large language model, Sora exhibits amazing 3D consistency when learning from the physical world.

The speed with which large AI models have developed has far exceeded people's expectations. Witnessing the progression from ChatGPT to Gemini and now to Sora, we have identified two key trends:

- **Trend 1:** The shift from pure NLP models to multimodal models has brought a

corresponding shift in training data. Training datasets now encompass a mix of text, images, audio, and video, rather than text alone. This means that the volume of data required to train today's most advanced models is a staggering 10,000 times greater than was required for earlier models (as shown in Figure 1).

- **Trend 2:** Computing power, algorithms, and data are the three core elements of large AI models. Through the stacking of computing power, aggregation of data, and parameter scaling (from hundreds of billions to trillions, and even up to 10 trillion), complex behaviors emerge within the framework of deep learning algorithms. For example, in a video released by Sora, we witness a stylish woman strolling down a street. As she moves, the perspective on the street scenes behind her seamlessly adjusts. As she walks by, objects and other pedestrians are momentarily obscured, and

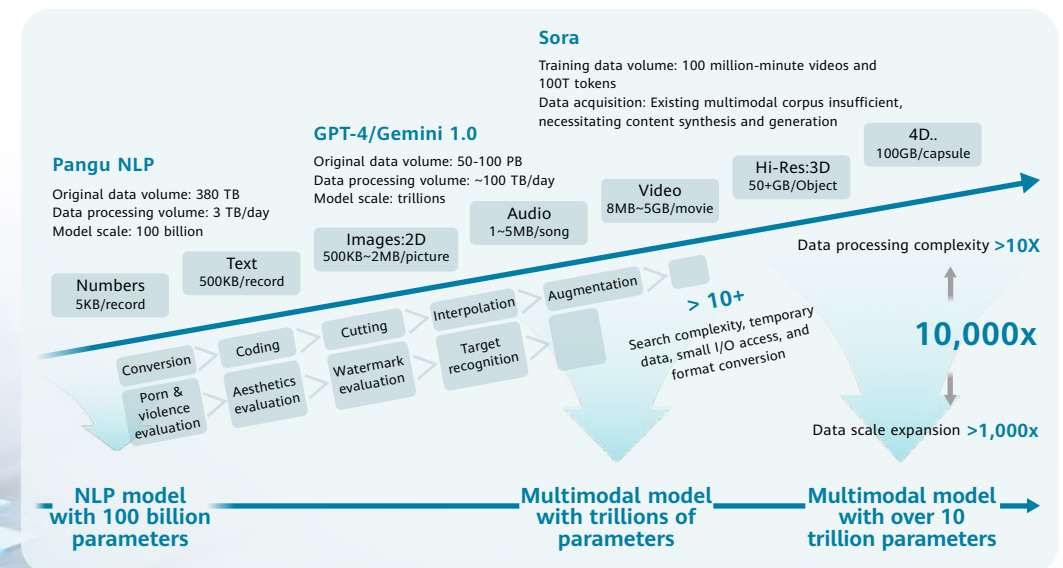


Figure 1: Multimodal models for exponential growth in the volume of training data

then reemerge, maintaining impeccable 3D consistency and object permanence, faithfully replicating how the world is perceived through human eyes.

The transition from unimodal to multimodal models will create a deluge of data, and a solution is urgently required that can integrate scattered data resources. As a way of presenting the real world, data serves as the cornerstone of training large AI models. With deep learning algorithms, the quantity and quality of data determines how far AI training will go. However, most data owners prioritize the efficient access of data by service applications, often overlooking where data is stored. Similarly, most data administrators focus on whether data is effectively stored, and may neglect data ownership and types. As a result, data ends up siloed in different data centers. One carrier we spoke to has seen the total amount of data they need to store soar to hundreds of PBs in recent years, and hundreds of TBs of new data is generated each day. However, this data is scattered across multiple data centers.

To provide as much data as possible for model training, the carrier's technical department has to migrate or replicate the data from scattered data silos across regions. As a result, the data preparation process is prolonged, accounting for over 50% of the entire model training duration.

Therefore, the biggest challenges and primary considerations for constructing large AI model infrastructure revolve around eliminating data silos and effectively gathering data in one place, swiftly transforming collected datasets into model training materials, and ensuring

that AI computing power has efficient access to training materials.

New requirements for data storage and management

The ideal AI data infrastructure should be able to support the key phases of large AI model training, such as data ingestion, data preprocessing, and model training, to provide high-quality data services. To achieve this goal, at least two layers of data infrastructure should be considered: the storage device layer and the data management layer.

Storage device layer

When dealing with a large amount of heterogeneous data from multiple sources, especially in multimodal AI training scenarios, the ideal storage device layer should feature multi-protocol interworking, high I/O performance, and easy scalability to address various challenges and support the following pivotal phases of training large AI models:

- **Data ingestion:** In this phase, data is often scattered across different silos, stored in various formats, and accessed through different protocols. To efficiently centralize heterogeneous data from different sources, data storage devices need to support various data formats and access protocols, and deliver high write bandwidth. Additionally, storage solutions need to be both flexible and scalable, while also remaining cost-effective, in order to handle new data sources that may be incorporated into training at any time. While data formats

The ideal AI data infrastructure should be able to support the key phases of large AI model training, such as data ingestion, data preprocessing, and model training.

and access protocols may vary during the data ingestion phase, servers only access data through the file interfaces during the training phase. Therefore, ideal storage hardware should support multi-protocol interworking to ensure that underlying data can be accessed by different protocols or interfaces, thereby eliminating the data replication caused by protocol conversion.

- **Data preprocessing:** Data preprocessing involves the cleansing, transformation, augmentation, and standardization of diverse data to yield high-quality training materials from massive volumes of raw data. In this phase, a large amount of temporary data is generated due to the diversity of preprocessing tools, resulting in data expansion. Therefore, storage devices should offer extensive shared storage capacity while also delivering high read/write bandwidth and supporting random access to expedite preprocessing.
- **Large AI model training:** In the model training phase, the performance of storage devices in tasks such as training data

loading and checkpointing directly affects training efficiency. Although the volume of pre-processed training data is small, high file access performance (OPS and IOPS) and low latency are required to ensure quick data loading and avoid wasting GPU/NPU computing power. Saving checkpoints during the training process is critical to ensure that, in the event of an interruption, training can be resumed rather than started from scratch. Therefore, it requires high write bandwidth for fast and frequent archiving and more stable training.

Data management layer

Based on the flexible capacity expansion and high hybrid-load performance provided by the storage device layer, the data management layer further provides advanced data management capabilities for AI training. It helps data owners and administrators maximize data value in a more efficient manner from three dimensions: visibility, manageability, and availability.

- **Visibility:** Data asset owners and administrators should have a comprehensive understanding of their data, including data storage location, data volume, and data types, akin to having a data map. Considered from this dimension, owners and administrators can quickly identify which data needs to be collected based on this map.
- **Manageability:** After determining the data to be collected, a mechanism is required to implement policy-based data flow. For example, a policy may be used to define the source and target of data flow, start and end time window, maximum rate limitation, and minimum rate guarantee, facilitating the management of data.
- **Availability:** Raw data needs to be preprocessed and converted into training data. While there are numerous tools for data preprocessing, it is still important to have a data preprocessing framework that works in synergy with the

storage device layer. Such a framework should help users streamline data preprocessing and enhance processing speed to improve data availability.

Core capabilities of AI data infrastructure

The ideal AI data infrastructure should have the core capabilities outlined in Figure 2.

To sum up, three key features are indispensable:

- **High performance**
High-performance data infrastructure underpins and accelerates every phase of large AI model training, from data ingestion and preprocessing in the early stage to training data loading and periodic checkpointing in the model training phase. High performance in this context is comprehensive and includes

both high OPS and low latency for random access and high read/write bandwidth for sequential access.

- **High capacity**
The data ingestion and preprocessing phases involve the intensive reading, writing, and generation of temporary data, which creates storage challenges. Data infrastructure therefore needs to offer flexible and non-disruptive scale-out and tiering capabilities. This ensures a balance between cost and capacity, while delivering the necessary bandwidth, IOPS, and OPS to handle sequential and random data access patterns.
- **Exceptional ease-of-use**
Data infrastructure should feature global data management, efficient data mobility, AI platform and toolset collaboration, and optimization and enhancement tailored for large AI model training. This would accelerate AI training and learning by ensuring that data is visible, manageable, and available.

AI data lake solution

Huawei has been actively engaging in large AI model training with customers in diverse industries, including carriers. Over the years, we have gained extensive expertise in AI data infrastructure. Recently, we have launched an AI data lake solution to help customers solve the problems encountered in the deployment of data infrastructure for large AI model training. This enables our customers to focus on model development and training. The architecture of the Huawei AI data lake solution, as shown in Figure 3, consists of three layers: a data storage layer, a data fabric layer, and a data service layer.

Data storage layer

In this layer, data is stored in different data centers.

In each data center, data is intelligently stored in the hot and warm tiers. The hot tier is

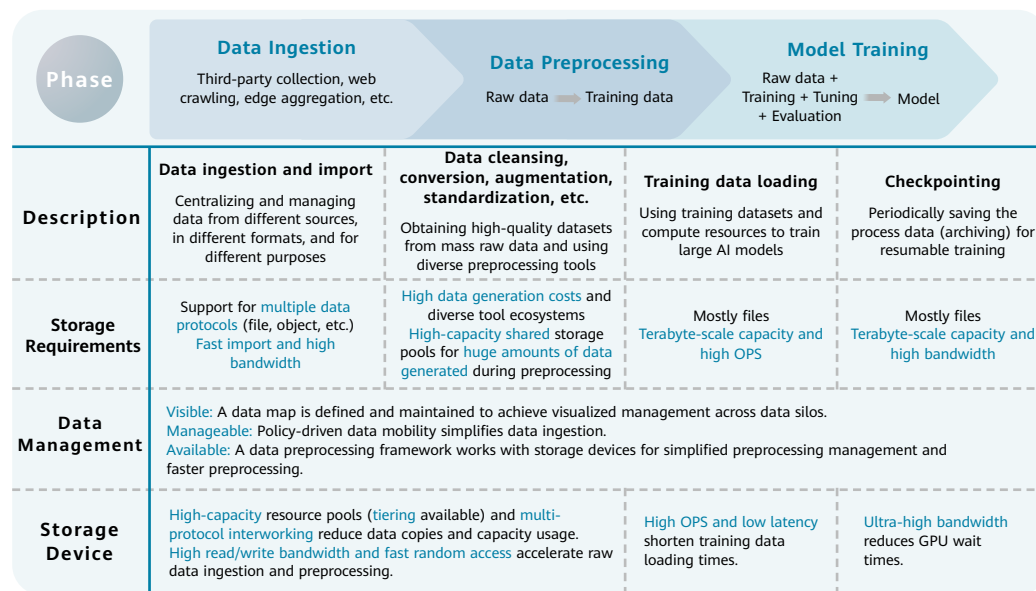


Figure 2: Core capabilities of AI data infrastructure

The architecture of the Huawei AI data lake solution consists of three layers: a data storage layer, a data fabric layer, and a data service layer.

Intelligent tiering can be implemented between the OceanStor A series and the OceanStor Pacific series. The two tiers are combined to externally present a unified file system or object bucket. Internally, data is intelligently and automatically tiered.

powered by the OceanStor A series, which is Huawei's high-performance storage specially designed for large AI model training. It can scale out to thousands of nodes. The warm tier uses Huawei's OceanStor Pacific scale-out storage to handle mass unstructured data. Intelligent tiering can be implemented between the OceanStor A series and the OceanStor Pacific series. That is, in the same storage cluster, the OceanStor A series nodes form a high-performance storage tier, and the OceanStor Pacific series nodes form a high-capacity storage tier. The two tiers are combined to externally present a unified file system or object bucket that supports multi-protocol interworking (one copy of data can be accessed through multiple protocols). Internally, data is intelligently and automatically tiered to achieve an optimal balance between capacity, performance, and cost.

A data replication relationship can be created between storage clusters across data centers to facilitate reliable and on-demand data mobility between data centers. This ensures the data device layer is fully prepared to support data ingestion for large AI model training.

Data fabric layer

The data fabric layer creates a seamlessly interconnected network to facilitate data mobility. It makes data visible, manageable, and available, helping large AI model training derive value from data.

Huawei uses a software layer powered by Omni-Dataverse to achieve data visibility, manageability, and availability. Omni-Dataverse is an important component of Huawei's Data Management Engine (DME). It provides the unified management of metadata on Huawei storage devices across different data centers to form a global data asset view. Omni-Dataverse also invokes interfaces on storage devices to control data movement. Omni-Dataverse executes actions based on user-defined policies, and can implement GPU/NPU-storage passthrough and intelligent file prefetching on demand to ensure training data is readily available, minimizing computing power wait times.

This makes data ingestion and model training much more efficient, improving cluster utilization.

Data service layer

The Huawei AI data lake solution provides commonly used service frameworks at the data service layer, including data preprocessing, model development, and application development frameworks.

The data preprocessing framework performs data cleansing, conversion, augmentation, standardization, and other preprocessing operations. Large AI model customers can integrate their algorithms and functions into this framework for simplified preprocessing management. Furthermore, customers can choose an alternative framework.

Similar to data preprocessing, model development and application development are the other two frameworks provided by Huawei for customers'

convenience, which they can select on demand.

The Huawei AI data lake solution is built on years of experience and expertise in large AI model training. It helps enterprises eliminate data silos, achieve smooth data mobility, and implement a data fabric between data applications and storage devices, making data visible, manageable, and available. As large AI models continue to evolve from unimodal to multimodal, the ever-increasing volume and diversity of data will inevitably lead to a non-linear increase in management complexity and performance requirements. An AI data lake solution running on a three-layer architecture can effortlessly address these challenges to facilitate the development of large AI models, accelerate the emergence of intelligence in model training, and push AI innovation and development to new heights. **IT**

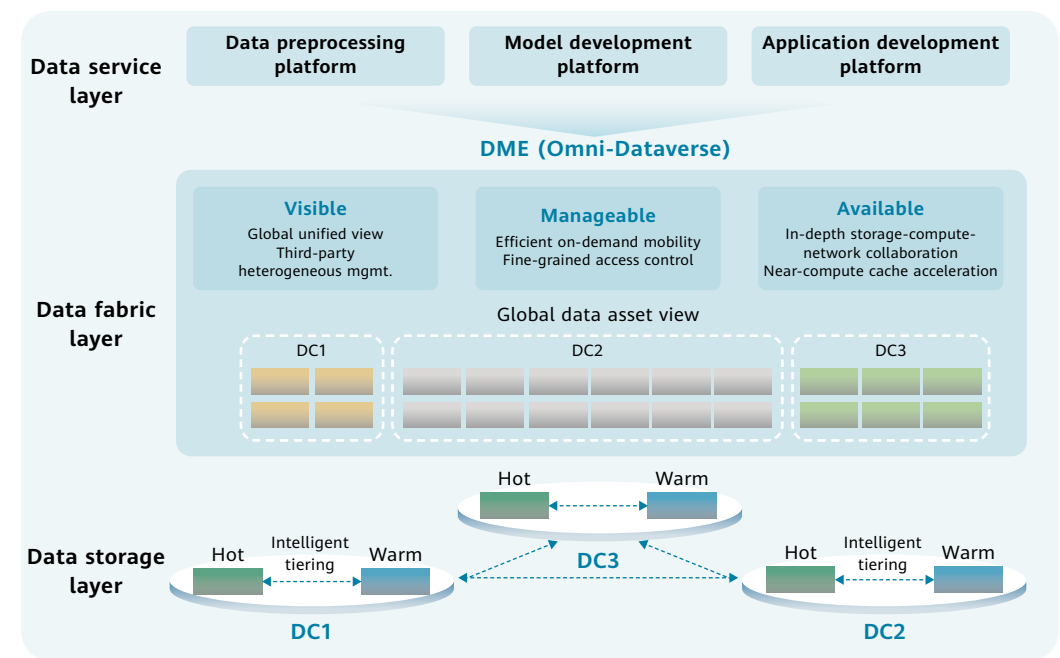



Figure 3: Architecture of Huawei AI data lake solution



03. Intelligent Transformation

Accelerating Intelligent ICT with a Four-layer Framework and Transformation in Three Areas

This article provides a detailed analysis of innovative AI applications in business, telecom operations, network-cloud synergy, and infrastructure. It discusses how Huawei uses a four-layer framework and transformation in three areas to help the industry go intelligent faster and help carriers seize new opportunities and overcome new challenges. This demonstrates Huawei's vision to work with partners to promote technological innovation.

Wang Su

Director, Integrated Solution Marketing Dept, Huawei



Rapidly evolving AI technology is bringing unprecedented opportunities and transformation potential to telecom carriers. At the business level, AI applications can help carriers increase revenue by developing smart and easy-to-use services. For telecom operations, AI can also help

carriers manage and optimize network resources better and improve O&M efficiency. At the same time, carriers need to explore and apply ICT solutions to support AI by providing stronger data connection, storage, transmission, and computing capabilities.

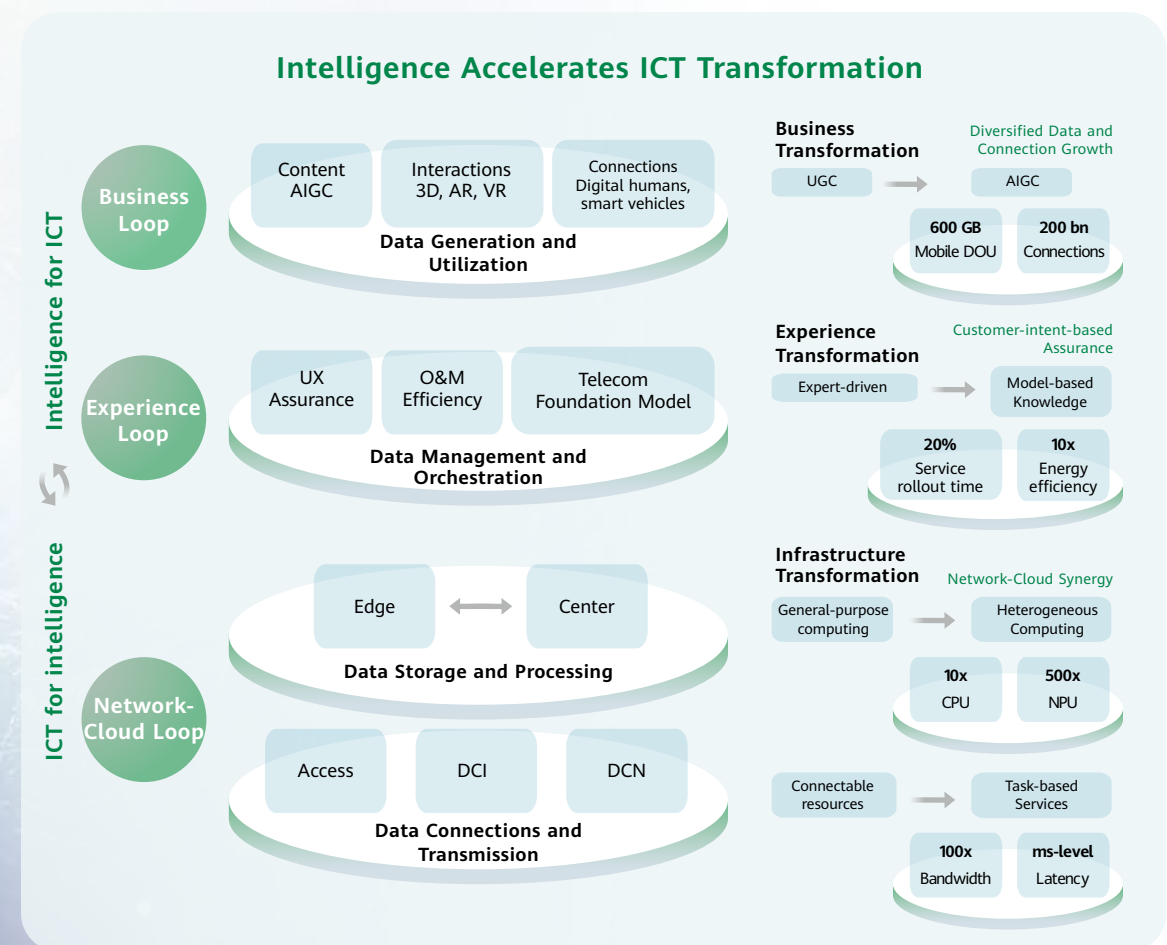


Figure 1: Four layers and transformation in three areas

New digital content will increase network traffic by over 10 times and increase the number of connections from tens of billions to hundreds of billions, meaning new business opportunities for carriers.

Centering around data, Huawei has proposed a four-layer framework and transformation in three areas for the intelligent era (Figure 1). The four layers are service, network management, cloud infrastructure, and network infrastructure. The three areas of transformation are business, experience, and infrastructure. Intelligent technology has changed the way data is produced, managed, applied, stored, processed, and transmitted on the four layers, and carriers need to adapt to transformed business models, O&M-based experience, and cloud-network infrastructure in the intelligent era.

Business transformation: A new chapter for AIGC and business value

New ways of producing data are driving business transformation. For example, AI is learning to create new worlds through AI-generated content (AIGC). This will potentially take AI to a new level—artificial general

intelligence (AGI)—which will drive transitions to new business models. In the text-to-image and text-to-video fields, foundation model algorithms and applications are maturing, with diffusion-model-based algorithms already being widely used for image generation. The Transformer architecture used by text-to-video models like Sora follows scaling laws and realizes video generation abilities.

AIGC is dramatically transforming content generation and seeing fast adoption in fields like picture generation, short video creation, and movie production. AIGC is also transforming connectivity from connections between people to comprehensive connections between people, objects, digital humans, smart vehicles, and drones. These new connections involve not only manual operations, but automation and machine-to-machine communications. In addition, AIGC is transforming interactive experiences by combining AI, 3D, VR, and AR

technologies to enable a leap from content creation to holographic interactions.

Business (or business cycle) transformation means new ways of producing data are increasing content production efficiency (machines, not just humans, as producers), content complexity (HD, 3D, and holographic content) and distribution speeds (faster interactions with and ubiquitous access to content). This is generating massive amounts of data and connections. It is predicted that by 2030, new digital content will increase network traffic by over 10 times and increase the number of connections from tens of billions to hundreds of billions, meaning new business opportunities for carriers. New ways of producing and transmitting data are also transforming business models to enable innovation in personalized products and services and improve user experience and loyalty.

Experience transformation: A future with intelligent network O&M

In the field of network O&M, intelligent technologies are changing service rollout, resource configuration, and O&M processes. These all center on transforming user experience. The intelligent transformation of network infrastructure will follow three trends:

First, from event-driven networks to intent-driven networks. Traditional event-driven networks respond only to specific events, while intent-driven networks can understand user intent and adjust network configurations to user needs rapidly and efficiently.

Second, from empirical decision-making to AI-assisted decision-making. Traditional network management decisions are made based on expert experience and analyses. With the help of AI, network management personnel



Huawei's innovative elastic transport networks support task-based data express services for users.

can make decisions more efficiently using AI algorithms and data analytics, which also improves network performance and reliability.

Third, problem-solving evolution from scenario-specific models to foundation model generalizability. Traditional AI models are created to solve specific problems in specific scenarios. Foundation models can be used to create more generalizable problem-solving systems. When used alongside scenario-specific models, they can solve complex problems through analyses that streamline different network scenarios and processes.

Telecom networks use the following three layers of solutions to make network O&M intelligent:

- **Network element (NE) intelligence:** More real-time sensing components and AI inference capabilities can be added to network equipment. AI-native hardware will enable networks to be capable of finer-grained sensing and real-time synchronization.
- **Single-domain intelligence:** An intelligent management, control, and analysis platform can be used to create digital models of networks that associate

discrete data related to network resources, services, and status, and provide digital twins adapted to different domains. This platform's single-domain, high autonomy capabilities, from data collection, sensing, and analysis to simulation, decision-making, and control, can ensure guaranteed network connection quality and timeliness. For example, Huawei provides an IP-based digital map with six visualized layers—physical, network, slicing, routing, service, and application—as well as capabilities such as congestion view, experience view, and fault view.

- **Cross-domain intelligence:** Collaboration can be performed across domains. Huawei provides intelligent platforms such as NCE-Super, ADO, and SmartCare to support collaboration across domains, such as fault demarcation through coordination between IP and optical domains, and intelligent orchestration for cross-domain services.

Telecom foundation models will be key to realizing intelligent network O&M. Carriers' complex O&M processes in all manner of scenarios across domains will require the

capabilities of telecom foundation models and single-domain autonomy to implement cross-domain collaboration and the autonomous completion of complex processes. Huawei's Telecom Foundation Model, for example, consists of three layers: the foundation model, telecom industry models, and applications. The foundation model layer provides multimodal, computer vision, natural language processing (NLP), scientific computing, and prediction based on Huawei's Pangu models and third-party open-source foundation models. The telecom industry model layer provides industry-specific capabilities using high-quality industry corpuses and efficient toolchains. The application layer provides role-based copilots and scenario-based agents to improve employee efficiency and customer satisfaction. Huawei's Telecom Foundation Model provides out-of-the-box applications and supports local deployment and incremental training by carriers to deliver intelligent technologies that meet different user needs.

O&M experience transformation uses intelligent technologies like foundation models, and the three-layer intelligent telecom network solutions to bring intelligence to network infrastructure and build AI-native, intent-driven, and digital twin capabilities. These features can help carriers guarantee user experience, agilely roll out services, optimize resource configuration, and improve O&M efficiency, accelerating network evolution towards autonomy.

Infrastructure transformation: The network foundation for the AI computing era

The nascent AI computing era is seeing a major leap in computing technology from general-purpose computing to AI computing. This means the number of foundation model parameters and demand for computing power will grow significantly. Trillion-parameter foundation models will require AI computing centers that deliver 10,000-GPU computing power and efficient data management. This is making coordination across computing, storage, networks, management, and efficiency a key trend. Networks that support AI service growth are critical to realizing ubiquitous computing power and efficient data flow. End-to-end collaboration between devices, edge, and cloud across all manner of scenarios will improve overall computing efficiency.

Two technologies are key to achieving network-cloud synergy.


The first is hyper-converged DCNs. AI computing is characterized by small data flows with

large packets, which can easily lead to load imbalance, decreasing network throughput and reducing computing efficiency. Hyper-converged Ethernet DCNs can meet data center service requirements in different phases of development and different scenarios. Ethernets that integrate general-purpose computing, storage, HPC, and AI improve deployment and O&M efficiency and reduce maintenance costs. Huawei's innovative network scale load balancing (NSLB) algorithm, for example, optimizes traffic paths to achieve global load balancing, delivering 20% higher Ethernet efficiency than the industry average and 10% higher than IB.

The second technology is elastic data center interconnect (DCI) and data center access (DCA) networks. AI computing requires fast transmission of massive amounts of data for training and inference. Traditional private lines are constrained by their lack of service elasticity and high costs. Superior to the current common practice of shipping hard drives via express delivery, Huawei's innovative elastic transport networks deliver end-to-end 400GE, SRv6, and elastic high throughput. These networks use AI-powered data flow identification and

path orchestration, as well as high throughput and high-reliability elastic 10GE private lines, to provide users with elastic, task-based data express services (same-hour, same-day, or next-day delivery).

Traditional cloud infrastructure urgently needs upgrading in areas such as data center power supply and cooling, hardware infrastructure clustering, and the integrated deployment and collaborative O&M of computing, storage, and network resources. Key requirements for network infrastructure in the cloud and intelligence era include ultra-high bandwidth and ultra-low latency, moving storage and computing from devices to the edge, high security and reliability, and task-based network services.

The four-layer framework and three-area transformation were proposed to systematically outline the carrier opportunities presented by the intelligent era and help carriers build the required capabilities to seize these opportunities. The convergence of networks, cloud, and intelligence is already a major trend, as new opportunities are created by intelligence powering ICT, and demand for infrastructure upgrades is generated for ICT to boost intelligence. Huawei looks forward to working with customers and partners in exploring new architectures, applying new solutions, and jointly developing new technologies to reap the value of intelligence for data and traffic. 

EM 2.0: A Road for the Digital Intelligent Transformation of African Carriers

The EM 2.0 model is a partnership between Ethio Telecom and Huawei that has accelerated digitalization in Ethiopia and provides a valuable roadmap for transformation in emerging markets worldwide.



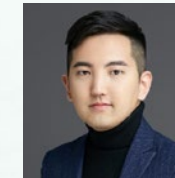
Chris Meng

Director, Northern Africa ICT Marketing & Solution Sales Dept, Huawei



Wang Jie

Director, Northern Africa Marketing Dept, Huawei



Liu Jifan

CEO, Huawei Ethiopia

The light of digitalization is now shining on Africa, transforming the continent into one that is brimming with dynamism and a growing digital economy.

The Emerging Market 2.0 (EM 2.0) model – a result of successful collaboration between ethio telecom and Huawei – represents the marriage of technology and innovation. EM 2.0 is more than a technology upgrade. It also exemplifies a profound change in mindsets and models. By optimizing digital infrastructure, building a digital cloud foundation, deploying digital O&M platforms, and enabling diverse digital services, EM 2.0 has established a new blueprint for Africa's digital economy.

With 5G and other innovative initiatives pushing Africa to the forefront of the global digital economy, EM 2.0 is a model for success that provides valuable experience and inspiration for other emerging markets.

Digital and intelligent

According to the *White Paper on Global Digital Economy (2023)* released by China Academy of Information and Communications Technology (CAICT), the digital economy already accounts for 46.1% of GDP in major economies, making it an increasingly important engine for economic growth worldwide. The digital era is presenting conventional telecom markets with rare opportunities to transform, and world-leading carriers have started exploring ways to grow through transformation. They are employing innovative ICT solutions to transform themselves from traditional network operators that provide only connectivity services into digital solution providers that offer diversified services and that can even serve as enablers of national digital transformation.

Africa is an emerging market with huge demographic dividends — on average, it gains

1.5 new mobile Internet users per second. The continent is expected to have 620 million home broadband users by 2025, nearly double that seen in 2021. At the same time, 40% of the African population uses mobile payment services today. According to GSMA, the total number of registered mobile money accounts worldwide reached 1.75 billion in 2023, with Sub-Saharan Africa accounting for over 70% of total growth. As the digital economy takes the African continent by storm, the most strategically important issues facing African carriers involves how to transform business operation models, address new market needs during digitalization, and achieve new business success.

ethio telecom's successful EM 2.0 transformation

With a 129-year history, ethio telecom is a state-owned telecom carrier and the second largest carrier in Africa after MTN. It has 69.5 million mobile users and dominates Ethiopia's

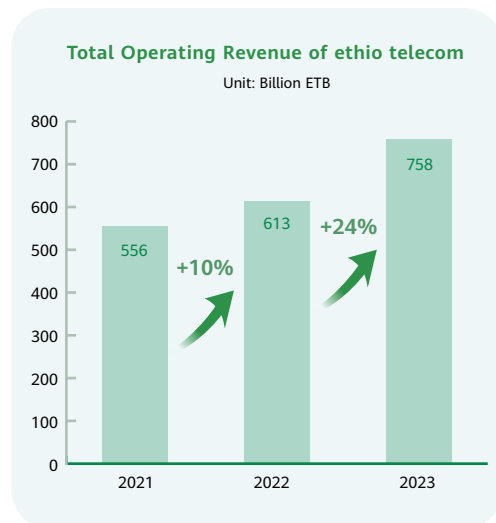


Figure 1: Revenue increase over the past three years

telecom market. The emergence of new telecom carriers has intensified competition in the local market, leading ethio telecom to carry out a range of transformation initiatives that ensure sustainable business growth and empower it to seize opportunities presented by Ethiopia's digital economy.

The carrier has worked with Huawei to drive digital intelligent transformation through the EM 2.0 model by building high-quality digital infrastructure, creating cloud infrastructure, deploying a digital O&M platform, and developing innovative digital services. This has helped ethio telecom transform from a traditional telecom network operator into a digital and intelligent solution provider that has achieved enormous business success (Figure 1 and Figure 2).

Unlike EM 1.0, which focuses on connectivity, EM 2.0 (Figure 3) uses cloud-based digital O&M to enable digital services, forming a future-

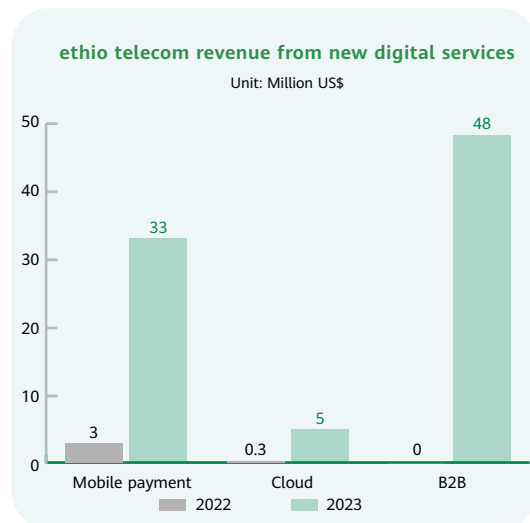


Figure 2: Growth of three new digital services

oriented business model that embraces network, cloud, and intelligent transformation. From an architecture perspective, local industry policies and the carrier's basic networks remain as the basis of the EM 2.0 model.

The model reshapes infrastructure by integrating the cloud platform through multi-cloud convergence and cloud-network synergy, enabling robust O&M, ensuring a good user experience and higher efficiency, and supporting the development and operation of innovative and diversified digital services that create a new growth curve.

The new networks, platform, and services support one another and create a ladder for continuous progression.

Strategy, collaboration, and capabilities

How did ethio telecom find its path forward and succeed in digital and intelligent transformation?

In 2021, the Ethiopian government licensed a new telecom carrier, shaking up the once calm market. This meant that ways to maintain robust growth in the face of intense competition moved to the top of the list of concerns for the CEO and other top executives of ethio telecom, an established carrier that mainly provided traditional connectivity services. At that time, the government was preparing a strategy to develop the county's digital economy. ethio telecom analyzed the local market environment and trends, and planned to implement digital

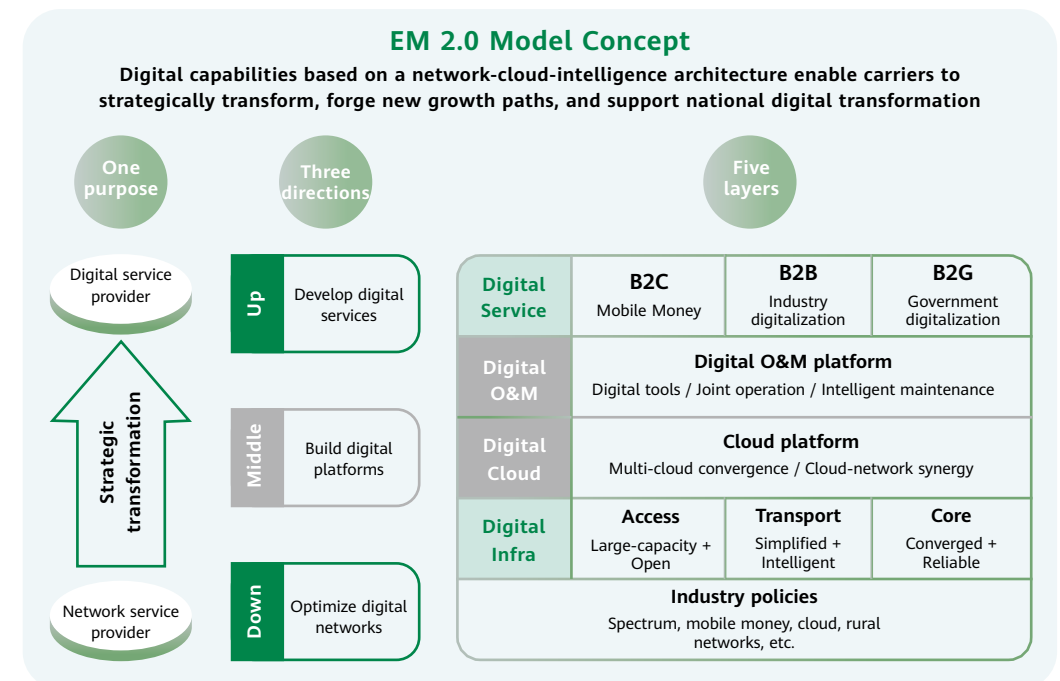


Figure 3: Conceptual framework of the EM 2.0 digital transformation business model

and intelligent transformation based on its business development and operation model. To achieve this, the carrier built two foundational capabilities: mobile payments and a cloud platform. It planned to develop diversified and innovative digital services and solutions for consumer and enterprise markets, expand service boundaries, and compete at a higher level.

In May 2021, with the support of Huawei products and services, ethio telecom successfully launched the country's first carrier mobile money brand "telebirr" after just five months of development, which was key to its transformation. By the beginning of 2024, telebirr had over 40 million users, generating a transaction volume of 91 billion birr (about US\$16 billion). As a result, the service received the gold award from Future Digital Accreditation Institute in 2023 in the best mobile money supply category.

However, these changes were accompanied by a number of challenges.

The first challenge was a lack of revenue despite an increase in users. Within a year of its launch, the mobile money product's user volume quickly increased to 22 million, but service revenue remained low. The service's actual usage rate was low due to the carrier's inexperience at operating new services, a lack of a specialized team, and the service's limited functions (deposits, withdrawals, and transfers). These factors combined led to weak profitability. As a result, the product generated only around US\$3000 in revenue per month, barely contributing to the company's overall business and putting enormous ROI pressure on ethio telecom. The carrier responded with measures to overcome this challenge.

By the beginning of 2024, telebirr had over 40 million users, generating a transaction volume of 91 billion birr.

First, the company adjusted its organizational structure (Figure 4). It established an independent Mobile Money business department headed by the CMMO, optimized the functions of the CMO's department, and formed teams for marketing, business operations, sales development, and technical support. This new organizational structure aimed to better support the company's operations from strategy to execution.

operations, and strengthened ethio telecom's capabilities in market development and the operation of new services. With the support of Huawei's expert team, ethio telecom quickly identified telebirr's service operation issues and market limitations due to a lack of diverse offerings. To address these issues, the carrier planned a product development roadmap that added mobile payments to the existing mobile money service, upgraded mobile finance, and expanded the telebirr product series by introducing value-added services to meet local market demand and increase service revenue and profits. In August 2022, ethio telecom launched a mobile finance product, pulling them out of the revenue downturn within less than a year.

Second, ethio telecom and Huawei planned to develop the mobile money service, aligning on strategic areas such as the future transformation roadmap and cooperation model. They agreed to add mobile finance to the mobile money service in order to develop a super app that would run on a full-stack service model, and then open up market opportunities by launching fintech cloud, public cloud, and government cloud services in three steps. This would in turn support the country's digital transformation in public services, industries, and government services. The partners also launched the joint operation of the mobile money service.

In June 2023, the monthly revenue of telebirr increased 500-fold to US\$1.5 million, restoring ethio telecom's confidence in business transformation.

The second challenge was a financial product that failed to meet market needs. A lack of specialized experience in financial services coupled with an incomplete understanding of potential needs regarding financial services had resulted in slow user growth and a high bad debt rate. Huawei assigned a specialized

The collaboration framework leveraged Huawei's global experience in mobile money services, increased Huawei's involvement in service

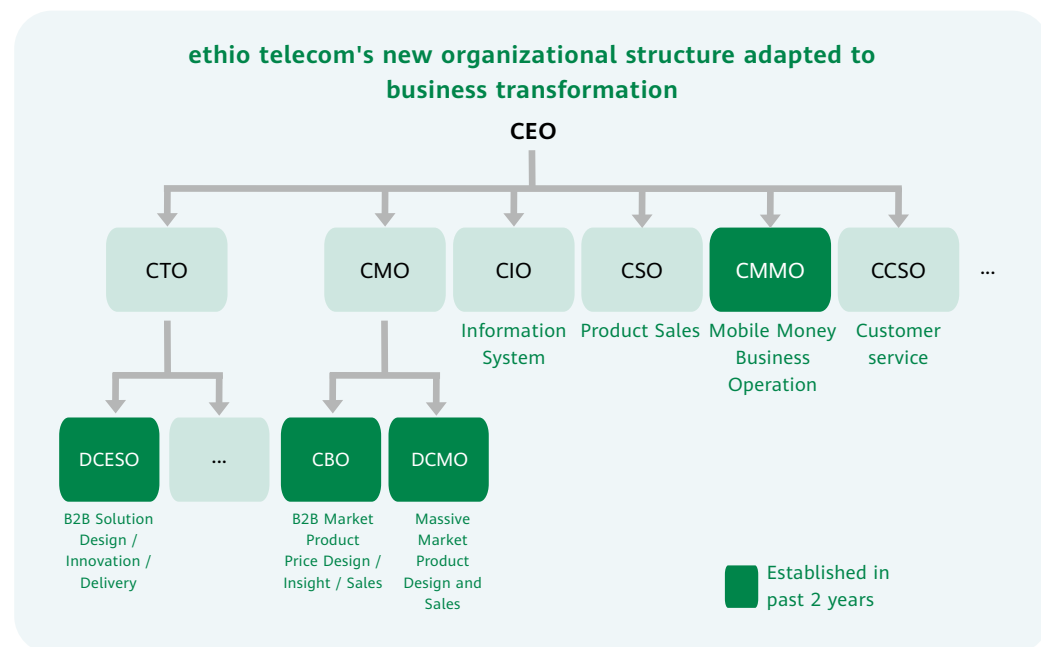


Figure 4: ethio telecom's new organizational structure adapted to new services

team of senior financial experts to join the joint operation team to help solve the problem. The team analyzed the product in terms of risk control design, tariff pricing, and user experience, and provided ethio telecom with a dynamic pricing solution for its mobile finance product, optimizing over a dozen user experience metrics in the process. Quickly seeing market returns, the bad debt rate dropped by 66% and profit margins were much higher, laying the foundation for future growth.

The third challenge was the overall business hitting a ceiling. Through several marketing improvement measures, telebirr's business scale increased significantly, with monthly transactions reaching US\$200 million (about 2% of Ethiopia's GDP). However, growth quickly encountered a bottleneck, and the company's performance was a far cry from the government's expected inclusive finance target.

In response, Huawei assigned a development team of more than 30 engineers to work with ethio telecom. They scanned potential mobile payment scenarios, such as public transport, individual taxes, daily services, health services, and tuition fees, and developed services for the selected scenarios. They also launched a super app with 15 new services in two weeks, a process that would have previously taken two months. This helped the carrier shorten TTM and expand service scope.

Specialized data analytics based on the super app helped ethio telecom identify a high-value mobile payment application scenario — gas station payments. This allowed the carrier to work with both the government and gas

station operators, causing mobile payment transaction volumes to surge. By March 2023, telebirr's monthly transaction volume had reached US\$3 billion, accounting for nearly 30% of Ethiopia's GDP.

This meant that the carrier had successfully broken the business ceiling.

While developing mobile finance services, ethio telecom also rolled out cloud services. In October 2022, the carrier released the "telecloud" brand, which leverages a Huawei Cloud solution to implement a digital transformation strategy based on both mobile payments and cloud. This service is also positioned as a 'national sovereign cloud' to help support the government's digital transformation.

Telecloud's first batch of cloud resources was snapped up five days after its launch, demonstrating great market enthusiasm. Within ten months of launch, ethio telecom had encouraged multiple state organizations, such as the High Court, to migrate their IT systems to the cloud, and was officially awarded the national digital ID project by the government.

In the enterprise market, the carrier promoted cloud-based industry solutions. To date, more than 90 enterprises have become the carrier's targets for market expansion.

ethio telecom had integrated its fintech cloud, public cloud, and government cloud services based on its multi-cloud convergence strategy. With this strategy, the carrier works with ecosystem partners to develop innovative digital solutions for education, agriculture, healthcare,

and other scenarios, empowering numerous industries and accelerating Ethiopia's digital transformation.

Another key factor driving ethio telecom's successful transformation was xCare (Figure 5), a digital O&M platform tool that matches the EM 2.0 architecture. xCare maximizes the potential of data using AI modeling and inference based on big data from networks, users, and the market, providing the eyes, ears, and brain for market expansion, service operations, and O&M efficiency improvements. For example, Huawei has developed FinCare, the world's first mobile finance operation support tool, to help with key issues like managing mobile money users, agents, and credit. In terms of user development, FinCare identifies high-value information, such as user consumption habits and transaction behaviors, to provide accurate marketing strategy suggestions for acquiring new users

and increasing user engagement. ethio telecom has already applied FinCare across a range of scenarios, such as coffee shop promotions, Christmas promotions, and loan collections, improving the marketing conversion rate by 60%. FinCare can display the locations of agents and mobile users, allowing carriers to accurately deploy and manage agent outlets. It also provides location-based, real-time marketing capabilities that send SMS messages to users near merchants, encouraging users to develop mobile payment habits.

In multiple ways, FinCare helps quickly improve carrier capabilities to develop and operate mobile money services.

xCare provides a variety of other functions. It offers scenario-based solutions, such as UserCare, 2ndSIMCare, RuralCare, and RoamingCare, to overcome typical problems

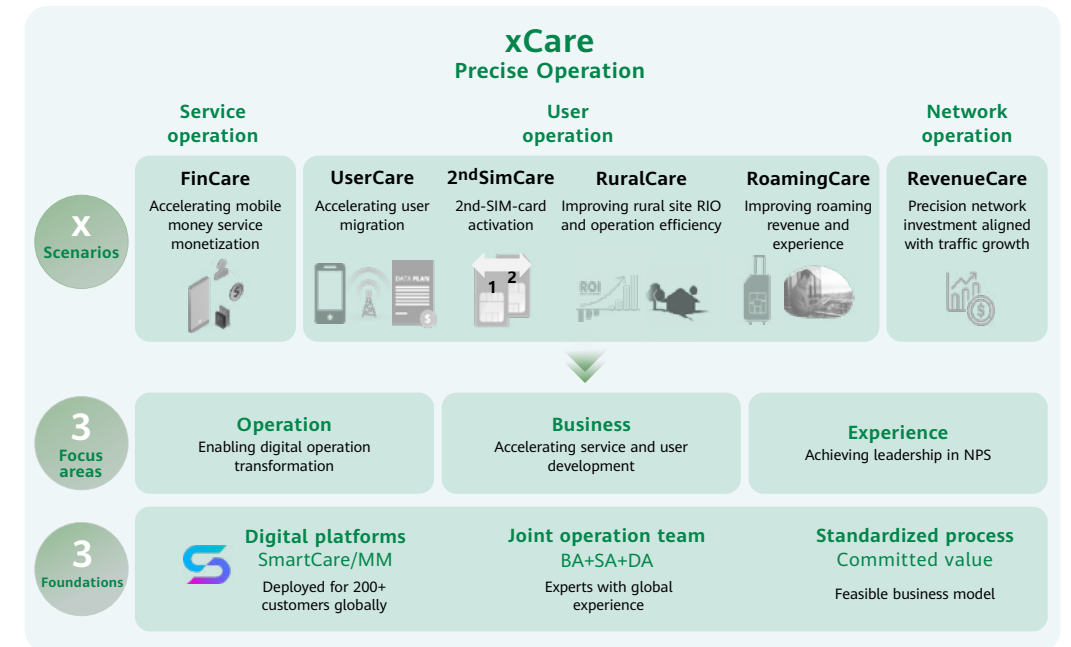


Figure 5: Overview of Huawei's xCare digital platform solution

facing the development of mobile broadband services in Africa. These include 4G user migration, improving rural coverage, activating second SIM card slots, and international roaming operations and management.

In just three years, ethio telecom has achieved its preliminary strategic goal of digital transformation and has become a digital and intelligent solution provider.

However, it has continued on the transformation journey. In December 2023, ethio telecom and Huawei expanded the carrier's business scope by leveraging its current digital capabilities. For the B2C market, they planned a range of Internet services, such as social media, e-commerce, streaming, and gaming. For the B2H market, they planned one-stop smart home services with a premium network experience (5G FWA, premium Wi-Fi, and FTTR-H). The partners also planned to add digital ID and eKYC atomic capabilities to

existing cloud and mobile payment capabilities to create the one-stop telecom application store Tele-Gallery, forming a unified gateway for users, content, and data traffic.

This would allow the carrier to monetize its role as a platform and drive Ethiopia's digital economy forward.

Lessons from the EM 2.0 model

Twenty out of nearly 30 countries in Northern, Western, and Central Africa have developed national top-level digitalization plans, with more than 80% of these countries already licensing mobile money services and over 20 carriers planning to deploy cloud platforms, paving the way for a digital Africa. As a pioneer of digital transformation in the region, ethio telecom serves as a valuable reference for other carriers across the continent. Important lessons have been learned in the following areas:

Clear strategies: Carriers should formulate digital transformation strategies based on the stages of digitalization and policy environments in a given country. Decisions on digital transformation can be driven by either national strategies or commercial factors.

Optimal service development paths: Carriers should understand the local conditions for developing innovative digital services and, as they seek ways to transform, choose those that best fit the local market (e.g., mobile money, cloud, and B2B).

Suitable organizations and talent: To ensure the robust development and operation of new services, carriers should consider adapting organizational structures to provide the specialized talent required by services.

In-depth joint operation: Carriers often face challenges when marketing new digital services, including insufficient expertise and a lack of market experience. Joint operations, with the introduction of industry experts, agile development teams, and digital O&M platforms, can make up for capability deficiencies and help carriers achieve business success.

In just three years, ethio telecom has transformed into a digital and intelligent solution provider.

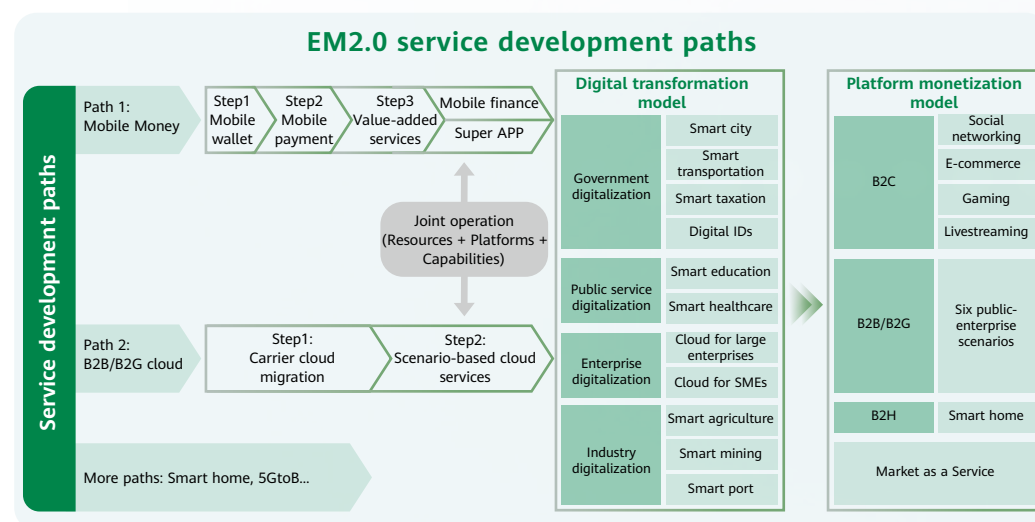


Figure 6: EM 2.0 service development paths

Accelerating Intelligent Transformation with Application-driven Industry Collaboration



Li Changwei

Chief Strategic Marketing Expert, Huawei

ChatGPT rocked the world in 2023, and now is the time to take a hard look at the current trends dominating AI development, innovation, evolution, and competition. Clearer insights into feasible directions and asymmetric competition strategies will be critical for China's AI industry as AGI draws near.

ChatGPT 3.0's launch in December 2022 was like the big bang for AI. Since then, development has continued at breakneck speeds, with new foundation models, including GPT-3.5 and GPT-4, shortly following and artificial general intelligence (AGI) just over the horizon. AI models now have more parameters than human brains have neurons, enabling models to make inferences in similar ways as humans do.

Breaking bottlenecks: AGI is approaching a critical point

ChatGPT is propelling AI development forward using a model called the "data flywheel," which uses statistical algorithms rooted in calculus and probability theory. This leaves ChatGPT to deal with two challenges: accuracy and training costs. ChatGPT 3.5 was trained on 1.8 trillion parameters, costing nearly US\$5 million per training cycle, but its accuracy bottlenecked at the 90th percentile (Figure 1).

AI has become the key to competitiveness and leadership for the ICT industry in the intelligent era.

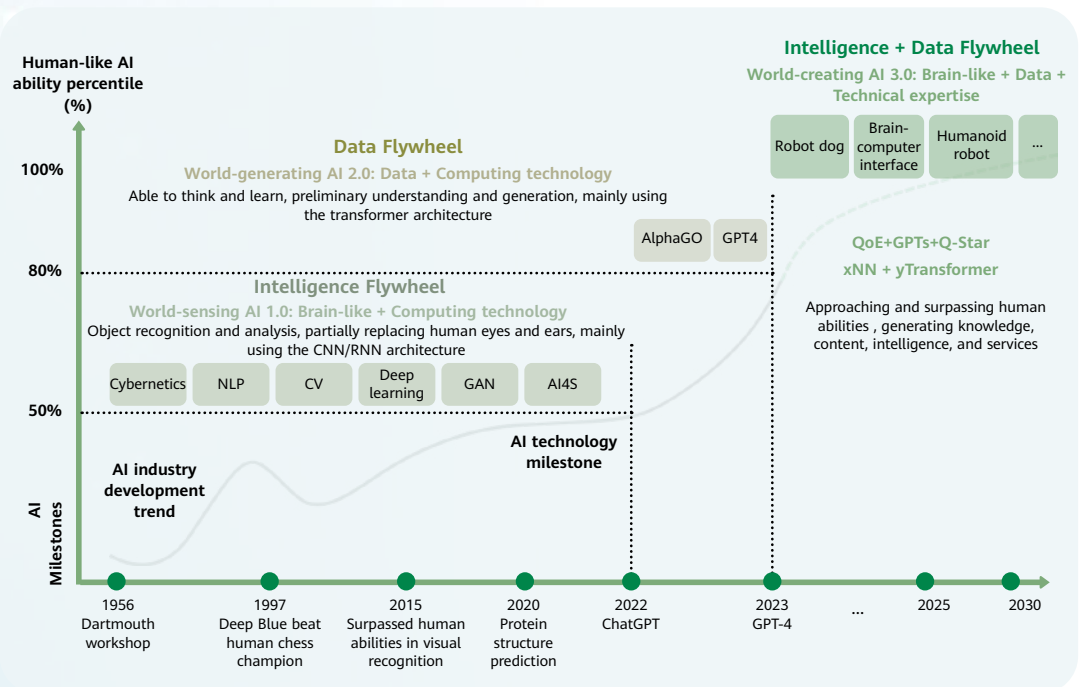


Figure 1: The data + intelligence flywheel approach is driving the evolution from AIGC to AGI

China's AI development has focused on strategy and business.

To break through this bottleneck, OpenAI has adopted a new three-step strategy: First, adding specialized AI, or a Mixture of Experts (MoE), which means linking 16 specialized AI models with 110 billion parameters to the general AI foundation model which will improve its accuracy. Second, providing open ChatGPT APIs to encourage ecosystem partners to generate their own specialized GPTs based on GPT and develop specialized intelligent agents. Third, developing, integrating, and absorbing new algorithms, such as Q-Star, to get ever closer to AGI.

Tempted by the enormous market potential and the low market threshold, many major tech companies are increasing investment in AI. In 2023, Google introduced its multimodal model Gemini, which will help it move from AIGC to AGI. The Google DeepMind team also recently launched their FunSearch algorithm, which works by pairing a pre-trained large language model (LLM) designed to output creative solutions in the form of computer code with an automated "evaluator" that guards against hallucinations and incorrect ideas. By iterating

back-and-forth between these two components, initial solutions "evolve" into new knowledge, gradually approaching what's required for AGI.

Once achieved, AGI will set the stage for AI to be truly integrated into enterprise operations. Microsoft's Copilot, for example, was launched as part of its Office apps, with ChatGPT pre-integrated. Enterprise users have flocked to this new service, despite its US\$30 per month subscription fee, making it a new growth engine for Microsoft. Other current industry applications include Tesla's humanoid robot Optimus and autonomous driving system FSD V12, as well as Huawei's Pangu Weather foundation model, which have made breakthroughs in critical indicators, like industrial experience and accuracy, to shape new business models that boost efficiency across industries.

The turning point for AGI is almost here, according to many. NVIDIA CEO Jensen Huang anticipates the advent of AGI within the next five years, while Tesla CEO Elon Musk believes that AGI will come within three years.

AI is already changing the paradigm of economic growth. According to China's Ministry of Industry and Information Technology, growth in AI computing scale can effectively drive growth in the digital and intelligent economy. They found that every 1-point increase in the

Computing Index in 15 key countries correlated to a 3.5‰ and 1.8‰ increase in national economic growth and GDP, respectively. This new AI-driven economic development model will begin to impact global strategic competition in 2024. (Figure 2)

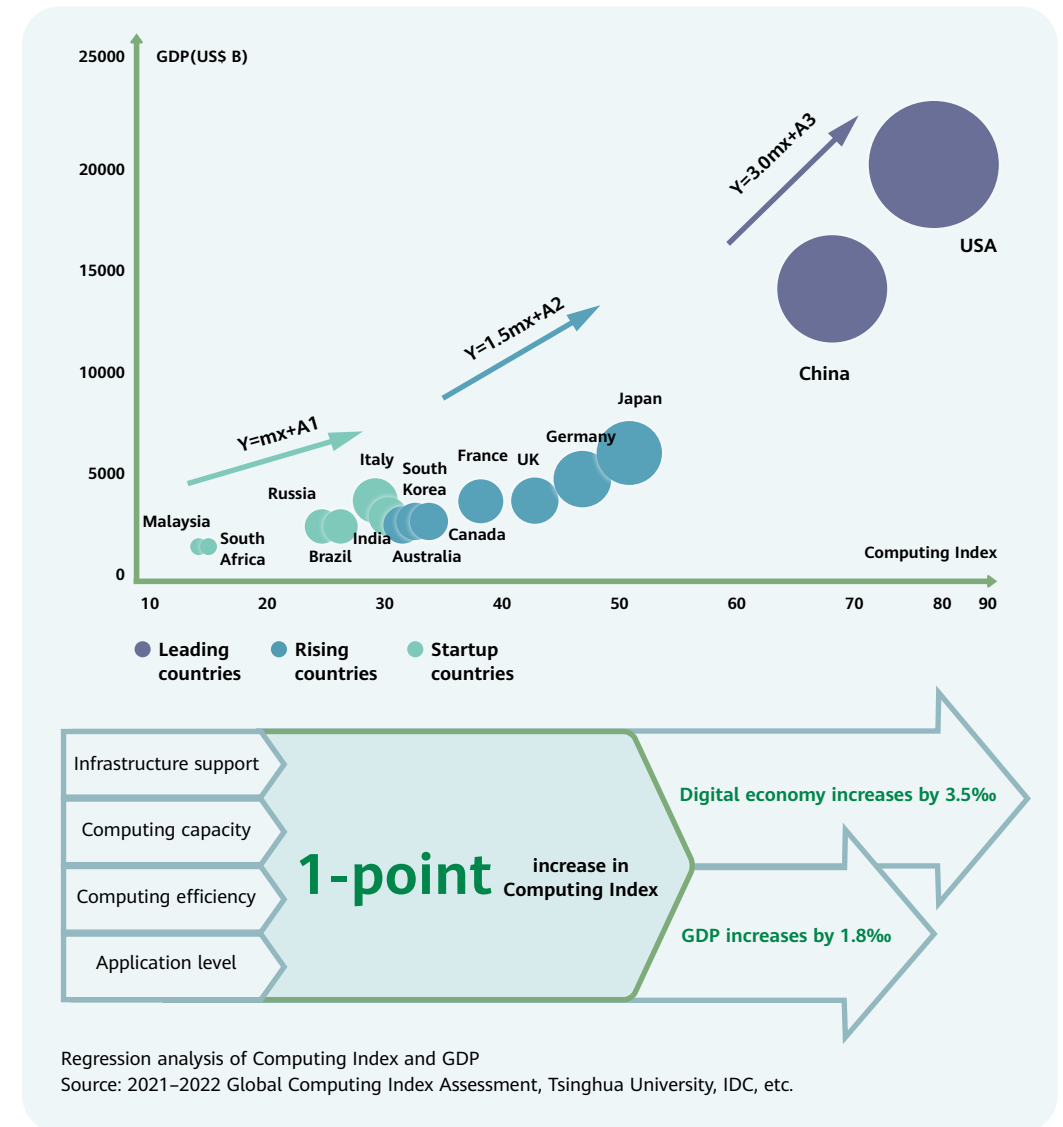


Figure 2: AI computing power is closely related to GDP

**Differentiated leadership:
The emerging path forward for AI**

AI used to be seen as an open-source and non-profit technology. However, as it matures, it is increasingly found at the center of closed and business-oriented competition. Tech giants that used to collaborate with each other have become divided and focused on competition.

Major tech companies now either want to work with OpenAI, the biggest name currently in the game, or follow in its footsteps. Microsoft works extensively with OpenAI, outperforming everyone else in LLMs. Copilot was jointly launched by the two companies.

Anthropic, which has a different approach to AI safety, is now one of the biggest competitors of OpenAI. Google, whose search business faces the greatest risk of being replaced, has launched its own AI services including Bard and Gemini. Google has also taken advantage of YouTube's service data and algorithms to solidify its leadership in multimodal technology. The FunSearch algorithm recently launched by the Google DeepMind team now allows Google to potentially outperform other players in AGI. Tesla launched AI systems that are closely linked with its electric vehicles, such as FSD V12 and Optimus, to establish differentiated leadership. Meta AI has integrated more than 20 new AIGC methods, focusing on improving user experience related to search, social discovery, advertising, and business communication across its platforms,

including Facebook, Instagram, Messenger, and WhatsApp. Apple is using AI to improve Siri and is expected to launch generative AI functions for its iPhones and iPads in 2024 with the release of iOS 18 and iPadOS 18. Finally, Amazon has leveraged its advantages in cloud to launch Titan AI, which includes text models for content generation and embedding models that can create vector embeddings for efficient search functions. Amazon also launched the CodeWhisperer, an AI coding assistant that is free to users.

Each of these tech giants has chosen a different way to develop AI. But together, they represent the convergence of AI, security, cloud, services, and devices, as all focus on innovation in scenario-based application experiences and value integration, which ultimately enables differentiated leadership.

Core AI players have added algorithms, like specialized AI, new Q-Star, and FunSearch, to data statistical analysis methods to maximize their strengths and move from driving exponential growth in AI computing demand to driving logarithmic growth.

China's approach: Strategy- and business-driven AI development

International tech giants are not the only ones pursuing AI development. China is doubling down in this field. The country's AI strategy is seen as critical to future breakthroughs in both the ICT industry and growth of the nation's digital economy. Generally, China's AI development has focused on strategy and business.

In terms of top-level strategy design, national resources and those of individual enterprises are being organized along specified development paths from the top down to achieve collaboration and boost industry revenue (Figure 3).

Early on, the government played a leading role through preferential policies, strategic investment, and data sharing. It started by setting digital security rules and application requirements for the government's own digitalization and intelligent application scenarios. Then, it turned to state-owned

enterprises to promote the construction of new AI infrastructure by integrating network, transmission, computing, intelligence, and security elements. This is helping the nation to build an AI computing platform that will serve as critical infrastructure and promote both corporate and academic AI research and development. The aim here is to build a natural momentum that will drive development.

The government has also prioritized ecosystem integration and policy collaboration at the industry and enterprise levels to prevent excessive competition in AI foundation models

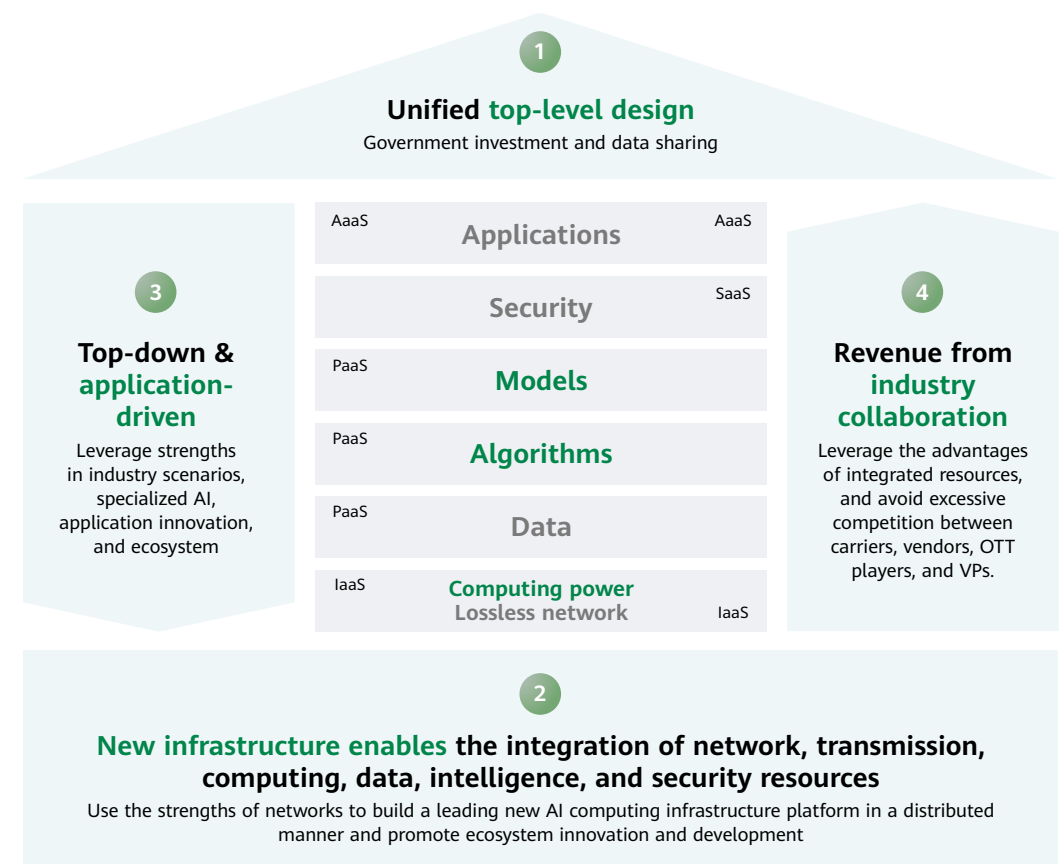


Figure 3: Government-led top-level design + Carrier-led new infrastructure + Application-driven + Industry collaboration

at the PaaS layer and preserve breakthrough opportunities at the SaaS and application layers.

This top-level design is supposed to ensure an application-driven, top-down AI development path. Commercial breakthroughs are expected to be made thanks to China's diversity in low-level application scenarios and AI expertise in some domains. This will further encourage the development and quick adoption of smaller models, specialized AI applications, and innovative solutions (Figure 4).

However, AI infrastructure is prioritized over applications in China so as to leverage the country's network advantages at the IaaS

layer. AI computing clusters can make up for insufficient single-point computing power and integrate scattered computing resources to power the upgrade and iteration of developing and verifying AI foundation models. Carriers are prime representatives for state-owned enterprises (SOEs) here, and are expected to replicate NVIDIA's AI Foundations strategy while building new AI computing centers. They will be able to turn a profit by enabling technology companies and scientific research institutions and by leasing AI computing resources.

By expanding and strengthening AI computing centers, they will be able to implement "inside-out" service enablement, where

Every 1-point increase in the Computing Index in 15 key countries correlated to a 3.5% and 1.8% increase in national economic growth and GDP, respectively.

carriers upgrade their primary commercial services through AI-based service integration, while achieving real growth in new market segments.

New MaaS: The AI collaboration strategy for carriers

For carriers, intelligent digital platforms play a key role in integrating telecom services with AI. This is how they create new capabilities and competitiveness in networks, operations, and O&M. As the foundation of intelligent digital transformation for carriers, cloud and intelligent technologies make networks, computing power, and service configurations more flexible, efficient, and affordable. This has increased their focus on a large number of multi-service scenarios (Figure 5).

When it comes to AI collaboration, carriers are now focusing their strategies on integrating their vertical industry value chain resources (security, data, and networks) to form a new Model as a Service (MaaS). Under it, layer-1 XaaS technologies and capabilities are provided and supported by leading technology vendors. Carriers, on the other hand, focus on the second layer where they develop middleware to integrate XaaS at each layer. This new MaaS encapsulates and integrates

Microsoft has simplified complex service management capabilities through AI-based integrated operations, which can be a big help here. For example, network automation can help carriers simplify network management and improve customer services. SKT in South Korea has set a benchmark for this kind of strategic AI transformation. In November 2022, SKT CEO Ryu Young-sang announced the company's plan to become an AI company. It has since worked with partners to develop a Korean large language model, which is similar to GPT-3. In May 2022, SKT also launched a Korean AI chatbot "A.", attracting more than 1 million users. At MWC 2023, SKT showcased numerous intelligent services based on foundation models. These services include companion AI, data AI, visual AI, medical AI, and urban air mobility (UAM). SKT's key advantages over other carriers here is that it possesses both telecom know-how and proprietary local data that OpenAI does not have.

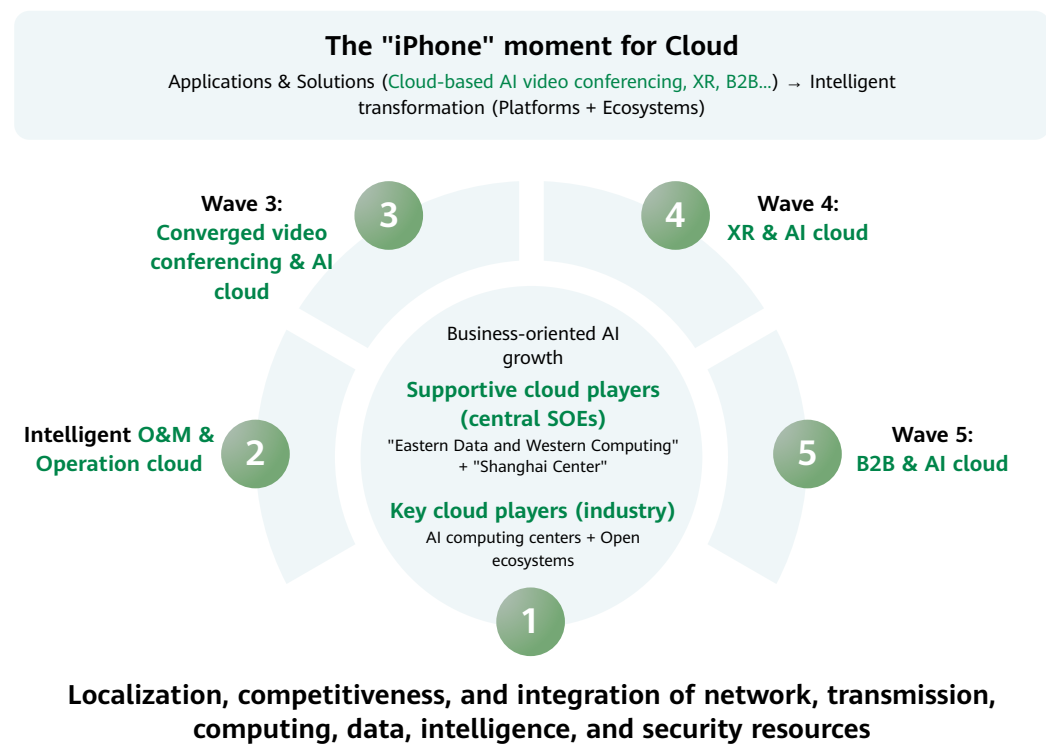


Figure 4: AI foundation model services + New service planning

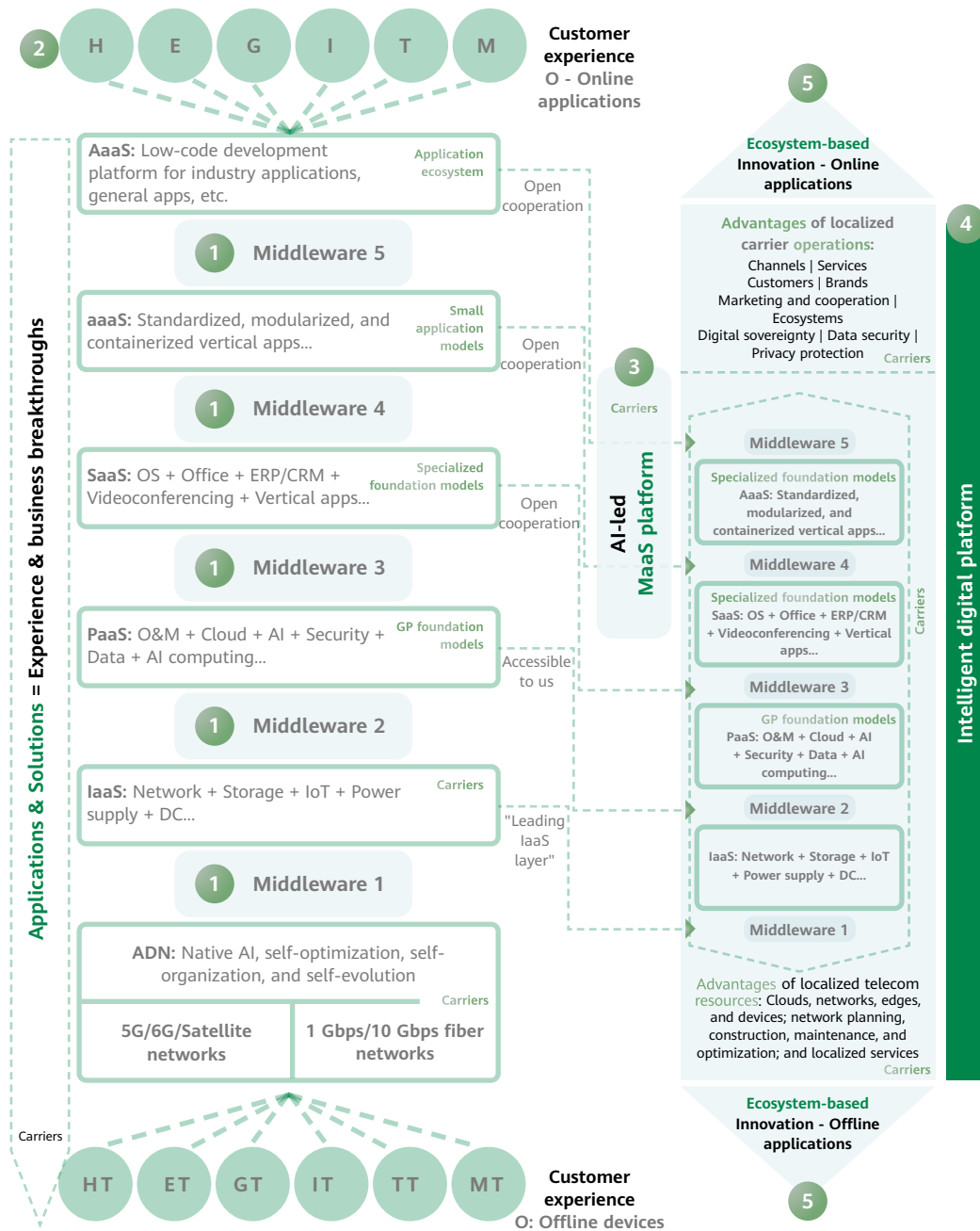


Figure 5: Transforming MaaS to build an intelligent digital platform and ecosystem leadership

vertical elements throughout the value chain leveraging operation and infrastructure advantages.

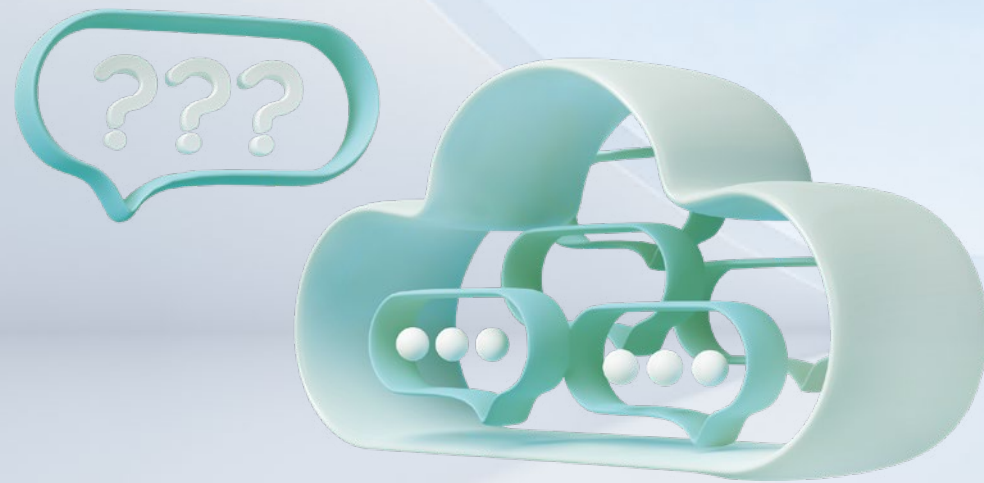
MaaS will serve as the broad infrastructure that underpins the AI era, providing secure, efficient, and low-cost model usage and development support for downstream applications. Carriers will direct this usage, as they will provide the key technical capabilities and industry resources needed, preventing excessive competition at the PaaS-layer, streamlining the vertical ecosystem, accelerating breakthroughs in applications and solutions, and achieving leadership in both business and scale. This will shape a positive cycle of technology iteration driven by applications and businesses.

The accelerated integration of AI into industry triggered by ChatGPT has made it clear that

we must aggregate limited industry resources through top-level design and strategic planning. This will be the only way to continue forward with the asymmetric competition strategy that has defined the mobile Internet era. We should start with smaller models and specialized AI, and collaborate with verticals to achieve joint innovation and breakthrough. Iterative growth and mature AI software, hardware, and foundation model technologies will be achieved through the application and business success of smaller models.

These efforts will finally enable effective growth in new market segments and allow countries, industries, and enterprises to build their competitiveness in the new phase of vertical industry integration, innovation, and transformation. 

AI-ready Cloud Drives Intelligent Digital Transformation for Carriers



ChatGPT: Igniting AI models globally

ChatGPT provides a complex mix of answers that can be amazing, amusing, or utterly wrong.

GPTs have put AI, an industry with a history of over 70 years, under the spotlight again. The emergence of ChatGPT is regarded as a turning point for AI, transitioning it from a technology that understands the world to one that shapes it. The integration of technologies, such as scaling laws, reinforcement learning, and human feedback, has given ChatGPT human-like inference capabilities.

OpenAI went on to launch a series of innovative products, including GPT-4 (1.8 trillion parameters), GPT-4 Turbo, Assistants API, ChatGPT Enterprise, and GPTs for ChatGPT in 2023, as well as Project Q* and Sora, which were just disclosed. Other players on the industry value chain are also accelerating innovation: Humane launched AI Pin, Microsoft

Wang Xiaobin

Chief Architect, ICT Computing Products and Solutions (Carrier Domain), Huawei



To achieve business success through digital and intelligent transformation, carriers need to be AI-ready in terms of corporate architecture, business models, and infrastructure.

released the generative AI office suite Microsoft 365 Copilot, and Google DeepMind launched Gemini. These are accelerating the application of foundation models to industry and individual markets and guiding AI development towards multimodality and intelligent agents.

In China, many major tech companies have joined the AI race by launching their own foundation models. These include Baidu's Ernie Bot, Alibaba's Tongyi Qianwen, Huawei's Pangu, and 360's Smart Brain, Kunlun Tech's Skywork, JD's ChatRhino, iFLYTEK's Spark, Tencent's Hunyuan, and SenseTime's SenseNova. In just six months, the number of foundation models in China increased from around 100 to more than 200. Many SMEs have also released vertical models, including Trip.com's Xiecheng Wendao for

tourism, NetEase's Ziyue for education, JD Health's Jingyi Qianxun for healthcare, and Ant Group's finance AI model. These are facilitating the evolution of AI capabilities from perception to cognition and from identification to generation, while expanding the scope of AI applications from general-purpose to industry-specific. AI is expected to be applied to over half of core industry scenarios over the next two years.

We can already see the potential of AI for reshaping all services in the carrier industry. Since the marginal cost of foundation models approaches 0, applying AI to internal operations can help carriers reduce costs and improve efficiency. Leveraging AI in B2C, B2H, and B2B services will allow carriers to generate higher revenue. For example, AI-powered CCTV solutions can create a 15%-

In the future, foundation models will provide ubiquitous intelligent services through tremendously diverse business models.

plus premium over traditional solutions, and the AI-powered 5G New Calling service can generate more than 10% higher ARPU. AI is accelerating carrier digital transformation with carriers around the world evolving cloud-based transformation to cloud and AI-based intelligent digital transformation.

Lower costs, better efficiency, and higher revenue

To seize the enormous opportunities presented by AI, carriers require advancements in three areas to be AI-ready.

Making AI part of the transformation strategy and creating an AI-ready technology architecture

Carriers must make AI a key part of their digital transformation and overall transformation strategy, with specialized teams in place to implement AI strategies and build AI-oriented enterprise architectures and capabilities. TOGAF is an enterprise architecture framework that bridges enterprise strategic planning and IT solutions, and serves as the core of enterprise digitalization. This

framework consists of four components: business architecture (BA), information/data architecture (IA), application architecture (AA), and technology architecture (TA). Without an enterprise architecture that maps strategic objectives to execution, digital transformation is unlikely to succeed. Therefore, it is recommended that carriers integrate AI into their enterprise architecture so that AI can power their strategies, businesses, data, and technologies.

At the same time, the major initiative of digital transformation may encounter large setbacks during execution. Involvement and promotion by top executives will be key to making carriers AI-ready. Huawei, for example, announced its AI strategy at HUAWEI CONNECT in 2018. In that same year, the company's founder Ren Zhengfei issued a resolution on increasing investment in AI and using AI to improve corporate efficiency, establishing the AI Enabling Dept headed by Ren himself. Today, Huawei has already made AI part of every business process in every department and business unit. To date, Huawei has used AI to create more than 600 intelligent applications in over 80 business

settings, build nearly 7,000 AI models, and create over 20,000 digital employees that serve Huawei organizations in more than 170 countries and regions around the world. Huawei significantly reduces operation costs with a company-wide foundation model-based intelligent agent that gives all employees access to an AI assistant.

In terms of products, the AITO M9, which is equipped with Huawei's intelligent driving solution, caused a sensation in the new energy vehicle market in 2023. The voice assistant Celia, which is pre-installed in every Huawei smartphone, improves user experience. AI-enabled energy-saving algorithms allow Huawei's wireless base stations to consume over 20% less energy than competing products, in turn driving up carriers' revenues. In the carrier industry, China Telecom and China Mobile also made "AI+" part of their group strategy in 2023.

Using platforms to create AI-ready business models, promote industry collaboration, and foster a thriving ecosystem

There is no lack of companies or other players who are actively exploring applications for foundation models. Explorations into new AI applications for industries will likely result in high trial-and-error costs and even failure. Although carriers can participate in creating foundation models, their greater strength lies in ubiquitous cloud-network infrastructure that can serve as supporting and monetization platforms for foundation models, as well as existing industry customers that can form a potential market as carriers look to commercialize foundation models. Carriers' computing and O&M capabilities and credibility in local communities give them the potential to provide professional platforms for local AI markets. With such platform capabilities, carriers can engage

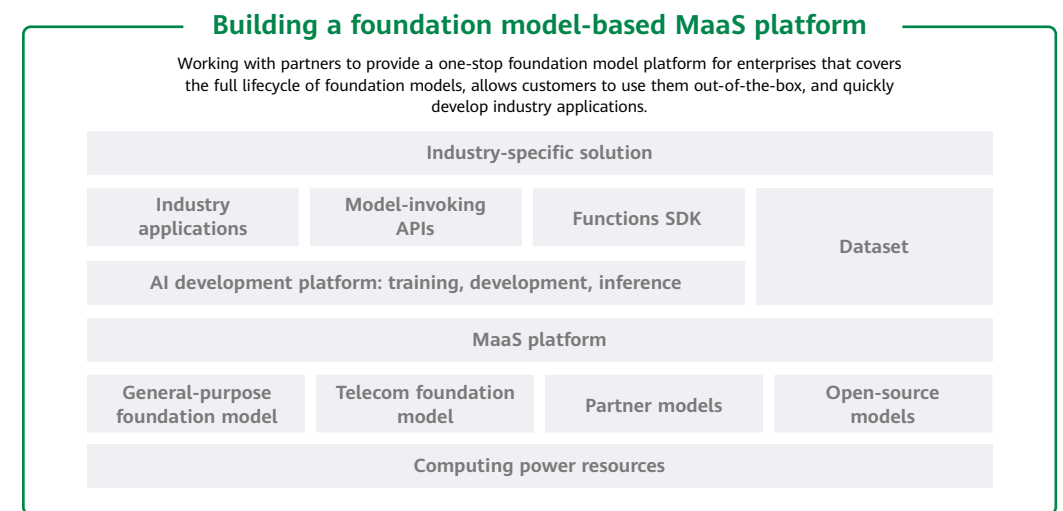


Figure 1: Foundation model-based MaaS platform capability framework

and aggregate foundation model providers for business exploration and provide the Model as a Service (MaaS) business model (Figure 1). This will enable monetization, encourage more partners to sell their model products on the platform to form a long tail market, prevent carriers from having to make foundation models themselves that compete with their partners, and reduce trial-and-error costs for uncertain models.

China Telecom, for example, launched a foundation model ecosystem cooperation alliance in July 2023 to help implement China's cloudification and digital transformation strategy and accelerate the construction of ICT infrastructure centered on general computing,

AI computing, and supercomputing. This alliance publicly engaged influential foundation model partners across the industry. At the Digital Technology Ecosystem Conference in November 2023, China Telecom and its partners launched the first batch of 12 industry-specific foundation models for trial commercial use, which involved sectors like education, construction, finance, and mining. These foundation models are all embedded on the Xingchen MaaS ecosystem service platform. By integrating networks, cloud, intelligent computing, AI, and partner models, this business model is driving the intelligent upgrade of Telecom China's cloud services and is already generating commercial returns.

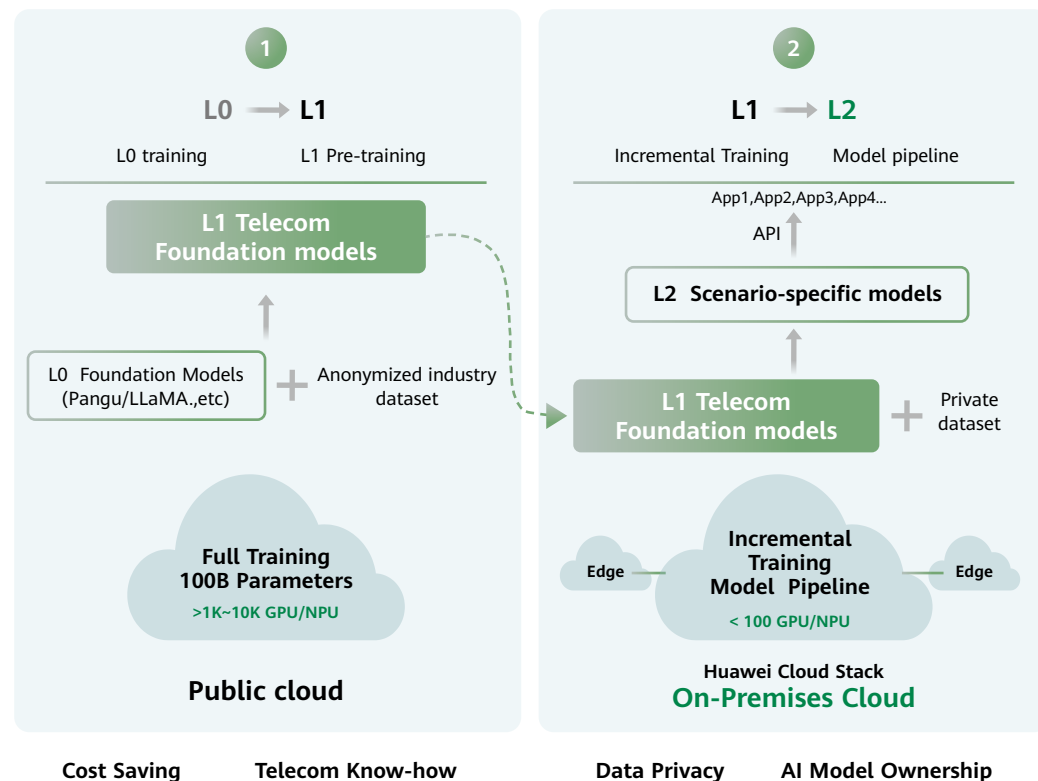


Figure 2: Customized models based on collaborative training on public and private clouds

Carriers need to be AI-ready in terms of strategy, organization, capability, and infrastructure.

For individual model developers, carriers should consider taking the lead to develop a foundation model ecosystem community similar to Hugging Face to maximize the ecosystem value of foundation models across society. Reasonable industrial division, platform building, and ecosystem development should be key objectives of carriers in the AI foundation model era.

Building AI-ready infrastructure with AI-powered clouds and data

In the future, foundation models will provide ubiquitous intelligent services through tremendously diverse business models. These foundation models can be embedded in a variety of software and hardware systems, such as intelligent vehicles and robots, to make their way into commercial markets in the form of intelligent products. They can also be deployed in the cloud to provide cloud services based on commercial foundation models and transform existing business logic. Cloud and network infrastructure are the pipes through which AI can reach end users (consumers, businesses, and homes). AI can only reach where cloud and networks are available, so an AI-ready intelligent cloud is needed for carriers to enter the AI industry. For carriers that are engaged in the AI busi-

ness, cloud will be essential for AI in the early stages, but AI will become a key part of cloud further down the road. Over the past few years, China Telecom has created a nationwide e-Cloud service. In March 2024, China Telecom's Chairman Ke Ruiwen said at the company's 2023 Annual Results Announcement, "Without AI, cloud does not have a future." He also claimed that e-Cloud will accelerate the upgrade to intelligent cloud to become a leading foundation model computing service provider in China.

Carriers should build AI-ready intelligent cloud solutions based on the three key elements of AI: computing power, data, and algorithms. An AI-ready cloud platform should feature one architecture and two capabilities:

- **Computing power: Distributed multi-layer cloud architecture for ubiquitous computing power**
Distributed multi-layer cloud architecture requires collaboration between the public cloud, on-premises cloud, and edge cloud. Full training is performed on the public cloud, incremental training and centralized inference on the on-premises cloud, and other inference on the edge clouds (Figure 2). This architecture brings three benefits for carriers to deploy foundation models.

- **Reduced costs:** The pre-training of foundation models requires huge amounts of AI computing power (thousands of GPUs) within a certain period of time (weeks or months). For example, the LLaMA 2 70B model was trained using a distributed supercomputing cluster with 2,000 NVIDIA A100 GPUs, while the Falcon 180B model used 4,096 NVIDIA A100 GPUs. Most carriers, especially those outside China, do not need to purchase computing power in advance for this purpose. They can instead lease public cloud computing power. This allows them to perform full training with the 10,000-GPU computing power already deployed on the public cloud and publicly available industry datasets. Pre-

trained foundation models can be deployed on the on-premises cloud where incremental training can be performed based on a small amount of private data of carriers or specific industries. This exponentially reduces the required hardware investment (dozens of GPUs). This model featuring full training on the public cloud and incremental training on on-premises clouds means a 100-fold saving on investment in computing power.

- **More secure data:** Laws and regulations in many countries prohibit data (such as network information and traffic information), which usually is high-quality datasets for model training, from being transferred out of the local network or abroad. On-premises clouds and edge clouds can meet these requirements. Foundation models trained on such clouds are already built on privacy and sensitive data alongside inherent knowledge. This means they run locally and privately, ensuring data security.
- **Better service experience:** Edge clouds can bring AI inference closer to end users and provide lower-latency experiences.

- **Data: Cloud-based data production line for efficient data processing**

While data is a core asset of carriers, it is scattered in independent IT systems and does not flow. Not only are there hardware costs for storing data, its value remains untapped. Local cloud services require full-lifecycle one-stop data governance capabilities (Figure 3), including pool-based data storage and cross-domain collaborative data scheduling and management. Data lake-warehouse convergence can help create a logical data

Converging data warehouses and lakes with AI

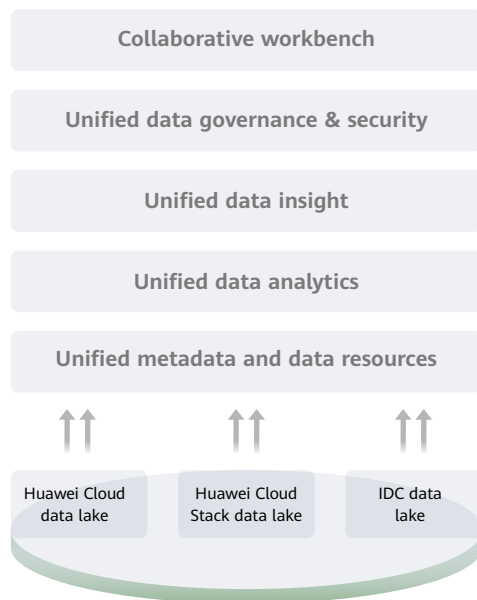


Figure 3: Data-AI convergence: An efficient cloud-based data foundation for the AI era

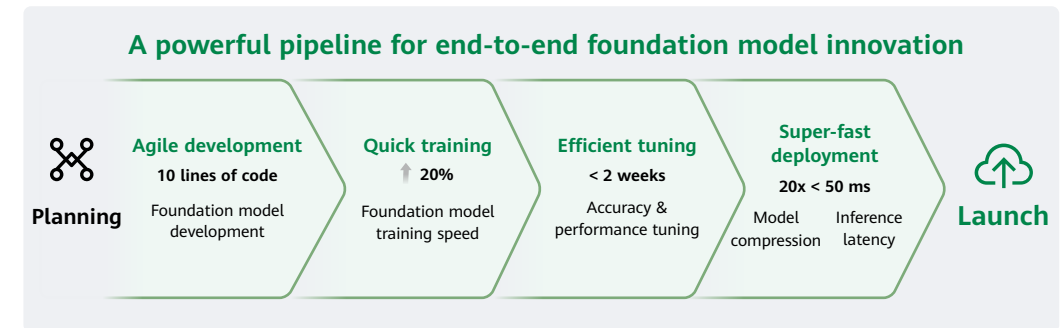


Figure 4: A powerful pipeline for end-to-end foundation model innovation

lake that brings together scattered data lakes and warehouses. This makes it possible for one copy of data to be shared by multiple data analytics engines and AI engines. It also enables the development of model training-oriented data processing capabilities, such as the efficient integration, cleaning, filtering, and labeling of datasets.

- **Algorithms: Cloud-based model production line enabling lifecycle management of foundation models**

The lifecycle of a foundation model includes key phases such as development, training, fine-tuning, and deployment. Tools and capabilities adapted to these phases should be deployed on the on-premises cloud as services. Simply put, the on-premises cloud should enable the development, training, fine-tuning, and deployment of foundation models as services to help carriers manage models (Figure 4). When it comes to ecosystem compatibility, cloud services should support

major open-source models. Carriers need to build independent model-related capabilities that can ensure data security and the manageability and controllability of models.

We are stepping faster into an AI era as the fourth industrial revolution approaches. Carriers need to be AI-ready in terms of strategy, organization, capability, and infrastructure. They should use their cloud-network strengths, position themselves appropriately in the industrial division of work, and build their ecosystem platform capabilities to provide Model as a Service. This will allow them to bring together more industry partners and succeed in intelligent digital transformation. ■





04.
Innovation
for the Future

Jointly Defining ICT Architecture for 5.5G: The Key to Unlocking New Opportunities

Dang Wenshuan

Chief Strategy Architect, Huawei



For carriers to maximize the value of their infrastructure and stimulate new demand, they need to redefine ICT architecture, evolve their ICT infrastructure, embrace autonomous networks, rethink their public cloud strategy, and reassess their approach to data storage.

Technologies (ICT) like 5G, cloud computing, and AI are seeing broad adoption on a global scale. In addition to revitalizing user experience, these technologies are also bringing exciting changes to the telecom industry, like new 5G business models, services, and devices, new forms of home connectivity, AI applications, and carrier-operated public cloud services. These new developments will give carriers the opportunity to maximize the value of their infrastructure and stimulate new demand.

But to get there, carriers also need to adapt: They'll have to redefine ICT architecture, evolve their ICT infrastructure, embrace autonomous networks, rethink their public cloud strategy, and reassess their approach to data storage. These efforts will put carriers in a better position to seize new opportunities and raise the entire industry to new heights.

Six megatrends in telecoms

Trend #1: New business models are paving the way for 5G growth

Driving the commercial success of 5G is top of mind in the industry. Carriers around the world have been exploring a number of different business models for 5G, and along the way have made huge headway in monetizing their network capabilities.

Some notable examples include: The Finnish carrier Elisa differentiates 5G packages based on preferred downlink speed, which has driven more users to upgrade and ultimately increased Elisa's 5G ARPU by more than €3.

China Unicom Guangdong launched special 5G livestreaming packages, which provide subscribers with greater uplink speeds, network priority levels, and data limits for a smoother livestreaming experience. By the end of 2023, these packages brought in 500,000 new subscribers, increasing the carrier's ARPU by 76% compared to regular 5G service packages.

Thai carrier AIS now provides on-demand 5G boost packages, while Three in Hong Kong offers 5G packages specially designed for stock market

The six megatrends in telecoms will stimulate new demand and present new opportunities.

tracking and gaming. Both carriers charge their subscribers based on customized network priority levels, latency, and traffic volume.

China Mobile launched 5G service packages specifically for delivery couriers. In China, delivery workers go through tons of data and minutes to manage logistics and contact their customers. In addition to providing custom 5G packages with more minutes and data, China Mobile also offers road accident insurance and dedicated lounges in their service centers, where couriers can take a rest, get something to drink, and charge their phones. These diversified offerings have attracted a number of high-value subscribers.

Moving forward, new 5G business models that monetize different combinations of service capabilities will keep emerging as carriers work to better understand behavioral changes in the digital age. And by providing more targeted communications services, carriers will be able to generate more lucrative returns and set the stage for a healthy 5G business.

Trend #2: New 5G services and applications are flourishing

In 2023, we saw incredible progress in new 5G services and applications that the industry has been looking for.

China Mobile Jiangsu launched its New Calling service in September 2023 and attracted 1.5 million subscribers in only three months. Within the first six months of launching its cloud phone service, China Mobile signed up 11 million users. Globally, 5G fixed wireless access (FWA) is now serving more than 160 million users, attracted by 5G's enhanced experience and the growing affordability of 5G customer premises equipment (CPE). To date, there are now 155 carriers offering commercial FWA services around the world, double the number in 2021.

Carriers outside China are also rapidly scaling their 5G business in the B2B market. Between 2022 and 2023, 5G private networks for enterprise customers also doubled to 222, spanning

industries like media, healthcare, education, ports, manufacturing, oil & gas, and mining.

New 5G services and applications are undoubtedly creating new growth momentum in the industry.

Trend #3: Novel smart devices are multiplying

Smart devices are the primary vehicle through which consumers make use of network capabilities, and they are also a major driver of industry development. A number of novel smart devices hit the market in 2023. So far, there are already five smartphones powered by 3CC carrier aggregation (CA) – a capability that's fundamental to delivering the 5.5G experience. The availability of these smartphones suggests that device vendors have outpaced network providers in the upgrade to 5.5G.

Spatial computing and AI technology are giving rise to innovative applications like glasses-free 3D, MR devices, and intelligent connected vehicles, as well as incredible new AI devices – all of which are creating new possibilities for the telecom industry.

Trend #4: Innovations in FTTH and FTTR are speeding up the arrival of F5.5G

The Fiber to the Home (FTTH) market is booming. In 2023, more than 50% of carriers around the

world provided 1-Gbps connections for roughly 200 million home users. 1 Gbps has effectively become the new standard for home broadband, driven by nonstop innovation in passive optical network (PON) and optical distribution network (ODN) devices. The cost of FTTH build-out can be halved by using dual-port Building PON and transparent adhesive fiber.

We are also seeing an exponential increase in the number of homes with fiber to every room. Twenty-two carriers have launched Fiber to the Room (FTTR) services as of 2023. Now, FTTR is benefiting 12 million home users and bringing better connectivity to 300,000 small- and medium-sized businesses.

In November 2023, the European Telecommunications Standards Institute (ETSI) released the F5G Advanced Generation Definition in a well-timed move to propel F5.5G forward.

Trend #5: Foundation models are enabling intelligent network O&M

There are currently more than 300 AI applications for the telecom industry. And yet, these days, carriers deploy nearly 80% of their O&M engineers to work on things like network monitoring, troubleshooting, and complaint handling.

This is where an "AI for Telco" strategy can help, harnessing the power of AI to address unique pain points in network O&M. Role-specific AI copilots and scenario-based AI agents – built on foundation models and digital twins – can help to better equip O&M engineers and further enhance customer satisfaction.

AI for Telco applications are growing in number and coming into wider use. For example, AI copilots are helping call centers greatly increase their first call resolution (FCR) rate, a metric that represents the percentage of customer inquiries resolved in a single interaction. With the support of AI copilots, field maintenance engineers can also vastly cut the time it takes to perform fault recovery.

AI agent technology is valuable for tasks like optical path maintenance: identifying faults within one minute, locating them within three, and recovering within five. Deploying AI agents in high-value scenarios also helps decrease the number of serious customer complaints.

With foundation models built specifically for the telecom industry, carriers can streamline O&M in entirely new ways.

Trend #6: Carrier-operated public clouds are on the rise

When it comes to cloud services, over the years a vast majority of carriers have gone the cloud reseller route, while only a few have built and operated their own public cloud platforms. But the tides are turning. In Africa, for example, Ethio Telecom released its own cloud computing service platform in the third quarter of 2022. Currently,

this platform offers more than 60 localized SaaS applications, exceeding the number of its globalized SaaS offerings.

As digitalization continues to sweep across the globe, localized digital services like mobile payments, online shopping, food delivery, ride hailing, and digital government are flourishing. This presents huge opportunities for carriers with local service advantages to develop public cloud services. These opportunities are unfolding in Asia-Pacific, Latin America, and the Middle East, where more than 20 carriers are launching or have plans to launch their own public clouds. The time is ripe for carriers to rethink their public cloud strategy.

Defining ICT architecture for 5.5G and evolving ICT infrastructure

To seize the new opportunities brought about by these megatrends, we need to address a number of shared challenges in the industry, including defining our overall ICT architecture, its key features, and implementation paths in the 5.5G era.

Our proposed ICT architecture for 5.5G has six defining features: ubiquitous 10-Gbps access, 400G/800G transport networks, core networks with new user planes and intelligence planes, L4 autonomous networks, carrier-operated public cloud services, and AI-oriented unified data storage. (See Figure 1.)

Feature #1: Ubiquitous 10 Gbps access

"Ubiquitous 10-Gbps access" includes mobile, home, and campus scenarios.

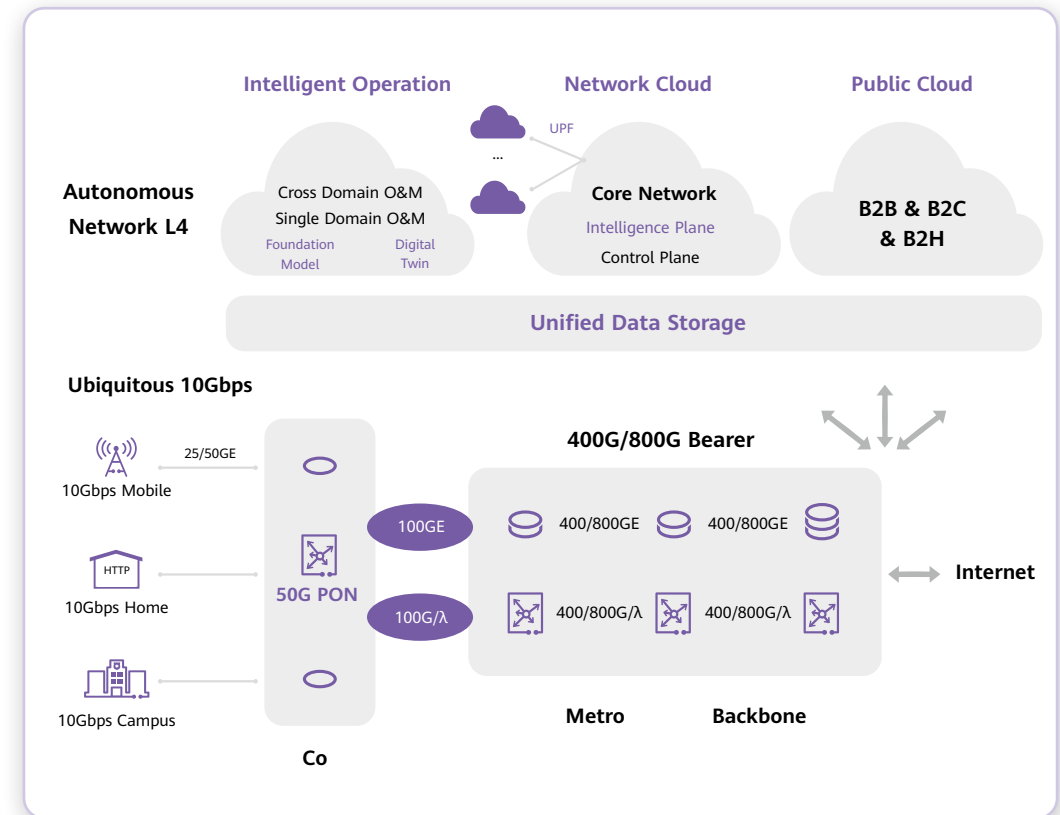


Figure 1: Proposed ICT architecture for 5.5G

For mobile users, extremely large antenna array (ELAA) technology and the availability of more frequency bands can help deliver a first-class experience with 10-Gbps downlink and 1-Gbps uplink.

Energy efficiency will also be a key metric of future wireless network performance. To ensure "0 Bit 0 Watt", where the lowest possible amount of energy is consumed when no bits of data are transmitted, we need to constantly push the limits of our kit. Huawei has launched a series of products to help carriers build leading 5.5G wireless networks for all scenarios, including the industry's only TDD 128T MetaAAU with a 12

Gbps capacity, the industry's only dual-band 64T MetaAAU that supports a 5 Gbps experience, FDD tri-band Massive MIMO sites, RRUs, mmWave high-frequency AAUs, and LampSite X.

At the same time, Wi-Fi 7, FTTR C-WAN architecture, and 50G PON can be used to bring a 10 Gbps experience into people's homes. To help accelerate the adoption of 10G PON, Huawei has launched a next-generation FTTR product, the iFTTR F50 series; our brand-new MA5800T, the only OLT platform in the industry that supports GPON, 10G PON, and 50G PON on one port; and our scenario-specific AirPON solutions.

For campus users and enterprises, Wi-Fi 7 and upgraded switches will help deliver a 10 Gbps experience across the board. To support this evolution, Huawei has launched its next-generation series of AirEngine Wi-Fi 7 products and CloudEngine campus switches.

Feature #2: 400G/800G transport networks

As mobile, home, and campus networks move towards 10-Gbps connectivity, metro and backbone transport networks will need to support end-to-end 400G and smooth evolution to 800G. At the IP layer, capabilities such as E2E SRv6 and network slicing will be necessary to support multiple services at the same time. These can help carriers satisfy transport requirements for mobile and home scenarios, while seizing opportunities in the enterprise market, such as multi-cloud migration. At the optical transmission layer, applying pooling architecture to metro networks, and leveraging

E2E OXC in both metro and backbone networks, will deliver ultra-low latency, ultra-large capacity, and optimal total cost of ownership. In this regard, Huawei has launched a series of routers and optical transmission products that promise an industry-leading performance and advanced functionality.

Feature # 3: Core networks with new user planes and intelligence planes

Core networks will need new user planes and intelligence planes to enable 10-Gbps access, deliver a more intelligent and personalized experience, and enable more innovative services. Huawei can help with its Intelligent UDG, a new user plane product that supports a 10-Gbps experience. To support new intelligence planes, Huawei has also introduced its Multi-modal Communication (MMC) and Intelligent Personalized Experience (IPE) solutions.

Feature #4: L4 autonomous networks

As networks grow in complexity, intelligent O&M is a must, so key technologies like telecom foundation models and digital twins will be needed to support evolution to L4 autonomous networks. Building on its Autonomous Driving Network (ADN) solution, Huawei has released a series of role-specific AI copilots and scenario-based agents, including copilots for field maintenance engineers, call centers, home broadband installation and maintenance engineers, and marketing managers, as well as AI agents for optical path maintenance, high-value scenario assurance, network risk management, and user impact event management.


Feature #5: Carrier-operated public cloud services

Finding the right partner to help seize new opportunities in the public cloud market should be a key consideration as carriers build out and operate their own cloud service platforms. Huawei supports several different partnership models to help carriers thrive in this market, such as deploying Huawei Cloud Stack on-premises, or running Cloud On Cloud using Huawei Cloud's local nodes. Building on years of experience in public cloud services, Huawei also provides joint marketing support, as well as operation and ecosystem-centric enablement to help carriers make the most of their local cloud markets.

Carriers have a number of advantages in local markets, including their infrastructure, their brand, as well as business development and delivery capabilities, so they are well-poised to find a development model that helps maximize their opportunities in public cloud.

Feature #6: AI-oriented unified data storage

As AI gains traction around the world, data will become an even more important driver of productivity. This will raise new requirements for data storage. The size of AI training datasets has grown from billions of parameters to hundreds of billions of parameters, and databases need to perform millions of reads per second, rather than just tens of thousands per second. The bar is rising for bandwidth as well. To support unified data storage in the age of AI, Huawei has launched its OceanStor series of products to greatly improve AI training efficiency and GPU utilization. OceanStor can also be used in key application scenarios, such as billing and CRM, to support unified data storage.

A journey of a thousand miles begins with a single step. Huawei looks forward to working with all industry players to identify key trends and changes in the industry, jointly define ICT architecture for 5.5G, and drive the evolution of ICT infrastructure. Together, we can achieve shared success and raise the telecom industry to new heights. 

The six defining features of our ICT architecture for 5.5G will drive the evolution of ICT infrastructure.

AI: The Bridge to 6G

In the age of AI, the core carrier of 6G services and applications will shift from the mobile Internet and smartphone apps to AI agents across various sectors. This means AI will serve as a bridge to 6G.

Progress in foundation models and AI is promoting digital transformation across industries, while laying a foundation for the future of communications technology and playing a vital role in the shift towards 6G.

In June 2023, the International Telecommunication Union (ITU) completed the recommended framework for the 6G vision, which answers the question "What is 6G?" from two aspects^[1]. First, 6G will continue to evolve mobile communications and expand in three usage scenarios, i.e., enhanced Mobile Broadband (eMBB), Ultra-Reliable and Low-Latency Communication (URLLC), and massive Machine Type Communication (mMTC), to provide immersive and deterministic communication experiences and support massive connections. Second, 6G will go beyond the scope of mobile communications to achieve integrated sensing and communication, the integration of AI with communications, and ubiquitous space-air-ground integrated connectivity. These advances will allow us to "observe" the physical world in ways that



Tong Wen

Huawei Fellow, CTO of Huawei Wireless



Ma Jianglei

Technical Vice President of Wireless Research, Huawei



Zhu Peiyong

Huawei Fellow, Senior Vice President of Wireless Research, Huawei



Chen Yan

Senior Expert of Wireless Research, Huawei

exceed human limitations and create digital twins in the virtual world. The 6G vision is an embodiment of global consensus and a key milestone on the path towards a globally-unified 6G standard.

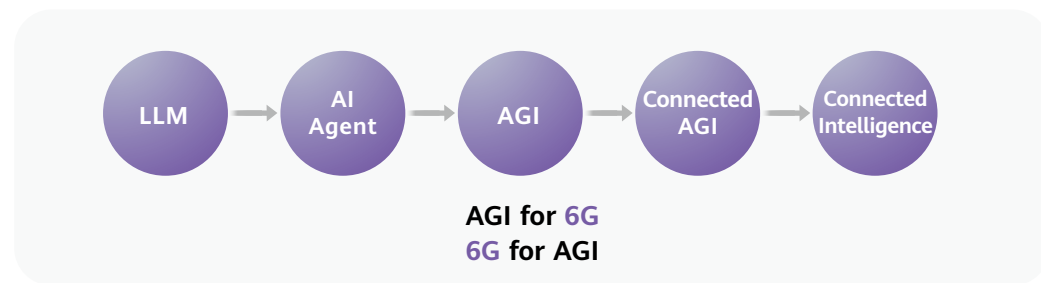
Within the three scenarios beyond communications, integrating AI with communications mainly focuses on how 6G can be designed to natively support massive AI services and applications in the future. Over the next five to ten years, 99% of all development, design, and administrative tasks is expected to be done by AI. In the near future, foundation models will even replace manual architecture design chip. This future trend will overlap with the window

for 6G deployment. Technological innovation in 4G LTE brought us into the mobile Internet age, in which smartphone apps were the main carriers of applications and services. In the age of AI, the core carrier of applications and services will shift from mobile apps to AI agents.

AI agents can sense and proactively take action. Capable of sensing, learning, and acquiring knowledge, they can set action objectives based on the environment and constantly improve their capabilities. The recent success of foundation models has taken AI agent capabilities to a new level, going beyond just generative AI, to creating interactive AI capable of complex dialogues and decision-making. Therefore, in the

6G era, networks will power not only AI agents, but artificial general intelligence (AGI). Huawei's vision of Connected Intelligence (Figure 1), proposed in 2019, assumed support for native-AI capabilities and involves two aspects: AGI for 6G and 6G for AGI^[2].

This article covers both of these aspects, with a particular focus on 6G for AGI. As shown in Figure 2, 6G for AGI looks to explore areas like how to design communications capabilities like eMBB+, URLLC+, and mMTC+, and how to use networks' sensing capabilities to better support AI and make 6G into the neural center that connects future AI agents and a key part of AI learning, training, and inference. If 6G is to succeed in these areas, the 6G system needs to be designed with an architecture that expands beyond connectivity and must integrate the four basic functions of AI agents: sensing, cognition, decision-making, and action. This architecture should use efficient, intent-based communications to closely integrate the physical and digital worlds and thus influence how the physical world operates.



AGI for 6G

In the 6G era, the basic model of communication for AGI will be based on effectiveness communication^[3], as proposed by Warren Weaver, or intent-based communication. Such a framework would go beyond Claude Shannon's model of communication which involves the transmission of only bits. Bits do not represent understanding and are not intelligent, which essentially outlines how AGI-enabled 6G communication differs from traditional communication.

AI agent-powered 6G communications can be broken down into four types:

1. Human-to-human system-1 and system-2 communication
2. Machine-to-machine intent-driven communication
3. Human-to-machine ultra-reliable low-latency communication
4. Machine-to-human spatial-computing-based metaverse communication

Figure 1: Connected Intelligence = AGI for 6G + 6G for AGI

To effectively support an AI-agent-powered 6G communication framework, 3GPP standard design must consider uplink channels that support both sensing and learning, as well as downlink channels that support inference, low latency, and metaverse applications. The remainder of this section will focus on the first two types of 6G communication. The other two types are similar to the first two in terms of how they use the AI-agent-based framework, but differ in terms of specific communication requirements, depending on scenarios.

AI-agent-assisted human-to-human communication

In the future, foundation-model-based communication will organically integrate communications in both the physical and digital worlds. This will give rise to a post-Shannon-model AGI communication architecture. Human-

to-human communications, for example, are grounded in two core concepts:

- First, everyone can use a foundation model, which is a generative pre-trained transformer (GPT), as an agent.
- Second, each person's foundation model can use a GPT and spatial computing to generate a multimodal virtual real-time response, representing a proxy response of the person's deterministic behavior. In the digital world, a communication system for such GPT-based agents is known as system 1. It is worth noting that foundation models cannot accurately learn and model non-deterministic behaviors such as emotions. A communication system that transmits such information in the physical world in real time is known as system 2, which can also be designed based on GPT foundation models.

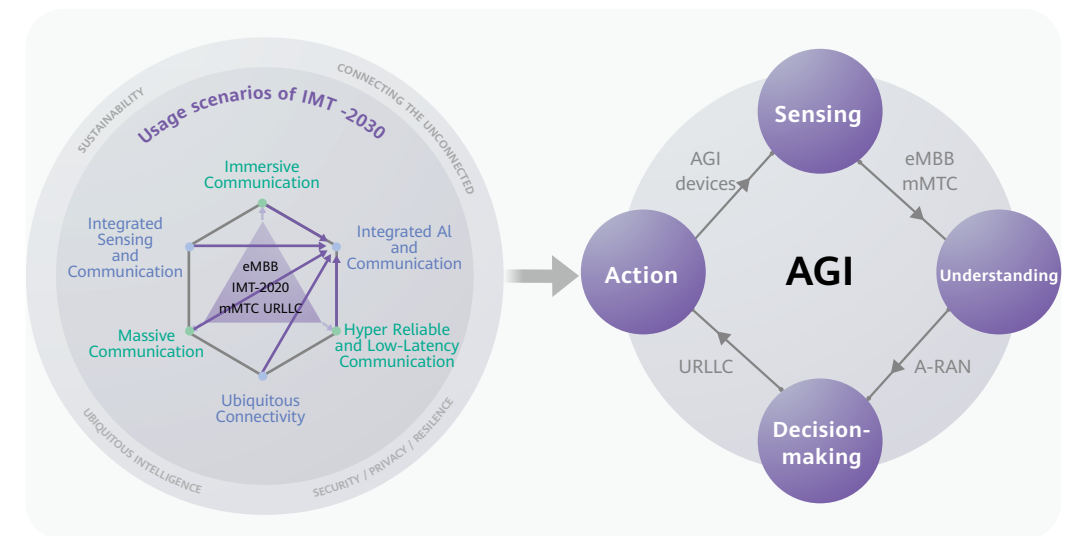


Figure 2: 6G's AI-native capabilities and the 6G for AGI framework

Each person can release their own foundation model, allowing people to access each other's models before communicating. This facilitates system-1 communication between foundation models, which is essentially local communication that does not use wireless communication resources. Wireless channels will be used for communication in cases where foundation models fail to generate what's necessary for system-2 communication.

6G communications between AGIs (shown in Figure 3) include internal-channel communications based on the Shannon model and external-channel communications between neural networks, between foundation models, and between agents. It should be noted that radio air interfaces are increasingly being powered by GPUs, at the cost of higher power consumption without higher performance.

Human-to-human communications assisted by AI agents are an advanced method of interaction that uses powerful AI capabilities

to enhance and optimize the communication process. Within this framework, everyone will have two major GPT foundation models: system 1 is used for local intelligent processing, and system 2 for physical communications.

First, everyone will need to train the GPT foundation model used by their system-1 AI agent, as well as the GPT foundation model used for system-2 physical communication. Such training will primarily involve supervised offline learning. Training based on a general-purpose foundation model in a broader sense can enable continuous updates and ensure humans are kept in the loop, making the resulting GPT model more accurate and powerful. Second, an emergence detector will be required to detect whether system 1 is working properly. Parts that system 1 cannot learn or model will be distributed to system 2 for learning.

These will create an architecture that combines both fast and slow communications. System 1 is fast communication that facilitates local closed-

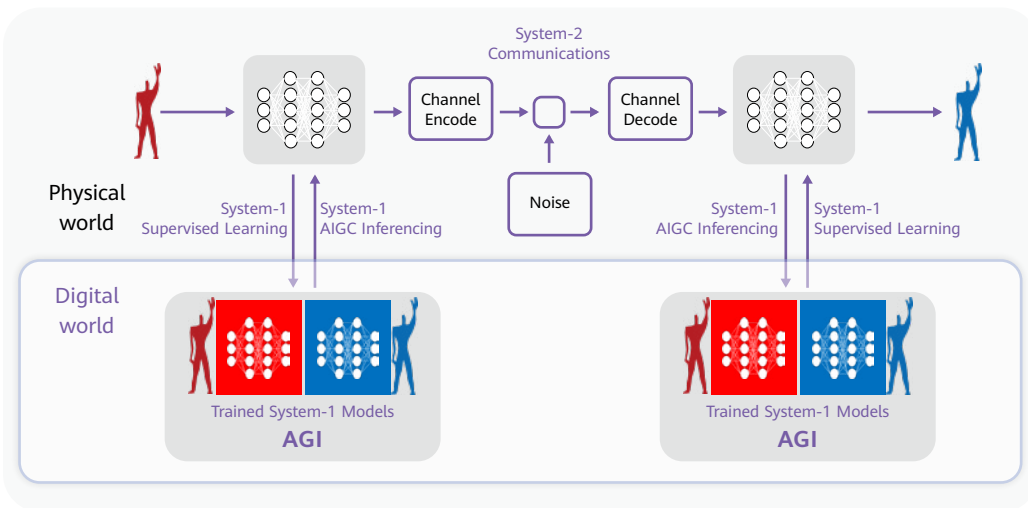


Figure 3: Post-Shannon-model AGI communication architecture based on GPT models

loop communication between the AI agents of both ends without occupying communication resources (shown in Figure 4). System 2 is slow communication and can use wireless channels (shown in Figure 5). It should be noted that the use of an AGI-based intent-driven

communication mechanism allows system-2 communication to reduce data traffic by 100-fold or even 1,000-fold compared with direct video communication. Furthermore, the AI agent of both system 1 and system 2 can constantly update the general-purpose foundation model.

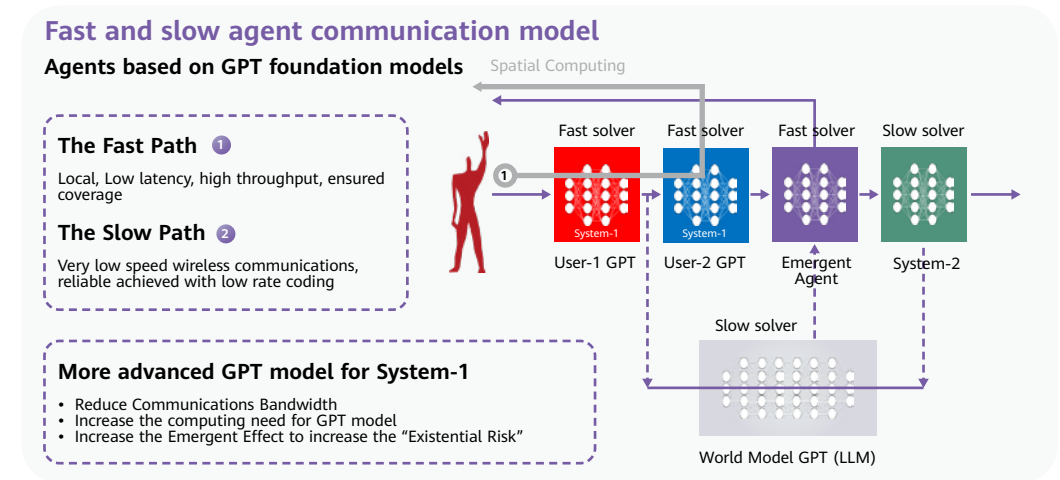


Figure 4: System 1 based on a GPT foundation model – Fast local communication

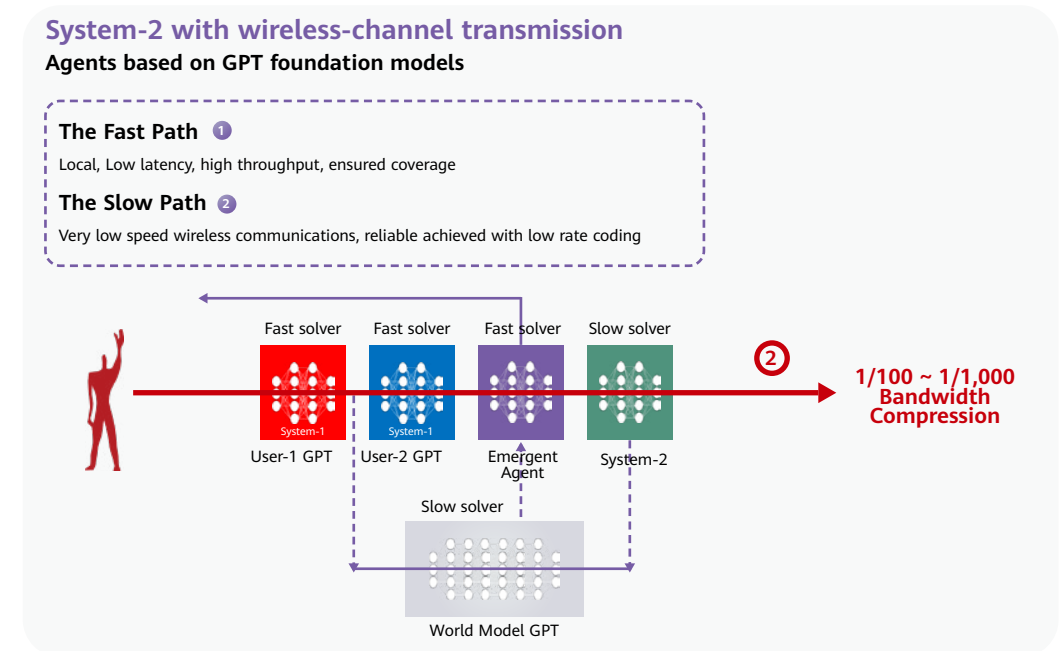


Figure 5: System 2 based on a GPT foundation model – Slow physical communication

6G is essentially about integrating communication, AI, and sensing to create a neural center for numerous AI agents.

AI-agent-assisted machine-to-machine communication

In terms of machine-to-machine communication, uploading visual-sensing results (e.g., complete videos and point clouds) to support foundation-model computing on the edge or cloud will result in a huge amount of uplink traffic, limiting the number of machines that can be supported. To combat this, a primary AI agent can be used on devices for the purpose of real-time token alignment with foundation models on the cloud through wireless channels to facilitate AI agent collaboration across devices, pipes, and cloud, thus realizing massive machine-to-machine communications (shown in Figure 6).

Specifically, an AI-agent-based post-Shannon-model communication framework uses intent-driven communication according to the following steps:

- Step 1: Use an AI agent on devices to perform primary preprocessing and analysis of the scenario, which is also known as goal-oriented filtering, in order to clean sensing data in real time.
- Step 2: Perform embedding in the transformer foundation model on the extracted objects to obtain the simplified mathematical descriptions (tokens) of intents.
- Step 3: Transfer data back to the edge or cloud through wireless channels to align intents (represented by tokens) on both ends in real time, thereby realizing efficient machine-to-machine communication with device-pipe-cloud synergy.

Compared with direct video transmission, this transmission mechanism can reduce data traffic by 100-fold or even 1,000-fold, increasing the number of communicating users the system can support by an order of magnitude.

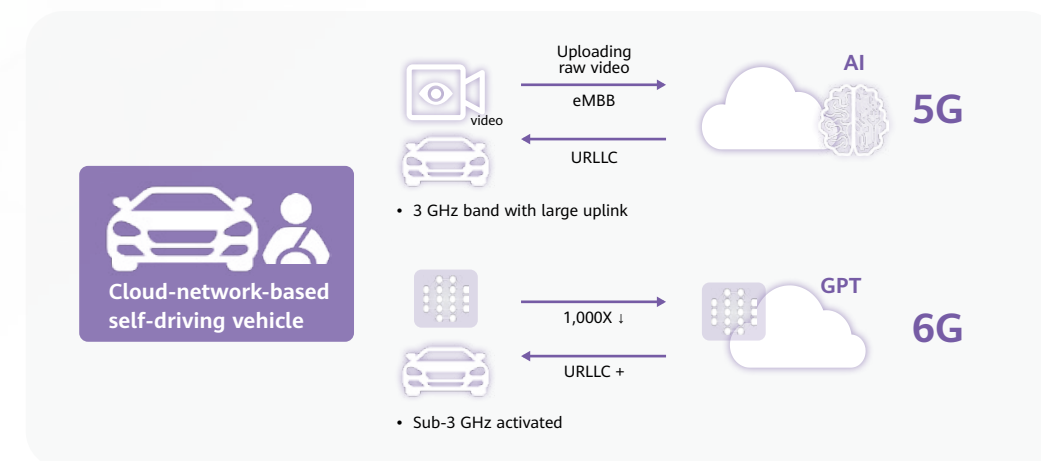


Figure 6: 6G-powered, intent-driven machine-to-machine communication

6G for AGI

6G sensing provides big data sources for AI learning

6G's integrated sensing and communication (ISAC) is another unique advantage of applying AI agent services on 6G networks. ISAC brings new opportunities to wireless communication systems—providing wireless sensing services while supporting communications. Native convergence of sensing and communication enables mobile base stations and devices to obtain a larger sensing scope and higher sensing accuracy through collaborative sensing, without additional spectrum or increased equipment costs. With shorter radio wavelengths, broader spectrum resources, and larger antenna apertures, 6G can support the highly-accurate, real-time reproduction of the physical environment as a service. This capability

can also help significantly reduce transmission power consumption while enhancing wireless transmission performance.

Data extracted from 6G sensing can be used for modeling the physical world in areas the network can reach, as well as providing a source of big data for AI learning (shown in Figure 7). People, machines, vehicles, buildings, materials, and even weather can be objects of 6G sensing. Wireless sensing can provide big data on the environment through parameter estimation, imaging, and even mass spectrometry, all of which are transmitted over radio waves. Attention and study are both essential for sensing across the entire communications spectrum, including centimeter wave, millimeter wave, and sub-THz bands. THz technology has the potential to see wide adoption in high-precision sensing.

6G-based intelligent and inclusive A-RAN and A-Core

Foundation models for natural language processing, represented by ChatGPT, will match and exceed human capabilities in the near future. However, such human-like intelligence will be possible only when supported by the computing power provided by supercomputing clusters. For example, 500 billion neural network parameters require over 10 million watts of power supply to run. In the next 10 years, it is unlikely that such foundation models will be able to run on mobile devices.

6G networks must be able to deliver inclusive intelligent services for all people and all things, anytime and anywhere. This requires 6G networks to adopt a C+A+S (communication + AI + sensing) architecture powered by foundation models. The convergence of communication, AI, and sensing is a key feature of 6G. The post-

Shannon-model communication architecture that supports AGI (i.e., the C+A+S radio access network) is called A-RAN (shown in Figure 8).

6G networks can also be built with AI agents, with each serving as a logical network element (NE). Huawei's 2012 Laboratories proposed the concept of application-driven networks in 2015. The core idea behind this is automatically generating customized networks for customers' application requirements, and cancelling such networks when the applications are no longer needed. The infrastructure is like a unified computing platform that stretches across the entire network. It is a task-based network architecture and the prototype of 6G A-Core (as shown in Figure 9).

Creating a real-time twin world with 6G AGI

In the digital world, virtual copies of physical entities are built based on application intent in order to simulate and analyze real-world

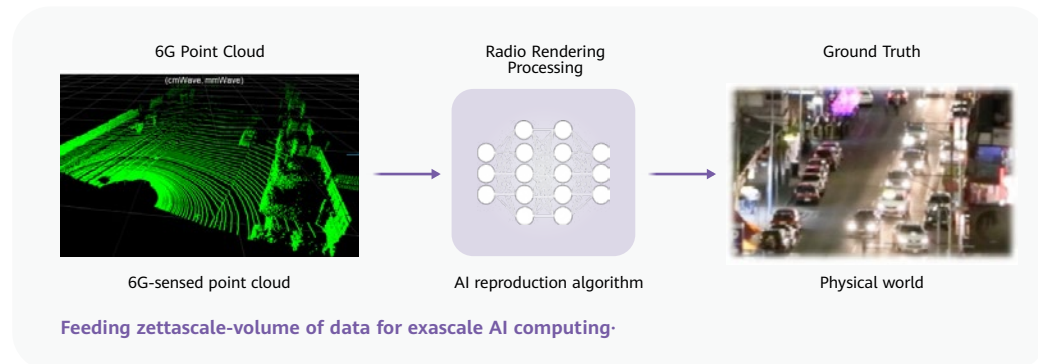


Figure 7: 6G sensing is a major data source for future AI foundation models

behavior and performance, in what we call digital twins. This section covers two technologies related to digital twins. The first is building digital twins that can mirror the physical world in real time through intent-driven communication. The second is performing accurate spatiotemporal inference about the physical world based on the real-time twin world. Both technologies have the potential for wide application across numerous scenarios like self-driving vehicles, robotics, smart industrial production, and telemedicine.

6G devices can be placed in a small physical space, such as a small room, to support integrated sensing and communication provide the AI computing power required to

run a small-scale AI model. These devices can sense the physical world through wireless signals. By processing the sensing signals, the devices can generate a point cloud that both depicts and describes the physical world.

Then, with a given intent, the physical world is selectively reproduced through AGI communications between an AI foundation model in the network and the small AI model running on the 6G devices. The twin world in question does not have to completely describe the physical world, but is capable of identifying things relevant to the intent. This reduces communication overhead, while protecting personal privacy and keeping information about the location confidential.

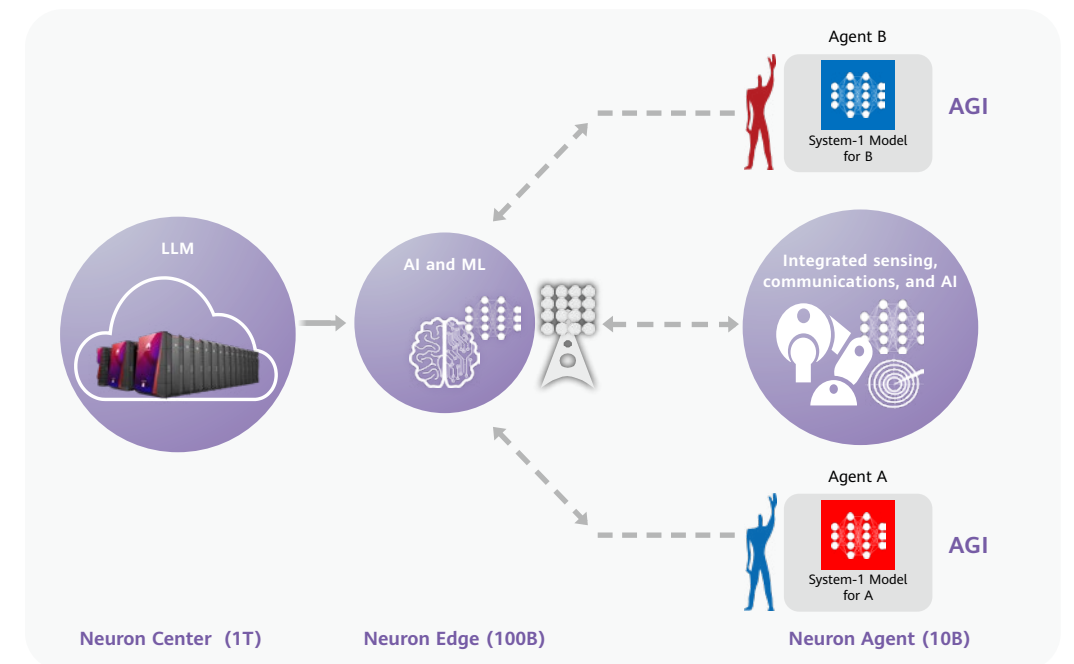


Figure 8: A-RAN architecture: Communication-AI-sensing converged access network


The sensing system senses point-cloud information, which is lattice representations of objects in the three-dimensional space. To reduce the amount of data to be transmitted and protect user privacy, we upload semantic spatiotemporal information to the cloud for fusion only when the semantic target matches the original sensing information. That is why we have developed a unique semantic spatiotemporal fusion and prediction algorithm that can fuse semantic information to form a digital twin. The algorithm enables a digital twin to mirror the information of concern in almost real time, while reducing the required uplink transmission bandwidth by orders of magnitude.

Object positioning and human-pose tracking can already be driven by natural language through a real-time twin world. Unlike

ChatGPT, such a system can sense the physical world in real time, perform semantic, temporal, and spatial inference, and then present the results in the form of natural language. For example, the positions of an object and person can be displayed in the digital twin in real time in the form of a rendered animated object and figure. A 6G AGI system can use the sensed point cloud information to identify actions relevant to preset intents (e.g., reading a book), and locate the spatial positions of such occurrences in real time. Continuous improvements to the sensing system will further increase inference accuracy.

In conclusion, we believe that 6G is essentially about integrating communication, AI, and sensing to create a neural center for numerous AI agents. We also believe that the following elements of 6G native-AGI communication

will be essential in the future foundation model era:

- Connected Intelligence = AGI for 6G + 6G for AGI
 - AGI for 6G: Effectiveness communication powered by the post-Shannon-model communication architecture
 - 6G for AGI: An inclusive intelligent neural center that integrates AI learning, training, and inference
- 6G A-RAN which integrates communication, sensing, and AI
- Task-based 6G A-Core that is built with agent NEs 

References:

[1] Recommendation ITU-R M.2160-0, Framework and overall objectives of the future development of IMT for 2030 and beyond.

[2] Wen Tong and Peiyong Zhu, "6G: The Next Horizon – From connected people and things to connected intelligence," Cambridge University Press, May 2021.

[3] C. E. Shannon and W. Weaver, The Mathematical Theory of Communication. The University of Illinois Press, 1949.

6G networks must be able to deliver inclusive intelligent services for all people and all things, anytime and anywhere.

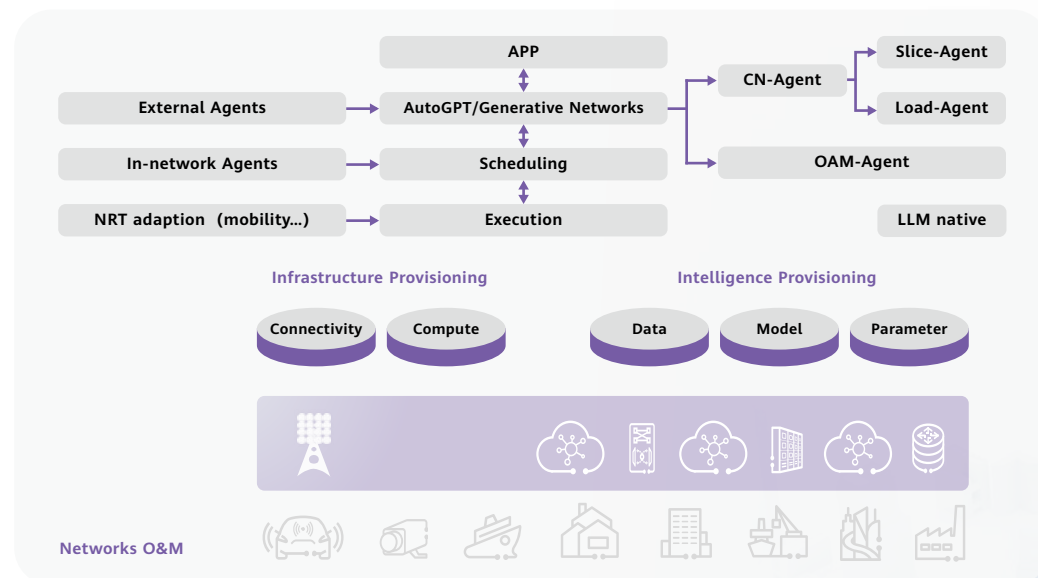


Figure 9: A-Core: Task-based network architecture

The Evolution Toward 5.5G and New Telco Business Models

Contents

- The broadband roadmap to 5G-A
- Future application demand
- Potential new business models for broadband service providers



Download PDF
*This is an Omdia report.

ELITE FWA Club

Enhancing User Experience,
Monetizing 5G,
Facilitating Business Success

Industry Support:



Industry Partner:

4G
5G FWA Forum



For more information



Tech for Nature

Digital technology enables smarter
biodiversity conservation

