ZTE GSMA Intelligence

B5G Technology White Paper

CONTENT

01

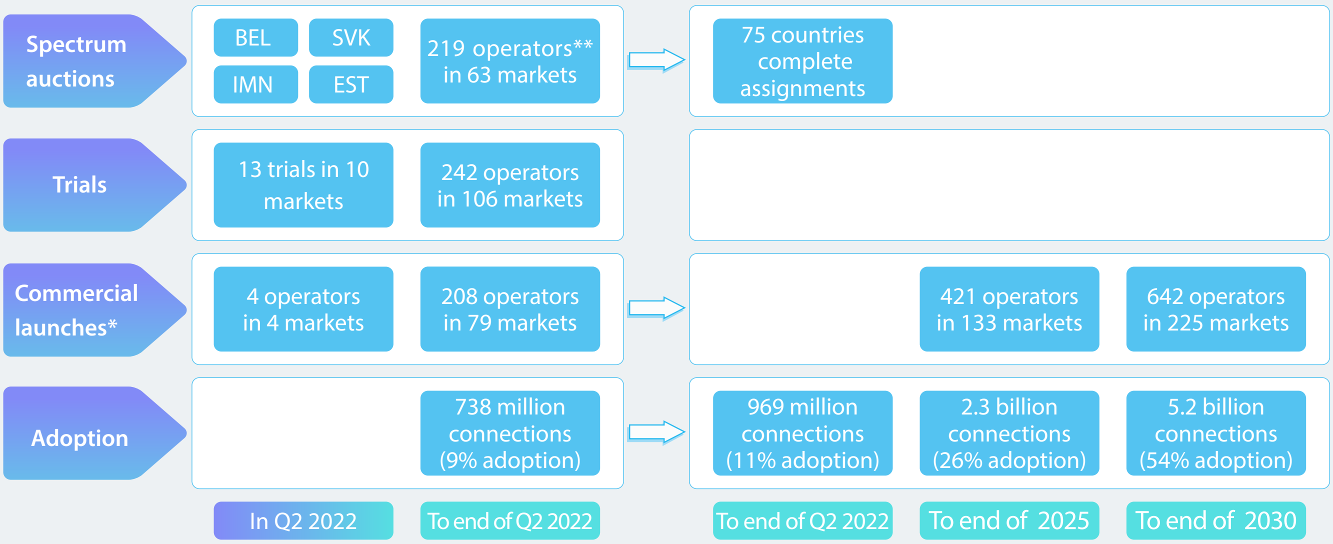
Welcoming the beyond-5G era:

from 5G-Advanced through to 6G

At GSMA Intelligence, our analysts spend their days working closely with operators to understand and guide their business strategies, demands and direction. As part of this remit, we have been talking about the state of the 5G market for several years already.

We have tracked 5G trials and commercial deployments, enablers such as 5G spectrum deployments, and the impact of 5G on operators' supplier decisions. We have carried out this work because 5G is the latest generation of mobile broadband technology, but also because of the high hopes around 5G in terms of addressing ever-increasing demand for data and helping operators enable (and create value from) the digital transformation of enterprises.

5G at a glance: global outlook



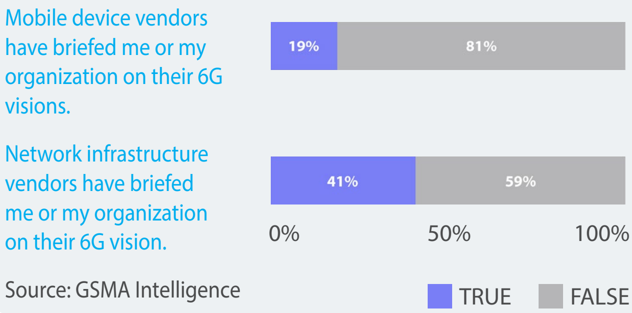
* Launches of commercial mobile and FWA 5G services
** Excludes regional US and Canadian operators

Data correct to 30 June 2022. For updates, see gsmaintelligence.com
Source: GSMA Intelligence

After years of planning for 5G's arrival and working through its early phases, we are now firmly in the 5G era. Yet, as is often the case, when one milestone is achieved, the industry begins to ask what is next.

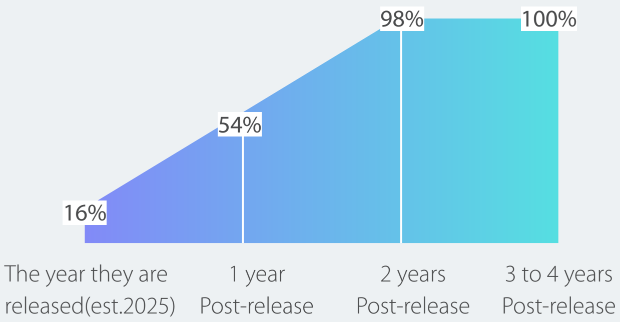
5G-Advanced standards, for example, have progressed to the point where there is line of sight to what the next major iteration of 5G may deliver within the next few years. Vendors and operators have been busy releasing 5G-Advanced and 6G messaging, communicating their visions, expectations and R&D plans for beyond 5G (B5G). Meanwhile, network infrastructure and mobile device suppliers have been communicating B5G visions and R&D roadmaps to their customers – the mobile operators that will eventually deploy these technologies.

6G R&D: vendor outreach to operators - Q3 2022



Source: GSMA Intelligence

5G-Advanced timelines: how long following standards release do operators expect deployment?



Source: GSMA Intelligence

Industry messaging efforts are about more than just sales and marketing. They are a very real consequence of operator interest, customer demand and network deployment plans. Interest from operators around 5G-Advanced can only be categorised as intense, with more than half of those surveyed expecting to deploy the technology between one and two years following the release of standards. This isn't particularly realistic. Regardless, it points to the recognition of – and extreme focus on – B5G technologies by operators as they look to continue the momentum of 5G but also achieve the business gains that 5G promised. For many operators, after all, 5G deployments have not resulted in real revenue uplift.

Vendor messaging, combined with over-eager operator interest, does not necessarily mean that B5G technologies are important in the here and now. However, market trajectories and trends make the value and importance of B5G technology very clear.

While we may take them for granted, today's realities all point to the critical nature of 5G, as well as the need to seriously plan for what's next. These dynamics include the following:

- **Unending data demand:** more data, in more places, supporting more services
- **6G use-case development timelines:** a long-term proposition
- **5G in support of B2B:** the core industry promise requiring added support
- **5G in a 6G world:** an enduring legacy
- **The three S's:** security, sustainability and spectrum.

Against this backdrop, B5G technologies are more than something it would be nice to begin thinking about; they represent fundamental enablers of the industry's health and its ability to continue delivering critical, life-changing services. This means it cannot be too early to begin exploring their potential and thinking about them holistically, asking what we hope to achieve and how to think about them as part of a broader technology evolution.

ZTE's research and whitepaper is an important part of that process. With myriad potential B5G candidate technologies, the work helps to explain how some of the key technologies operate and what they promise. Just as importantly though, taking a step back to ask what we hope to achieve in the B5G era is critical and part of making sure we exploit the full potential of new B5G technologies.

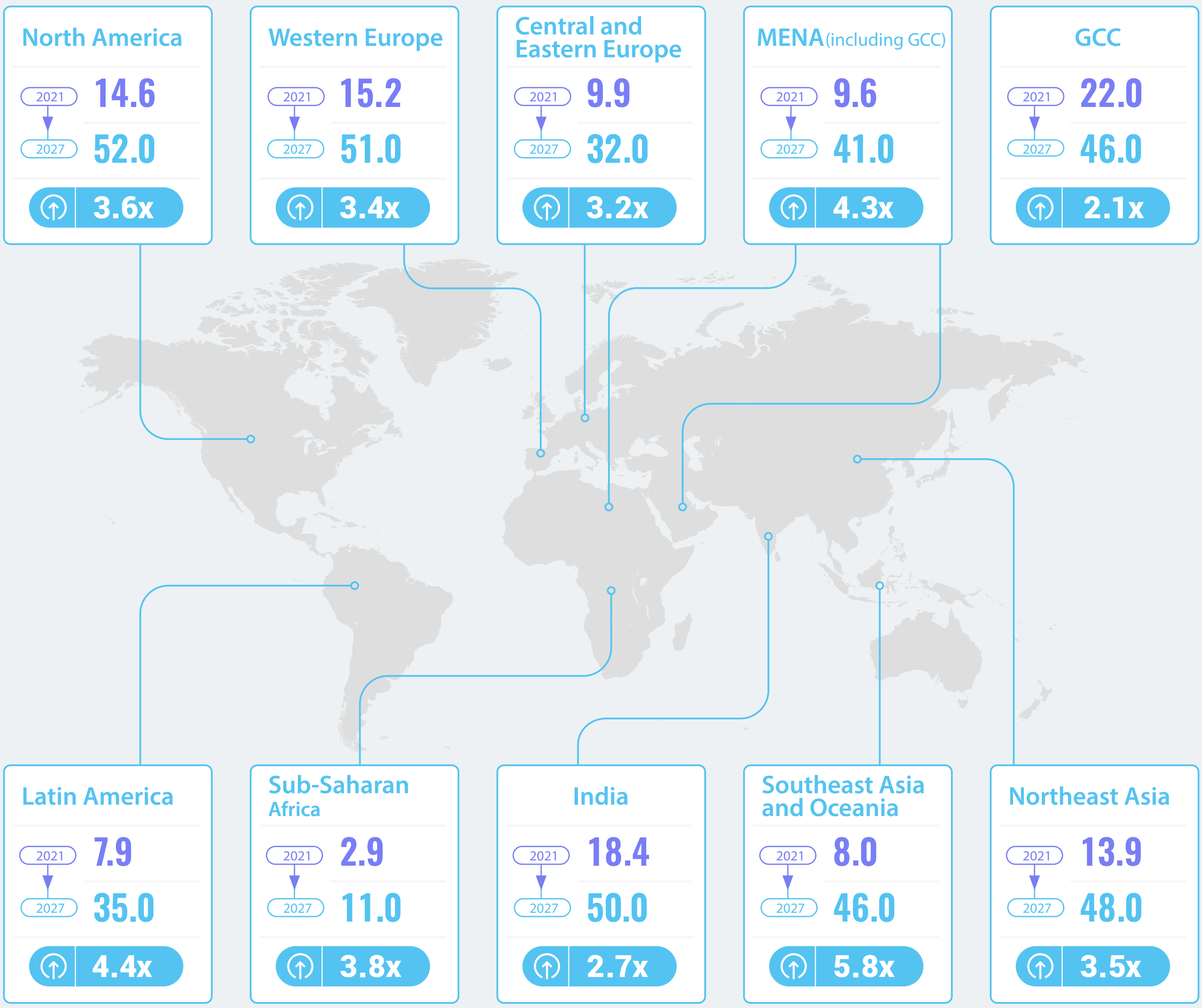
1.1 Why B5G matters

Mobile broadband technologies will continue to evolve. They will need to evolve in order to meet increasingly demanding enterprise, consumer and societal requirements. Why do B5G technologies and innovations matter in the near term then? The market dynamics outlined earlier provide a solid foundation for explaining why B5G matters today and why it is not too early to begin thinking about and planning for it.

1.1.1 Unending data demand

Between 2021 and 2027, mobile data usage across the globe will grow by 3.6× to reach 41 GB per smartphone, per month. Growth dynamics will vary between markets, but every operator in every country will need to deal with incredible demands on their networks and service infrastructure.

Mobile data traffic,2021-2027(GB per smartphone, per month)



Source: GSMA Intelligence, Mobile Economy 2022 https://www.gsma.com/mobileeconomy/

That mobile data traffic will continue to grow over the mid-term is not a new insight; demand for data has been on an inexorable climb since the introduction of mobile data services.

However, the evolution of mobile data demand over the next decade will be about much more than traffic demand and data usage.

Just as the mobile email and web browsing services that drove early data usage gave way to interactive video, gaming and user-generated content, tomorrow's dominant mobile broadband services will bring new requirements. For example, if expectations around the success of the metaverse come to pass, it will require massive data capacity improvements but also latency, coverage quality (cell edge, in particular) and uplink bandwidth improvements over the capabilities of today's 5G networks.

If the only material change in requirements involved added data traffic, it might be reasonable to expect operators to meet customer demand by simply doing more of the same. 5G rollouts, combined with legacy network decommissioning and radio access network (RAN) densification, could allow network capacities to scale incredibly. But, putting aside the costs involved and their impact on end users, current technologies will simply not accommodate the new latency, coverage and uplink-versus-downlink realities – along with other requirements we have yet to even realise.

These requirements demand new technologies and mobile network innovations. And, since tomorrow's use cases will be arriving sooner than many people think, we need to begin planning early.

1.1.2 6G use-case development timelines

The term 'killer app' has been around for decades, cited over the years by tech luminaries including Steve Jobs and Bill Gates. However, in many ways, the concept is something of a fallacy. For new technology generations at least, there is rarely a single, novel application that drives the industry's success. Just as importantly, few people have ever been able to accurately predict the new 'killer app' that eventually went on to drive traffic and usage of the next mobile generation.

Consider some of the potential 6G use cases being discussed today. The NextG Alliance outlined an impressive set of them in its '6G Applications and Use Cases' report from earlier this year:

- field robots for hazardous environments
- immersive gaming
- mixed-reality telepresence and immersive education
- untethered wearables
- remote data collection powered by terrestrial and non-terrestrial network synergies.

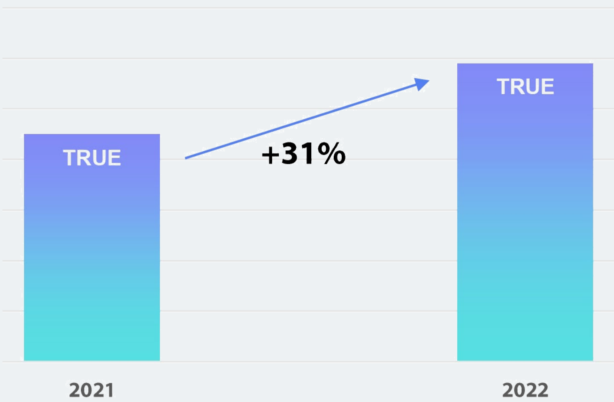
In a 6G world (2030 or later), fully executing on these use cases will require support for stringent performance criteria in terms of bandwidth (uplink in particular), network capacity, latency, coverage, security, costs and more. However, much of this feels very familiar in the near term.

Robots are already being used in industrial environments, supported by 5G private networks. Gaming is already shifting from consoles to mobile devices, thanks to mobile network innovation. Interactive, immersive communications took a huge leap forward thanks to the Covid-19 pandemic and are now attracting attention as part of metaverse aspirations. Meanwhile, multi-network connectivity integrating terrestrial and non-terrestrial networks is garnering plenty of attention as mobile device vendors, and mobile device software vendors, rush to integrate direct-to-device satellite support into their solutions.

This isn't to imply that 6G won't bring new capabilities and amazing new applications. Rather, it is a reminder that tomorrow's use cases will build on what we have already developed and are planning for in the near and medium term. That means we can gain insight into what is needed from B5G technologies well before those technologies and use cases arrive. The fact that we can see the seeds of B5G use cases in today's applications signals that we will need today's 5G networks to evolve in order to meet market demand.

Recent GSMA Intelligence research suggests operators were 31% more likely to see 6G as 10+ years away in 2022 than they were in 2021. This underscores a realization that 6G may take longer to arrive than once expected, and that 5G networks must continue evolving as the run-up to 6G's arrival extends into the long-term.

Mobile operator viewpoint: commercial 6G is at least 10 years away?



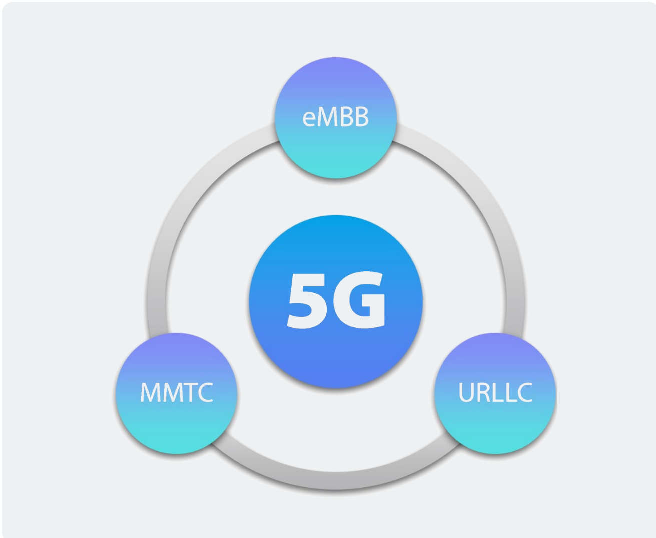
Source: GSMA Intelligence

1.1.3 5G in support of B2B

From the earliest days of 5G (well before it was even commercialised), simultaneous support for mass-market consumer applications along with more demanding enterprise use cases was the major technology and business innovation.

Enhanced mobile broadband (eMBB) would offer higher speed connections with lower latencies, which would benefit consumers and enterprises alike. But massive machine-type communications (MMTC, or scaled IoT) and ultra-reliable, low-latency communications (URLLC, or mission-critical communications) would provide operators with an opportunity to target new market segments in the business-to-business (B2B) space – and help them to grow their revenues in the process.

5G was never just about faster data or more network capacity. It was about serving very demanding and less demanding performance requirements from a common network. This is why 83% of operator CEOs in a GSMA survey claimed that enterprise and government segments were the most likely source of 5G revenue upside. That shouldn't be surprising when you recognise that the consumer segment accounts for about 70% of revenues across the market's top operators, and it helps to explain why so many conversations around 5G still reference those three core use cases.



To their credit, operators are making steady progress in expanding their revenue base beyond the traditional consumer demographic. However, if operators are to execute on the broad array of enterprise demands and support a diverse set of enterprise requirements, there is still more to be done.

Some of this is simply about continuing on the journey that 5G began and improving the performance of mobile networks: higher network capacity and connection speeds, along with further latency improvements that open up new use cases. But it also means extending coverage while improving uplink and cell edge connection performance to provide enterprise users with the confidence that networks can meet the demands of myriad devices, regardless of where they are operating. An evolving enterprise service environment embracing cloud environments, new security architectures and advanced network slicing business models will only add further demands on today's networks. Finally, some new requirements have nothing to do with individual device, service or business requirements but rather the fate of the world. That might sound like hyperbole, but GSMA Intelligence estimates that digital transformation driven by mobile broadband technology can support 30–40% of the decarbonisation goals of critical, energy-intensive industries such as transport or power & energy.

To execute on the foundational enterprise digital transformation promise of 5G, driving operator and societal goals in the process, an evolution beyond today's 5G is necessary.

1.1.4 5G in a 6G world

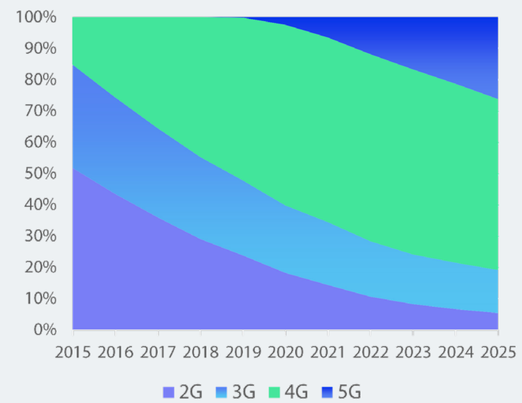
In 2015, 2G and 3G accounted for nearly 85% of global cellular connections. Five years later, that figure fell to about 40%, with 4G having scaled to take 58% of the market. Yet, in 2025, 2G and 3G will still account for about 20% of global connections, with 4G maintaining a relatively constant share of connections and 5G growing steadily.

While legacy technology generations will naturally decline in use as new ones scale, there is an important implication here. 4G delivered a critical set of mobile broadband capabilities, giving birth to an app economy that drove mobile technologies and smartphones to be a part of everyday life. As a result, 4G now dominates mobile broadband and will continue to do so in the near term. Yet, legacy technologies continue to live on; phasing out older networks and usage is difficult and time-consuming.

It is therefore natural to expect that a decade from now (after 6G should have arrived), 5G will still be a dominant technology for delivering mobile broadband.

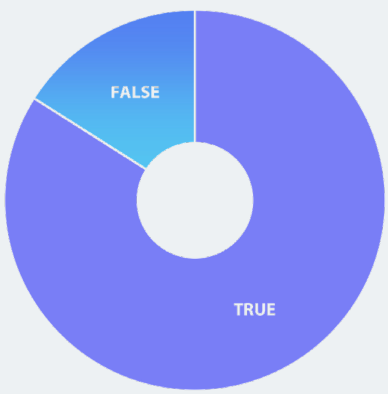
Against that backdrop, there are two realities to consider. First, we need to continue focusing on how 5G will evolve and how its performance will improve over the medium term. Second, given the importance of 5G, we need to ask how 6G technology innovations and visions can support 5G networks and services in the interim. This is a large part of what B5G technologies promise. Indeed, per GSMA Intelligence research from mid-2022, this is exactly what many operators expect.

Mobile technology generations: share of connections



Source: GSMA Intelligence

Mobile operator viewpoint: 6G R&D will improve 5G networks



Source: GSMA Intelligence

1.1.5 The three S's: security, sustainability and spectrum

Addressing the importance of B5G technologies, we've already discussed the key aspects of an operator's business and the way they will be looking at their business going forward, including:

- new business models built on enterprise digital transformation and the financial, as well as societal, gains to be had by addressing them
- new network and service capabilities – including capacity, coverage and latency improvements – that will be required to support customers old and new
- the pace of network innovation which dictates that the 5G technologies operators are investing in today will need to continue evolving and serving users well into the 2030s

Moving past these basics, three cross-cutting network and service dynamics are front of mind for operators today and need to play a role in how they think about, and plan for, B5G technologies: security, sustainability and spectrum.

When GSMA Intelligence asked operators in mid-2022 about their most important network transformation priorities, sustainability and security dominated the top three positions. Sustainability (and energy efficiency) was highlighted as very or extremely important for 84% of operators around the world. Network security saw 72% of operators make the same claim, with end-user security being very or extremely important for 69% of operators around the world. These ranked ahead of oft-cited strategic focus areas such as automation, open networking technologies, supply-chain diversity and the upskilling of staff.

Spectrum, though, might not capture operator attention in the same way. New spectrum allocations, mmWave deployment and spectrum refarming were actually near the bottom of operator 5G RAN investment priorities per our survey. Yet, the attention given to spectrum allocations, ongoing network sunset activity and the amount of money spent at spectrum auctions all highlight how important spectrum truly is to operators.

And what do spectrum, sustainability and security all have in common?

They represent a foundational set of priorities that transcend the marquee business initiatives that drive tectonic network shifts. At the same time, they are fundamental to those business initiatives. Success in the enterprise, for example, will require secure networks. Profitably supporting increased data demands, in turn, will require energy efficient, spectrally efficient networks. This is why, in one way or another, operators have been clear that the three S's are critical to how they view the future of their networks.

A focus on security, sustainability and spectrum was core to the way 5G technologies were developed. Security was baked into 5G standards from the outset. Sustainability has been a central component of the 5G value proposition to the extent that 5G can deliver more data for a given amount of energy than previous technology generations. Like every new mobile technology, 5G was also designed to be more spectrally efficient than previous generations, as well as more spectrally flexible and capable of supporting diverse spectrum bands.

As 5G evolves and gives way to 6G, the need to be secure, sustainable and spectrally efficient will not wane. In fact, today's operator priorities suggest the need will only intensify. 6G will play a role in meeting those needs, and we should begin exploring how. 5G, in turn, will continually evolve in order to execute on these needs and improve on existing network capabilities.

1.2 How to think about B5G: efficiency, extension, enhanced capabilities

One major complication in looking forward to future mobile technology generations, as well as the evolution of existing technologies, is where to focus our attention. There are many aspects that help to characterise a given 'G' and determine its development trajectory:

technology innovation – the new air interfaces, antenna technologies, transmission protocols, architectural revolutions (and more) that enable a new generation of network and service capabilities

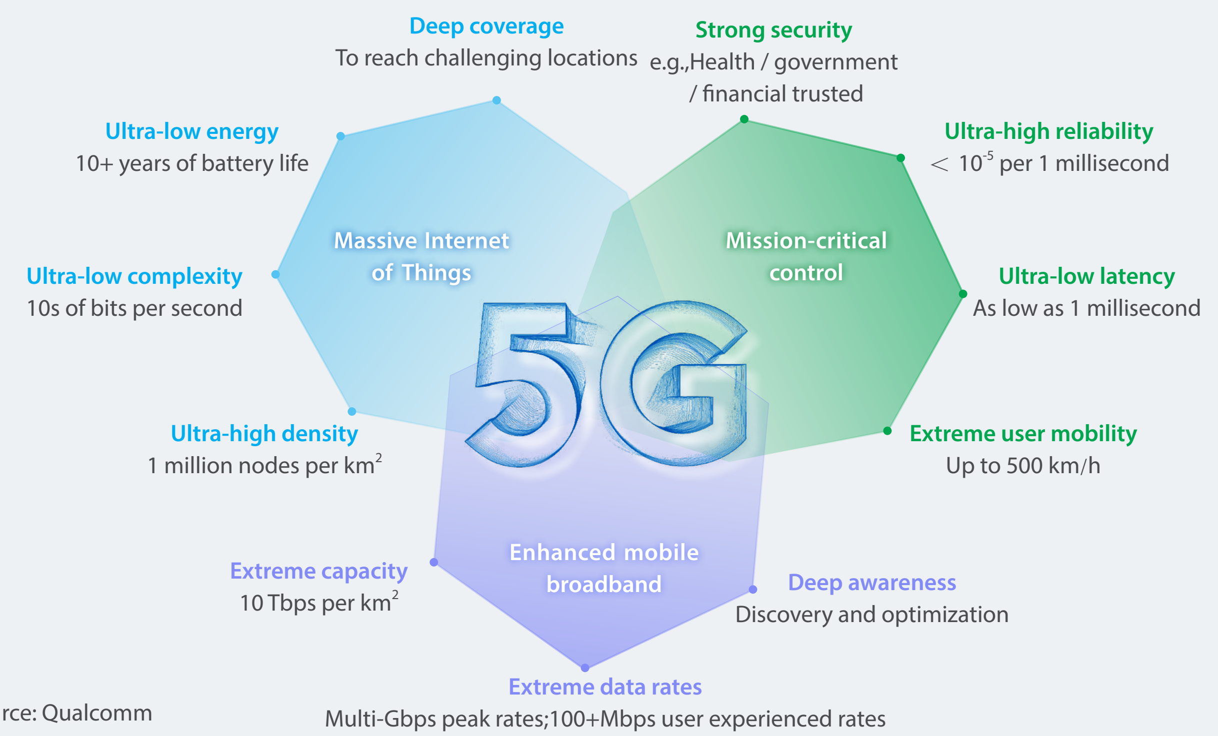
standards – the technical specifications adopted by standards development organisations which turn technology innovations into scalable, interoperable solutions

organisations and study groups – national and international research groups that bring together operators, suppliers, academics and governments to drive innovations which often feed into industry specifications and standards

use cases – the applications, services and specific situations in which the technologies could be put to use, and which will define their performance requirements

societal demands – the broader goals for a new mobile technology that stretch beyond network or service performance, such as security, sustainability and inclusion

You can see this dynamic in how some industry leaders and innovators have tried to capture the multiple dimensions of 5G use cases, breaking down each use case into performance requirements and even the technologies or standards that enable that performance. Adding specific business or consumer applications to be supported can add another layer of completeness – and complexity.

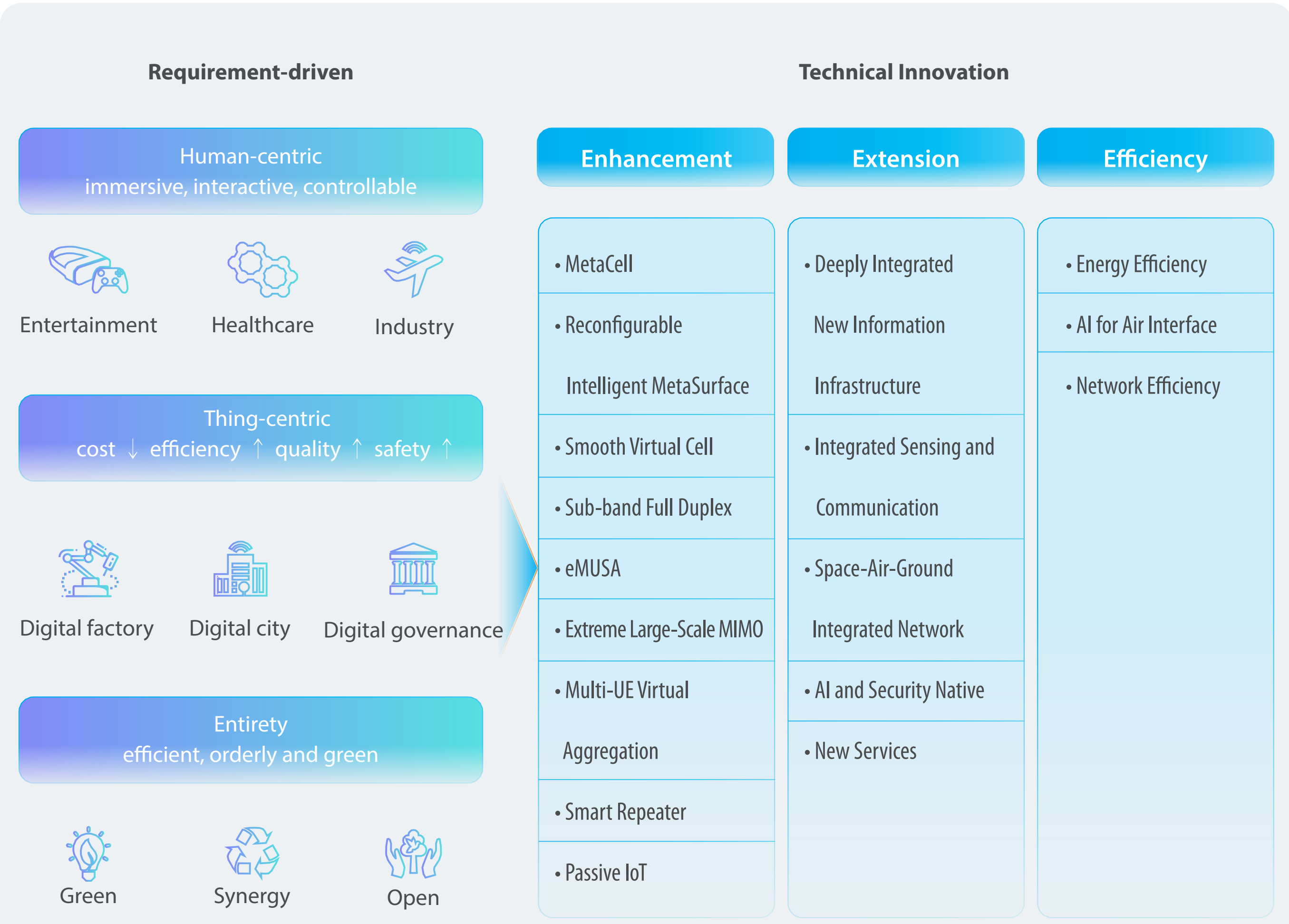


Source: Qualcomm

On top of this, national and regional dynamics add a layer of complexity and completeness in the search to understand future technologies, because so much of what goes into creating a new technology generation will differ across markets. The spectrum allocations and regulatory regimes in Europe may differ from those in Asia or North America. The services and applications driving uptake in pioneering markets such as Japan, South Korea, China or the US will differ from those in Latin America or Sub-Saharan Africa. User behaviour patterns; societal demands; market structures impacting fixed-mobile service convergence; deployment maturity and whether operators in some regions will be able to begin their 5G journey with 5G-Advanced or will look to deploy it tactically. All of this will vary by market – across and within regions – impacting technical choices that go into the creation of new technology.

In reality, all of these are important considerations; R&D innovations, use cases, standards, study groups, societal demands and regional differences all feed into the shape of tomorrow's 5G and 6G technologies. No attempt to characterise this development with a single vision will be perfect.

Against this backdrop, ZTE's Beyond 5G Vision attempts to represent some of these varied realities in a clear, logical manner.



Source: ZTE

Beginning with the notion that future use cases will need to address both human-centric and thing-centric application requirements (no different from today's 5G use cases), ZTE reminds us that societal and network requirements must also be taken into account. Whether that means a focus on sustainability, security, open networking or synergies across use cases, it's clear how this future B5G vision carries forward today's operator demands and 5G thinking.

Building on these requirements, ZTE's discussion of enhancement, extension and efficiency lays out the innovations that could help support B5G requirements. Enhanced network and device capabilities deliver on future coverage, capacity and quality demands. The extension of wireless networks to new architectures and domains ensures they benefit from innovations taking place across communications ecosystems. Focusing on efficiency – including AI and automation tools required to drive efficiency gains – recognises that new use cases, technologies, networks and services will still need to be rolled out in a cost-effective manner while delivering solid user experiences. The process by which specific technology innovations and vendor R&D translate into standards and future technology generations is not a simple one; standards organisations do the difficult work of building specifications from innovation, balancing user demands with the capabilities of technology. We are currently seeing that as the 3GPP continues work on Release-18 and 5G-Advanced while already beginning early work on Release-19.

ZTE's contribution to the discussion, via its in-depth and technical whitepaper, is both an attempt to frame a way of thinking about B5G technologies and a menu of technical innovations and advancements that could bring those technologies to life. More importantly, it highlights questions we need to be thinking about. Which technologies will enable in-demand capabilities and which need to be prioritised? Where will we see connections between disparate parts of the communications industry? How should operators connect themselves into new revenue sources? To this end, the whitepaper is an important component in planning for what comes next in mobile – what lies beyond today's 5G networks and services.

1.3 Why beyond 5G tech matters now

ZTE's whitepaper may be a valuable contribution to the discussion of what future mobile networks and services will look like. To understand why it is worth our attention we must answer one last question.

Why now?

With 4G and 5G networks continuing to support mobile broadband demands in the here and now, and 6G not expected until 2030 or later, why should we be focusing on beyond 5G in 2022? The fundamental question we need to answer before turning the first page of any B5G whitepaper is why B5G technologies matter today. Luckily, we've already laid out the parts of an answer:

5G is mature. While 5G networks and services have not yet come to every country around the world, we can no longer consider 5G to be in its infancy. As of mid-2022, some 200+ operators had launched commercial 5G services in almost 80 markets. Bolstered by rapidly falling device prices, uptake has been solid, with 1 billion connections forecast by year-end.

Network and service demands are not abating. Many of today's mobile services and use cases require capabilities that networks from a decade ago would not support. Tomorrow's services and use cases will require capabilities that today's networks may struggle to keep up with. Metaverse visions and business models provide a strong example. They have the power to transform societies but will require new capabilities that touch on network performance as well as broader goals in terms of security, sustainability and societal goals such as inclusivity.

6G represents a long-term solution. Recognising that new technology generations arrive with roughly a 10-year cadence, 6G marketing and thought leadership activity began gaining momentum in 2020. Underscoring these efforts was an acknowledgement that developing a new technology generation would take many years and the support of operators, governments and solution suppliers alike.

Operators need long-term and mid-term solutions. Whether or not commercial 6G solutions arrive in 2030 (and operators are beginning to express doubts), evolving market demands mean today's 5G networks will need to evolve if they hope to keep pace. As enterprise digital transformation moves forward, metaverse business models take shape and operators attempt to balance commercial priorities with a focus on sustainability, security and spectrum, mid-term solutions need to arrive sooner rather than later.

Summing up these dynamics is a well-worn quotation: "With great power comes great responsibility".

Of course, the assertion that power and responsibility are linked is not a novel concept. In the context of the 5G era, however, it provides a critical reminder.

The 5G era has demonstrated how mobile broadband technologies have the power to transform business, society and daily life. The expectations of them going forward are therefore bound to be high. They will need to execute on their promise and continue to meet increasingly demanding requirements. Against this backdrop, we need to ask what we want from tomorrow's technologies, how we will put them to use, and which innovations to prioritize. It is the responsibility of the industry to ask, and answer, these questions.

GSMA Intelligence has not written this study or conducted any of the analysis included herein. Regardless, we know that messaging and positioning papers from industry leaders like ZTE are an important part of driving the industry to think about its future and believe you will find it insightful.

02
Enhancement Technologies

2.1 Extremely Large-Scale MIMO

In the future, Massive MIMO will continue to focus on large-scale antenna arrays and evolve into extremely large-scale MIMO. As we know, some emerging IIoT applications have higher requirements for uplink services. For example, the application of machine vision in modern factories, or simultaneous uploading of HD videos by a large number of users in high-density areas such as stadiums. To provide better communications services and value-added services, extremely large-scale MIMO can be applied to significantly improve system performance, including broader network coverage, larger throughput, higher degree of spatial freedom, more connections, and ultra-high sensing and positioning accuracy. Especially for the uplink, extremely large-scale Rx antenna arrays can greatly improve uplink coverage and increase the number of transmission layers and throughput, which is very helpful for meeting uplink requirements. The upcoming WRC-23 will allocate frequency bands between 6 GHz and 10 GHz, which feature short wavelength, high propagation loss, and large usable bandwidth. To meet the requirements for uplink and downlink coverage and throughput performance of the frequency bands higher than 1 GHz, larger-scale antenna arrays need to be considered for higher antenna gain yet lower Tx power consumption and network costs.

Extremely large-scale MIMO can be implemented in a distributed or centralized manner. For distributed MIMO, the multiplication of nodes means larger antenna quantity and a higher ratio of antenna quantity M to UE quantity K. The ratio may reach 10 and even 100. For centralized MIMO, an extremely large scale can be achieved via appropriately enlarged antenna aperture, or more intensive antenna deployment, which will significantly increase the number of antenna elements. In addition, with the maturity of mmWave communications and the rise of the Terahertz technology, the number of antenna elements will increase sharply with the equal aperture, even reaching 100–1000 times of the current value.

Whether distributed or centralized, extremely large-scale MIMO can

always bring ultimate spatial resolution and transmission efficiency. Meanwhile, it enables channel hardening, simplifying system design and saving overheads, such as reducing the complexity in resource allocation and scheduling as well as the pilot density in the frequency domain. However, extremely large scale also brings huge challenges. To fully unleash the potential of extremely large-scale MIMO, very narrow Tx/Rx beams are needed. This leads to high reliance on high-precision real-time CSI, which requires a large quantity of pilot and feedback overheads, the full use of the sparse channels for dimension reduction, and better design of hierarchical measurement and feedback. Both centralized and distributed schemes face the problem of large overheads. Therefore, joint measurement and feedback for multiple nodes also need to be implemented based on channel features to efficiently reduce overheads.

Extremely large-scale MIMO is sensitive to Time Alignment Error (TAE). Nanosecond-level TAE affects the gain of both beamforming and precoding. For a centralized scheme, TAE can be optimized through internal calibration of the gNB to reduce the differences between antennas. While for a distributed scheme, TAE changes quickly due to different environment factors. This problem may be solved via UE-assisted calibration. Even minor remaining TAE will degrade network performance in multi-user multiplexing. Especially in airports, railway stations, stadiums, and other scenarios where a large number of users are close to each other, it is challenging to achieve extremely high throughput.

An extremely large scale also means high costs and power consumption. Currently, antennas usually require a large amount of hardware such as dividers, combiners, and phase shifters. Instead of using the current architecture that will greatly increase costs and hinder commercial use, a novel low-cost architecture needs to be adopted, lower-cost phase shifters also need to be used, and low-cost hardware architectures in RIS may be used in Tx antennas. In addition, electromagnetic lenses can also support extremely large-scale MIMO at low cost, but their performance needs to be further optimized, and they are applicable only to high-frequency scenarios.

An electromagnetic field is a vector field, however, an electric field is regarded as a complex-valued scalar field for simplified design in the existing wireless communications systems. In traditional antenna theory and multi-antenna communications systems, further approximate treatment is done through the definition of far-field region (Fraunhofer zone), radiative near-field region (Fresnel zone), and inductive near-field region. The boundary of near-far field is $2L^2/\lambda$, where L is the antenna length and λ is the wavelength. Both the 4G and 5G systems are designed based on the far-field plane wave propagation assumption, which has been considered as a reasonable approximate assumption for a long time. However, with the evolution of communications technologies as well as the combination of high-frequency and extremely large-scale MIMO, the traditional model based on the far-field electromagnetic propagation assumption is highly likely to fail. When the antenna aperture is doubled, the near-field range will be quadrupled, and when the communications frequency is doubled, the near-field range will be doubled. The Rx end is highly probable to be located in the radiative near-field region, while the near-field region is obviously different from the far-field region.

Major technological changes will therefore be triggered. How to match the channel characteristics of the near-field region needs to be reconsidered in the multi-antenna related designs, including antenna topology, spatial resource allocation, precoding, and channel information feedback. Moreover, some non-communications applications based on extremely large-scale MIMO, such as sensing, positioning, and wireless power transferring, also need to be deployed in the near-field region. It should be noted that distributed extremely large-scale MIMO can also be regarded as a large array of huge-aperture antennas, so UEs are all in the near-field region. At present, the research into near-field region generally only focuses on the restoration of the approximation of distance. But in fact, if the near-field region is regarded as a vector field, its characteristics will be much more complicated than imagined. This is another big challenge in the future, which needs to be tackled via the review and optimization of the existing communications design based on the underlying electromagnetic characteristics.

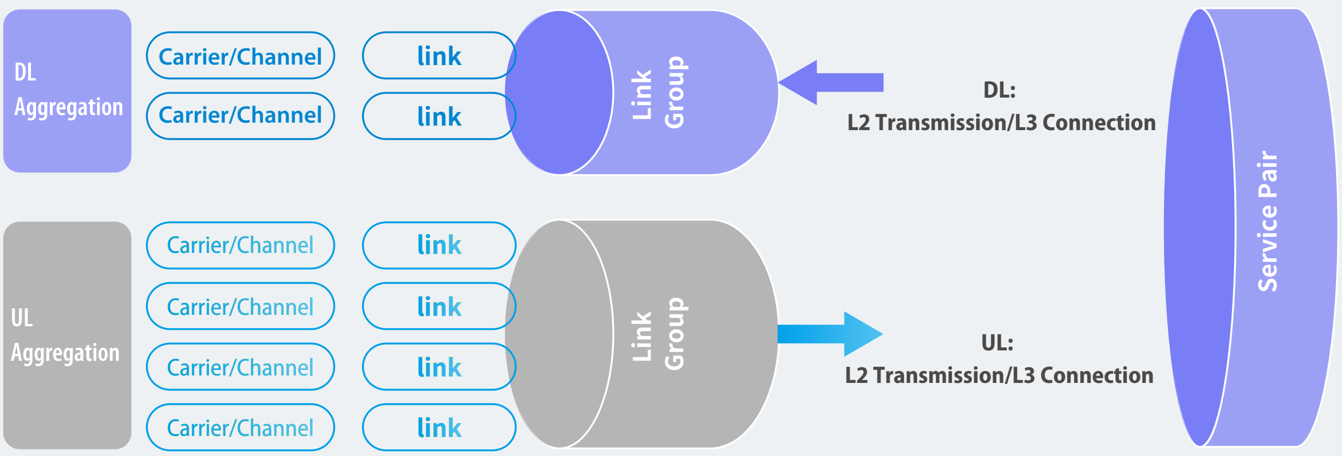
2.2 Full-Band Integrated Networking Based on MetaCell

With the increasingly diversified requirements for 5G and 6G services, wireless communications will use higher frequency bands, for example, terahertz bands. The legacy chimney-type architecture is too tightly coupled to meet the diversified and ever-changing requirements of networks. Against such backdrop, MetaCell is developed. It is a new network architecture that maximizes spectrum value and improves spectral efficiency based on a service-centered spectrum allocation strategy, providing superior quality of experience. With MetaCell, energy saving and low-carbon targets are achieved through higher energy efficiency, O&M is simplified via multi-network integration, and industry-based customization and on-demand orchestration becomes possible for an ecosystem encompassing hybrid B2B and B2C networks.

MetaCell supports uplink/downlink decoupling and service-centered full-band integrated networking. In B2B scenarios, more carriers can be aggregated to improve the performance of high-speed uplink services. With full-band integrated networking, MetaCell also supports inter-band carrier aggregation or on-demand carrier allocation. Meanwhile, via the pooling of uplink and downlink carrier groups, MetaCell enables joint resource management and scheduling of multiple carriers. In this way, it can improve the spectral and energy efficiency of full-band networking, more effectively utilize the low-frequency spectrum bands, especially the scattered refarmed bands, lower the barriers to allocating higher-frequency bands to hotspots, while strengthening system robustness.

For full-band integrated networking, MetaCell pools the System Information (SI) of multiple carriers into one SI Group and broadcasts it through one of the carriers. Obtaining the information of the current frequency band through cross-band common channels can effectively reduce overheads, because, for example, Synchronization Signal Blocks (SSBs) and System Information Blocks (SIBs) are not required for each frequency band. In this way, network O&M is simplified, overheads of configuration, scheduling, and messaging are lowered, while the power consumption of gNBs is reduced.

Regarding UEs, after receiving the SI of multiple carries through one carrier, UEs can flexibly select one or more access channels and carriers in accordance with the service or network strategies. This means that uplink and downlink carriers can be decoupled, since they are not necessarily from the same frequency band.



Full-Band Integrated Networking Supported by MetaCell-Enabled UL/DL Decoupling

2.3 Smooth Virtual Cell (SVC)

Coordinated Multi-Point (CoMP) is to solve the interference among gNBs through avoidance, coordination, and conversion. The current 5G standard already supports some CoMP functions, such as incoherent joint transmission and dynamic cell selection, and there will be continuous enhancement in the upcoming standards.

SVC is essentially an ultimate, user-centric CoMP technology. Compared with traditional geographical location–based cells, SVC has overcome the constraints of physical cells to form a user-specific cell by dynamically adding or deleting some Access Points (APs). Without being affected by the boundary of a physical cell, SVC can ensure highly consistent user experience. At the same time, because appropriate APs are selected as service nodes, the SVC architecture has more freedom in handling interference, thereby improving wireless network performance for users.

SVC aims to provide unified and efficient services for users, and guarantee seamless switch when users are moving, regardless of their location. Compared with the traditional centralized antenna array, SVC consists of a large number of distributed, low-cost APs with a small number of antennas. Distributed APs are connected to one or more CPUs through fronthaul networks. Because APs are distributed in the entire space, the large-scale fading diversity gain is much larger than that of centralized APs. Different from the legacy network-centric coordination modes such as CoMP and Network MIMO, SVC serves each UE with only the nearby APs, eliminating the influence of cell edges through user-centric mode. SVC can be used not only in large-capacity scenarios but also in low-latency and mission-critical scenarios. Served by multiple APs, UEs have lower requirements for storage and computing capacities, because the large number of APs can form a distributed computing architecture for network-intensive computing. In scenarios with a small quantity of service requirements, some APs can be shut down to further reduce power consumption.

The existing distributed AP system adopts the BBU/P-Bridge/pRRU architecture. IF0/IF1 interface is used between pRRU and P-Bridge. The digital signals of multiple pRRUs are combined in a P-Bridge, but this mode is not always suitable for high-frequency bands. Due to large path loss and small coverage of pRRUs, ten times the original number of pRRUs will be needed to cover the same area. If a large quantity of pRRUs need RF synthesis, the uplink sensitivity will be sharply reduced. In addition, the IF0/IF1 fronthaul costs will be greatly increased due to the larger bandwidth of high-frequency bands. The IF3 interface is an alternative, which reduces the fronthaul pressure and the impact on sensitivity brought by the uplink RF synthesis of multiple pRRUs.

Compared with the centralized antenna schemes, SVC can significantly improve the spectral and energy efficiency by balancing the centralized and distributed processing, especially in the indoor and hotspot scenarios. On the other hand, further research into SVC is required, such as initial access, power control, codec distribution, resource allocation, channel model, downlink channel estimation, coexistence with existing standards, and prototype design.

SVC involves key technologies such as coordination sets selection, power control, user scheduling, and beam training, which can be modeled as an optimization problem. Optimization can be divided into continuous optimization and combinatorial optimization in

accordance with whether the variables are continuous or discrete.

The selection of coordination sets, user scheduling, and beam training are all combinatorial optimization where the best or preferable combinations are selected from the limited sets. Power control determines power parameters based on the known coordination sets and scheduled users, and it is continuous optimization.

For combinatorial optimization, to reduce the difficulty in balancing the complexity and performance, a feasible engineering solution is to add some constraints or apply special policies to restrict the number of combinations, and to obtain preferable results instead of best results after comparison.

For continuous optimization, the concave and convex of functions need to be considered. If it is a convex function, the local optimal value is the best value; if it is a non-convex function, a new penalty item can be introduced to convert the original function into a convex function.

Power control in SVC is a typical non-convex problem. In the existing research papers, it is sometimes converted into a convex optimization problem. The relationship among large-scale parameters and power control parameters is also studied. Heuristic algorithms and Deep Neural Network (DNN) are applied to function solving.

In summary, heuristic algorithms can be applied when the best results are not the target; the data-driven AI solutions can be applied when the model is particularly complicated or the solution domain is extremely large. In addition, the exploration of the association among the corresponding feature parameters and the target functions can further simplify the model. SVC requires high-precision synchronization and correction among cells, which set high requirements on engineering.

SVC is applicable to indoor and hotspot scenarios, while centralized antennas are recommended in outdoor scenarios. With SVC, coverage is expanded, requiring more distributed RF channels supported by the baseband.

2.4 Full Duplex (FD) and Sub-Band Full Duplex (SBFD)

FD is a new technology to improve the network throughput and spectral efficiency. To meet the requirements of large bandwidth and low latency in the future, FD will be used in unpaired spectrum. By overcoming the mutual exclusion of downlink and uplink on resource use, FD can enhance resource utilization and reduce downlink and uplink transmission latency.

On the other hand, it is challenging to implement FD. gNBs and UEs must be able to handle self-interference, sub-band interference, and adjacent-frequency interference, to support simultaneous Tx and Rx. In the early stage of 5G-Advacned, the focus will be placed on SBFD, and then gradually expanded to FD networks applicable to gNBs and UEs.

FD must support the cancellation of network interference and self-interference. The elimination of self-interference includes the following three steps:

- **Step 1 Propagation Domain:**
  Antenna isolation and circulator-based offsetting

- **Step 2 RF Domain:**
  Creation of signals with phase opposite to that of self-interference, adjustment to latency and attenuation, and combination of cancellation signals and self-interference signals.
  Overall cancellation requirement in the propagation domain and RF domain: Transmit output power Pout – (adaptive gain control threshold – peak-to-average ratio)

- **Step 3 Digital Domain:**
  Cancellation of linear and non-linear interference, multi-path cancellation, and digital cancellation: (adaptive gain control threshold – peak-to-average ratio) – receiver noise

For SBFD and FD, the cancellation requirements are the same in the analog and propagation domains. However, the situation for FD is more complicated because it cannot provide frequency domain isolation as SBFD.

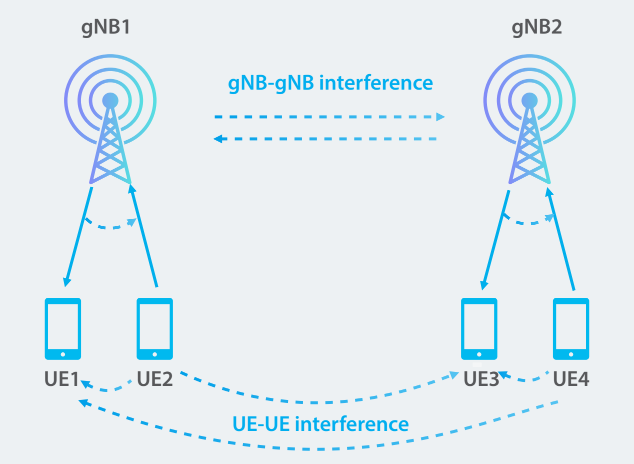
It is feasible to support FD in single-station scenario, but Tx-Rx isolation is required, which means doubled number of antennas, larger antenna size, and increased deployment difficulty. In addition, there are other prerequisites. For example, the output power per channel of the gNB cannot be too high (now 37dBm), ACLR cannot be too high (now 45dBc@20 MHz), LNA is not saturated, and NF does not deteriorate. Signal cancellation includes physical isolation of antennas (such as 50 dB), analog cancellation, and digital cancellation. The indicators of each part are added and coordinated for overall cancellation.

It is more difficult to implement analog cancellation for MIMO models: Each Tx channel will leak signals to each Rx channel, therefore each Tx channel needs to be coupled to a separate cancellation module, and then combined to a Rx module for corresponding cancellation. The number of modules used for cancellation is n^2, so MIMO models have higher requirements for antenna isolation. If the Tx and Rx antennas are separated, the channels are asymmetric, increasing the difficulty in channel estimation.

The above is the analysis of self-interference cancellation, that is, signals sent by Tx are received by Rx channels. For self-interference cancellation, because Tx signals are known, the difficulty of cancellation is relatively low. According to some research, relatively strong cancellation capabilities have been achieved. However, in addition to self-interference, Cross-Link Interference (CLI) is also a problem for FD or flexible duplex. This problem is easy to be overlooked, but it has a great impact on networking, and may restrict the applicable scenarios and scope of FD.

CLI usually includes gNB-UE, inter-gNB, and inter-UE interference, as shown in the figure below:

1. Inter–sub-band CLI of gNB or UE
2. Intra–sub-band CLI of gNB or UE
3. CLI between adjacent bands of gNB or UE



gNB-gNB/UE-UE Interference

3GPP Rel-16 introduces Layer-3 CLI measurement and feedback for inter-UE CLI, including SRS-RSRP and CLI-RSSI. The gNB can instruct a UE to measure and feedback inter-UE CLI for coordination and scheduling. In addition, via the exchange of TDD configuration and pilot signal information, inter-gNB CLI can be measured for coordination and scheduling. For SBFD and FD, further study is required for 5GB on the interference cancellation/coordination solutions for the above types of interference.

Intra-frequency FD, widely studied by academics, can greatly improve spectral efficiency in theory. SBFD can allocate downlink and uplink resources in the same frequency band, which can flexibly configure more uplink resources, and thus reduce uplink and downlink latency, and improve uplink coverage and capacity.

However, it is challenging to realize the above potential. In some researches, frequency-domain or space-domain methods are used to deal with interference. The spectral efficiency will be increased if space-domain resources are used to eliminate CLI, but in fact, the space-domain resources are not directly used to increase capacity and the actual network capacity decreases in many cases. Therefore, the FD technology is applicable to scenarios where gNBs are deployed sparsely.

SBFD and FD have slightly different requirements for digital cancellation, but both have high requirements for interference cancellation in the analog and propagation domains.

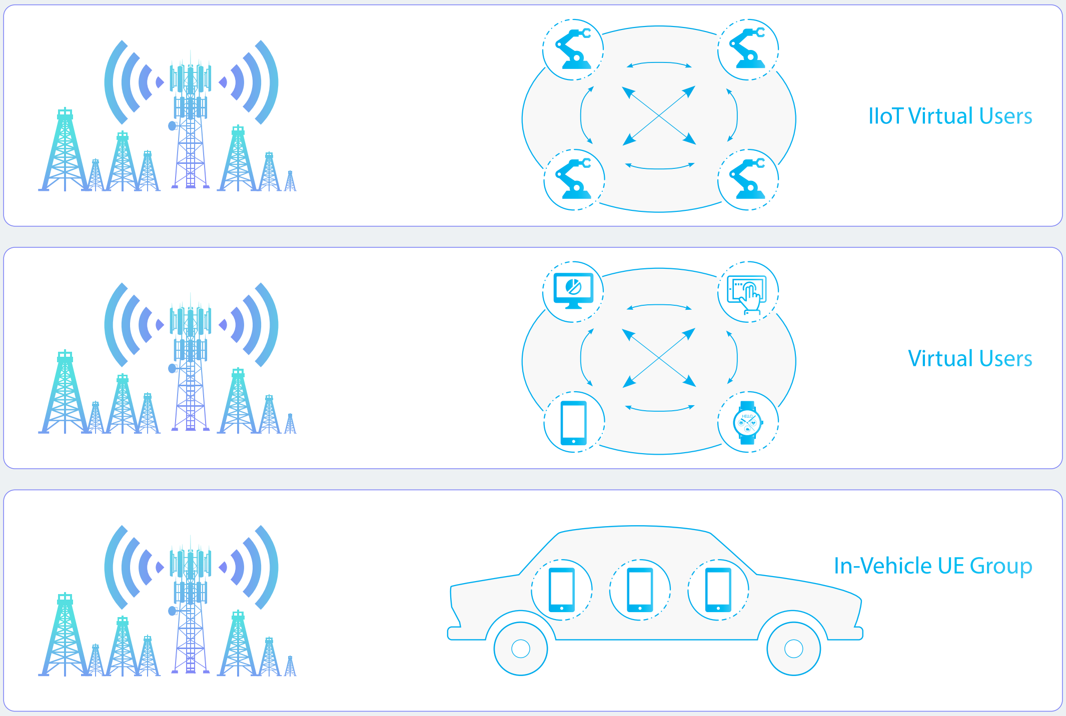
FD implementation is complicated and the corresponding hardware cost is high. Even if only self-interference is taken into consideration, the implementation difficulty is high for Massive MIMO models. In reality, the multi-antenna technology is mutually exclusive with FD, which is also one of the important limitations of the FD technology. To improve spectral efficiency, FD is a choice, which may double the spectral efficiency, but at a cost of higher complexity and less antennas. Another choice is TDD with more antennas to increase spectral efficiency several or even tens of times, and eliminate interference naturally. Therefore, FD is not necessarily the optimal solution to increase spectral efficiency. FD shows advantages only in networks with sparsely deployed gNBs, low interference, and low latency.

In conclusion, FD, especially SBFD, has advantages and limitations. Most researches start from models with fewer antennas and SBFD, evaluate the actual gain, and then apply the FD technology in stages. Although SBFD reduces the requirements for internal interference cancellation of gNBs, the inter-UE interference remains serious and needs to be solved by efforts from the industry.

2.5 Multi-UE Virtual Aggregation

Some emerging IIoT applications of 5G NR have higher requirements for uplink services. For example, machine vision applications in modern factories, and simultaneous uploading of HD videos by a large number of users in high-density areas such as stadiums. In such scenarios, the requirements for uplink capacity by cells may be much higher than those for downlink capacity. In extreme cases, there may be a Gbps-level or 10 Gbps–level throughput, and the requirements for latency is high, which are difficult to satisfy in the current NR design.

One solution is to enhance the existing NR uplink directly, for example, more antennas or MIMO layers, MU-MIMO that supports more users, and more flexible carrier allocation and aggregation. These enhancements require that the device initiating uplink transmission must be powerful enough. However, the capability and Tx power of a single UE are limited. In this case, the idea of cloud computing may be applied, namely, to improve the UE capability through user virtualization and collaboration. Multiple adjacent devices can form a temporary group of virtually aggregated UEs, and the devices can implement virtual MIMO or carrier aggregation through shared antennas, Tx power, and carriers to improve transmission quality. In addition, to better support FR2 dense networks, a UE is designed to support simultaneous multi-panel uplink transmission. If a UE supports FD and simultaneous multi-panel Tx and Rx, it can communicate with multiple TRP nodes at the same time, effectively improving the transmission performance and robustness of networks. We believe that these technologies can effectively support uplink enhancement, and help apply 5G-Advanced in more scenarios with high requirements for uplink capacity.



Multi-UE Virtual Aggregation

2.6 Reconfigurable Intelligent metaSurface (RIS)

With reconfigurable/reprogrammable wireless channel at its core, RIS reveals new horizons for future wireless networks, with its advantages of low costs, high energy efficiency, high reliability, and large capacity.

RIS improves the coverage, throughput, and energy efficiency for the users at cell edges, primarily through the following ways:

1. Providing a reflection propagation path for Non-Line-of-Sight (NLoS) to avoid coverage holes;
2. Beamforming for target users, to fully utilize space diversity and multiplexing gains;
3. Null-forming for interfering users to suppress inter-cell interference.

RIS can also be used in sensing and positioning, which can reduce electromagnetic radiation. In the long term, RIS can be used to build a new intelligent network paradigm. For future networks, both the design of Tx and Rx ends as well as the optimization of the wireless environment in systems design should be taken into consideration, so that networks can be more intelligent and flexible. Traditional networks are designed and upgraded to adapt to the constraints of wireless channels, for example, using beamforming to counter transmission loss and interference, and applying multiple modulation coding and the MIMO technology to increase capacity. However, due to the ever-increasing demands for wireless communications and the complexity of networks, the increasing network construction costs (including power consumption costs) has become a crucial issue for future wireless communications. Although RIS opens new horizons, the deeper research into RIS has showed that many technological difficulties still need to be tackled for its commercial use, especially for large-scale deployment. Thus, more research and development are necessary to gain a profound understanding of the practical application scenarios and commercial use of RIS.

RIS： Adaptively Reconfigure Radio Channel

2.6.1 RIS Architecture and Hardware Implementation

In principle, metasurfaces are artificial 2D materials (for example, planar surfaces) that can manipulate incident electromagnetic waves, including ideal absorption and targeted reflection. Compared to traditional metasurfaces that achieve only fixed functions, RIS is configurable (or programmable) in real time due to its electromagnetic features. RIS contains a lot of units (usually at sub-wavelength level). By controlling and modifying the setting of an individual unit, RIS is able to dynamically change the electromagnetic features of each cell, thereby controlling the amplitude, phase, polarization, and frequency of the electromagnetic waves.

Current research on hardware implementation includes the following fields.

PIN Diode / Varactor Diode

In this implementation method, the manipulation of electromagnetic waves is mainly achieved through adjustment of capacitance or inductance of the equivalent circuit. For example, adjustment of multiple phases can be achieved through more diodes or multiple stages of bias voltage. Additionally, load impedance can be adjusted to change the amplitude of the electromagnetic waves. This type of RIS responds to electromagnetic signals in an extremely short time, with low reflection loss and power consumption. It is also a very mature technology, with a relatively low cost.

Micro-Electromechanical System (MEMS)

MEMS is used to control the electromagnetic units, for example, adjusting the signal transmission latency, rotating electromagnetic units, and tuning capacitance through piezoelectric effect, thereby manipulating the electromagnetic waves. This type of RIS has good linearity, and enables low power consumption and high integration.

Liquid Crystal

Through the change of parameters, nematic liquid crystal can display birefringence under changing offset voltage, thus achieving electromagnetic wave absorption. Yttrium Iron Garnet can be introduced into the substrate to adjust the permeability of the substrate via change of the magnetic field. Graphene is considered another material for RIS. The introduction of metal can reduce the equivalent impedance of a surface, and thus improve its capability to manipulate the amplitude, phase, and frequency of the electromagnetic waves.

2.6.2 RIS Algorithm

The acquisition of RIS channel information is critical for beamforming, secure transmission, passive information transmission, as well as sensing and positioning. Passive RIS contains a large number of units with a gap of less than half the wavelength between each other, and the elements of units are passive and not capable of processing complex signals, making it difficult to acquire segmented channel information. To effectively perform channel estimation, it is necessary to reduce the complexity of channel estimation and to avoid complicated signal processing operations. A feasible solution is to replace passive units with active units that have sensing and signal processing capabilities. The number of active units and arrangement patterns will be determined based on the estimation of segmented channels and transmission performance. The position-based RIS channel estimation reduces complexity by obtaining the angle of arrival and other key information via the utilization of such characteristics as the fixed positions of gNBs and RIS as well as the RIS arrays. Based on the sparsity of high-frequency channels and the programmability of RIS, methods such as compressed sensing and matrix completion can be applied to RIS channel estimation.

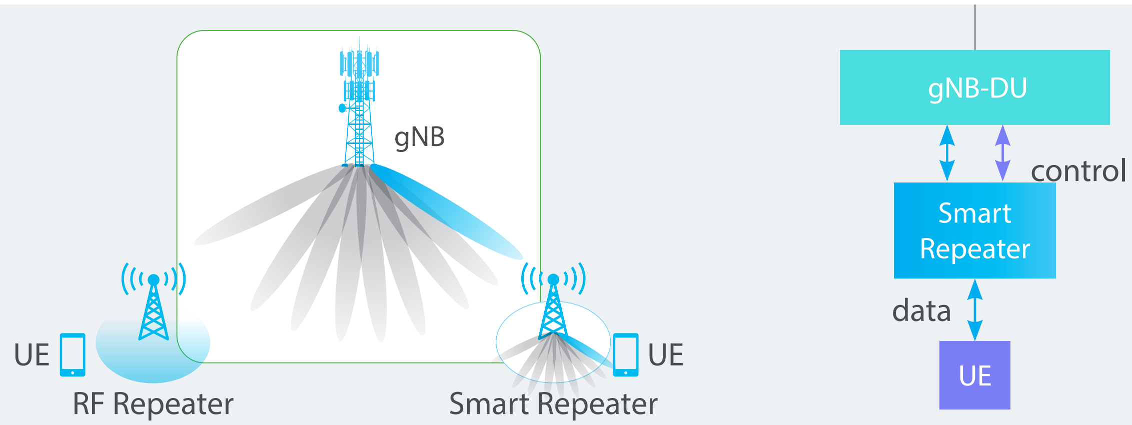
The cooperative solution of channel estimation and beam matching coordination can be used for RIS channel estimation. Segmented transmission will impose challenges on traditional codebooks. Thus, it is urgent to design codebooks for gNB-RIS channels and RIS-UE channels as well as the corresponding codebook matching methods, so that random access design and beam selection can be achieved. The solution for downlink channel estimation should take into account the constraint of limited channel feedback, and comprise between the accuracy of channel estimation and non-ideal channel feedback. To effectively solve problems about channel estimation when the channel model is unknown, AI may be introduced to reduce the dependency on channel modeling while maintaining high robustness and noise immunity.

Currently there are many studies on RIS algorithms. Although the channel estimation process has been simplified in many algorithms, channel sensing is still needed, at least partially. This is further aggravated by the fact that RIS has a very large number of units, which far exceeds that of Massive MIMO. Even with sparsification, it will still greatly increase hardware complexity, which has to be tackled in future research.

2.6.3 Smart Repeater / Network-Controlled Repeater (NCR)

A smart repeater (also known as NCR), similar to RIS, aims to improve the coverage of high-frequency networks by adding available antennas and by controlling the transmission of electromagnetic waves. Smart repeaters mainly use power amplifiers to increase coverage and reduce coverage holes, thereby increasing channel ranks.

Traditional RF repeaters only support amplify-and-forward, without considering TDD duplex, beamforming, and scheduling-based switch. They amplify all signals in all directions during uplink and downlink time slots, which causes self-interference between mutually coupled Tx and Rx units, and amplifies noise and interference. Thus, repeater gain and coverage improvement are limited.



Differences Between RF Repeater and Smart Repeater

Smart repeaters/network-controlled repeaters will be introduced for B5G. Based on the architecture of traditional RF repeaters, repeater control functions will be increased. The potential control functions are as follows.

Beam Information

With beam information available, NCR can switch the beams to UEs based on the UE status, by taking the control instruction from gNBs.

Timing and TDD Pattern

Based on the TDD pattern, NCR can selectively receive and transmit signals in certain directions.

On/Off:

gNBs can switch NCRs on/off to save power and reduce interference.

Power Control:

gNBs can control and change the transmit power of NCRs to reduce interference.

Assumed Scenarios

- NCRs in Rel-18 is transparent to UEs, namely, they do not affect UE design and support Rel-15/16/17 UEs, which will increase NCR compatibility. However, this curtails the beam sweeping capability and it is difficult to support beam management between gNBs and RIS.
- Key scenarios include FR2 with TDD, and outdoor or outdoor-to-indoor scenarios
- Only one-level NCR is considered, namely, neither multi-level (cascading) NCR nor mobile NCR is considered.

2.6.4 Application and Commercial Deployment of RIS

RIS is essentially a distributed space diversity technology. With multiple antennas, it improves the wireless network performance by better manipulating the space electromagnetic field. Also, SVC, smart repeater, and relay are all distributed space diversity technologies. By contrast, Massive MIMO is a centralized space diversity technology.

In terms of space resource utilization, Massive MIMO is a digital method, and also the optimal method; SVC is also digital and considered second to Massive MIMO under strict synchronization; NCR is analog and second to SVC; RIS is analog and has the least efficient utilization of space resources. However, the above order is just the reverse in terms of hardware cost. RIS, due to its low cost, is easier for large-scale deployment.

The industry already has a full understanding of the advantages of RIS. With the research going deeper, the application scenarios and commercial deployment of RIS need to be evaluated in a more objective manner. Currently, it is generally accepted that, in initial stages, RIS may be used for high-frequency applications and to reduce coverage holes. The corresponding technical challenges such as trade-offs between coverage and capacity gain as well as multi-user interference are to be overcome by the joint efforts from both the industry and academics.

2.7 Enhanced Multi-User Shared Address (eMUSA)

Some emerging IoT use cases show higher requirements for networks. For example, event-driven sensor monitoring networks not only have high requirements for the number of connections, but also require efficient data reporting. In a smart grid, some advanced intelligent charging devices are densely connected, so the corresponding networks must have wide coverage, high data rate, low latency, and high reliability. The application scenarios include the following two categories.

Application scenario 1: massive UEs (mMTC or ultra mMTC)

The number of UEs is very large, and vertical coverage needs to be considered.

There may be requirement bursts, but the requirement for transmission latency is not high. The UEs require low data rate, low costs, and low power consumption.

Application scenario 2: critical MTC

In the future, the critical MTC scenario will be a compromise between mMTC and uRLLC. The number of UEs and the requirements for transmission latency and reliability are between those of mMTC and uRLLC. Although with the uRLLC technology, ultra-high reliability and low latency become possible for a small number of UEs, further research is needed to work out a solution for efficient support in the critical MTC scenario where the number of UEs is huge. A typical use case of this scenario is Enhanced Vehicle to Everything (eV2X).

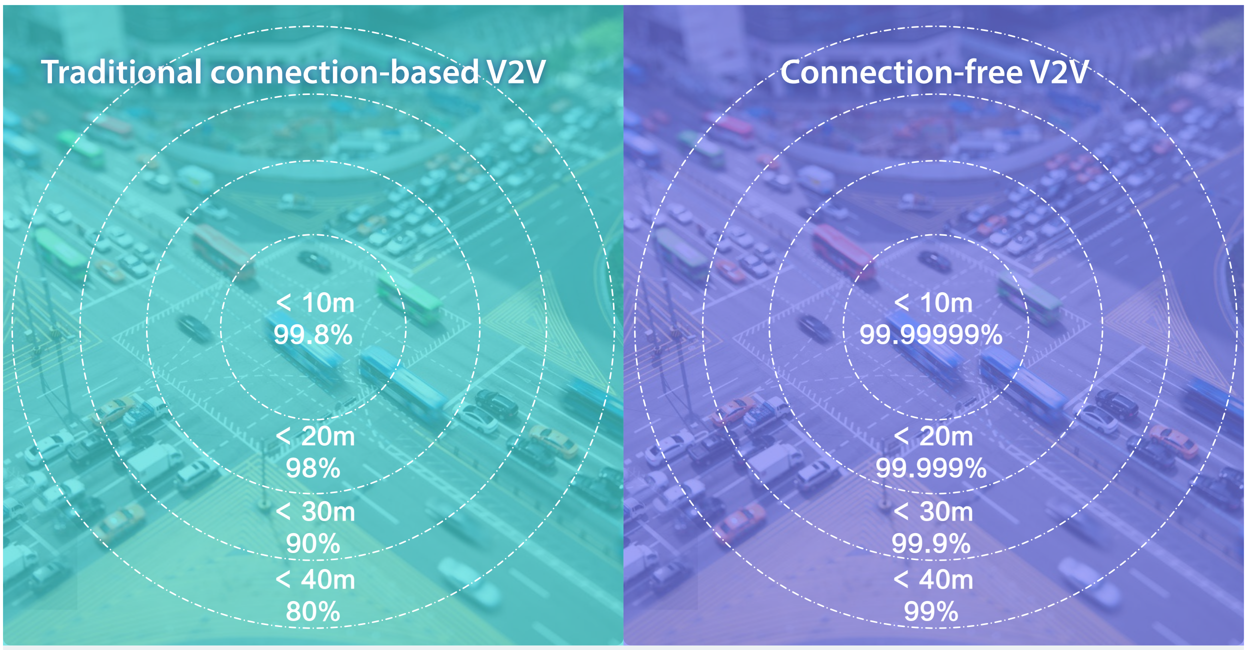
Based on the NR-based massive IoT, access modes and data transmission is simplified, such as 2-step RACH and small data packet transmission in inactive state. Furthermore, in the B5G era, the non-orthogonal connectionless transmission technology is very promising. Connectionless means that UEs can stay in idle/deep sleep status for most of the time to minimize system loads and power consumption. Traditional dynamic grant-based transmission requires UEs to enter connected state in advance, and then apply for transmission resources or Grant to transmit data. Such procedure will significantly increase the overheads on signaling, power consumption, latency, etc. In addition, a large number of nodes mean that collision and blocking are highly likely to occur during random access and application for transmission resources, which will inevitably reduce the transmission efficiency.

In the hundreds of billions of IoT scenarios, if only the one-dimensional power domain (near-far effect) is used, efficient transmission is hard to be achieved in connectionless contention-based grant-free transmission. The code domain and spatial domain need to be exploited efficiently to support multi-user transmission.

In the connectionless contention-based grant-free transmission in the code domain, long sequences, such as long PN sequences, have a certain overload rate. When the overload is very high, the complexity of Successive Interference Cancellation (SIC) will be exposed. Further, large time-frequency spreading will increase the transmission complexity of UEs. For short sequences, if they are optimized so that the overload rate can approximate to that of long sequences, they will have an advantage in reducing complexity. When the traditional PN sequence is shortened, the user overload rate will decrease significantly, because the traditional PN sequence is a binary real sequence. After the sequence is shortened, the low correlation of sequence sets randomly generated cannot be guaranteed. In some NOMA solutions, the contention-based NOMA performance of the code domain is improved through the optimization of the symbol spreading sequence. The eMUSA technology uses multi-level code sequences in the complex number field. Such sequences can keep relatively low cross-correlation even if they are short, for example, 8 or even 4 complex numbers. By virtue of the excellent cross-correlation of the eMUSA sequences and the efficient SIC-based multi-user detection receivers, the eMUSA can support high overload.

On the other hand, in the traditional MU-MIMO system, the gNB can estimate the spatial domain channels of different users through the orthogonal reference signals and then separate them through the spatial domain. However, in connectionless transmission scenarios, Reference Signals (RS) are also user-defined, that is, RS collision may occur. High overload means high probability of collision. Once the RS collides, the gNB cannot separate the users through the spatial domain. To reduce the collision and pollution of reference signals, and estimate wireless channels, Time/Frequency Offset (TO/FO), the overhead of reference signals is multiplied, and the detection complexity of RS is increased significantly.

It is worth mentioning that connectionless transmission can also be used together with other technologies in a wide range of scenarios. For example, the high-efficiency connectionless transmission technology can be used with the FD technology, allowing a large number of vehicle UEs to directly exchange information without listening first. In addition, it avoids the problems of missed reception and hidden nodes of traditional technologies, and finally achieves direct V2V communications with ultra-low latency and ultra-high reliability. The comparison results are shown in the figure below. In the high-density vehicle scenario, the transmission latency of this method is only 1/5–1/10 of that of traditional methods, equaling an increase of reliability by 1–3 magnitudes.



Reliability Comparison Between Traditional Connection-based V2V and Connection-free V2V

2.8 Passive IoT

In the future, there will be more diversified IoT application scenarios. Devices in many of the applications are expected to feature lower power consumption, smaller size, simpler structure, and lower cost, so it may be difficult or even impossible to integrate conventional batteries into these devices. Services of some applications will also be simplified. For example, a device only needs to respond to query with a variable frequency. To meet ever-changing requirements, passive IoT is introduced and becomes a great trend in IoT evolution. Compared with traditional IoT devices with battery, passive IoT devices can be battery-less and maintenance-free with a long life span. Additionally, the number of devices connected may be tremendous and their power consumption is very low. Passive IoT is therefore especially applicable to the scenario where there are many devices that are hard to be managed but only a small amount of data needs to be collected.

> Different from traditional IoT devices, passive IoT devices need to support key technologies such as energy harvesting and backscatter communications.

The energy harvesting technology converts received RF signals into electric energy. Our major focus are device nodes that can harvest energy from RF signals in the environment and send and receive data as well as information. Such a device can be called energy storage device, which includes the following modules:

- Energy harvester
- Low-power RF transceiver
- Power management module

Main problems in energy harvesting and storage for rechargeable batteries are as follows:

- Efficiency of converting RF energy to DC remains low.
- Energy harvesting depends heavily on the distance between the device and the RF source.

Backscatter technologies enable special energy collection and processing, that is, energy is used once it is collected. Therefore, IoT devices that support the backscatter technology can be truly passive. The device nodes supporting the backscatter technology (also called backscatter devices) do not need to store energy. The backscatter transmitter directly reflects the received RF signals by adjusting the antenna impedance, modulates the signals, and sends the signals to the backscatter receiver. There are three types of communications systems that support the backscatter technologies:

Monostatic Backscatter Communications System (MBCS)

There are only two main components: A reader and a backscatter transmitter. The reader can be considered as an integration of an RF source and a backscatter receiver. The typical application of this type is the RFID system.

Bistatic Backscatter Communications System (BBCS)

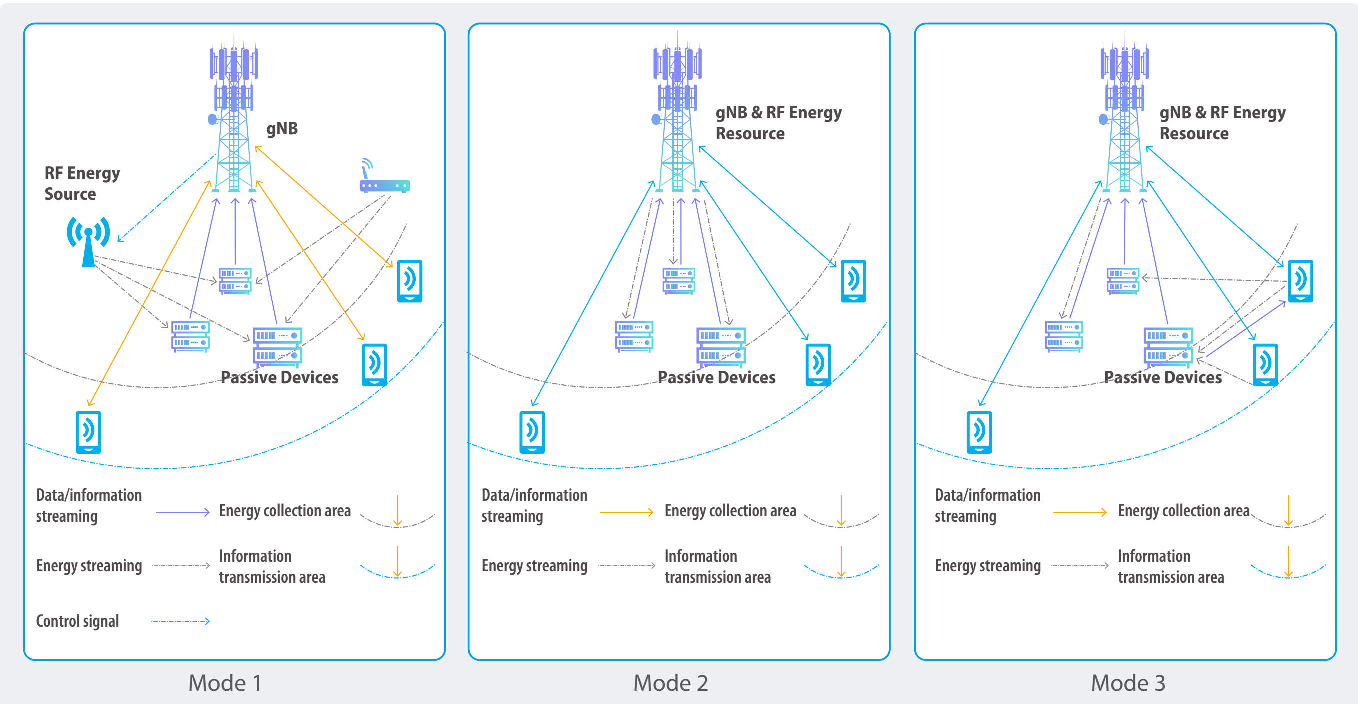
The RF source, i.e., carrier emitter, and the backscatter receiver are separated. Therefore, a BBCS can avoid the round-trip path loss in the MBCS. In addition, the performance of the BBCS can be further improved by placing the carrier emitters at optimal position.

Ambient Backscatter Communications System (ABCS)

Unlike the BBCS, the RF source in the ABCS is not dedicated. It can be various RF signal sources in the environment, such as a TV tower, cellular gNB, or Wi-Fi access point. Therefore, there is no need to deploy and maintain dedicated RF sources in the ABCS system, thus reducing the overall cost and power consumption. Furthermore, by using the existing RF signals, it is unnecessary to allocate new spectrum for ABCS, which can improve the utilization of spectrum resources.

Although passive IoT shows advantages in power consumption and device simplicity, it also has disadvantages in terms of coverage, transmission performance, multiple access, and mobility, in which traditional cellular networks have advantages. Therefore, the combination of traditional cellular networks and passive IoT devices has become a research focus in the industry.

Based on the deployment locations of RF energy source nodes, cellular network-based passive IoT can be deployed in several modes as shown in the following figure.



Mode 1:

The RF energy source node is deployed as an independent network element, which can provide RF energy for passive IoT devices so that they can backscatter the energy and transmit data or information to gNBs. The RF energy sources can be further divided into the following:

It can be a dedicated RF energy source, which can be controlled by the cellular network. For example, the operation frequency band and transmit power can be set.

It can also be a node in another non-cellular network, such as a TV and broadcast tower and WiFi access point.

Mode 2:

An RF energy resource node can be merged with a cellular network node (such as deployment on a gNB or relay node).

Mode 3:

When the RF energy source node is integrated with common UEs or passive IoT devices, the passive IoT devices can obtain energy from closer devices, thus improving energy collection efficiency. Through the multi-hop operation, the passive IoT devices can send data to adjacent UEs, and also help expand the communications coverage.

The combination of traditional cellular networks and passive IoT devices also poses certain challenges on standard protocols and system implementation. The potential challenges and major research directions for the physical layer, high layer, and core network are listed as follows:

Technical challenges for the physical layer:

Uplink: coding and modulation, multiple access, and backscatter channel design.

Downlink: design of downlink energy transmission signal, and mechanism for simultaneous transmission of energy and data.

Others: positioning method design and coexistence with existing wireless systems.

Technical challenges for the high layer:

Protocol stack (protocol stack integration and simplified protocol stack functions)

Access and scheduling (efficient access mechanism and resource scheduling optimization)

Connection control (simplified connection control and mobility redefinition)

Security (identification of new security requirements and air interface security)

Technical challenges for core network:

Management of massive passive devices, device registration, authentication, optimized transport network management and QoS guarantee, and mobility management and optimization.

03

Extension Technologies

With the explosive demands from various industries, 5G application will be further extended. In the first phase of 5G, typical scenarios such as eMBB, uRLLC, and mMTC defined by 3GPP have been realized. During the B5G era, new scenarios and new demands from industries raise higher requirements for networks. In particular, the extension technologies, which are converged with the various existing technologies horizontally and vertically, are required, for example, the deeply converged information infrastructure, Integrated Sensing and Communication (ISAC), space-air-ground integrated network, AI native, and security native.

3.1 Deeply Integrated New Information Infrastructure

In the future, ubiquitous connectivity will give rise to the combination of multiple scenarios, and the operators' communications network will be greatly supplemented with the B2B network, Internet of Things, Internet of Vehicles, etc. The network models will be more complicated, bringing more challenges to network planning and implementation. Generally, the model for next-generation networks is predicted based on the model for the previous generation. However, such method will hardly be applicable to new networks. Multiple types of networks will be combined to be a complex network or a sub-network in another network. To cope with these uncertain changes, cloudification and pooling are still the best solutions. Unified orchestration, scheduling, and coordination of the terminal, network, and cloud is a must for the evolution of the next-generation networks. That is, the computing power needs to be integrated, and the terminal, network, and cloud need to be converged deeply. It should be noted that the "network" here is not only a transport network, but also a wireless gNB.

3.1.1 New Architecture

Then, how should we define the new architecture to deal with the above challenges? Operators have proposed various solutions, expecting to build a new information infrastructure that integrates multiple elements such as AI, Blockchain, Cloud, Data, Network, Edge, Terminal, and Security (ABCDNETS). They aim to promote the computing power to become a "Click and Use" social service that is easy to use and available anywhere anytime, similar to water and electricity supply. As a result, we propose the in-depth convergence of terminal, network, and cloud, and the unified orchestration for computing power. Specifically, the focuses are as follows in terms of infrastructure, unified cloud platform technology stack, and unified orchestration management:

3.1.1.1 Building a Heterogeneous Hardware Pool by Layer and Scenario

For central scenarios, the computing power is supplied on demand in a centralized manner, with large-scale centralized O&M, which is low-cost and low-carbon. For edge scenarios, the performance and integration of data edge offloading are equally important, and edge security and plug-and-play deployment are also required. For access scenarios, the data is processed in advance, with lower requirement for performance but higher priority for integration. In addition, the differences in hardware are shielded in the virtual layer, so that consistent resources and capability services are provided for the application layer.

High integration and low power consumption:

- **Prefabricated full modules in IDC:** short construction period and low PUE.
- **Edge server:** short chassis and integration of the computing and network.
- **All-in-one cloud-network cabinet or hyper-converged equipment:** pre-installation before delivery, fast-deployment in the edge equipment room.
- **Built-in board:** integration with the IT BBU/OLT access devices.

High performance and strong security:

- **GPU/DPU:** meets the special application requirements such as image processing.
- **Smart NIC:** improves network forwarding performance.
- **Fully uninstalled smart cloud cards:** fully improves system performance, security, and integration.
- **Dedicated wireless hardware:** dedicated wireless hardware for high real-time processing of services by wireless gNBs.

3.1.1.2 Technology Stacks for the Unified Cloud Platform

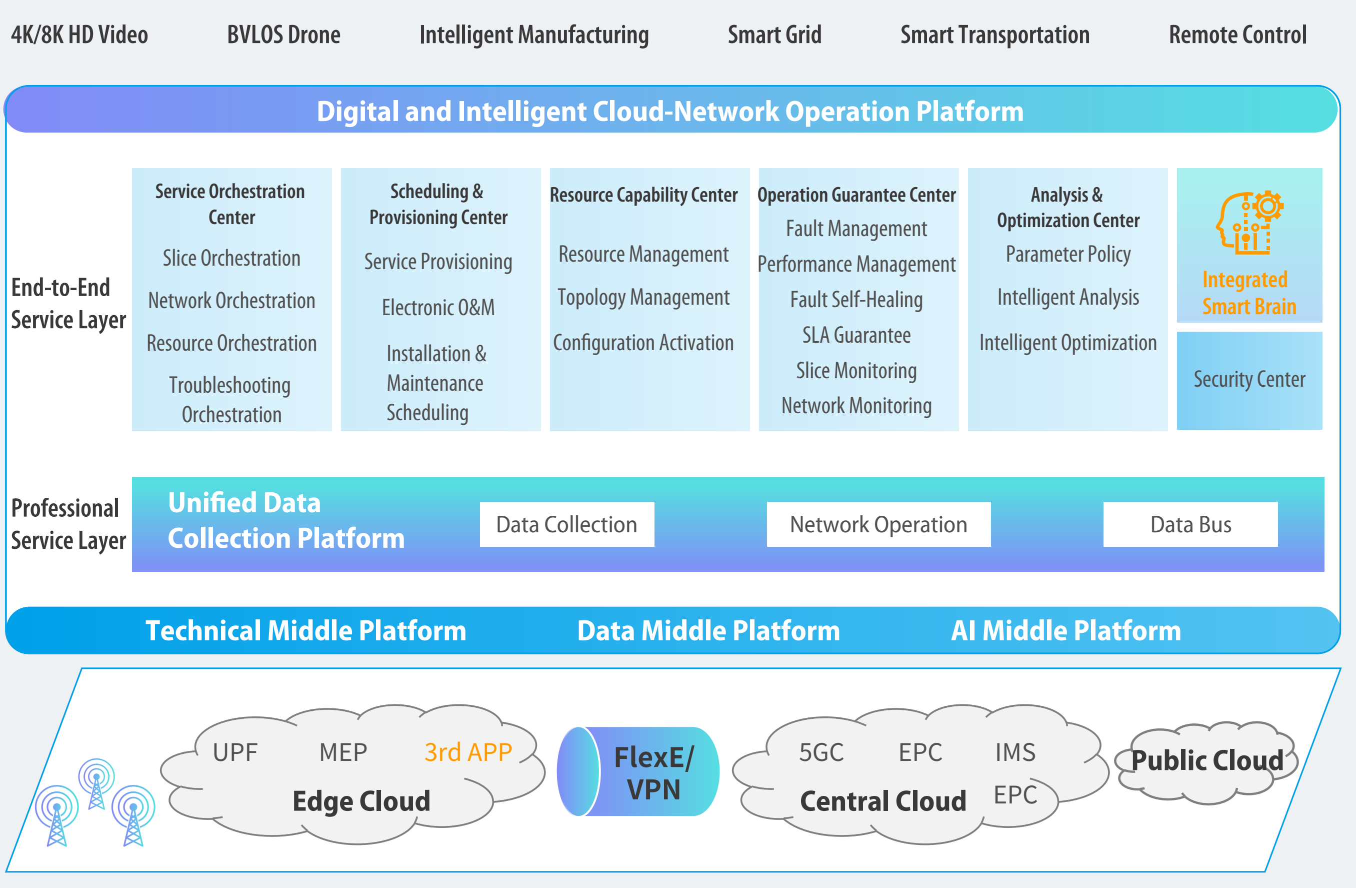
- **IaaS layer:** Enhanced infrastructure, including the dual-core engine of VM containers, lightweight solution with hardware acceleration technology and security native.
- **PaaS layer:** Platform capabilities that can be tailored as required, including the service capabilities for video, IoT, vertical industries, enterprises, and telecom operators. The big data middle platform provides the capabilities of big data and AI, and these capabilities are open to external parties.
- **SaaS layer:** Agile development environment customized on demand, and applicable to telecom and industry customers.
- **Full-scenario deployment capabilities:** Support large-scale deployment at central cloud for multi-cluster smooth capacity expansion, and component tailored on demand and hyper-integration, supporting lightweight deployment at edge cloud.
- **Precise match to next-generation applications:** In-depth convergence of OpenStack/K8S, flexible supply of VMs/containers/bare metals, strong middle platform capabilities to enable agile development of applications; security native, which ensures the security and reliability of systems; automatic O&M/intelligent O&M to improve platform management efficiency.

3.1.1.3 Orchestration Management

The Digital and Intelligent Cloud-Network Operation Platform enables the collaborative orchestration of the cloud, edge, and terminal, and provides open APIs. On this basis, the higher-level unified orchestration center of the converged cloud and network can be developed to support the unified orchestration and coordination of the cloud, network, edge, and terminal.

Service orchestration:
Unified collection and control center, supporting full data resource collection crossing professional networks and unified management of O&M data; full-scenario, slicing orchestration, and edge capability orchestration; interconnection of business, production, maintenance, and operation data.

Resource orchestration:
Unified cross-domain orchestration, supporting unified life cycle management of heterogeneous resources, end-to-end slicing orchestration, service orchestration of cloud & network, multi-cloud collaborative orchestration, and cross-cloud migration; middle-platform architecture, enabling flexible orchestration of underlying O&M capability.



Digital and Intelligent Cloud-Network Operation Platform

3.1.2 Wireless Convergence

According to the analysis and practice by the industry, if a gNB has the micro-service and cloud native capability, and is deeply converged with the cloud and transport network, it can release a large amount of computing power, bringing significant benefits for networks:

Efficiency improvement of computing power cloud of gNBs: Multiple gNBs provide distributed computing power to improve the efficiency in some scenarios.

Acceleration of function innovation: Algorithm iteration and function upgrade can be carried out on a daily basis, thus accelerating innovation.

Open platform as well as management and control: The computing power platform and O&M management capabilities are open for the third-party apps.

Enterprise-oriented service customization: B2B and B2C services can be provided separately to meet the requirements of enterprises for service capability and operation efficiency.

Service pooling to improve capacity: Services running on multiple gNBs can be scaled flexibly to improve the capability of coping with burst traffic, and support the smooth processing of requirements at peak traffic time.

A gNB can be equipped with the capability of cloud-based orchestration and scheduling, and a unified computing power network can be formed and converged with the edge cloud and centralized cloud in the operators' networks. Specifically, the following aspects are included:

A computing power cloud can be formed between multiple gNBs, to schedule the load flexibly and meet the requirements of different service models.

A computing power cloud can be formed by combining multiple gNBs with the cloud server on the centralized convergence side, and the multiple gNBs can share the processing resources on the centralized convergence side to meet the requirements brought by traffic bursts or service model changes on the site.

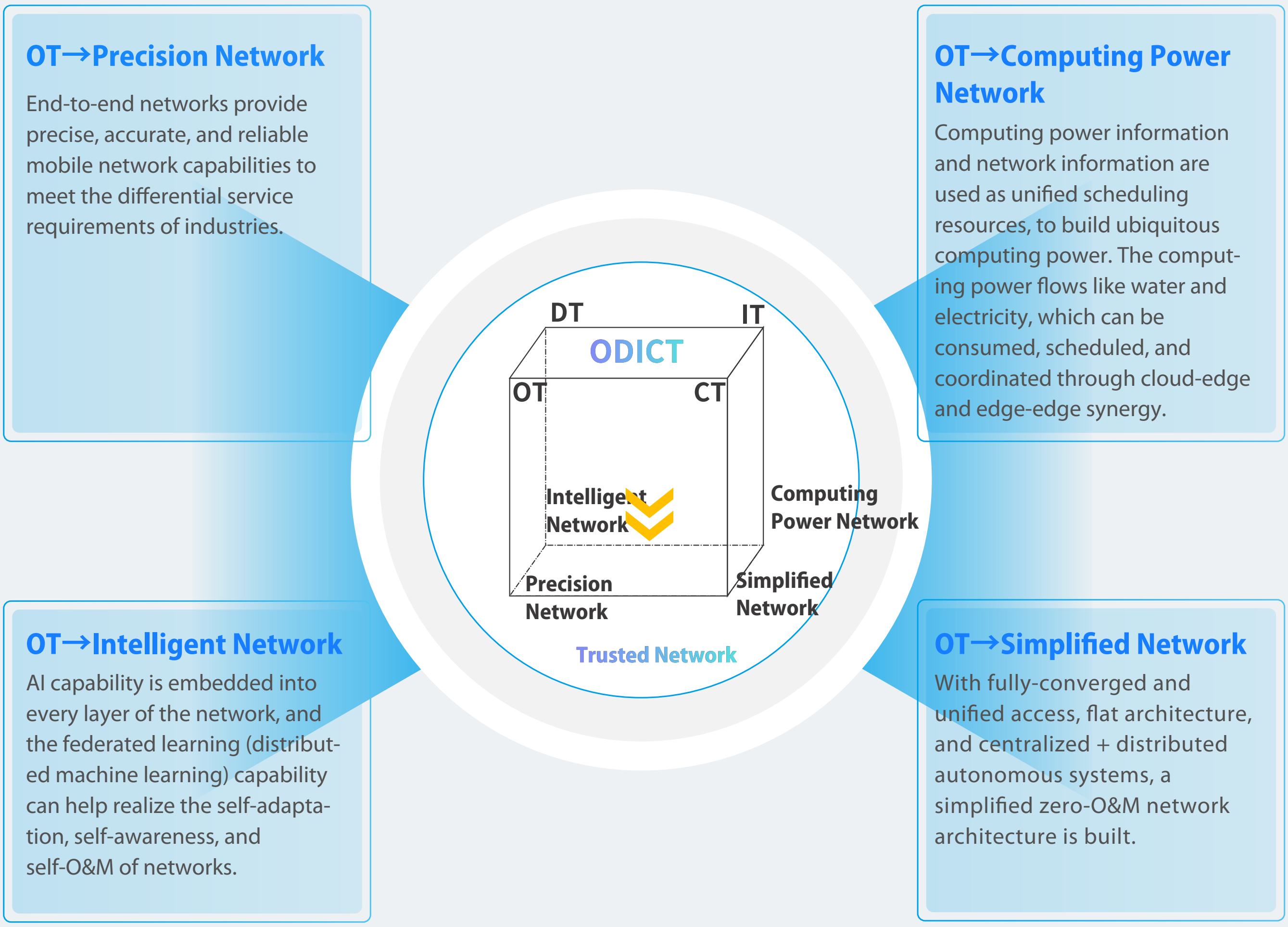
The idle resources from multiple gNBs can be used for the computing and unloading of services on the centralized convergence side, and the large computing power of gNBs can form a distributed computing power network, to improve the capability of simultaneous service processing.

The accelerators or acceleration cards of the gNB, core network, and transport network can achieve unified resource utilization and orchestration.

When the standard interconnection interfaces of general servers cannot meet the complex interconnection requirements, a unified high-performance forwarding software package is required to realize the data delivery between the computing power networks.

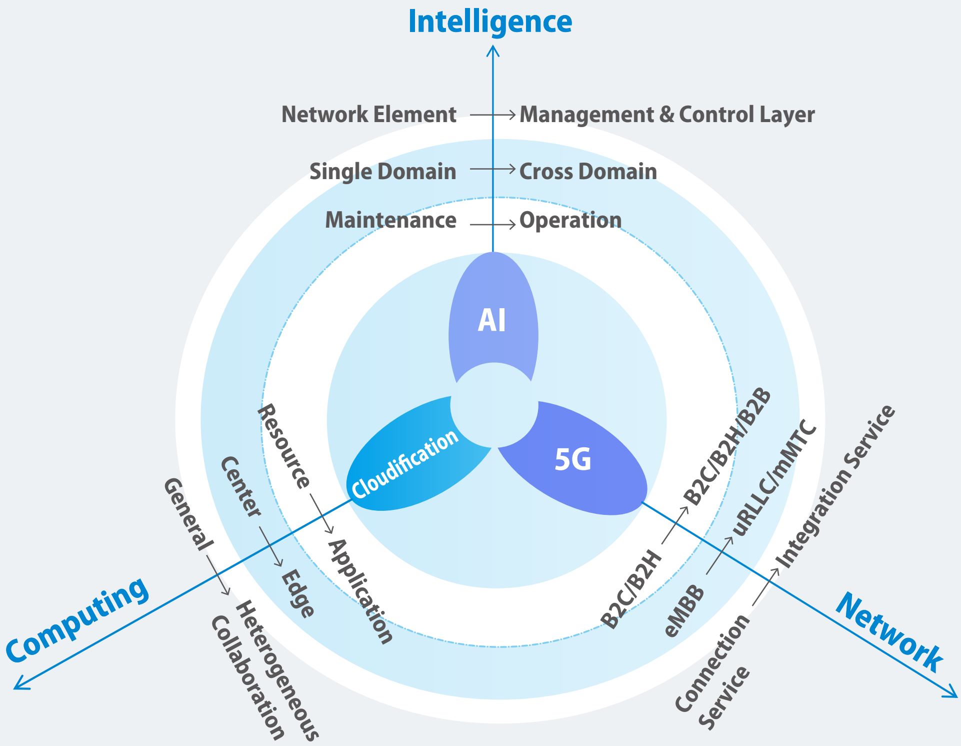
3.1.3 B5G Core Network: ODICT Convergence

The development of wireless core network is a process of continuous convergence and expansion of technologies in multiple fields. In the 2G and 3G era, the dedicated CT equipment played a leading role in supporting the development of the voice service and value-added service of the core network. In the 4G era, the core network with full IP technology promoted the mobile Internet service, while the convergence of ICT was at the initial stage. Entering the 5G era, the core network introduces such technologies as orchestration, VM, container, and Service-Based Architecture (SBA), and convergence of ICT is further advanced. At the same time, Network Data Analytics Function (NWDAF) is a preliminary attempt of network intelligence based on data analysis, which marks the convergence of DICT. In the B5G era, the development of the 5G core network is further enhanced and the deterministic Operation Technology (OT) is introduced. **The convergence of ODICT is a major technical evolution trend of core networks in the future, as illustrated below:**



ODICT Convergence

Based on the trend of ODICT convergence, the B5G core network will continue to evolve in three directions, namely, cloudification, intelligence, and service capability, as is shown in the following figure:



Three Dimensions: Intelligence, Computing, and Network

3.1.3.1 Cloudification

SBA-based 5G core networks accelerate the iterative upgrade of the technical architecture, and can better serve the differential requirements of 5G verticals. Cloudification reflects the trend of software-based, flexible, extensible, and self-adaptive network management. The B5G core network will continue to evolve toward cloud native from the following three aspects:

From resource cloudification to cloud-native application:

Based on the virtualization technology, 5G core networks enable resource layers to be deployed on cloud through VMs and containers, to build a unified cloud telecom infrastructure. B5G core networks will further focus on the cloud-native evolution of the application layer, and form a unified PaaS to rapidly build the network function based on the micro-service architecture and the public application component sinking. At the same time, the introduction of cloud native technologies such as BaaS/FaaS and Serverless turns the focus of B5G core networks to innovation in the business logic, thus achieving a leap from the elastic network to flexible network.

From centralized convergence to edge extension:

The forwarding plane of B5G core network will further sink to the user site or will be close to the BTS. As a result, B5G core network features large-scale and super-distributed deployment, taking up the downlink cloud services and uplink terminal computing tasks together with the nearest computing power provided by MEC. In this way, the in-depth convergence of cloud, edge and terminal can be achieved, and the delay of data interaction and consumption of network transmission bandwidth can be significantly reduced.

From general computing to heterogeneous collaboration:

To achieve truly lasting and immersive experience of meta-universe in the future, the demand for computing power could be increased by at least one thousand times. In the post-Moore era, to address the performance bottleneck of general CPUs, the high-performance computing power of GPU, FPGA, and DPU needs to be introduced, and the cloud-based general and heterogeneous architecture should be used to accelerate hardware-software synergy. Software is used to adapt to the frequently changing demands, while heterogeneous hardware is used to meet the requirements of high computing power.

3.1.3.2 Intelligence

With increasing business scenarios of 5G core network, network functions and management are becoming more and more complicated. Traditional networks use static and manual configurations and management, which bring high management costs. Therefore, intelligence is needed to help improve the service capabilities and quality of network functions, network maintenance, and network operation. B5G core networks will evolve to be more intelligent from the network element layer to the management and control layer, from single-domain to cross-domain operation, and from maintenance to operation.

Intelligence from Network Elements to Management and Control:

As the engine of data analysis and AI, the Network Data Analytics Function (NWDAF) has been introduced by 3GPP SA2, and features standardized analysis capability, high real-time performance, and closed-loop, flexible deployment. NWDAF is widely used in scenarios such as user experience optimization, network performance optimization, resource optimization, and UE exception detection and identification. B5G core networks will be further enhanced to support more application scenarios and the federated learning framework, and networks centering on intelligent decision-making will be implemented in the future. At the same time, the Management Data Analytics Function (MDAF) is also introduced in the network management layer by 3GPP to provide data processing and analysis of networks and service-related events, and enable intelligent and automatic network service management. For example, the MDAF provides slice coverage, slice availability, and slice prediction information for the optimization of network resource utilization efficiency, thus saving energy.

Intelligence from Single-Domain to Cross-Domain Operation:

Up to now, some automatic scenarios have been realized, such as fault location, alarm association, and service optimization, in the single domain of 5GC, IMS, and cloud platform. In the future, to guarantee end-to-end user experience and cross-domain collaboration, multiple domains can be interconnected with upper-layer services through intent-based network interfaces, to complete the network lifecycle management. In this way, a user-centric network is built. To achieve cross-domain privacy-preserving computation, a federated learning framework will be introduced to allow multiple participants or multiple computing nodes to jointly carry out efficient model training without sharing raw data.

Intelligence from Maintenance to Operation:

At present, business operation is the focus in the operation of 5G core networks. In the future, AI capability will be introduced in B5G core networks to achieve all-round operation of services and resources. The business operation layer provides zero-fault products for end-users and business partners, and enables self-planning and self-marketing through AI. The service operation layer realizes the business implementation, quality management, and problem management, and enables self-healing, self-optimization, and self-management through AI. As for resource operation, independent resource management is implemented, and the self-distribution and self-guarantee of physical, virtual, and logical resources is enabled through AI.

3.1.3.3 Service Capability:

The B5G will strengthen the service capability in terms of XR-based immersive communications, Metaverse, ISAC, and high Precision, Reliability, Openness, and Security (PROS).

Further Evolution to Full-Service Scenarios:

The service of B5G core networks will evolve from the current concept of "Human-Human Communication" and "Internet of Things" to B2X (B2C, B2H, and B2B), namely, intelligent Internet of everything in all scenarios. The B2C business will provide ultimate experience for consumers, and evolve from audio or video communications to multimedia communications. The dual-channel (VoNR and DoNR) will integrate with the new data channel technology to provide interactive experience. The ultra-broadband will enable multisensory coordination, XR packet perception, and multi-stream QoS control, to implement XR-based, immersive, interactive, and multisensory communications, and even the Metaverse communications in the future. On the basis of highly precise and highly resilient technologies, namely, TSN, URLLC, and 5G LAN, the B2B business empowers various industries such as manufacturing, energy, health care, and transportation, and achieves full connection of individuals, machines, and things through the enhanced machine-machine communications and human-machine communications.

Evolution from Large Bandwidth to Low Latency, High Reliability, and Massive Connections:

B5G core networks will continuously evolve from eMBB to URLLC and mMTC, realizing lower latency, higher reliability, and wider connections. With enhanced coordination with DetNet and TSN for backhaul, B5G core networks achieve the evolution from local-area determinacy to end-to-end determinacy and wide-area determinacy, and better supports the long-distance services with low latency and high reliability, such as remote surgery and remote control. Integrating with industrial applications and NR, and matching the scheduling timeslot of downlink to reduce NR transmission time, B5G core networks can be further applied to services requiring extra-low latency (< 2ms), such as industrial control. Evolving from low-speed IoT (CIoT) to medium-speed IoT (RedCap), B5G core networks enable the access to low-power and high-speed wearable devices and industrial wireless sensors.

Evolution from Connection to Integration of Connection, Computing Power, and Capability:

B5G core networks provide computing power network services integrating connection, computing power, and capability. The computing power network includes three supporting technologies: computing awareness, scheduling, and metering. In terms of computing awareness, computing power resources on distributed MEC nodes are released to the mobile network through the UPF. In addition, as the first entrance for user packet perception, the UPF can perceive the types of user packets, identify the specific demand for computing power resources, and then forward them to the best load node in the mobile network.

In a word, B5G core networks will continue to evolve into simplified networks, intelligent networks, computing power networks, and precision networks based on the convergence of ODICT in three directions: cloudification, intelligence, and service capability.

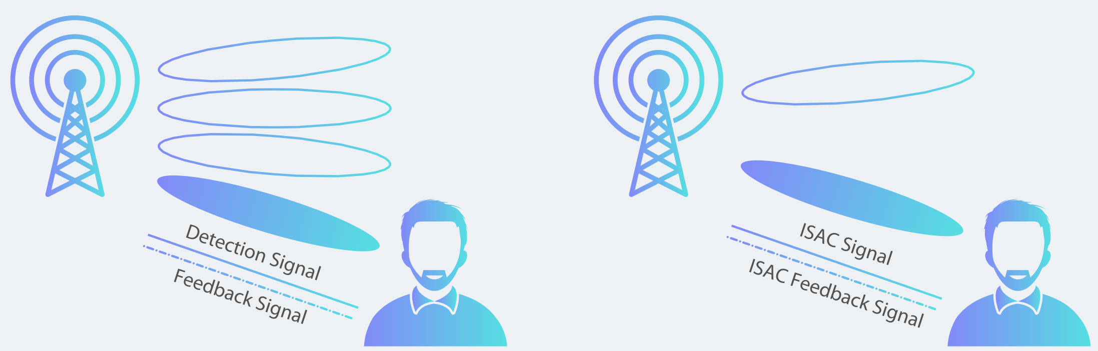
3.2 Integrated Sensing and Communication (ISAC)

Integrated Sensing and Communication (ISAC) means the integration of communications and sensing, which enables the communications system to have both functions of communications and sensing with the two interacted each other. For example, the sensors can sense the physical characteristics of the surrounding environment by perceiving and analyzing the characteristics of the radio channel while communicating. In this way, the mutual enhancement of communications and sensing is achieved. In fact, ISAC is not a new concept. In the 3GPP standards, ISAC includes the following initial applications:

Positioning:
To provide the positioning service for communications users, 5G-Advanced will introduce such technologies as sidelink-based ranging, multi-carrier aggregation, carrier phase, and AI to improve the positioning performance. Through channel fingerprint learning, AI can be used to solve the positioning problems in NLOS scenarios. There has been ongoing research in the industry on using ultra-large-scale antennas and large-bandwidth technologies to separate multi-paths in angle and delay domains, to reduce the NLOS impact and the requirements for the number of positioning sites, and to improve positioning accuracy to the centimeter level.

Sensing:
The sensing function provides services for users not using SIM cards, including location information and map construction. It requires the gNBs to have large bandwidth and multiple antennas, to provide better sensing performance.

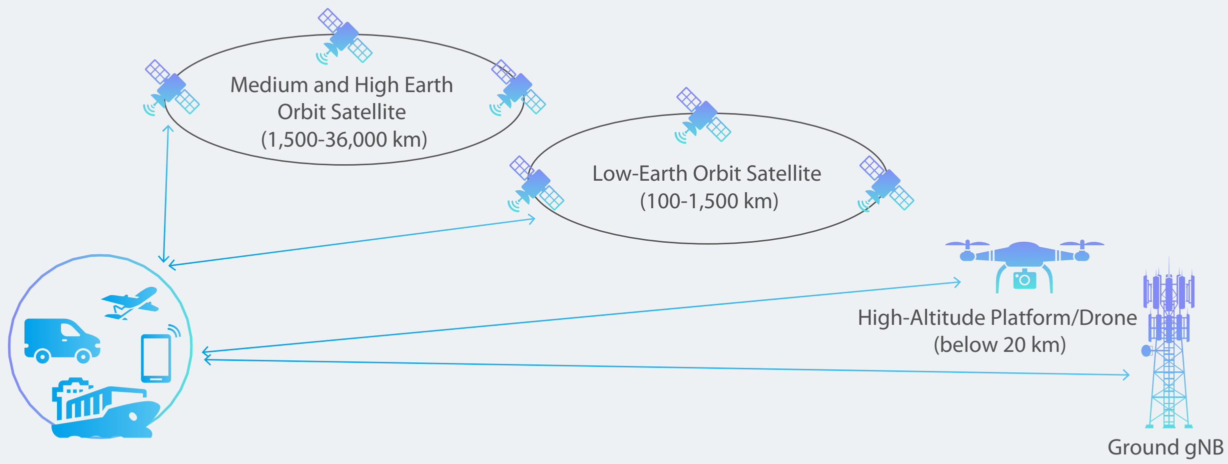
The main sensing application specified in the 3GPP standards for terminal users and vertical industry users is positioning. However, the development of positioning still depends on low-power and low-cost UEs. High-precision wireless sensing depends on large bandwidth and array antennas, while high-frequency communications technologies such as millimeter wave have not been deployed widely yet. This restricts the market expectation on the high-precision ISAC application in a short term. Currently, the standardized ISAC application is still focusing on indoor and short-distance scenarios, while the killer ISAC applications based on cellular network sensing still needs to be explored. In the 5G-Advanced phase, the positioning accuracy and performance will be further improved. The wireless communications system adopting RIS, high frequency, large bandwidth, and large-scale MIMO technologies has the potential of providing high-precision sensing. At the same time, the research and evaluation of ISAC application based on cellular networks are being conducted, and more ISAC applications (B2B, B2C, and cellular network



ISAC improves the communications performance including beam tracking.

3.3 Space-Air-Ground Integrated Network

After nearly 30 years of development, terrestrial wireless communications systems have provided a relatively complete network coverage in most parts of the world, serving 80% of the world's population. However, due to factors such as economic costs, technologies, and natural conditions, terrestrial wireless and wired networks are currently not available in remote areas with low population density, such as deserts, forests, and oceans. The effective solution to these problems is a space-air-ground integrated network, which can make up for the shortage of terrestrial communications by making full use of the advantages of flexible deployment and wide coverage of low-altitude communications and satellite communications. The main application scenarios of space-air-ground integrated network are uninhabited areas such as oceans and deserts. The space-air-ground integrated network is a supplement, rather than a replacement, for ground networks.

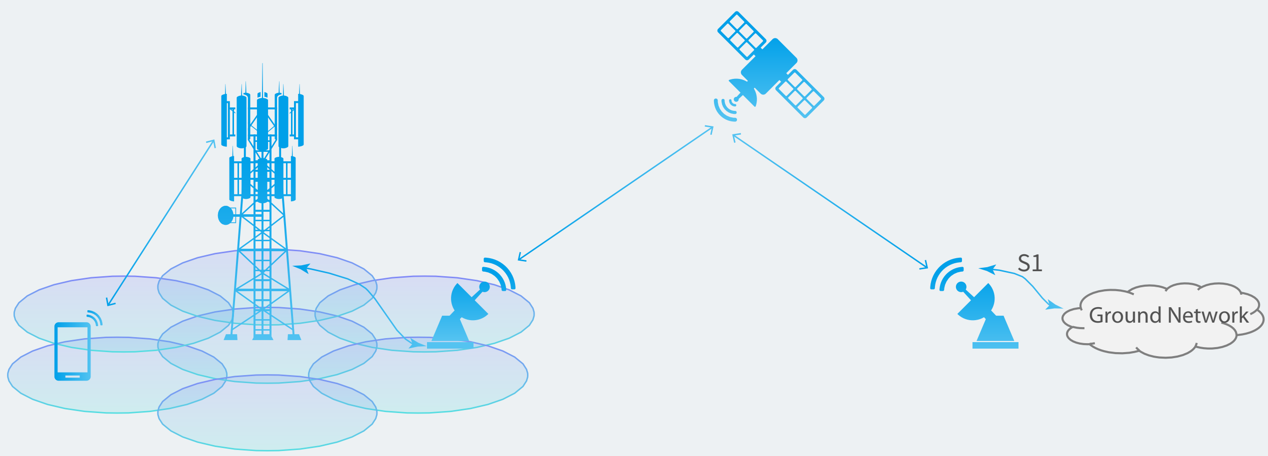


Layer of Space-Air-Ground Integrated Network

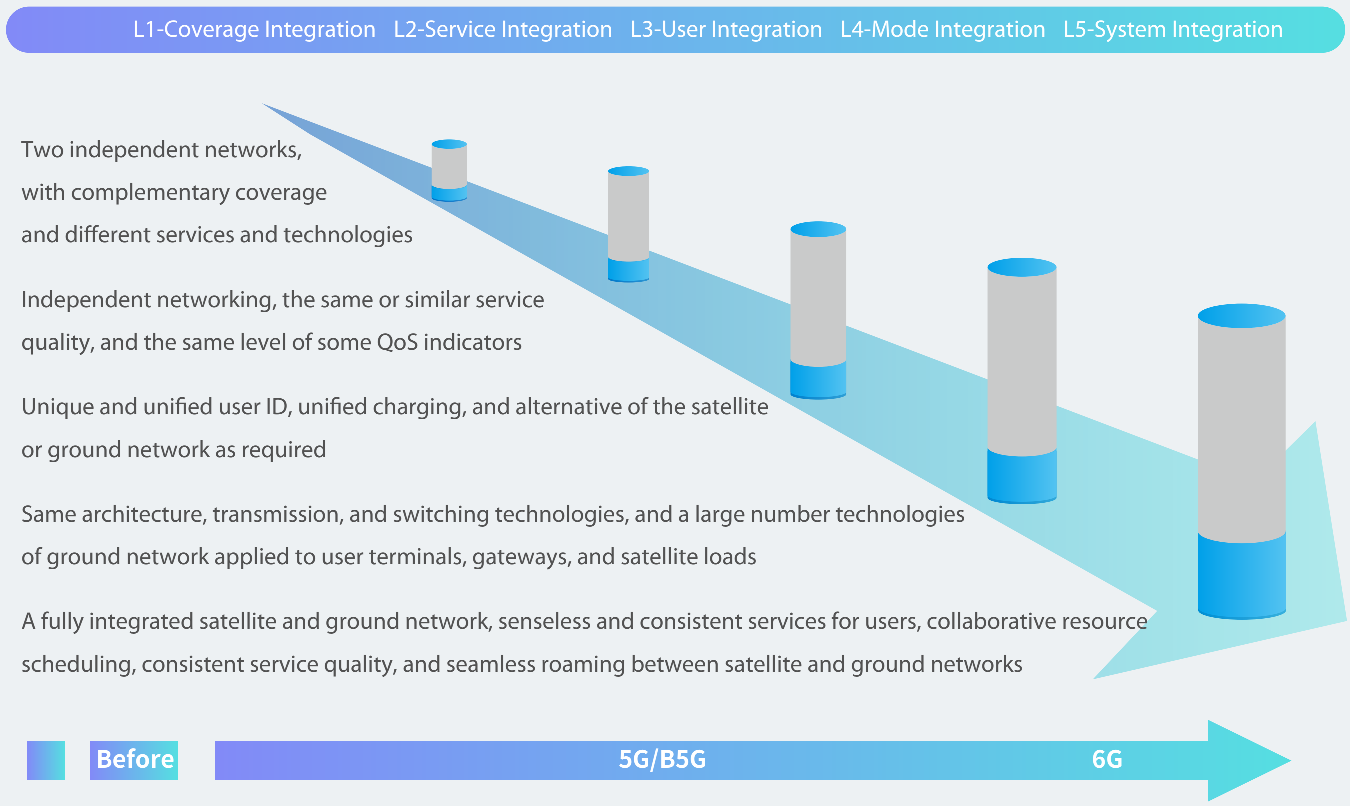
For a space-air-ground integrated network, "space" usually means 100 km above ground, and the network nodes include the low/medium/high earth orbit satellites; "air" means the 20 km below the stratosphere, and the network nodes include a High Altitude Platform System (HAPS) and drones; and "ground" means the surface of the earth. Among the three altitudes, "air" and "ground" are highly similar in communications scenarios and technologies, and therefore their integration is relatively simple. In addition, there has been many successful applications in recent years. For example, above a flooded area, the communications networks can be recovered immediately by carrying gNBs on a drone or a hot-air balloon. On this basis, the integration of "space" and "ground" is the main direction of the future, that is, the integration of satellite network and ground network.

The satellite networks and ground networks have been developing respectively for more than 30 years. During this process, the relationship between the two has gradually evolved from competition, complementarity, to integration. Especially in the past five years, the constellation of low orbit broadband Internet satellite has attracted extensive attention in the industry; the number of satellites in a single constellation reaches thousands or even tens of thousands, which can provide the transmission rate of tens of Mbps. Such a rate is close to the performance of the 4G/5G mobile communications networks on the ground. Therefore, the idea of replacing ground networks with satellite Internet has emerged. Although the satellite networks can achieve seamless coverage around the globe without being restricted by terrain conditions, the communications distance between satellites and ground users may reach hundreds of kilometers, which is usually hundreds of times of the ground network. The link loss caused by satellite networks is more than 20 dB higher than that caused by the ground networks. To narrow this gap, an effective way is to increase the antenna gain and transmit power, but for UEs, especially handheld UEs, this is almost impossible. In addition, a beam of a satellite can be regarded as a gNB. A ground gNB usually covers an area with a radius of hundreds of meters or even several kilometers, while a beam of a satellite can cover an area with a radius of dozens of kilometers. If satellite networks are applied to urban, suburban, or even rural areas, the number of users covered by a beam far exceeds the user capacity of a beam. Only from the above two simple perspectives, the idea of replacing the satellite Internet with the ground network is obviously impossible.

Satellite networks and ground networks cannot be replaced by each other. The integrated development of the two networks is the only answer. Although at this stage, the two only realize interconnection through the integration of terrestrial gateways, that is, the traditional satellite communications can provide trunk transmission and backhaul services for the ground network. With the development of technologies, the future space-air-ground network will be a three-dimension layering, integrated, and collaborative network, which will be based on the ground network and be expanded through the space network. These networks will adopt unified technical architecture, technical system, and standards and specifications. During the process, the space-air-ground network will undergo integration of different levels, and finally achieve the integration of modes and systems.



Backhaul Network



Integration Levels of Satellite and Ground Networks

To achieve this goal, the international organizations such as 3GPP have carried out related work based on 5G and from the perspectives of application scenarios, network architecture, and air interface technology. At present, the main focus lies in two application scenarios: eMBB and mMTC, aiming to solve the following four major problems.

In the future, operators can provide 5G commercial services in areas where the ground network is inaccessible or the ground network infrastructure is underdeveloped.

The "ubiquitous" coverage of 5G networks can be achieved, to enhance the 5G service continuity for user terminals or mobile platforms (such as high-speed railways, steamships, and airliners).

The "ubiquitous" availability of 5G services can be enhanced, especially 5G-based systems for emergency, railway, maritime, and air communications.

The "scalability" of 5G networks can be achieved, and efficient multicast/broadcast data transmission can be implemented through the 5G network edge (5G MEC) and even directly through the 5G user terminals.

With the evolution of B5G and 6G networks, enhancement design is indispensable for achieving in-depth integration.

In terms of network architecture, the future space-air-ground integrated network will evolve from a single-layer network to a multi-layer network. For example, as for macro network architecture, low-earth orbit satellites and ground gNBs are used as space-based and ground-based access networks respectively, and high-earth orbit satellites are used as space-based backbone networks. The access networks and backbone networks together form a multi-layer integrated network. At the access network level, it is expected to see new architectures such as CD/DU, Integrated Access Backhaul (IAB), and multi-link.

In order to achieve the integration under a unified technical system, the space-air-ground network integration also requires in-depth research on key technologies, including such key technologies for the physical layer as new coding and modulation access, spectrum resource management, mobility management, AI, and satellite Internet protocols.

3.4 AI Native

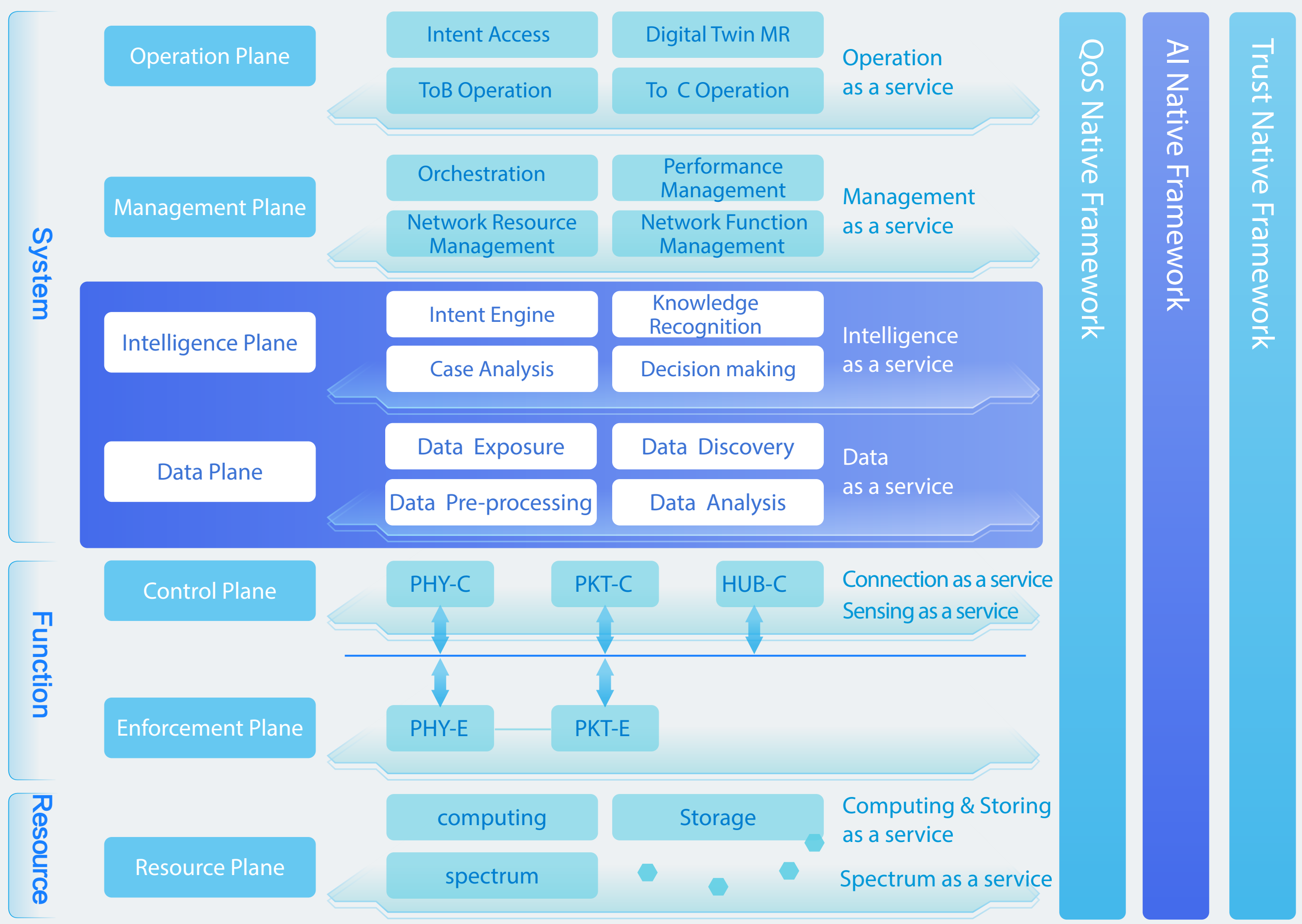
Based on the deep integration of AI computing capabilities into the communications process, AI native improves the intelligent learning and scenario adaptation capabilities of different layers (physical layer, MAC layer, and network layer) in an all-round manner, and supports the continuous evolution of personalized intelligent service capabilities. In addition, the logical intelligent plane and data plane are being further formed, and integrated into the network architecture design and the implementation of NEs and interfaces, so that the future intelligent networks can be self-adapting, self-learning, self-correcting, and self-optimizing. Moreover, AI native enables the unified intelligent orchestration and scheduling of AI resources and services, and continuously guarantees AI service quality through in-depth integration of intelligence, computing, and networks, enabling networks to achieve intelligent autonomy and industries to build a ubiquitous intelligent ecosystem. Integrating with intent network and digital twin technologies, the wireless intelligent orchestration network will evolve further to become a service- and demand-centric intelligent network.

The main features of AI native are as follows:

1.It provides complete AI capabilities, including data, computing power, and algorithm capabilities, in the communications networks defined by the 3GPP standards.

2.In terms of the communications network architecture, it fully considers the application requirements of future intelligent planes from the function and logic aspects.

3.In terms of future protocol, protocol stacks and interfaces are designed to adapt to the deployment of AI applications.

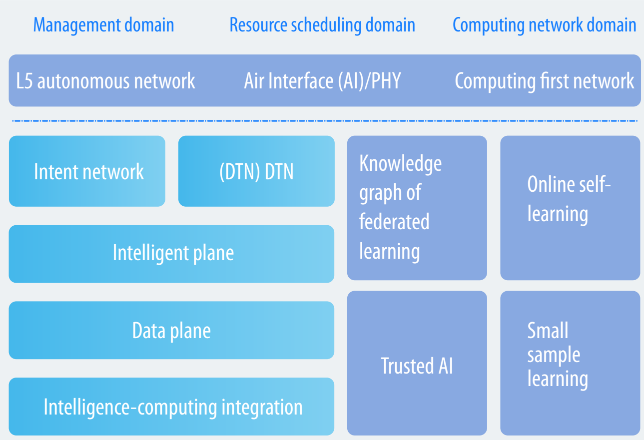


System architecture of AI Native

Based on AI native and its evolution, as well as the development of AI, such as automated machine learning, trusted AI, and model pre-training, the key technical graph of the future B5G AI native networks is worked out, which is shown in the figure below:Based on AI native and its evolution, as well as the development of AI, such as automated machine learning, trusted AI, and model pre-training, the key technical graph of the future B5G AI native networks is worked out, which is shown in the figure below:

In the B5G era, the key technologies involved in the AI-native network architecture are:

1. Definitions of the intelligent plane and data plane of AI native
2. Distributed federated learning technology
3. Intent network and intent engine technologies
4. Single-domain autonomy in the autonomous network and cross-domain synergy technologies
5. Integrated intelligent and computing networks and intelligent orchestration and scheduling technologies
6. Digital twin network and simulation technologies



Key AI Native Technologies

3.5 Trust and Security Native

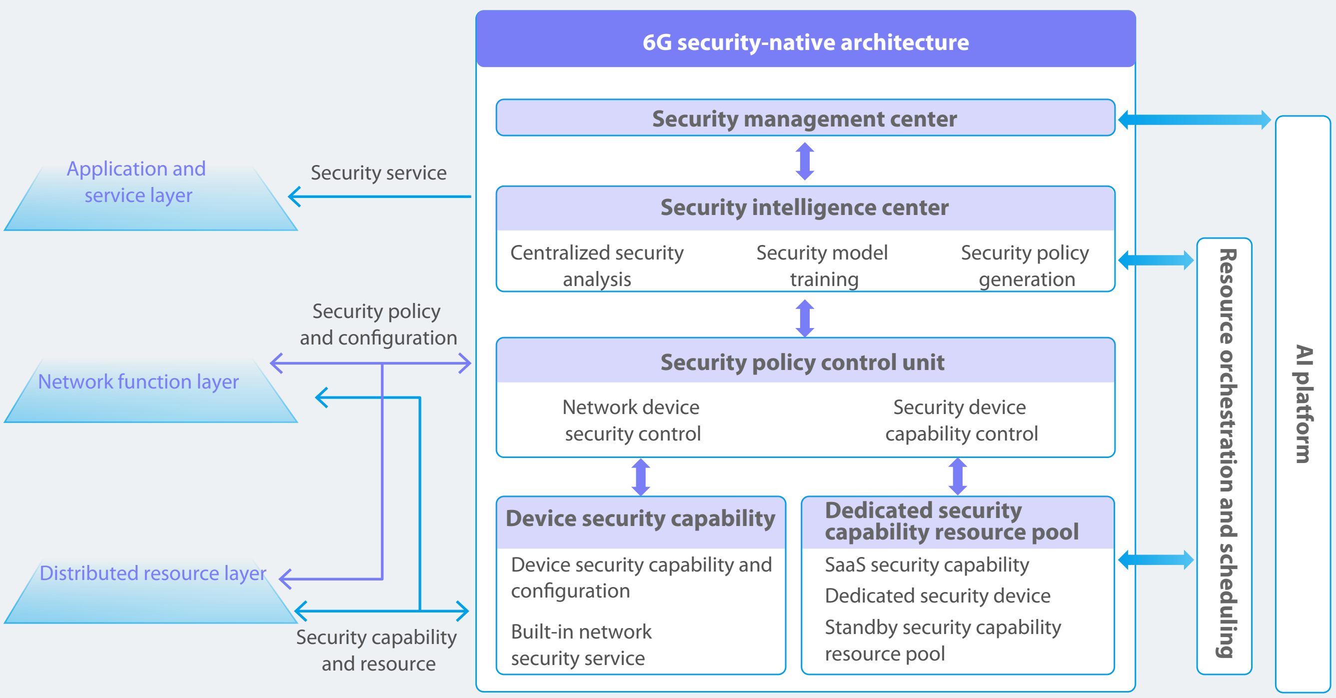
In the construction of a security-native immunity system for 6G, the first step is to design and construct a unified trust-based consensus identity and trust evaluation system for 6G. Building deep trust -based consensus identities that cannot be repudiated, forged, or tampered with is the key and basis. People and devices must have the same identities, and must be associated and managed in real time. Based on the trusted identity and multiple factors such as behavior extraction, a trusted identity evaluation system can be established. The 6G consensus identity and trust system is deployed in both centralized and distributed manners, to better support ubiquitous access and service operation, as well as efficient security processing, such as tracking.

Furthermore, based on the 6G trust-based consensus identity and trust evaluation system designed at all network layers, NEs, and boundaries, an integrated, synergetic, intelligent, and computable security immunity system is built for 6G, in which security and trust should be embedded. This is called trust and security native. It includes:

Convergence and governance: It means that 6G networks aggregate different security protocols and security mechanisms under the background of technology integration and service integration, to implement network security governance. The access-side security in 6G networks is guaranteed through convergence and governance, to safeguard the security-native 6G network.

Self-reliance: It means that the security protection of 6G networks should be self-driven, to synchronously and even prospectively adapt to network changes, thereby developing robust intrinsic defense capabilities of 6G networks. The 6G networks features the inward-to-outward security and stability, which is an important aspect of "self-reliance."

With the goal of "on-demand, flexible, intelligent, and simplified networking," a 6G security-native architecture is designed, which consists of a security management center, a security intelligent center, a security policy unit, and a security capability layer. It coordinates with the AI and resource orchestration capabilities to form a systematic architecture.



6G Security Native Architecture

3.5.1 Distributed Multi-Domain Trust Model

The distributed multi-domain trust model adopts the security and trust model with the coexistence of cluster, decentralized, and third-party centers. Meanwhile, it provides security authentication and access service for various heterogeneous B5G/6G networks, such as air-space, air-ground, and space-air-ground integrated networks. As a result, users do not need to perform multiple authentication during cross-domain switching Adapting to the distributed deployment of network services, this model can provide efficient and fast security authentication for communications services distributed in different locations such as centers, regions, and edges. Enhanced multi-party security and trust models can be designed to accommodate operators, users, vertical industries, and other stakeholders.

3.5.2 Blockchain

The blockchain technology can be used to build a multi-party security and trust model. It can provide operators with a decentralized platform to store and transfer data securely and efficiently, and improve network trust. The blockchain technology can enable fast, secure, stable, and decentralized data storage in 5G-Advanced to ensure the reliability and transparency of network sharing. Some key data (such as cell resource usage and spectral efficiency) on the network side can be uploaded to the blockchain platform in a secure and trusted manner through the shared gNB of the operator The operator can optimize the system resource configuration in accordance with the actual resource usage on the blockchain platform. In addition, the hosting operator can provide services for the customers of the operators participating in network sharing, and send the fault reports of these customers to the blockchain platform with the same priority, while the operators participating in network sharing can obtain customers' fault reports through the blockchain platform, achieving secure and trust-based network O&M.

3.5.3 Active Defense Immunity

Active defense immunity uses the trusted technology to provide internal security capabilities such as trusted startup, security measurement, and trusted management for network infrastructure and software components, to ensure that software and hardware can operate as expected.

Security is integrated into the communications protocol design. The cryptographic application technology and key management mechanism that match the B5G/6G networks are adopted. The encryption/decryption and signature mechanisms that are applicable to high-speed and high-concurrency communications are also adopted. The data exchange mechanism with the privacy security protection capability is designed.

Through cross-domain security capability scheduling, the intelligent coordination of multi–security capabilities can accurately perceive the security situation and forecast risks of the entire network, and implement distributed response strategies for security attacks to guarantee agile response to security risks and maintain service continuity regardless of abnormal communications components.

3.5.4 Extensible Self-Evolving Architecture

Security capabilities are flexible and scalable. They can be flexibly combined and orchestrated to meet the multi-scenario and multi-domain security requirements of B5G/6G.

With capabilities such as digital twinning, the validation of security components and the intelligent training and simulation of security policies can be performed in the virtual world, to facilitate efficient and intelligent evolution of security in the physical world.
The security framework evolves with network design, and flexibly adopts the latest security technology in the industry.

3.6 New Business Empowerment

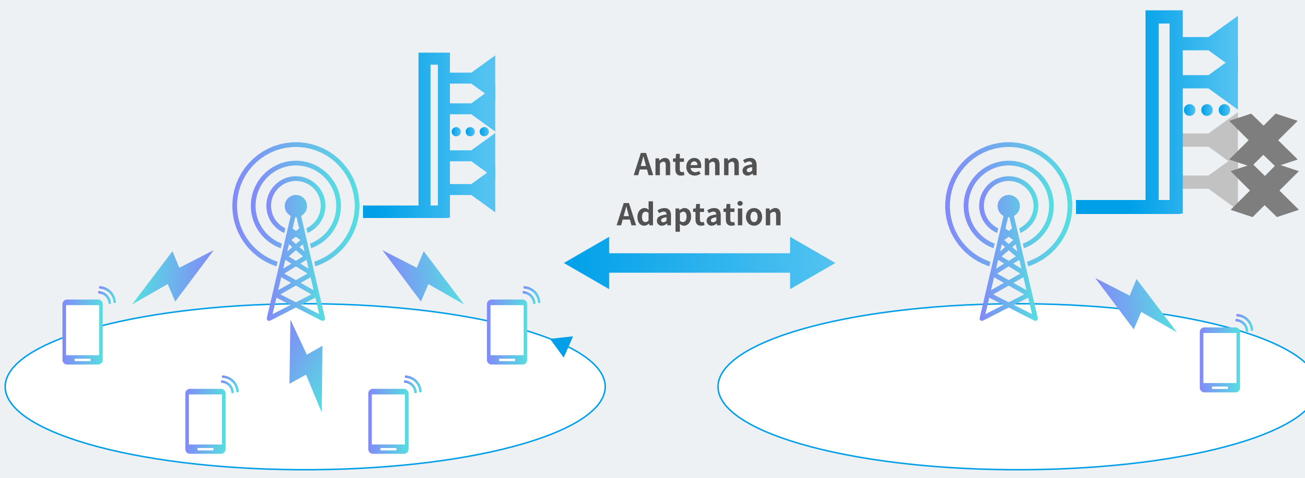
In some new business, such as Metaverse, Internet of Vehicles, smart home, and smart agriculture, has raised new technical requirements for networks. In Metaverse, XR is one of the most important scenarios. It has been used in many fields, including engineering assistance, professional training, telemedicine, social communications, and gaming, and new business models has been developed as a result. XR requires 1 Gbps bandwidth for immersive user experience, while the real-time holographic image transmission rate can reach 1 Tbps. To prevent dizziness and fatigue in user experience, the network is required to achieve extremely low latency. Time-Sensitive Networking (TSN) is one of the major network technologies of Industry 4.0. Based on Precision Time Protocol (PTP), the core principle of TSN, a time plan is created and distributed among network devices to first transmit the planned frames. B5G is interconnected with TSN via network bridge and provides ultra-high reliability, deterministic communications, and extremely low latency in the core production workshops of the factory, to precisely control the production behaviors of industrial robots.

The new WRC-23 spectrum, mmWave, and Terahertz provide large bandwidth and high data rate for industries. SVC turns the inter-TRP interference into useful signals to improve communications reliability and capacity. The SBFD, FD, flexible frame structure, and short frame technologies effectively reduce the communications latency and improve the resource utilization efficiency. The enhanced MIMO technology improves cell capacity, and supports more XR services at the same time. In short, B5G provides a solid technical basis for the development of new business.

04
Efficiency Improvement Technologies

4.1 Power Efficiency

B5G uses higher bands as signal carriers, with a data rate of terabits per second. At the same time, emerging demands for new business, such as holographic communications, intelligent interaction, sensory interconnection, digital twinning, and communications perception, bring a surge of hundreds and thousands times of network traffic. To reduce power consumption as much as possible when traffic increases, a new generation of green equipment should be developed based on highly integrated chipsets, high-efficiency devices, upgraded processes, and strong computing algorithms, so that extremely deep sleep and extremely short response time can be achieved. On the basis of hardware with low power consumption, multiple technologies, including intelligent power saving technology, power consumption reduction by multi-frequency coordination, power efficiency improvement by traffic navigation, are applied to empowering all industries to co-build a green ecosystem,



Network Power Saving by Antenna Adaptation

4.1.1 Deep Sleep

The core idea of power saving technology for wireless software is to reduce redundant network capacity and network power consumption by shutting down some wireless capabilities to match the traffic load distributed in networks. In the future, modular hardware design, scalable software architecture design, and delicate hibernation scheme will be used to continuously improve the shut-off depth of devices. This enables hardware power consumption to gradually approach linear variation, minimizing power consumption under the condition of medium and low load.

Space domain:

As shown in the figure, the network can adopt the dynamic antenna adaptation method based on the service requirements, and reduce power consumption by shutting down part of the Tx/Rx antennas. In addition, the network power consumption can also be reduced by using dynamic Tx/Rx ports as well as panel and beam adaptation

Time domain:

As for the time domain, the focus is to activate some network components, such as cells and frequency bands. They can be turned into sleep mode to reduce the network power consumption. To guarantee user experience, more delicate turn-on and shut-off operations should be considered in the time domain.

In the current discussions, some companies also propose to use the wake-up mechanism to achieve network power saving. In terms of this method, the issues including the function of wake-up signal, specific function design, transmission mode, and how to avoid frequent wake-up signals still need to be further studied.

In the power saving mode, hardware response time is a key factor affecting network indicators and user experience. It will radically change from minutes to milliseconds, making possible power saving from only idle time to all day.

In particular, the power amplifier, as the core component of the radio frequency devices, has low efficiency under typical traffic load. Normally, the time to regulate a power amplifier is long, which has an impact on the user experience and network indicators. Based on these situations, we adopt power amplifier envelope tracking, dynamically monitor traffic load, and adjust power amplifier voltage in real time, to maintain power amplifiers in the best working condition with maximum efficiency at any time and under any traffic load.

4.1.2 Multi-Dimensional Synergy

Power saving of networks not only involves the power saving of gNBs, but also requires the multi-dimensional synergy of various frequency bands, networks, and UEs.

Multi-frequency synergy (PowerPilot):

The deep synergy of multi-frequencies and multi-systems reduces network power consumption, realizes dynamic matching of services and resources, and achieves power saving in multi-layer wireless networks without significantly affecting user experience. On that basis, a multi-layer AI platform is built, and the big data analysis and AI technologies are introduced to facilitate the evolution towards full intelligence. The intelligent self-configuration of power saving parameters will be realized by integrating multiple-domain information (spectrum utilization, user experience, traffic trends, etc.) to evaluate the network power efficiency, selecting power-saving strategies based on the specific scenario, and considering the information of the station type, network configuration, and traffic statistics and forecasts. According to the evaluation of power saving effect, the closed-loop optimization is carried out by tailoring strategies for each station, realizing the balance between power consumption and network indicators, and advancing the deployment of power saving technologies in operational networks.

UE synergy:

UE mobility, UE business models, and configuration requirements can be helpful for the obtaining of UE requirements, thereby adjusting configurations more accurately, and saving power while reducing the impact on user experience. For example, for UEs with low data transmission requirement, the number of antennas or bandwidth can be dynamically adjusted to save power for the network and UEs.

Network synergy:

Information can be exchanged between cells by network synergy, such as information on cell state (sleep state, etc.) and beam adaptation. Configuration can be adjusted according to the information obtained, so that power consumption can be reduced.

4.2 AI for Air Interface Efficiency

In the 6G communications system, AI/ML technology will also become indispensable, but the application of AI/ML technology in the physical layer of the wireless access network is still in the pre-research stage. It is still unknown whether we can push through the performance limits of traditional methods. However, AI/ML technology will provide certain possibilities for dealing with problems that traditional methods cannot solve. How AI/ML technology is combined with physical layer applications in future wireless communications system can be divided into the following five levels:

Specific functions:
At present, there are still some problems hard to tackle in the wireless communications system, such as interference detection, FDD uplink and downlink interchange, and channel estimation. These problems are often caused by nonlinearity or inaccurate modeling. When traditional methods are used, the performance is usually not good. But machine learning excels in these areas.

Update of the existing discrete modules:
The traditional physical layer design of the communications system is generally based on the linear model. Therefore, once the system is affected by strong nonlinear factors, the performance of the system will decline sharply. At the same time, if the receiver side intentionally introduces some nonlinear operations, the system performance will be improved. Many modules, such as the coding and decoding module, and modulation and demodulation module, as well as waveform design, pilot design, channel estimation, and MIMO detection, can help improve the overall performance of the system by introducing some nonlinear processing algorithms. Similarly, if non-model machine learning algorithms are introduced into the system, some unexpected effects may occur. Because of the application of deep learning, strong nonlinear processing capabilities are demonstrated in the field of image recognition. This has also provided a new solution for the physical layer of the communications system to cope with the interference of nonlinear factors.

Integration of AI/ML technology and traditional methods:
In the traditional communications system, we adopt the model-based method. Although the model is sometimes too ideal, it can describe the main characteristics of a process, and be used as a kind of prior information. If some

characteristics of the existing models are used as additional information in the design process of machine learning, some inherent defects of machine learning are likely to be overcome, such as the need for huge training data, underfitting or overfitting, and slow convergence rate.

Joint optimization among modules of the physical layer:
In the traditional design, the physical layer of the communications system is optimized by module. Although this design can ensure the optimal status of each module, it cannot achieve the optimal status for the whole system. For example, coding, modulation, and waveform are designed separately in the traditional system. Once the three factors are considered together, the optimization cannot be ongoing because of the high complexity of the receiver. However, for machine learning, it is not necessary to carefully design all kinds of coding schemes, or carefully think about various constellation graphs. Neural networks can be used to replace the module cascade method, so that the optimal end-to-end mapping method can be obtained through network autonomous learning. Use of machine learning for the joint optimization of modules in the physical layer is one exploration direction in the future.

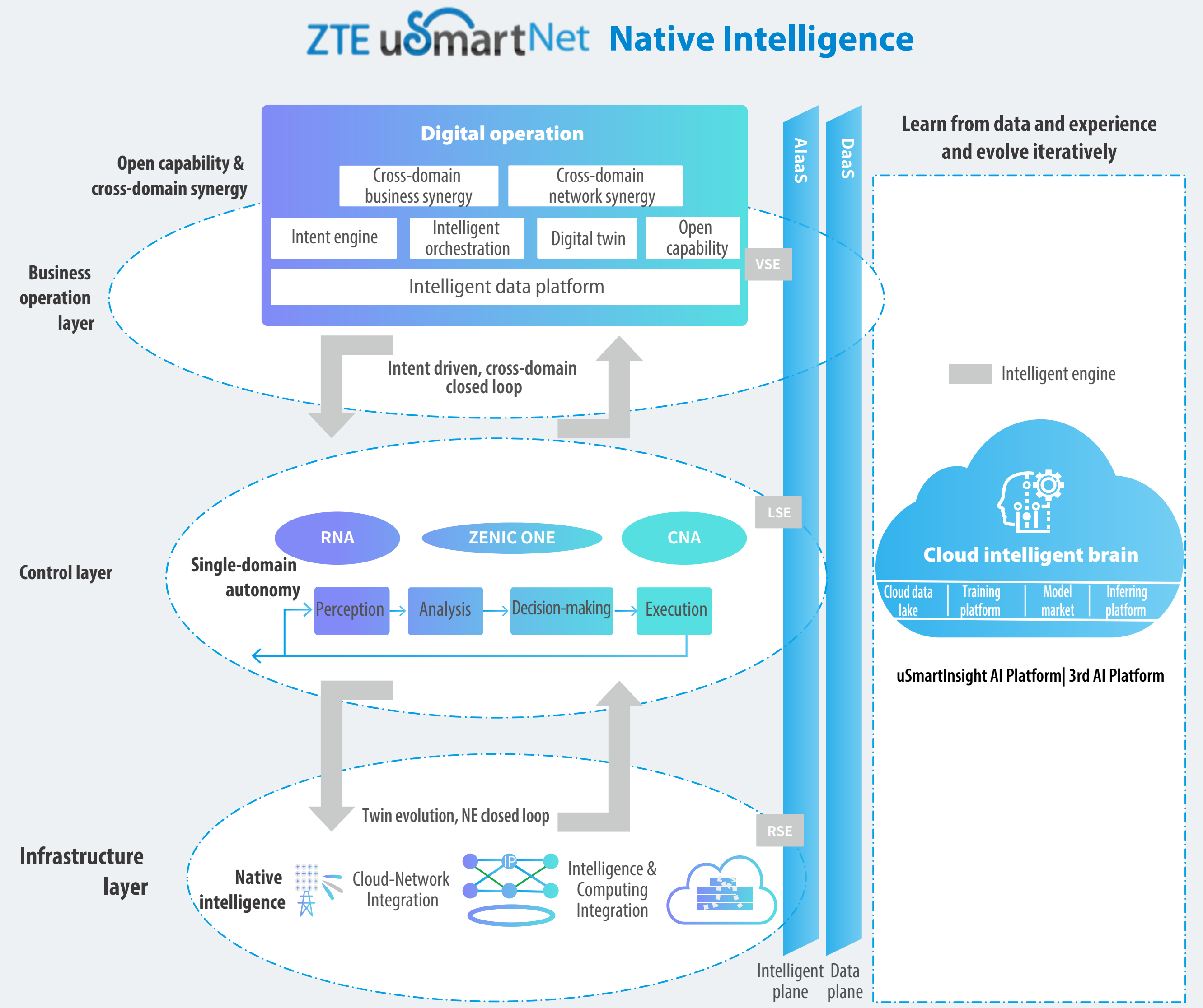
Structural machine learning at the physical layer:
Machine learning at the physical layer requires a functional entity that is structural and standardized. Therefore, various functions in the physical layer machine learning need to be abstracted as data collection, training, segmentation, and so on, so that various machine learning algorithms can be flexibly orchestrated to realize different physical layer functions.

4.3 Network Efficiency

AI and ML have attracted great attention from the whole industry in the past 10 years. B5G will continue to further enhance network intelligence. Due to the flexible design and complex topology of the NR system, there are a series of problems that make it difficult to construct the model. 3GPP RAN began to explore the application of emerging technologies such as AI/ML in communications networks, and started research on wireless AI/ML in Rel-17, to provide new ideas and development direction for the network automation and intelligence. According to Rel-18, corresponding standardization discussions will be conducted based on the research in Rel-17, to truly enable network automation and intelligent deployment based on wireless AI/ML.

To build an autonomous intelligent network, ZTE has put forward a self-evolving network (uSmartNet) solution by introducing AI in the different aspects of the network. The solution enables evolution in three directions: network evolution, O&M evolution, and operations evolution, promoting the continuous improvement of network intelligence and realizing the network functions and services as desired with simplified O&M.

In the future, the self-evolving network will evolve toward the ultimate goal of intelligent, comprehensive autonomy, and self-intelligence, gradually achieving optimal network investment efficiency and operating efficiency. Ultimately a network with ubiquitous services, computing, and intelligence will be built.



Autonomous Network Architecture

Here, we take wireless intelligent orchestration network evolution as an example to illustrate the evolution of the self-evolving network to L4/L5. To cope with the increasing scenario-based flexible service requirements, precisely meet the different value demands under different spatial and temporal coordinates of network development, the wireless intelligent orchestration network will be continuously developed from the following four aspects into an intelligent orchestration network with scenario-based intention and self-adaptation:

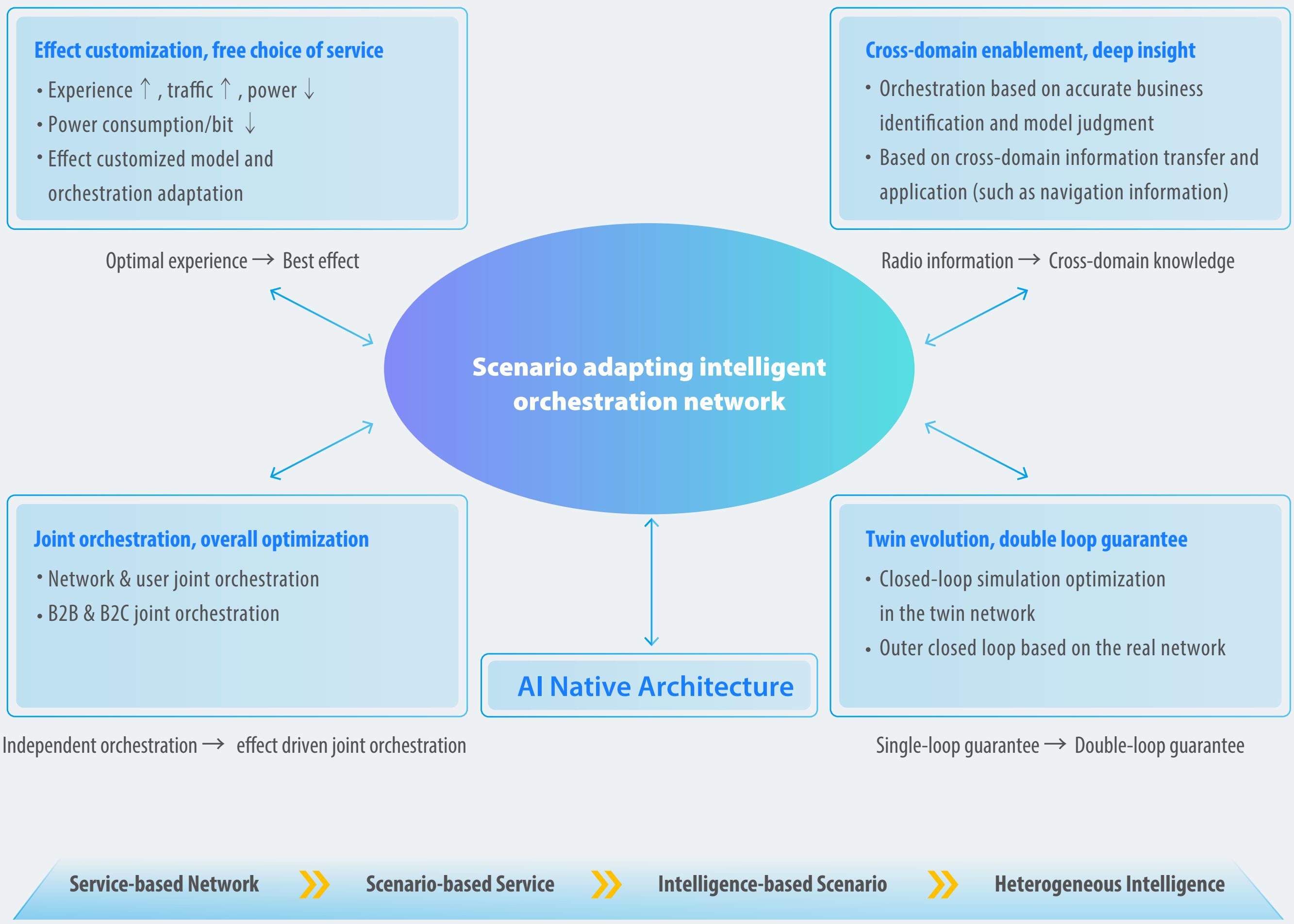
Experience-driven1 → intention-driven:

Aiming at the different value demands in different stages and different scenarios of network development, the capabilities of intention customization and orchestration are enabled to accurately match the network service objectives.

Inner-domain orchestration → cross-domain enablement:
From experience-driven single-domain orchestration to cross-domain orchestration based on knowledge-domain enablement, diversified intelligent service capabilities evolve from user experience to network value.

Independent orchestration → joint orchestration:
The joint optimum status driven by the intent network is realized through the deep synergy of network orchestration at the macro level and user orchestration at the micro level. Based on the integrated orchestration architecture of B2B and B2C and the evolution of multi-objective intelligent resource management capability, the optimal unification of service differentiation and capability consistency is achieved.

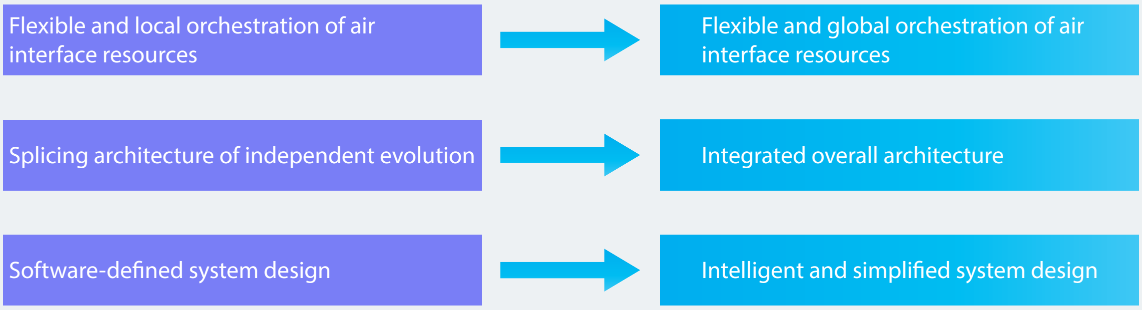
Single-layer closed loop → double-layer closed loop:
Relying on the high-precision digital modeling and simulation capability of digital twinning, efficient search of the best orchestration, effect prediction, and the closed-loop optimization of the virtual network are realized. Based on the optimal solution provided by the digital twin and the closed-loop optimization of the physical network, the optimal orchestration capability is provided at the lowest cost and the minimal risk.



Evolution of intelligent orchestration network

05
Paradigm Shift

The above technologies include some basic technical trends, that is, increasing computing capabilities, especially the bottom-layer computing capabilities, to improve resource utilization efficiency. With the development and popularization of 4G, 5G, and subsequent 6G, spectrum resources are becoming more and more scarce, especially the shortage of "good" spectrum resources. The single-link Shannon limit and noise limits are being approached. As a result, various SDMA technologies, duplex technologies, and non-orthogonal technologies have become necessary ways to further improve resource utilization. Due to low carbon and energy consumption, all of the above technologies impose higher requirements for hardware and chip capabilities. Therefore, the chips with higher integration, higher process, stronger computing capability, and lower power consumption are the key to next-generation mobile communications.



Shift of B5G Network Paradigm

The B5G network is emerging from the traditional network paradigm. In terms of capability enhancement, the local orchestration of air interface resources is migrating to global orchestration. In terms of boundary extension, the architecture is transiting from a combination of independent architectures to an integrated architecture. In terms of efficiency improvement, the focus of system design is changing from "software-defined" to intelligence and simplicity.

In terms of capability enhancement, the global orchestration of air interfaces changes the traditional pattern of orchestration within only local areas. Moreover, the frequency domain (full-band integration), air domain (enhanced Massive MIMO, smooth virtual cell), radio frequency domain (full duplex), UE domain (multi-UE virtual aggregation), or a combination of them (e.g., the smart repeater with a combination of air domain and radio frequency domain), are enhanced in an all-round manner.

In terms of boundary extension, the traditional network architecture is the combination of architectures with independent evolution in each network domain (e.g., access network domain, core network domain, and transport network domain), each capability aspect (communications and perception), or each deployment scenario (e.g., air, satellite, and ground). Instead, the integrated architecture takes all aspects into consideration, for example, the integration of ICT information infrastructure, ISAC, and space-air-ground integrated networks, end-to-end connection (e.g., B5G support for the meta-universe and TSN), and endogenous design (e.g., intelligence and security native).

In terms of efficiency improvement, intelligent and simplified system design, based on software-defined system design focuses more on using intelligence to simplify system O&M, and improve O&M efficiency, air interface efficiency, and power efficiency.