ZTE

Digital Infrastructure Technology Trends White Paper

Digital Infrastructure Technology Trends White Paper

Version	Date	Author	Remarks
V1.0	2023.05	ZTE	

Contents

1. Foreword **5**

2. Requirements and Challenges for Digital Infrastructure Technology **7**

2.1. Requirements and Constraints of Future Digital Infrastructure Technology 7

2.2. Challenges to The Traditional Technology Development Path 9

2.3. Overview of Future Technology Development Path 11

3. Connectivity **13**

3.1. Overview 13

3.2. Physical Layer (wireless): 5G-A&6G Requires More Spatial Multiplexing and Extended Frequency Bands 15

3.3. Physical Layer (optical): Single-wavelength rate improvement, Band Extension, and Spatial Multiplexing 18

3.4. Packet Layer: Packet Forwarding Chip Architecture That Takes into Account Both Capacity and Flexibility 19

3.5. Application Layer: Video Compression Efficiency is Further Improved with Neural Network-based Video Coding 21

3.6. Interconnection: Replacement of Electrial to Optical 22

4. Computing Power **24**

4.1. Overview 24

4.2. Chip Architecture: DSA&3D Stacking&Chiplet 25

4.3. Computing Architecture: The Integration of Computing and Storage 26

4.4. Computing Architecture: Peer-to-peer Computing 28

4.5. Network Architecture: The IP Network Technology That Supports The Convergence of Computing and Networks 29

5. Intelligence..... 32

5.1. Overview..... 32

5.2. AI Chip: Increase Computing Power/Energy Ratio..... 32

5.3. AI Algorithm: Evolution from Dedicated Small Models to General Large Model..... 34

5.4. AI for Network automation: Empower Autonomous Network to Higher Level..... 36

6. Conclusion..... 39

7. References..... 41

1. Foreword

Technological innovation is the core driving force behind productivity progress and industrial development. Klaus Schwab, founder of the World Economic Forum, pointed out in his book "The Fourth Industrial Revolution" that since the 18th century, humanity has experienced four industrial revolutions led by technological innovation.

The first industrial revolution began around 1760, marked by the invention and widespread application of the steam engine and railway, which transitioned humanity from manual labor to mechanized production. The second industrial revolution started in the late 19th century with the wide utilization of electricity, ushering in the era of mass production. The third industrial revolution emerged in the mid-20th century, driven by communication technology, computer technology, and the internet (referred to as information and communication technology or ICT), leading humanity into an era of automated production.

The current fourth industrial revolution is a continuation of the third, but with exponentially increasing speed, scope, and impact of technological innovation. It is primarily characterized by digitalization and intelligence, with iconic technologies such as the Internet of Things, big data, and artificial intelligence, progressively shifting society from digital to intelligent.

Efficient digital infrastructure serves as the fundamental cornerstone for a digital and intelligent society. New applications such as industrial interconnection, holographic communication, meta-universe, and autonomous driving have presented greater demands on information and communication technologies. However, it is important to note that the development of ICT technologies is built upon breakthroughs in mathematics and physics, such as electromagnetism, quantum mechanics, and information theory, achieved from the late 19th century to the middle of the 20th century. In recent decades, advancements in basic sciences have decelerated, posing formidable challenges for future technological progress. The traditional technological evolution route faces limitations imposed by Moore's Law, Shannon's Theorem, and carbon emission reduction. Therefore, there is a pressing need for fundamental innovations in basic theory, core algorithm and system architecture.

This white paper is an interpretation of the future technology development trends of digital infrastructure, which has been jointly prepared by ZTE's Technical Expert Committee. In contrast to typical industry white papers that focus on business models, application visions, and technology requirements, this technical white paper focuses more on the challenges confronted by technology development and the path towards technological realization to overcome them.

Chapter 2 outlines the technical requirements for future business scenarios, and proposes three key technical elements of digital infrastructure: Connectivity, computing power, and intelligence. However, the development of these three technical elements is facing the challenges posed by the Shannon limit, the slowdown of Moore's law, and insufficient cognition of the nature of

intelligence, which poses significant challenges to future technological advancements.

Chapters 3 to 5 describe the specific technical trends in three directions: Connectivity, Computing power, and Intelligence. Each technological direction presents the future technology challenges and solutions, alongside ZTE's technological innovations and predictions for future trends.

Chapter 6 provides a summary of the entire white paper along with some thoughts on how digital infrastructure capabilities can better serve all industries.

ZTE's technological innovation aligns with the technological development trend and industrial requirements. We will continue to work with our industry partners to promote technological innovation and contribute to the progress towards a digital and intelligent society.

2. Requirements and Challenges for Digital Infrastructure Technology

2.1. Requirements and Constraints of Future Digital Infrastructure Technology

Since the invention of the Morse code and telegraph in 1837, the development of ICT technology has significantly transformed human lifestyles and production methods. The scale of the global digital economy continues to rise. In 2021, the added value of the digital economy in 47 major countries worldwide reached 38.1 trillion US dollars, reflecting a nominal increase of 15.6% compared to the previous year, constituting 45.0% of the GDP[01] . In 2022, China's digital economy reached a scale of 50.2 trillion yuan in size, experiencing a year-on-year growth of 10.3%, surpassing the nominal GDP growth rate for 11 consecutive years[02].

Efficient digital infrastructure is a fundamental and essential capability of the digital economy. In the field of ToC and ToH, the outbreak of applications such as short video and live broadcasting, and the popularity of online education and remote working impose greater demands for network bandwidth and coverage; In the field of ToB, the in-depth expansion and integration from ICT to OT (production domain) also puts forward increased expectations regarding network performance, economic viability, security, and reliability.

Digital infrastructure encompasses three basic elements: connectivity, computing power, and intelligence.

"Connectivity" is the core feature of the Internet. Connection rates have increased from about one character per second in the initial telegraph to the current rate of "dual gigabit" access (that is, both wireless access and optical access reach gigabit) and dozens of Tbps for a single optical fiber in the backbone network infrastructure. Wireless communication networks are typically upgraded approximately once every ten years, increasing the rate by 10 times. For 2030 (6G), with the emergence of new services, including holographic communication and meta-universe, there is an expectation that the demand for connectivity in business will continue to increase by 1-2 orders of magnitude compared with the current technology(5G)[04].

In the digital society, "computing power" has become as essential as water, electricity, and gas. According to the assessment by IDC&Inspur&Tsinghua, the increase of computing power index by 1 point will increase the digital economy and the GDP by 3.5‰ and 1.8‰ respectively[05] . According to CAICT(China Academy of Information and Communications Technology), the total computing power of global computing equipment reached 615 EFlops in 2021, and is expected to reach 56 ZFlops in 2030, with an average annual growth rate of 65%[06].

With the breakthrough of deep neural network algorithm in the past decade, artificial intelligence

technology has become a driving force for societal advancement to move from digital to intelligent. The premise of digitization is to represent the physical world with mathematical models. Prior to the breakthrough of AI technology, there were a large number of complex systems in the real world that could not be represented by mathematical models. The essence of deep neural network is to use large-scale interconnected neural nodes to approximate the mathematical models of various complex systems (such as human cognitive systems or highly nonlinear physical systems), which greatly expands the breadth and depth of digital applications.

It is evident that connectivity, computing power, and intelligence are the fundamental technical requirements for future digital applications. The foundation of the future digital society lies in integrated computing-networking infrastructure and intelligent service systems. Under the influence of the data deluge, connectivity, computing power, and intelligence are mutually complementary, reflecting a closer relationship and a less distinct boundary.

Figure 2.1 illustrates the relationship between various application scenarios and the three technical elements in the future. These scenarios are derived from the future network application scenarios proposed by the ITU focus group FG-Net2030 in June 2020[07][08].

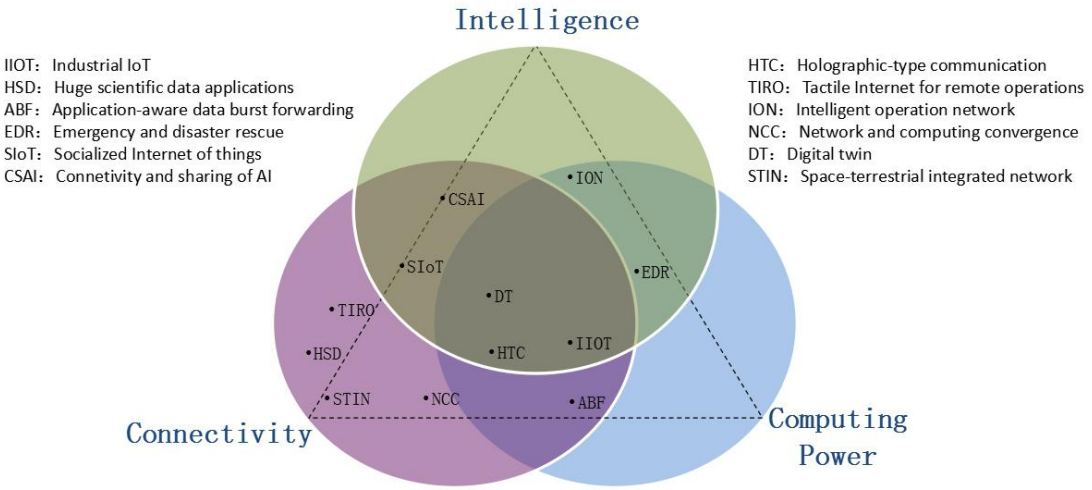

Figure 2.1 Mapping Between Future Scenarios And Three Technical Elements

At the same time, the goal of sustainable development imposes higher requirements for energy conservation and environmental protection. Information flow has the potential to enhance the efficiency of logistics and energy utilization, thereby reducing the overall carbon emissions associated with human activities. For instance, SMARTer 2030, a report by GeSI (Global Enabling Sustainability Initiative), indicates that ICTs (Information and Communication Technologies) are projected to contribute to a 20% reduction in global carbon emissions by 2030[09]. However, it is crucial to address the carbon emissions generated by the ICT industry itself. According to the aforementioned GeSI report, the carbon emissions from the information and communications industry are expected to account for 1.97% of global carbon emissions by 2030. Hence, future technological advancements must prioritize energy conservation and emission reduction as critical constraints.

2.2. Challenges to The Traditional Technology Development Path

From the late 19th century to the middle of the 20th century, human breakthroughs in electromagnetic, quantum mechanics, information theory and other scientific theories are the basis of modern information and communication technology. The three major elements of ICT, connectivity, computing power and intelligence, have their own development paths, but also show mutual support and synergistic progress.

The main technology to improve communication data rate is to develop better algorithms (modulation and demodulation, shaping and compensation, forward error correction, etc.) to approach the Shannon limit. Advanced algorithms bring the increase of computational complexity, and must rely on the progress of microelectronics technology to get stronger digital signal processing capabilities.

The microprocessor developed in accordance with Moore's Law has improved its performance by more than 1 billion times in the past fifty years, and the progress in semiconductor technologies has also driven the development of other chips, including digital signal processors (DSPs), network processors, and switching chips for communication. Advances in chip technology have enabled more complex communication algorithms.

AI algorithms have an unprecedented demand for increasingly powerful chips, and distributed AI computing puts forward high requirements for network bandwidth and delay.

In turn, the increase of bandwidth requires the support of AI technologies, such as physical layer optimization and lossless network parameter optimization.

It can be seen that the development of connectivity, computing power, and intelligence relies on the support of other factors. and conversely, the stagnation of any technological direction will affect the development of other technological directions.

The future paths of these three technology elements all face their own difficulties.

(1) The communication algorithm is approaching the Shannon limit.

The Shannon theorem ($C/W=log_2(1+S/N)$) reveals the relationship between the spectral efficiency (the maximum data rate that can be transmitted per unit bandwidth) and the signal-to-noise ratio. In communication practice, the minimum SNR tolerance is often determined by the pre-defined data rate and channel width. This SNR tolerance represents the limit of the signal quality required to achieve error-free transmission.

In 2001, a research paper pointed out[10] that the LDPC coding algorithm used in wireless communications achieves the SNR tolerance 0.31 dB, while the Shannon limit in this scenario is 0.18 dB, resulting in a difference of only 0.13 dB. This means that the actual signal-to-noise ratio tolerance is only 3% higher than the Shannon limit ($10^{0.013}\approx1.03$).

Furthermore, based on test data from ZTE, the current optical transmission algorithms using 4~16QAM modulation achieve a signal-to-noise ratio tolerance that is approximately 1 dB away from the Shannon limit. This implies that future algorithms can only increase the transmission distance by 25% or improve the spectral efficiency by 0.33 bps/Hz.

As the system approaches the Shannon limit, the performance benefits of increasing algorithm complexity become diminished. In many cases, several times the computational workload is required to achieve only a marginal improvement in performance. Therefore, even if future algorithms continue to approach the Shannon limit, their demand for computing power will far exceed the levels achievable through Moore's Law.

(2) Microelectronics is approaching the boundaries set by physics.

Moore's Law, which represents the advancement of microelectronics, is also encountering greater difficulties.

Prior to the 28nm process, the industry increased the number of transistors per unit area by reducing transistor size, such as gate length. However, transistors can not continue to shrink in size due to quantum tunneling, parasitic capacitance and other issues (a silicon atom is 0.2 nm in diameter, and the gate length of 20 nm is only about 100 silicon atoms). Therefore, innovative transistor structures, such as FinFET and GAA, have been introduced. However, these complex structures come with higher costs and power consumption, which become limiting factors for advancing process nodes.

It appears that the technological benefits derived from mankind's fundamental theoretical breakthrough in the microscopic world (such as quantum mechanics) have nearly reached their limits. The current technological revolution primarily utilizes quantum phenomena from a macro statistical perspective, while the observation and manipulation of physical matter still rely on macroscopic means such as current, voltage, and light intensity.

In order to fully release the potential of quantum, it is necessary to carry out accurate control and observation of microscopic particles such as photons, electrons and cold atoms and their quantum states. Scientific research in this area is still in its initial stage, and there are significant uncertainties regarding the future development path, methods, and goals.

(3) Intelligence technology lacks the guidance of cognitive science

Research on artificial intelligence began in the 1950s, shortly after the birth of computers, and the real breakthrough came after the success of deep neural networks in 2006. However, the neural network-based artificial intelligence algorithm is a superficial simulation of the physiological structure of the human brain. The deep working mechanism of human intelligence, which falls within the scope of cognitive science, has not yet been fully understood or achieved a significant breakthrough.

Current deep learning technology heavily relies on large-scale computing power and data.

However, in light of the slowdown of Moore's Law and the increasing need for energy conservation and emission reduction, this technology path is difficult to sustain in the long run. At present, the growth rate of computing power for artificial intelligence is much higher than that of Moore's Law. Particularly with the emergence of large models like Transformers, the growth rate of computing power required for training models increases to 275 times every two years on average, significantly exceeding the 2x growth rate of Moore's Law[11]. It is estimated that AI will consume approximately 15% of the world's electricity in the next decade, placing a substantial burden on the environment.

In general, the development of ICT has approached the boundaries set by three fundamental theories: Mathematics (Shannon's theorem), physics (quantum mechanics), and cognitive science. Each step forward requires greater resources than ever before, posing a major challenge for the current trajectory of technological evolution.

2.3. Overview of Future Technology Development Path

How to break through technical bottlenecks and build a digital foundation of "connectivity & computing power & intelligence" is a major task facing us now.

Chapters 3 to 5 of this white paper outline potential paths for future technological development from three perspectives: connectivity, computing power, and intelligence..

In his book "The Nature of Technology", American thinker Brian Arthur proposes that the nature of technology is the collection of phenomena captured and utilized. Technology evolution is similar to biological evolution, and is a process of combinational evolution. A new technology is a new combination of existing technologies. We believe that in the future, in addition to exploiting the potential of the existing technology, another promising path lies in the collaboration of multiple technologies and the optimization of system architecture.

The architecture of ICT systems, be it computing or network architecture, is characterized by modularity, layering, and decoupling. For example, the Von Neumann computing architecture separates computing and storage, while network architecture employs protocol layering and interlayer decoupling. The advantage of separation and decoupling is that each module develops independently, facilitating innovation and maintenance. However, achieving optimal performance for specific services often requires the collaboration and fusion of modules when a single module encounters a performance bottleneck. This collaboration and fusion can lead to performance improvements and reduced power consumption.

In the following chapters, we describe potential exploitation of existing technical path, for example, new spectrums and channels are developed in wireless and optical communications. There are also coordination and integration of multiple technologies, such as optical-electrical integration, computing-memory integration, computing-network convergence, etc.

Table 2-1 provides an overview of the technical development path in three directions: connectivity,

computing power, and intelligence.

Table 2-1 Overview of Future Technology Development Path

	In-depth exploitation	Cordination and Integration
Connectivity	• Improve the spectrum efficiency to the Shannon limit. • Spectrum band extension • Space division multiplexing	• Optical-electrical integration; • Innovative Packet Forwarding Chip Architecture
Computing power	• More Moore: Continue to pursue higher transistor density with innovations on transistor structure	• Integration of computing, memory, and network • Peer-to-peer distribution system
Intelligence	• AI chip architecture innovation: Higher computing power/energy consumption ratio • AI algorithm evolves from the diversified small separation model to the general large model.	• Intelligent capabilities empower digital infrastructure, industries, and enterprises.

3. Connectivity

3.1. Overview

Improving connection bandwidth is a key objective of information and communication technology. Currently. Both wireless access (5G) and wired access (10G-PON) are capable of providing users with "dual-gigabit" access bandwidth. Furthermore, long-distance 400G single-wavelength transmission technology is being deployed in backbone optical networks. As described in Chapter 2, in the next 5 to 10 years, the demand for bandwidth is expected to increase by one to two orders of magnitude.

Enhancing network bandwidth involves not only improving the physical-layer transmission capacity but also enhancing data processing capabilities at the packet and application layers. Additionally, the bandwidth of the interconnection between racks and devices needs to be improved accordingly.

(1) Physical layer

The physical layer is based on electromagnetic theory, and the electromagnetic field expression is shown in the following formula:

$$\overrightarrow{E(x,y,z,t)} = \underbrace{\overrightarrow{e_p}}_{\text{Polarization}} \cdot \underbrace{F_d(x,y)}_{\text{Spatial distribution}} \cdot \underbrace{|A_m}_{\text{Amplitude}}\underbrace{(T_S)|}_{\text{Symbol period}} \cdot \underbrace{exp[j\varphi_n}_{\text{Phase}}(T_S)] \cdot \underbrace{exp[j(\omega_k}_{\text{wavelength}} t - kz) + \varphi_0]$$

According to the formula, there are five dimensions that can be multiplexed in communication: polarization, spatial distribution, amplitude + phase(QAM, Quadrature Amplitude Modulation), wavelength, and symbol period (baud rate). The polarization, QAM and baud rate are related to the single-wave rate. Therefore, the total transmission rate can be written as:

$$\text{Total transmission rate} = \underbrace{D}_{\text{Spatial multiplexing}} \bullet \underbrace{K}_{\text{Wave channels}} \bullet \underbrace{R_s}_{\text{Single-wave rate}} \leq D \bullet K \bullet \underbrace{B}_{\text{Single-wave bandwidth}} \bullet \underbrace{\log_2(1 + SNR)}_{\text{Shannon's theory}} = \underbrace{C}_{\text{Total capacity limit}}$$

Note: This formula is a simplified representation. The wireless spatial multiplexing is much more complicated.

The formula indicates that transmission capacity can be increased by enhancing the single-wave rate, expanding the bandwidth, and implementing spatial multiplexing. However, the single-wave rate is limited by Shannon's theorem. To improve the single-wave rate, on one hand, spectral efficiency can be enhanced through high-order modulation, polarization multiplexing, and other technologies, approaching the Shannon limit. On the other hand, the single-wave bandwidth can be increased by raising the baud rate. Band extension involves expanding the frequency band available for communication, while spatial multiplexing aims to increase the number of channels, resulting in a significant capacity boost.

Table 3-1 is a brief summary of the foregoing five dimensions in wireless communications and optical communications. For more detailed information, refer to Section 3.2 and Section 3.3.

Table 3-1 The status quo and future development of wireless and optical

Five Dimensions	Three Technical Methods	Wireless	Optical transmission (long distance)
Amplitude and Phase (QAM)	Single-wave rate	● Status quo: 1024QAM has been standardized, but has not been put into commercial use yet. ● Trend: Higher Modulation Order, Increasing Constellation Shaping Gain, and Coding Modulation Joint Optimization	● Status quo: Coherent 4~16QAM modulation is close to the Shannon limit. Baud rate 64~128GBd ● Trend: Continue to increase the baud rate, and improve the SNR by using the new optical fiber/amplifier
Polarization			
Baud rate			
Wavelength	Band extension /spectrum efficiency improvement	● Status quo: 200 MHz carrier aggregation; Sub-band full-duplex is being standardized. ● Trend: CA, full-duplex technology, millimeter wave/Terahertz	● Status quo: The 12THz spectrum of the C+L bands supports 80 waves *400 G/wave. ● Trend: Extension to the S+C+L bands
Space	Space division multiplexing	● Status quo:The 64TR/16 stream has been put into commercial use. NCR is standardizing ● Trends: eMIMO/Beam, distributed MIMO, ultra-large aperture ELAA, Cell-free, RIS, NCRs, etc.	● Status quo: Not in commercial use ● Trend: Multi-core fiber and few-mode fiber; Multi-core weak coupling may be commercialized first

(2) Packet Layer

Since the birth of the Internet, the packet technology represented by IP and Ethernet is the core of the network domain. The packet processing capability of network devices often becomes a bottleneck for improving network capacity and performance. Effective packet processing requires a balance between capacity and agility. The performance of packet processing depends not only on the progress of the chip technology, but also on the improvement of the packet processing chip

architecture. Section 3.4 describes the evolution of the packet forwarding architecture in the future.

(3) Application Layer

Efforts to improve the video compression ratio are closely linked to communication capacity. With the advancements in applications such as XRs and holographics, it is expected that video traffic will account for over 90% of the total Internet traffic by 2030. Section 3.5 describes the utilization of deep learning in video coding technology.

(4) Interconnection

With the increase in link bandwidth and port density, the interconnection buses of ICT devices may become a bottleneck. Optical interconnection offers significant advantages over electrical interconnection in terms of performance and power consumption. As CPO (Co-Packaged Optics) technology continues to mature, the trend of “optical replacing copper” may emerge within devices as well. Please refer to Section 3.6 for more details.

3.2. Physical Layer (wireless): 5G-A&6G Requires More Spatial Multiplexing and Extended Frequency Bands.

Since the 1980s, mobile communications have gradually evolved from 1G to 5G. Currently, 5G has been widely deployed worldwide, while the development of 6G is underway.

As discussed in Chapter 2, in response to future service requirements, 6G aims to achieve significant improvements of 1-2 orders of magnitude compared to 5G in core features such as bandwidth, delay, and reliability. The initial version of 6G by 3GPP is expected to be released in 2030. Prior to that, there will be three to four versions of enhanced 5G technology known as "5G-Advanced".

Regarding spectral efficiency, while low-order modulation and medium-order modulation have approached the single-link Shannon limit, there is still a gap to be bridged at high-order modulation. Additionally, 6G will focus on increasing bandwidth, improving bandwidth utilization, and enhancing spatial multiplexing capabilities. This includes technologies such as carrier aggregation, full-duplex transmission, utilization of higher frequency spectrum (beyond 6G hertz and terahertz range), Non-Orthogonal/Orthogonal Frequency Division Multiplexing (OFDM) and its variations, high-frequency waveform and sensing waveform, massive MIMO and extremely large-scale MIMO, Reconfigurable Intelligent Surfaces (RIS) technology, Network Controlled Relays (NCRs), and more. These technologies represent a fundamental trend wherein increasingly powerful computing capabilities are leveraged to achieve better resource utilization efficiency.

Several typical technologies are described below.

(1) **Higher-order modulation/constellation shaping/code modulation scheme**

Currently, modulation schemes can reach up to 1024QAM, allowing each symbol to carry 10 bits. To further enhance spectral efficiency, the modulation order may be increased to 4096QAM or even higher. However, in high-order modulation modes, the efficiency of the traditional square QAM constellation may not be optimal, leading to a situation where the higher the spectral efficiency, the further it gets from the Shannon limit. Therefore, higher-order modulation techniques based on geometric shaping or probabilistic shaping are expected to approach the Shannon limit more closely, especially in areas with high signal-to-noise ratio (SNR).

(2) Improving spectrum efficiency: Full duplex and subband full duplex

Full-duplex is a new technology that improves network data rates and spectrum utilization. For high-bandwidth and low-delay services in the future, full-duplex uses unpaired spectrum resources. By releasing mutually exclusive restrictions on the use of DL/UL resources, spectrum usage efficiency can be increased and transmission delay can be reduced. However, to implement full duplex, a base station or a terminal needs to process self-interference (SI) to support a transceiver function that is simultaneously performed. Implementation complexity and a hardware cost are still relatively large, especially for a Massive MIMO transceiver. Therefore, the multi-antenna technology is actually mutually exclusive with the full-duplex technology.

Current research primarily focuses on models with a relatively small number of antennas and subband full-duplex, where separate frequencies are allocated for uplink and downlink resources. This approach allows for flexible configuration of more uplink resources, thereby reducing uplink and downlink delays and improving uplink coverage and capacity. Although subband full-duplex reduces the requirements for interference cancellation capabilities at the base station, mutual interference between user equipment (UE) remains a significant challenge that necessitates industry-wide collaboration.

(3) Expand More Spectrum: Terahertz Technology

As a potential 6G basic technology, THz refers to 100 GHz~10THz frequency band resources with continuous available large bandwidth. It will help build a 6G short-distance and high-rate transmission system.

However, terahertz technology does have certain drawbacks. Compared to millimeter waves, terahertz frequencies experience significant propagation path loss, and outdoor communication is also susceptible to additional loss due to rain and fog. Moreover, limitations such as low power output of transmitter power amplifiers, high noise coefficients of low noise amplifiers, and challenges in designing and manufacturing high-gain antennas greatly restrict the transmission range of terahertz waves.

Terahertz technology can be combined with multi-antenna systems, enabling the use of extremely narrow beams to mitigate path fading and extend propagation distances. Additionally, the application of reconfigurable intelligent surfaces (RIS) in the terahertz frequency band is a future development trend, where the dense distribution of RIS both indoors and outdoors can have a

positive impact on terahertz coverage.

(4) More Space-Division Multiplexing: Extremely Large-Scale Antenna and Distributed MIMO

Extremely large-scale antennas can effectively enhance the uplink capacity and the coverage of new frequency bands.

For emerging industrial Internet applications, such as machine vision in modern factories, throughput requirements in the order of Gbps or 10 Gbps are necessary. Potential solutions include increasing the number of antennas or MIMO layers to support more uplink connections in the NR (5G air interface), enabling more users with MU-MIMO, and introducing more flexible carrier distribution and aggregation. The 5G-Advanced supports up to 24 orthogonal demodulation reference signal (DMRS) ports, allowing support for up to 24 users in common time-frequency resources if each user employs single-stream uplink transmission. Additionally, 5G-Advanced supports more powerful uplink terminals, with a single user capable of supporting up to 8 streams. This greatly improves peak rates and effectively enhances uplink throughput, particularly in dense network deployments.

The future trend in air division multiplexing emphasizes higher levels of distribution and larger equivalent apertures. It progresses from systems like MTP/eCoMP with a small number of access points (APs) to larger-scale heterogeneous distributed MIMO, and further evolves into cell-free networks with extensive AP scales. Large-scale distributed MIMO systems must address challenges such as time-frequency synchronization, forward bandwidth, and AP power supply.

(5) Improving Channel Coverage Quality: Reconfigurable Intelligent Surfaces (RIS)

Reconfigurable Intelligent Surfaces (RIS) is a wireless environment optimization technology characterized by its low cost, low energy consumption, high reliability, and large capacity. RIS enhances the coverage, throughput, and energy efficiency for users at the cell edge through the following approaches:

a. Provide effective reflection propagation paths to avoid coverage holes when the direct propagation paths are blocked.

b. Implements beamforming for target users, and makes full use of space diversity and multiplexing gains.

c. Zero-point beamforming is implemented for the interfered UEs to implement inter-cell interference suppression.

In essence, RIS is a distributed spatial multiplexing technology. In comparison, Massive MIMO is a centralized spatial multiplexing technology. Due to its low cost, RIS is easy to be deployed on a larger scale.

3.3. Physical Layer (optical): Single-wavelength rate improvement, Band Extension, and Spatial Multiplexing

High speed, large capacity and long distance are the most important requirements for optical transmission. The current 200G PM-QPSK (Polarization multiplexing four phase shift keying) system using Super C (ultra-wide C-band) has been widely commercially deployed, and 400G PM-QPSK is expected to be commercially available in 2023.

As discussed in Section 3.1, communication capacity can be enhanced through three technical approaches: single-wavelength rate improvement, band extension, and space division multiplexing. Single-wavelength rate improvement is the most cost-effective method for expanding capacity, while new band expansions, such as the L-band and S-band, effectively double the available spectrum. Furthermore, space division multiplexing has the potential to significantly increase the capacity of a single fiber.

(1) **Single-wave rate improvement**

According to Shannon's theorem, the way to increase the single-wave rate involves increasing both the spectral efficiency and the bandwidth/baud rate. Improving spectral efficiency requires higher signal-to-noise ratio (SNR) at the receiving end. The use of new fibers (such as G.654 fiber and hollow fiber) can reduce loss and nonlinearity, combined with amplifiers to reduce noise factors, thus supporting a doubling of the capacity of a single channel. Advanced coherent DSP chips employ high-performance modulation and demodulation techniques, along with high coding gain Forward Error Correction (FEC) coding. This approach allows the SNR tolerance to approach the theoretical value and the transmission rate to approach the upper limit of the channel capacity.

Regarding the improvement of single-wave bandwidth, it is necessary to enhance the bandwidth of chips and optical devices. This enables an increase in baud rate from 64GBd to 96GBd/128GBd, and will continue to evolve towards 180GBd+.

(2) Band extension

Band extension is the primary approach to increasing the capacity of single-mode fiber. Adhering to the principle of increasing single-wave speed without reducing the number of waves and doubling the capacity, the C4T, C6T, and C6T+L6T bands are employed for long-haul modes of 100G, 200G, and 400G, respectively. Currently, the commercialization of long-haul 400G relies on the C+L band. The next direction for capacity improvement will be the long-haul 800G, combined with the expansion of the S+C+L band.

Band extension relies on material technology for new band optical devices that can support a broader range of wavelengths. Examples include amplifiers utilizing Tm/Bi ion or substrate doping processes, the use of 128GBd+ TFLN (thin film lithium niobate) coherent modulators in optical modules, multi-band external cavity technology in ITLA (tunable laser), and multi-band

anti-reflection coating design in WSS (wavelength switching) devices.

(3) Space division multiplexing

Through space division multiplexing technology, which involves increasing the number of fiber cores and transmission modes, the capacity of a single fiber can be greatly improved. This technical approach can be categorized into multi-core weak coupling, multi-core strong coupling, few-mode weak coupling, and few-mode strong coupling. Among them, multi-core weak-coupling fibers/devices are relatively mature and capable of long-distance transmission. Due to their advantages in energy consumption and size/density, multi-core fibers show more promise in submarine cable applications. Few-mode weak-coupling fibers have limited transmission distance and may be used in data center interconnects (DCI). However, multi-core strong-coupling fibers and few-mode strong coupling fibers are not expected to be practical in the near future.

3.4. Packet Layer: Packet Forwarding Chip Architecture That Takes into Account Both Capacity and Flexibility

We believe that in the next ten years, the forwarding capability of packet chips will continue to be crucial for improving network bandwidth. Currently, industry has released chips with a processing capability of 51.2Tbps. Following the trend of doubling chip capability every 2 to 3 years, it is estimated that the processing capability of a single chip will reach 102.4Tbps by 2025 to 2026. By 2030, the maximum processing capability of a single chip is expected to reach 204.8Tbps.

Simultaneously, over the next ten years, packet chips will need to enhance their flexible service processing capabilities. This involves strengthening chip programmability to accommodate the innovation of new services and reducing chip forwarding delay to meet the low-latency requirements of emerging scenarios such as digital twins and metaverses. Based on these business needs, we believe that future chips will not only rely on technological progress but also require innovations in architecture design and algorithms.

There are currently two mainstream programmable forwarding architectures: (1) Parallel RTC (Run To Complete) architecture; (2) Serial pipeline architecture.

The parallel RTC architecture offers large capacity tables, a vast instruction space, and the ability to process complex services. However, this architecture has a higher forwarding delay and cannot meet the requirements of low-latency services. On the other hand, the serial pipeline architecture has relatively lower delay and deterministic jitter, but it has smaller forwarding tables and limited programming capability, making it unsuitable for processing complex services.

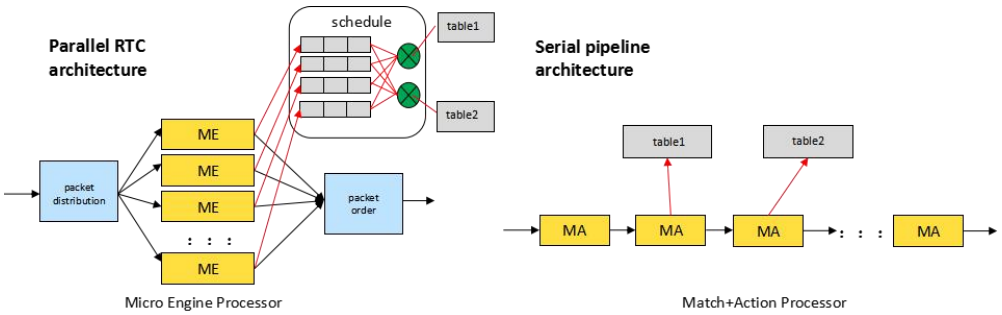

Figure 3.1 Parallel RTC Architecture And Serial Pipeline Architecture

We propose a new hybrid forwarding chip architecture that combines both parallel and serial architectures. This architecture dynamically allocates services with different characteristics to the appropriate forwarding architecture through orchestration, thereby meeting the requirements for future network performance, delay, and service expandability.

In low-latency scenarios, all low-latency services are processed by the serial pipelines, while a few complex services, such as those involving large-capacity forwarding table searches, are processed in parallel using the Run To Complete (RTC) architecture. In this scenario, since the services processed by the RTC are relatively simple and require fewer instructions, the architecture can still ensure a relatively low processing delay.

For other scenarios, such as general-purpose router scenarios, the services that need to be processed by the chip are highly complex and involve the search of multi-level large-capacity entries. In such cases, it becomes necessary to utilize the parallel RTC architecture to effectively address the limitation of programming capabilities in purely serial pipelines.

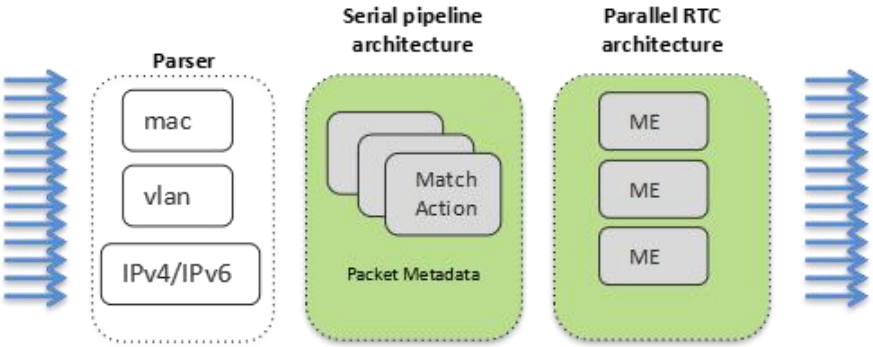

Figure 3.2 Parallel and Serial Hybrid Architecture Of the Packet Forwarding Chip

Based on our technical evaluation, if an appropriate service arrangement model is selected, the delay of the hybrid architecture is basically equivalent to that of the serial pipeline architecture in terms of forwarding delay, and is about 40% lower than that of the RTC parallel architecture.The chip area of the hybrid architecture is about 15%-20% less than that of the RTC parallel architecture and the serial pipeline architecture.The power consumption of the hybrid architecture is about 12%-20% less than that of the RTC parallel architecture and the serial pipeline architecture. We believe that the serial-parallel hybrid architecture can not only reduce the cost of chip development, but also significantly shorten the development and deployment cycle of new

functions, and quickly adapt to the continuously changing business needs.

3.5. Application Layer: Video Compression Efficiency is Further Improved with Neural Network-based Video Coding.

Video traffic has accounted for over 70% of Internet traffic in 2020. The pursuit of improved video content encoding quality and compression efficiency, while maintaining the same visual quality, is a key driving factor in video technology development. Video encoding aims to enhance compression efficiency within an acceptable range of information loss, thereby reducing video transmission bandwidth requirements. This represents another approach to address the limitations imposed by Shannon's theorem.

The Joint Video Experts Team (JVET), jointly established by ISO/IEC JTC1 SC29 and ITU-T SG16 VCEG, released the video coding standard H.266/VVC (Versatile Video Coding) [14] in August 2020. Under the traditional hybrid coding framework, H.266/VVC adopts predictive coding, transform coding and entropy coding techniques to reduce redundancy in the domain of spatial, time, frequency, inter-component and human visual perception. Compared with H.265/HEVC, H.266/VVC can achieve about 50% bitrate savings under the same visual quality.

However, the complexity of video coding algorithms is inevitably increasing with the more fine-grained block partition methods and coding modes, and with the more complex prediction and transformation technologies. It is becoming clear that it’s difficult to further improve the video compression efficiency only with traditional coding technologies. Deep learning technology has achieved great success in computer vision tasks such as image classification, target detection. In recent years, deep learning technology has defined a new structural paradigm for image/video coding frameworks, and significantly improved the performance of image and video encoder.

Neural Network-based Video Coding (NNVC) technologies mainly include: hybrid video coding technology, which combines traditional video coding and neural network coding, and complete end-to-end neural network video coding technology.

(1) Hybrid video coding technology

Hybrid video coding technology integrates deep neural networks into traditional video coding frameworks to further enhance compression performance. One approach in this category utilizes deep learning strategies to expedite the identification of numerous objects for block partitioning and prediction modes, thereby reducing search complexity and computational overhead. Another approach focuses on non-standard solutions that aim solely to improve compression efficiency, employing techniques such as super-resolution and post-processing filtering. The former involves performing super-resolution operations on the decoded image, resulting in high-resolution and high-quality reconstructions that effectively enhance coding efficiency. The latter endeavors to establish a direct value mapping between reconstructed pixels and original pixels, enhancing the

quality of reconstructed images through filter-based strategies.

(2) End-to-end neural network video coding technology

End-to-end neural network video coding technology leverages deep learning methods to handle the entire encoding and decoding process. By training neural networks on extensive datasets, these models learn the inherent knowledge required to remove video compression artifacts. The superior compression performance of end-to-end neural network video coding can be attributed to its powerful non-linear transformation and mapping capabilities. Furthermore, the end-to-end neural network encoder optimizes the entire coding loop, mitigating the issue of local optima encountered in manual design or independent optimization approaches used in traditional encoders. This overall optimization enhances the coding performance of the system as a whole.

Although neural network-based video coding can improve compression efficiency greatly, the high decoding complexity makes the implementation of such technologies face certain challenges in the short term.. Currently, industry manufacturers are actively studying the joint optimization of traditional video coding technology and neural network-based video coding technology. For example, the Exploration Experiments on Neural Network-based Video Coding (EE1) [15] and the Exploration Experiment on Enhanced Compression beyond VVC capability (EE2) [16] carried out by the JVET, take into account the compression advantages of traditional predictive transformation coding tools and the quality improvement advantages of deep neural network methods. The test results show that under the RA and AI configurations, the BD-rates of Y, Cb and Cr are saved: {-21.17%, -32.29%, -33.05%} and {-11.06%, -22.62%, -24.13%}, respectively, which indicates that enhanced video coding based on neural network has the technical potential to evolve into the next generation of video coding standards.

3.6. Interconnection: Replacement of Electrical to Optical .

Wider connections, for ICT equipment, mean higher interconnection rate and density with lower bit power consumption and bit cost. Optical interconnections have unparalleled advantages over electrical interconnections in terms of capacity and power consumption. Therefore, with the increase of data rate and connection density, the interconnection inside the equipment also shows the trend of optical in copper out.

Optical interconnections offer an additional advantage by greatly expanding the spatial interconnection distance of devices. This means that more switching boards and line cards can be interconnected within a 3-stage CLOS architecture, resulting in a low-cost, low-latency, and low-power solution for high-capacity information and communication equipment[17] 。

CPO (Co-packaged Optics) technology reduces the size of optical engines and co-packaging them with the main chip. It is a crucial technology for using optical interconnects into board-to-board and chip-to-chip interconnections. CPO will result in reduced power consumption, optimized signal integrity, reduced costs, and other benefits. Compared to pluggable optical modules on

panels (FPP), CPO significantly shortens the distance between the main chip and the optical components, resulting in significant cost and power savings.Taking 112G SerDes as an example, when the length of the SerDes PCB is reduced from 1000mm (CEI-112G-LR) to 50mm (CEI-112G-XSR), the power consumption is approximately reduced by 75%.[18]. For CPO in linear links, the elimination of internal DSP allows for even greater reductions in overall cost and power consumption[19].

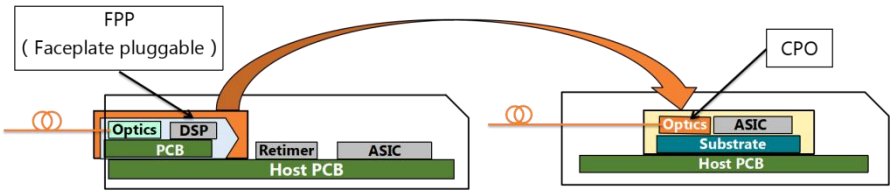

Figure 3.3 Illustration of the evolution from pluggable optical modules to CPO.

The co-packaging of low-power, high-density, and high-capacity CPO represents the future development trend for switch chips. In response to the pressures of power consumption, signal integrity (SI), and cost, industry stakeholders are actively promoting the standardization and industrialization of CPO. It is expected that switches with a capacity of 102.4T will serve as the starting point for large-scale CPO deployments. However, pluggable optical modules on panels (FPP) are continuously evolving and improving through various new technologies. Notably, Linear-drive Pluggable Optics (LPO) has garnered significant attention recently due to its advantages in power consumption and cost compared to non-linear direct-driven pluggable optical modules. However, fully covering the existing scenarios of incoherent optical modules by LPO remains challenging. LPO can be seen as a stepping stone towards CPO technology[21]. In conclusion, CPO and pluggable optical modules will coexist for a considerable period of time.

In HPC/AI networks and equipment, there is also significant pressure in terms of power consumption, cost, and latency. Optical I/O, as a specific form of CPO, is a promising technology in chip-to-chip interconnectivity between computing chips such as CPUs, GPUs, and XPUs. It is expected that under a 200G channel bandwidth in the future, an even lower power consumption of 0.1pJ/bit can be achieved[23]. Due to the recent popularity of ChatGPT, it is anticipated that CPO in the form of Optical I/O will be initially deployed at scale in commercial applications and in HPC/AI networks and devices.

The ultimate goal of CPO development is to achieve monolithic integration of optoelectronics. This aspiration represents the Holy Grail of optoelectronic integration, but it also entails significant challenges.

4. Computing Power

4.1. Overview

With the proliferation of new high-performance computing applications, such as artificial intelligence, privacy computing, AR/VR, and gene testing/biomedical research, the demand for computing power is rapidly increasing. For instance, the computational requirements of AI large models are growing at a pace that exceeds Moore's Law.

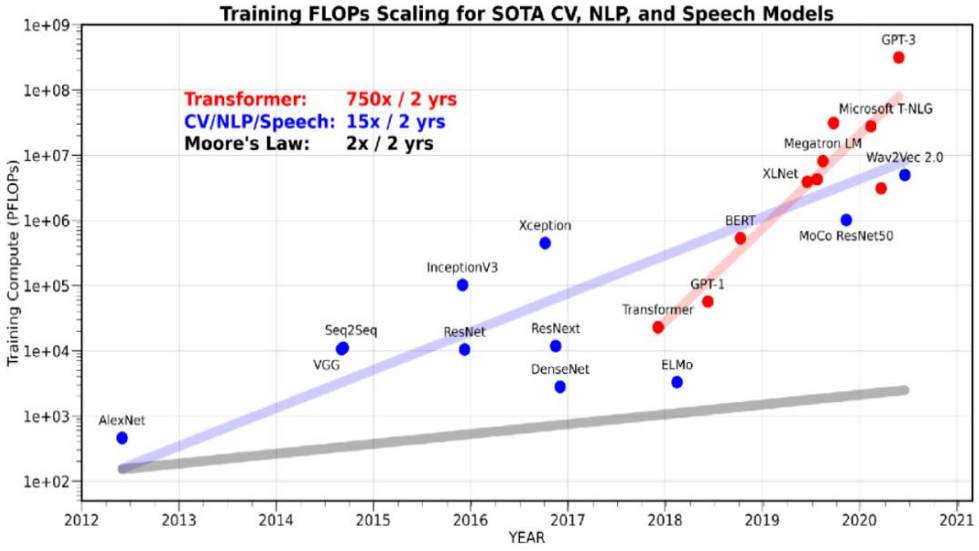

Figure 4.1 The computing power demand of AI large models is growing much faster than Moore's Law[11]

Since the advent of microprocessors, the growth in computing power has followed Moore's Law, which involves increasing the number of gates per chip area to enhance processor performance, while reducing cost and power consumption. However, in recent years, this approach has faced growing challenges. Simply relying on continuous miniaturization for performance improvement is no longer sufficient to meet the demands of modern applications.

In the post-Moore's Law era, continuous innovation in processes and materials provides opportunities to enhance the computing power of chips. There are two primary approaches:

- More Moore: The pursuit of higher transistor density by innovating transistor structures, such as FinFET and GAA. However, this path poses challenges in terms of cost and power consumption.

- Beyond CMOS: Exploring new materials and processes, abandoning CMOS technology. For instance, novel fabrication processes utilizing carbon nanotubes, molybdenum disulfide, and other two-dimensional materials, as well as transistors leveraging the quantum tunneling effect. However, this path is characterized by significant uncertainty and will require substantial time to mature.

On the other hand, architectural innovation plays a crucial role in enhancing computing power density and optimizing resource utilization, thereby enabling the continuation of Moore's Law. This chapter focuses on the following aspects:

- Chip level architectue: Domain-specific optimization through collaborative software and hardware design. Utilization of 3D stacking and Chiplet technologies to reduce chip design and manufacturing costs. (See Section 4.2)
- Computing system level: Introduction of new computing architectures and paradigms, such as computing in memory, to achieve energy-efficient computing. (See Section 4.3) Additionally, the adoption of the "peer-to-peer system" architecture optimizes computing, control, and data paths. (See Section 4.4)
- Network level: Innovations in network architecture and the integration of computing and networking to enhance the efficiency of computing power resource scheduling. (See Section 4.5)

4.2. Chip Architecture: DSA&3D Stacking&Chiplet

In their 2019 book, "The New Golden Age of Computer Architecture," Turing Award winners John Hennessy and David Patterson propose that as Moore's Law becomes less applicable, a software-hardware co-design approach known as Domain Specific Architecture (DSA) becomes dominant. This approach involves defining computing architectures specifically tailored to solve problems in a particular domain. Artificial intelligence (AI) chips and emerging DPUs (Data Processing Units) have emerged as typical examples of DSA technology.

DSA utilizes efficient architectures designed for specific domains, employing techniques such as dedicated memory to minimize data movement, optimizing chip resources for computation or storage based on application characteristics, simplifying data types, and employing domain-specific programming languages and instructions. Compared to Application Specific Integrated Circuits (ASICs), DSA offers similar performance and energy efficiency when utilizing the same amount of transistors, while retaining flexibility and versatility in the field. For instance, ZTE's customized chip architecture "Quark" in the field of artificial intelligence abstracts computing resources into tensor, vector, and scalar engines based on the computational characteristics of deep neural networks. It separates computation and control, efficiently scheduling various processing engine (PE) units through an independent control engine (CE), enabling efficient execution of various deep learning neural network computations. Due to the customized hardware and software design, DSA can achieve significantly higher performance, up to tens or even hundreds of times faster than traditional CPUs, while consuming the same amount of power.

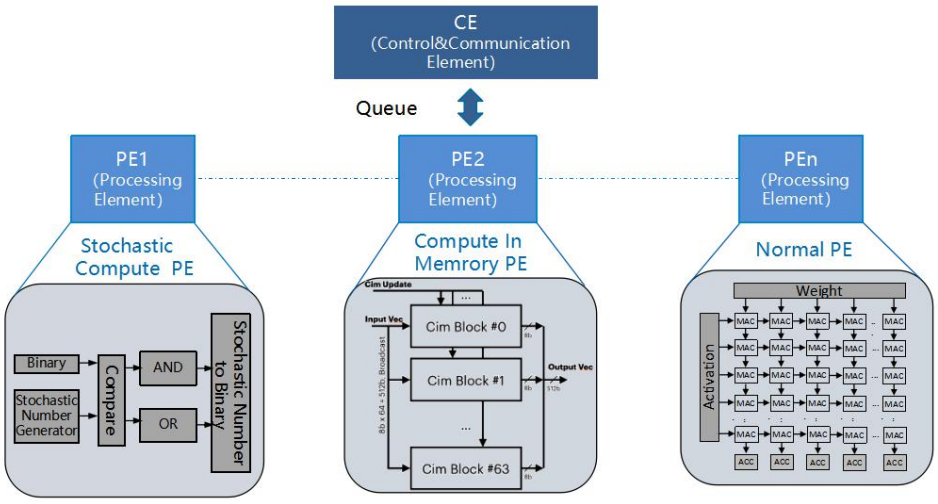

Figure 4.2 Customized Architecture of ZTE "Quark"

Moore's Law is primarily evaluated in the 2D space of chip manufacturing. However, as chip miniaturization becomes more challenging, 3D stacking technology has emerged as a crucial means to improve chip density. 3D stacking involves vertically stacking chips without changing the original package area. This chip design architecture helps address the memory wall problem by enabling better scalability and energy efficiency.

The Chiplet technology is considered to be a key technology to Moore's Law. The Chiplet technology modularizes chip design and miniaturizes large chips, effectively improving yield and reducing complexity. In addition, the Chiplet technology can manufacture different chiplets separately as required (for example, the core computing logic uses advanced process to improve performance, but the peripheral interfaces still use mature process to reduce costs), and then assemble them by using advanced encapsulation technologies, which can effectively reduce manufacturing costs.

Compared to traditional chip solutions, the Chiplet approach offers three key advantages: design flexibility, lower costs, and shorter time-to-market. The primary challenge associated with Chiplet technology lies in interconnection techniques. To address this, the UCIe Industry Alliance was founded on March 2, 2022. The maturity of the Chiplet industry and the establishment of a complete industry chain encompassing interconnection interfaces, architecture design, manufacturing, and advanced encapsulation are expected.

4.3. Computing Architecture: The Integration of Computing and Storage

The classic Von Neumann computing architecture follows a paradigm of separating computation and storage. However, if the memory access speed fails to keep pace with CPU performance, it can create a bottleneck known as the memory wall. Google conducted a study on the power consumption of its products and discovered that over 60% of the system's power consumption was attributed to read and write operations between CPUs and memories [21]. With the advancement of

big data and artificial intelligence, the conventional computing architecture is increasingly limiting the performance of emerging data-intensive applications, necessitating the development of a new computing architecture to address this challenge.

Computing-memory integration technology involves a collaborative design that optimizes computation and memory based on application requirements. Its aim is to reduce unnecessary data movement, increase data read and write bandwidth, and improve energy efficiency. By doing so, the limitations imposed by the memory wall and power consumption can be overcome.

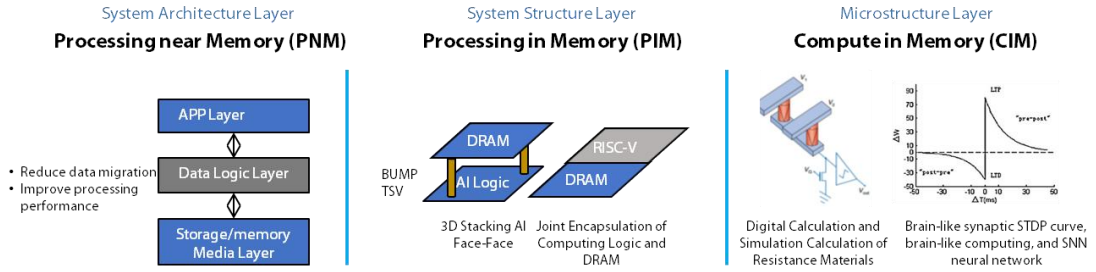

Figure 4.3 Three Architectures of Computing-Memory Integration

There are three forms of computing-memory integration architecture: Processing Near Memory(PNM), Processing In Memory(PIM), Computing In Memory(CIM).

(1) Processing Near Memory

Near-memory computing introduces computing power in the data cache location, generates local processing results, and directly returns the results, reducing data movement, speeding up processing, and improving security. As shown in FIG. 4.3, a data logical layer is added to an Data-Centric-type application, and cache processing is introduced to minimize data migration.

(2) Processing In Memory

PIM involves integrating the computing engine inside the memory, typically using DRAM. The objective is to perform simple processing directly while reading and writing data, without the need to copy the data to the processor. An example is the conversion between Celsius and Fahrenheit. Processing in memory essentially follows a computing-memory separation architecture, but with memory and computing closely integrated, thereby reducing the overhead caused by data movement. Memory manufacturers are driving its commercialization.

(3) Computing In Memory

CIM involves embedding a computation unit into memory, particularly suited for executing highly parallel matrix-vector products. It has promising applications in machine learning, cryptography, differential equation solving, and more.

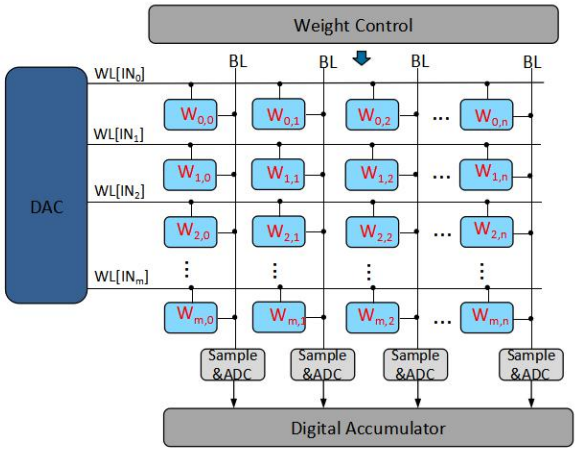

Figure 4.4 Computing in memory Architecture

CIM adopts a unified computing and memory design architecture. Taking the matrix-vector multiplication and addition operation in deep neural networks as an example, the architecture shown in Figure 4.4 is commonly used. It consists of an input DAC, unit array, output ADC, and other auxiliary circuits. The weight data is stored in the storage unit, and the input undergoes DAC conversion to perform read and write operations on the stored data. Using Ohm's law and Kirchhoff's law, the output currents of different storage units are automatically accumulated and then output to the ADC unit for sampling and conversion into the output digital signal, thereby completing the matrix-vector multiplication and addition operation.

4.4. Computing Architecture: Peer-to-peer Computing

The traditional computing system is built with the CPU as the center. The surge of business has higher and higher requirements for the system's processing power, and the data interaction between accelerators usually need to be transferred through the CPU. CPU is easy to become a bottleneck, the efficiency is not high.

Peer-to-peer systems based on xPU (data-centric processing unit) can establish a new type of distributed computing architecture. As shown in Figure 4.5, a peer-to-peer system is formed by interconnecting multiple nodes with similar structures. Each node has xPU as its core, which comprises various heterogeneous computing resources such as CPU, GPU, and other computing chips. The primary function of xPU is to access and interconnect the heterogeneous computing resources within the node and across other nodes. The general processor core inside xPU can manage and schedule computing resources within the nodes. The CPU is no longer the central component of the node, and the CPU, GPU, and other computing chips are placed on an equal footing. Tasks are allocated by xPU based on the characteristics and capabilities of each computing chip.

A new transmission protocol based on memory semantics is used within nodes and between nodes in a peer-to-peer system. Compared with existing transport protocols like TCP and RoCE, memory

semantics-based transport protocols offer advantages such as low latency and high scalability.

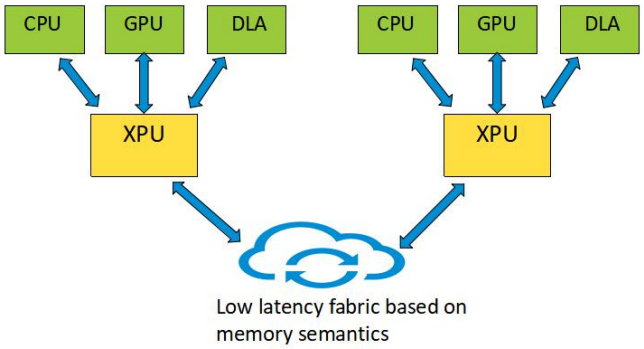

Figure 4.5 Peer-To-Peer Computing System

A server based on the peer-to-peer computing architecture can be seen as a "distributed computing system," which facilitates independent planning and development of each node in the industry chain, allowing them to leverage their advantages. By utilizing peer-to-peer memory semantic interconnections, the system can be smoothly expanded, treating the vast distributed computing power as a single "computer".

4.5. Network Architecture: The IP Network Technology That Supports The Convergence of Computing and Networks

With the development of edge computing, the computing power resources become deployed in a distributed manner. Considering that the growth of network bandwidth is restricted by Shannon's theorem, the improvement of computing power capability is restricted by Moore's law, and there is an increasing need for energy conservation and consumption reduction, it is inevitable to achieve effective scheduling of network and computing power resources and fine granularity system operation. By leveraging high-speed, flexible, and intelligent networks, the convergence of computing and networks can integrate distributed computing power nodes across different regions, provide open computing power services, and enhance the efficient utilization of computing and network resources.

The integration of computing and networks is driven by two main factors. Firstly, from the demand side, there is a requirement to coordinate the scheduling of computing power and networks to meet the unified demand for computing resources and network connectivity from various services. For instance, high-resolution VR cloud games necessitate not only computing resources from dedicated graphics processing units (GPUs) for rendering but also deterministic network connections to fulfill end-to-end latency requirements within 10 ms. Secondly, from the supply side, the deep convergence of computing and networks, leveraging the ubiquitous and distributed nature of network facilities, enables the distributed deployment of computing power resources to meet diverse application requirements in terms of latency, energy consumption, and security.

The convergence of computing and networks poses challenges to IP network technologies. From the perspective of the Internet architecture, “computing” is typically associated with upper-layer applications, while “network” is related to lower-layer connections. IP technology, positioned in the middle layer, plays a crucial role in connecting the upper and lower layers. The design of traditional IP networks follows a layered and end-to-end principle, which allows services to be developed independently from networks, reducing the threshold for service innovation and facilitating rapid service deployment. However, this principle also leads to services operating in a "best-effort" mode, decoupled from the underlying networks.

Consequently, for future IP networks, it is challenging to bridge the gap between services and networks to realize the coordination and fine granularity management of computing power resources and network resources. To address this challenge, ZTE proposes an innovative architecture "Service Awareness Network(SAN)" [25].

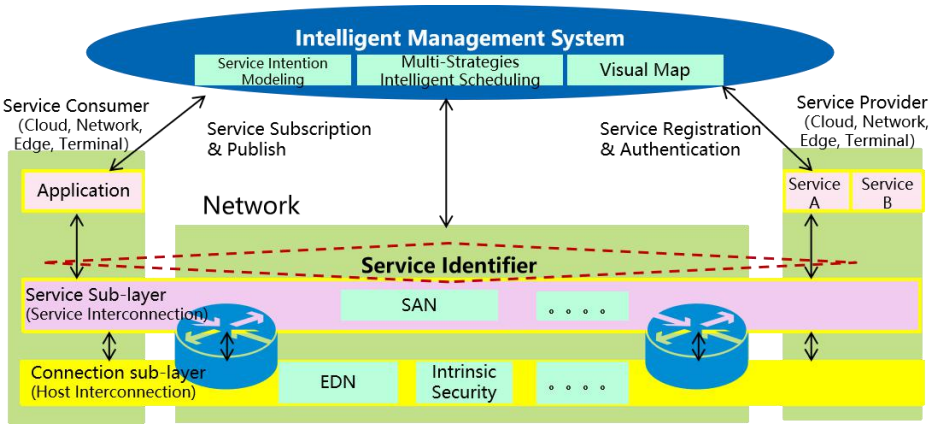

Figure 4.6 Service-Aware Network (SANs) Architecture

The architecture of a service-aware network is illustrated in Figure 4.6. The core concept is to encapsulate the computing power resources and network resources provided by service providers as "services" and assign a unique service identifier to each service. These services can be dynamically deployed anywhere within the network as needed. A service sub-layer is introduced at the IP network layer to enable service perception, routing, and scheduling. As a result, a service-aware network encompasses three core design elements:

(1) A service ID that can be recognized horizontally from terminal to network to cloud and vertically from applications to network facilities

A service ID can either identify a connection type of service (i.e., providing network resources to establish an end-to-end connection from a terminal to the cloud) or a computing type of service (i.e., providing computing power resources to compute) and it receives unified service governance from terminals, networks, and clouds. The application layer can directly use a service ID to initiate a location-independent transport layer connection, without the need for DNS domain name resolution. Such an operation greatly reduces the service response time and supports mobility by nature.

(2) A service sub-layer (3.5 layer) that is introduced at the IP network layer to implement network-centric service interconnection.

The introduction of an identifier-centric service sub-layer to traditional IP host routes offers several advantages. It enables the network to perceive the computing power requirements of service consumers and the computing power resource status of service providers. Leveraging service routing, service requirements can be adequately satisfied by available resources, promoting the evolution of network interconnection from host-centric to service-centric.

(3) A connection sub-layer with enhanced capability

The connection sub-layer enhances the basic capabilities of the network, including the ability to provide deterministic connections and intrinsic security. The connection sub-layer ensures that the network meets the connection-level QoS requirements of the service.

In summary, the service-aware network provides computing power services and network services in a unified manner and achieves efficient scheduling of computing power resources and network resources. This approach not only guarantees quality of service but also meets the requirements for energy saving and emission reduction.

5. Intelligence

5.1. Overview

AI technology serves as the driving force propelling humanity into the era of intelligence. Recognizing the significance of AI technologies in leading the new wave of industrial transformation, many countries around the world have actively promoted intelligent infrastructure construction and research across various fields.

The fundamental elements of AI technology encompass computing power, algorithms, and data. Data is intricately linked to specific business domains, and the establishment of an open, shared, and circulatable data resource system is crucial for a digital society. Computing power and algorithms form the foundational capabilities that digital infrastructure should possess.

The recent advancements in AI technology since 2016 have been remarkable. However, certain bottlenecks still need to be overcome to meet people's expectations. For instance, achieving powerful intelligent capabilities often necessitates complex algorithms and extensive computing power, resulting in high costs, energy consumption, and environmental pressures. While dedicated artificial intelligence for specific domains has demonstrated superior performance compared to human abilities, general intelligence is still in its nascent stages. As discussed in Chapter Two, breakthroughs in cognitive science are yet to be achieved, and there is a lack of theoretical guidance in the development of AI technology.

As the requirements for AI computing power surpass the scope of Moore's Law, the industry faces the important task of achieving more efficient AI chips. Section 5.2 delves into the innovative directions of AI chip architectures to attain higher computing power/energy ratios.

The success of ChatGPT has positioned large models as a promising research direction for Artificial General Intelligence (AGI). Section 5.3 highlights the trends in large model technology and its expanding applications, which may evolve into a new platform layer. Model-as-a-service (MaaS) emerges as a potential business model, offering universal AI capabilities for diverse scenarios.

Additionally, Section 5.4 examines Network Intelligence as a use case for intelligent infrastructure. The telecom industry has long been intrigued by how AI enables network operation and maintenance and facilitates the digital transformation of the network itself. With the support of more efficient AI computing power and new algorithms like large models, network intelligence is anticipated to progress from the current L2-L3 level to the L4-L5 level in the near future.

5.2. AI Chip: Increase Computing Power/Energy Ratio

As described in Chapter 2, the rapid increase in energy consumption in AI computing today will

place a heavy burden on the environment. Continuous research on more efficient AI chips is necessary.

There are two feasible directions for achieving high Tops/W (computing power/energy ratio) for AIs: spatial computation and approximate computation.

(1) Spatial computation

The power consumption of an AI chip is positively correlated with the distance that data is transmitted inside the chip. With innovative chip architecture design, the energy consumption of the chip can be significantly reduced by minimizing the distance that each operation's data needs to travel inside the chip.

Dividing a large computing core into multiple smaller computing cores can effectively reduce the average distance data needs to move, thereby reducing energy consumption. This has become the design trend for new AI chips. However, this kind of multi-core parallel computing introduces additional overhead, resulting in reduced computational efficiency. "Spatial computation" is a collaborative design of hardware and software architecture. It involves dividing a computing task into multiple subtasks, assigning these subtasks to different computing cores, and planning the data transmission path between tasks to minimize data movement distance. This approach aims to achieve optimal performance and the lowest power consumption.

To implement multi-core spatial computation, hardware and software need to be co-designed. In terms of hardware, the computing core can add hardware support for common communication modes of AI parallel computing, such as Scatter, Gather, and Broadcast, to optimize the topology structure and dynamic routing capability of the on-chip network. In terms of software, due to the complexity of spatial computation optimization, it cannot be solely borne by developers. The compiler needs to automatically divide tasks, assign tasks, and plan routes. The runtime should handle various anomalies, such as packet loss, disorder, and congestion.

An evolutionary path for future spatial computation is in-memory computing. In-memory computing can divide a macro computing core into tens of thousands of micro computing cores, rather than just hundreds of mini computing cores. In this architecture, the average movement distance of data is further reduced to the micrometer scale, and power efficiency can exceed 10 TOPS/W@INT8. For example, Untether AI's Boqueria chip has more than 300,000 processing elements at 30 TFLOPS/W@FP8[26].

Another evolutionary path to spatial computation is deterministic design. For example, Groq tensor flow processors (TSPs) use the deterministic hardware design [27], and the compiler can accurately schedule computing, memory access, and data transmission on each core to avoid access conflicts of shared resources.

(2) Approximate computation

One characteristic of deep learning models is that they do not require high precision. The errors

that occur during computation do not significantly affect the final outcome of the model. Approximation algorithms reduce memory usage and computational complexity, making computations more efficient.

Low-precision computing is an important technical direction for deep learning. Using low precision data types can reduce chip area and energy consumption. For example, multiplication and addition operations of INT8 consume only 1/30 and 1/15[28] of the energy of 32-bit floating point number (FP32). In the current hybrid precision training technology, an FP16-bit half-precision floating point number and an FP32 single-precision floating point number may be used together to complete model training.

Since the inference requires less precision, the model can be transformed into a lower-precision data type after training, a technique called model quantization. Currently, INT8 quantization technology has matured significantly, while INT4 quantization technology still faces some challenges.

Another type of approximate computation is sparse computation. It has been observed that the weights of deep learning models are sparse, meaning some weights are zero or very close to zero, especially in Transformer models where sparsity is more prevalent. Exploiting the sparsity of the model can eliminate unnecessary computations, thereby improving the efficiency of model computation. For instance, the 2 out of 4 sparse acceleration in Nvidia A100 GPUs can double the chip's equivalent computing power [28] while maintaining the same energy consumption. In the future, coordinated software and hardware approaches to sparse computation will remain a promising technology direction.

In the next 10 years, improving energy efficiency through manufacturing processes will become increasingly challenging. Spatial computation and approximate computation have significant potential to enhance the energy efficiency ratio of chips. Compared to current mainstream AI chips, these approaches can increase chip efficiency dozens of times, providing a powerful guarantee for the AI industry to achieve the dual-carbon goal.

5.3. AI Algorithm: Evolution from Dedicated Small Models to General Large Model

The nature of the AI algorithm is to provide a mapping between the real world and the digital world. The quality of the algorithm depends on the accuracy of the mathematical model used to represent the real problem. From the early statistical machine learning to CNNs, BERT, and Transformer, to the most recent GPTs, digital models are becoming more and more large and better matched with the real world. In particular, the emergence of GPTs, such as ChatGPT and GPT-4, has revolutionized the field of AI. Large models have become the development trend of artificial intelligence algorithms, and have kicked off the development of artificial general intelligence.

(1) Basic model behind AIGC: Transformer

In 2016, Google invented Transformer, a new deep learning model based on attention mechanism, which was initially only used for machine translation, but in 2017 BERT[30] divided training for a single task into two stages: Task-independent pre-training and task-related fine-tuning, making Transformer a universal model capable of handling multiple language tasks. In the same period, the Transformer-based OpenAI model uses a different pre-training idea from BERT , that is, only Transorformer's decoder part is used to pre-train the language model, which also proves its universality and achieves better results after expanding the data and model scale. In 2020,GPT-3 was born, which is the first 100-billion-level parameter model, triggering a computing power arms race.

(2) Multi-modal large model: CLIP

Transformer has become a universal natural language processing model in the language field, so can its versatility extend beyond language tasks? In 2020, ViT[31] demonstrated that Transformer could handle image tasks better than traditional convolutional neural networks (CNNs). The CLIP model proves that the same Transformer model can process data in both natural language and image modes. Subsequently, Chinese researchers also propose three-modal models. Applications such as text-to-image are emerging. In 2022, the open source StableDiffusion[32] based on the diffusion model can generate clear images with high resolution, further expanding the application scenarios of the AIGC.

(3) Reinforcement Learning from Human Feedback: ChatGPT

OpenAI took an in-depth look at its potential after the GPT-3 model was developed. In 2021, CodeX used source codes to replace natural languages as training corpus, so that the CodeX model (also based on GPT-3) can generate codes. In 2022, GPT-3.5 trained the model by using a mixture of natural languages and source codes, and made the model have the Chain-of-Thought capability. InstructGPT[33] uses human feedback to make the content generated by the model more in line with human values. All of these led to ChatGPT, which enhanced the ability to model historical conversations, capture user intent effectively, complete contextual understanding to achieve continuous conversations, and extract useful knowledge from massive amounts of data and apply it logically.

(4) Large models stimulate industry applications

With the release of GPT-4 large model and the performance leap, the large model is expected to usher in further applications in various fields. With its authenticity, diversity, controllability and composability, large model is expected to help enterprises improve the efficiency of content production and provide them with more diversified, dynamic and interactive content.

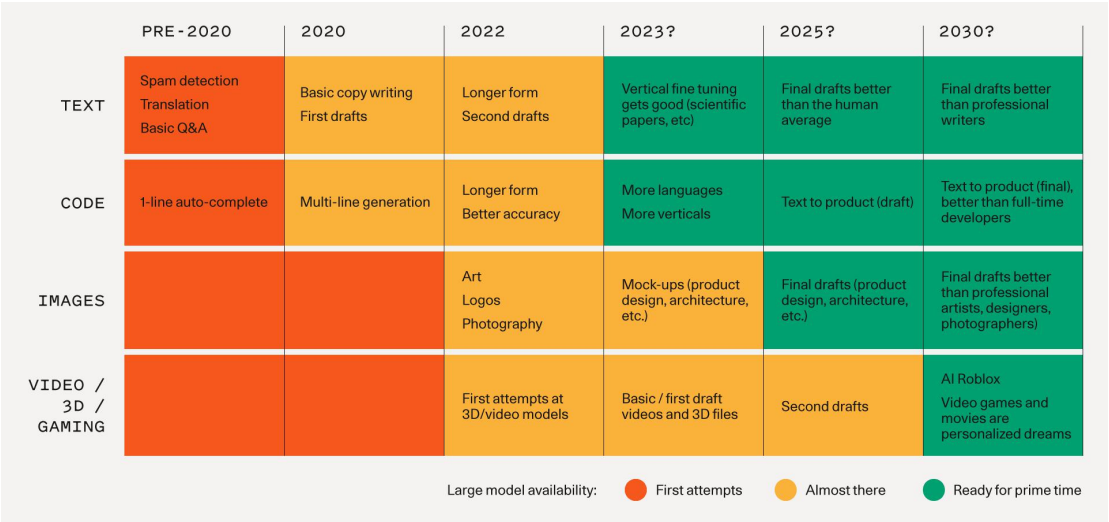

Figure 5.1 Timeline for large model progress and the associated applications[34]

The large model represents a breakthrough in deep learning technology, transcending different eras. Its most significant advantage over traditional deep learning algorithms lies in its exceptional universality. Unlike traditional models, which can only handle a single task, large models have the capability to perform multiple tasks. This has addressed the issue of fragmented artificial intelligence applications in recent years, reducing the cost associated with migrating across different scenarios. The universality of large models allows for training a single model to accomplish dozens or even more tasks, and their contextual learning ability enables them to acquire new tasks without requiring re-training. This universality positions large models as a new platform, empowering a wide range of applications at higher levels.

5.4. AI for Network automation: Empower Autonomous Network to Higher Level

Network automation refers to the implementation of automatic configuration, fault self-healing, and automatic optimization in networks, ensuring flexible service provisioning, high reliability, and high performance.

In 2019, TM Forum introduced the concept of Autonomous Networks in response to the communication industry's requirements. In 2022, they released the white paper "Autonomous Networks: Empowering digital transformation–from strategy to implementation." TM Forum has put forward the vision of "three-Zero three-Self," which aims to achieve three-Zero user experiences (Zero Wait, Zero Touch, Zero Trouble) by implementing three-Self capabilities (Self-serving, Self-fulfilling, Self-assuring) at the network O&M layer, as shown in Figure 5.2.

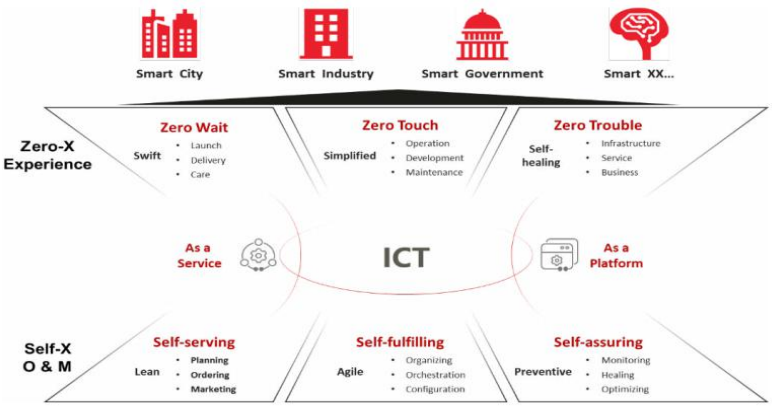

Figure 5.2 TMF Vision of Autonomous Networks

TM Forum has also proposed autonomy upgrading standards, categorized into six levels (L0 to L5) and six dimensions (Execute, Perception, Analysis, Decision, Intention/Experience, and Application), as depicted in Figure 5.3.

P: Manual AI: Artificial intelligence

Ddfine	L0: Manual	L1: Auxiliary	L2: Partial - Intelligence	L3: Basic-Autonomy	L4: High-Level-Autonomy	L5: Full-Scene-Autonomy
Execute	P	P+AI	AI	AI	AI	AI
Perception	P	P+AI	P+AI	AI	AI	AI
Analysis	p	p	P+AI	P+AI	AI	AI
Decision	P	P	P	P+AI	AI	AI
Intention/Experience	P	P	P	P	P+AI	AI
Application	N/A	Specific Scenarios				Full-Scene

Figure 5.3 Autonomous Network Levels

To reach the L4-level of an automous network, the key lies in integrating AI algorithms into scenarios such as network self-configuration, fault self-healing, and quality self-optimization.

(1) Intent-Based Closed-Loop to Support Network Self-Configuration

Currently, services are predominantly configured manually. With the maturation of intention technology, service parameters and system parameters can be automatically configured through customer intent perception, intent translation, and closed-loop verification based on AI algorithms. Intent-based closed-loop network self-configuration provides a superior zero-wait-zero-touch experience in both consumer (ToC) and business (ToB) scenarios.

(2) Multi-Dimensional Data Analysis to Support Self-Healing of Network Faults

At present, fault recovery primarily relies on aggregating and analyzing multi-dimensional data such as alarms, performance KPIs, logs, and service indicators to generate events and implement closed-loop fault handling based on intelligent event management. In the foreseeable future, the integration of multi-modal large models (e.g., NLP, network, and visual models), combined with

enhanced dimensional data, end-to-end training, and knowledge extraction technologies, will significantly improve O&M accuracy and expand the range of O&M scenarios.

(3) The algorithm tends to be interpretable to support network quality self-optimization.

The interpretability of AI algorithms is crucial for autonomous networks. An interpretable algorithm allows individuals to understand the decision-making process, including the reasons, methods, and content of the decisions made by the algorithm model. Autonomous networks play a vital role in the telecommunications field, directly influencing service quality. If an algorithm recommendation problem arises, it may lead to complaints. Algorithm interpretability helps users safely implement algorithm recommendations in production environments. Additionally, O&M personnel can comprehend the decisions made by the model and identify causes of deviations, optimizing and enhancing model performance.

In the long term, with the support of future communication technologies, big data, and computing power networks, autonomous networks will gradually evolve into full-stack L5 levels in an orderly manner, ultimately achieving complete self-X autonomy in autonomous networks and fulfilling the zero-X objectives of zero wait, zero touch, and zero trouble.

6. Conclusion

Since the 18th century, technology has been one of the core factors in production. Over the past 20 years, with the rapid development of emerging ICT technologies, such as cloud computing, artificial intelligence, and mobile communication, human beings have become increasingly capable of mining information and acquiring knowledge from massive amounts of data. Data, along with other production factors, will drive the high-quality growth of the digital economy and empower the construction of the digital society.

By 2030, the digital infrastructure of "connectivity, computing power, and intelligence" will serve as the foundation of the digital era. Expanded connectivity will enable new high-bandwidth applications, such as the metaverse and 3D holographic communication. Enhanced computing power will support the storage of vast amounts of data and enable real-time processing. Increased intelligence will inject powerful capabilities into the digital infrastructure, further promoting the evolution of communication networks towards intelligent networks, the digital economy towards an intelligent economy, and the digital society towards an intelligent society.

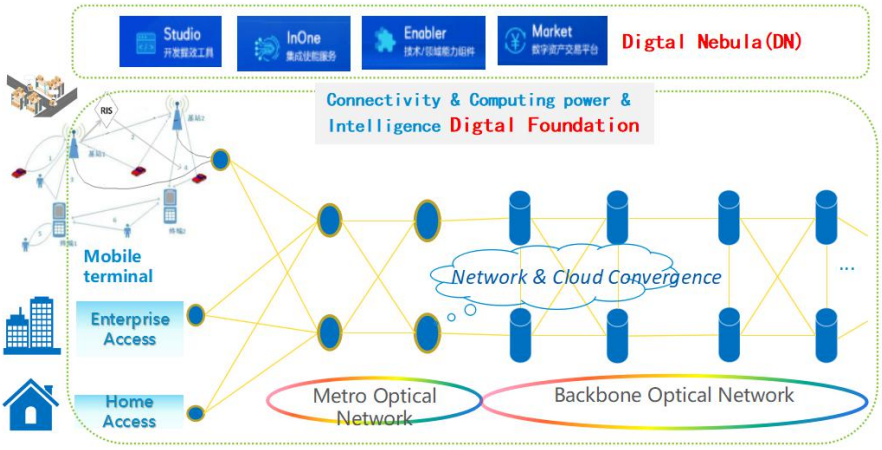

Figure 6.1 Digital Nebula Empower Digital Transformation

Expanded connectivity, enhanced computing power, and increased intelligence need to operate closely and mutually support each other. The key capability lies in forming a complex software system that can integrate, coordinate, and empower new ICT technologies as needed. In 2022, ZTE launched Digital Nebula (DN), and in 2023, DN 2.0[34], aiming to build a cloud-native, service-oriented, and data-driven digital solution and platform that fully integrates and leverages the digital infrastructure of connectivity, computing power, and intelligence. Industrial customers can utilize DN to further develop their own digital platforms, resolve the contradiction between diversified applications, unified governance, and efficiency improvement, and achieve resilient services, scalable systems, and cost reduction.

ZTE adheres to the positioning of being a "Driver of the Digital Economy" and embraces the

concept of "open and win-win." As a provider of digital infrastructure products and technologies, ZTE offers world-leading cloud, network, edge, terminal, software, and industrial products, and actively shares its core atom capabilities to assist carriers and large enterprises. Additionally, ZTE supports the rapid growth of SMEs and promotes coexistence and win-win relationships with ecosystem partners.

We expect that the release of this white paper will facilitate further in-depth communication and solicit sincere feedback on the technological development of ICT.

7. References

[1] China Academy of Information and Communications Technology: White Paper on Global Digital Economy (2022), December 2022

[2] China Academy of Information and Communications Technology: Report on the Development of China's Digital Economy (2023), April 2023

[3] Chinese government network: The Digital China Construction Overall Deployment Plan http://www.gov.cn/xinwen/2023-02/27/content_ 5743484 .htm

[4] Fang Min, Duan Xiangyang, Hu Liujun: 6G Technology Challenges, Innovation, and Outlook, ZTE Technology June 2020 Issue 3

[5] IDC&Inspur&Tsinghua: Global Computing Index 2021-2022

[6] China Academy of Information and Communications Technology (CAICT): White Paper on China's Computing Power Development Index (2022

[7] ITU-T FG-NET2030: Representative Use Cases and Key Network Requirements for Network 2030, January 2020

[8] ITU-T FG-NET2030: Additional Representative Use Cases and Key Network Requirements for Network 2030, June 2020

[9] GeSI: SMARTer2030-ICT solutions for the 21st Century, 2015

[10] Design of Capacity-Approaching Irregular Low-Density Parity-Check Codes, IEEE TRANSACTIONS ON INFORMATION THEORY, VOL.47, NO. 2, FEBRUARY 2001

[11] Amir Gholami https://github.com/amirgholami/ai_and_memory_wall/blob/main/imgs/pdfs/ai_and_compute.pdf

[12] Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2022). Efficient transformers: A survey. ACM C omputing Surveys, 55(6), 1-28

[13] STRUBELL E, GANESH A, MCCALLUM A. Energy and policy considerations for deep learning in NLP [EB/OL]. https://arxiv. org/abs/1906.02243

[14] ISO/IEC 23090-3 Information technology-Coded representation of immersive media-Part 3:Versatile video coding First edition 2021-02

[15] JVET-AB2023 EE1: Summary of Exploration Experiments on Neural Network-based Video Coding

[16] JVET-AB2024 Exploration Experiment on Enhanced Compression beyond VVC capability (EE2)

[17] Alexey Andreyev, Xu Wang, Alex Eckert, "Reinventing Facebook's data center network," MARCH 14, 2019

[18] M. LaCroix et al., "A 116Gb/s DSP-Based Wireline Transceiver in 7nm CMOSAchieving 6pJ/b at 45dB Loss in PAM-4/Duo-PAM-4 and 52dB in PAM-2," ISSCC, pp.132-133 , Feb.2021.

[19] ODCC-2022-0300 A, "White Paper on 112G Linear Optical Interconnection Solution," P7, 2022-09

[20] Rakesh Chopra, "Looking Beyond 400G" P5, TEF2021, January 25, 2020

[21] Janet Chen, Meta, Rob Stone, Meta, "Perspective on Linear Drive Pluggable optics", OIF

2023.123 .01,

[22] William Dally, "Accelerating Intelligence," P60,GTC China, and December 14, 2020

[23] LightCounting comments on CPO panel discussion at Photonics West, "Our industry is at a crossroads", February 2023

[24] A. Boroumand, et al., "Google workloads for consumer devices: Mitigating data movement bottlenecks", Proc.23rd Int. Conf. Support Program. Lang. Operating Syst., 2018

[25] ZTE Corporation: White Paper on Future Evolution of IP Networks (2.0), August 2022

[26] BEACHLER R, SNELGROVE M. Untether ai: boqueria [C]//Proceedings of 2022 IEEE Hot Chips 34 Symposium (HCS). IEEE, 2022: 1-19. DOI:10.1109/HCS55958.2022.9895618

[27] ABTS D, KIM J, KIMMELL G, et al. The Groq Software-defined Scale-out Tensor Streaming Multiprocessor: from chips-to-systems architectural overview [C]//Proceedings of 2022 IEEE Hot Chips 34 Symposium (HCS).IEEE, 2022: 1-69. DOI: 10.1109/HCS55958.2022.9895630

[28] HOROWITZ M. 1.1 Computing's energy problem (and what we can do about it) [C]//Proceedings of 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC). IEEE, 2014: 10-14. DOI:10.1109/ISSCC.2014.6757323

[29] POOL J. Accelerating inference with sparsity using the Nvidia ampere architecture and NVIDIA TENSORRT [EB/OL]. [2022-10-12]. https://developer. nvidia. com/blog/accelerating-inference-with-sparsity-usingampere-and-tensorrt

[30] Lee, J. D. M. C. K., & Toutanova, K. (2018). Pre-training of deep bidirectional transformers for la nguage understanding. arXiv preprint arXiv:1810.04805.

[31] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale [EB/OL]. [2022-10-12].https://arxiv.org/abs/2010.11929

[32] Borji, A. (2022). Generated faces in the wild: Quantitative comparison of stable diffusion, midjou rney and dall-e 2. arXiv preprint arXiv:2210.00586.

[33] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Tr aining language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

[34] Sequoia Capital: Generative AI: A Creative New World https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/

[35] Digital Nebula 2.0 https://www.zte.com.cn/china/about/news/20230419c9.html