ZTE Green 5G Core

White Paper

CONTENTS

1 Introduction

With the large-scale commercial use of 5G in the world, large-scale deployment of a large number of base stations, edge data centers, and large-scale data centers, the network traffic continues to increase, the energy consumption increases greatly, and the green transformation of 5G networks and applications is urgent. As the core of a 5G network, the core network (CN) needs to continuously explore energy-saving and consumption-reducing technologies, introduce intelligent elements, act as the brain during the green network transformation process, achieve end-to-end joint energy saving, and introduce new materials, new architectures, and new algorithms to improve resource energy efficiency.

This white paper analyzes the key factors that affect the green transformation of core network, identifies the direction and objectives of the transformation, and then focuses on the key energy-saving and efficiency-improving technologies and benefits at various levels such as server, cloud platform, NE, and O&M management. In addition, this white paper provides industry practice cases to provide theoretical support and reference for operators and industries when building green networks, promote the research and innovation of network energy-saving technologies, and contribute to the goal of carbon neutral and net zero for the green earth.

2 Green Development, Network Leading

2.1 Opportunities and Challenges of Green Development

According to the analysis from GSMA, the global temperature has been trending warmer since the last century, as shown in Figure 2-1, especially after the 1980s. If not controlled, the global average temperature will rise by 3℃ in 2100, making the earth no longer suitable for human living.

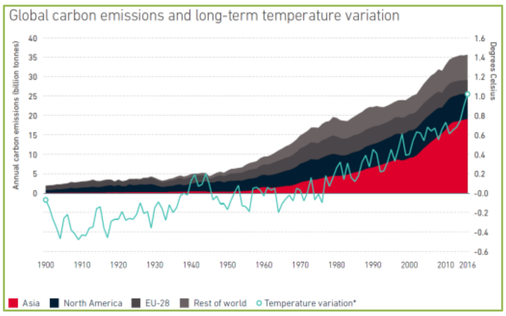


Figure 2-1 Global Carbon Emission and Temperature Change

Currently, climate change has a greater impact on living areas around the world. With extreme climate change, global economic development and human life are greatly affected. Climate warming has become a huge challenge for human beings. Therefore, carbon emission reduction has become the consensus of important countries around the world. By the end of 2021, more than 130 countries, 116 regions, 234 cities, and 683 enterprises had proposed carbon neutrality targets. China also committed to green development at the UN Climate Change Conference 2020. By 2030, China will reach its peak on carbon, striving to achieve its carbon neutrality goal by 2060. The world will enter the dual-carbon era, and the global economic development model is moving from energy resource dependent to energy technology dependent.

As China is in the middle and late stages of industrialization, there is still a big gap between the energy consumption of unit products in the industrial production process comparing with the international advanced level. There is huge potential for green transformation in these industries. At the same time, 5G has been put into commercial use for three years, and has made great achievements in network construction, user scale, and intelligent terminals. 5G ToB services have been developed and integrated with thousands of industries. As an engine for promoting digital and intelligent transformation of thousands of industries, 5G new infrastructure will bring huge opportunities to 5G networks through continuous promotion of green network construction, optimization of network energy-saving technologies, and green empowerment of industries.

With the transition of user services to multi-sensory, rich media, strong interaction, and high immersion, the network requires larger bandwidth and lower latency, resulting in a rapid increase of 5G network traffic and significantly exceeding 4G traffic. Compared with the same 4G coverage range, several times of 5G sites are required to meet user coverage requirements, resulting in several times more energy consumption than 4G networks. Therefore, how to build a green digital infrastructure through innovative technologies is a major challenge to 5G network development.

2.2 Green and Low-Carbon is the Basis of Future Network Evolution

In the face of the needs of social development in the future, domestic and international famous industry associations, such as IMT-2030 (6G) promotion group, NGMN, ATIS (Alliance for Telecommunication Industry Solutions), and European 6G flagship project Hexa-X have agreed that green and low-carbon is the basis of future network evolution technologies.

NGMN believes in the three core goals of next-generation networks: Social and environmental benefits, new and expanded and differentiated experiences, and operations of creating higher value. Social and environmental benefits require environmental sustainability, improving end-to-end environmental impact, energy efficiency, and digital inclusion are at the core of future technology considerations.

Future 2-2 Motivation and Drivers for 6G from NGMN

The US ATIS is committed to "driving the leading position of 6G mobile technology and future technologies in North America over the next decade", and believes that next-generation networks are equal to Green G, with sustainability being one of its six goals: Fundamentally changing the way that power is used to support next-generation communications and computer networks, while strengthening the role of information technology in environment protection.

Figure 2-3 Next Generation Network Target Released by ATIS

Hexa-X defines a brand-new 6G intelligent network architecture, and regards that sustainability is one of the three cornerstones of 6G. The goal is to optimize digital infrastructure, save energy and reduce power consumption, effectively reduce greenhouse gas emissions, and provide efficient and sustainable digital tools for industries, society, and decision makers.

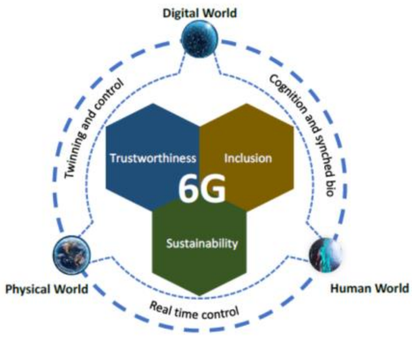


Figure 2-4 6G Network Cornerstones Defined by Hexa-X

3 Key Technologies and Practices of Green 5G Core

3.1 Analysis on Key Factors that Affecting Green Transformation of Core Network

As the core and brain of 5G network, CN (Core Network) needs to save energy and reduce emission. According to the analysis of GSMA reports, energy consumption of 5GC accounts for about 11% of the whole 5G networks. Because CN has multiple types and quantities of NEs, it is deployed separately, resource redundancy exists, and service tidal waves greatly affect resource usage, therefore, CN faces challenges of energy saving and consumption reduction. On the other hand, CN needs to play a role of a core brain, accurately perceiving information such as congestion status of uplink and downlink traffic, service status and user location, performing real-time scheduling and dynamic orchestration of network capabilities, and collaborating with other domains of the network, so as to empower various industries and achieve end-to-end energy saving and consumption reduction. Based on the above two requirements, the research and applications of CN energy saving technologies are indispensable. In accordance with the typical configuration, the control-plane energy consumption ratio of CN is 74%, and the user-plane energy consumption ratio is relatively small. Therefore, the key to CN green transformation is to save energy and reduce consumption on the control plane.

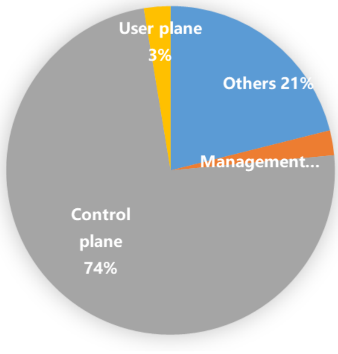


Figure 3-1 CN Energy Consumption Composition under the Typical Configuration

With the support of the NFV/SDN technology, the control plane of 5GC has been fully cloudified, so that 5GC can be deployed in a centralized manner to construct a unified physical network based on the data center, and resources can be shared to the maximum extent through software and hardware decoupling and hardware resource pooling, thus changing the traditional "silo" architecture. Resources are allocated and scheduled on demand through unified management to maximize resource utilization. Therefore, CN needs to be green transformed in three aspects:

1. Green transformation of infrastructure: Develop green servers and a green cloud platform as a green foundation.

2. Green transformation of network functions: Optimize the NE architecture to achieve green enabling.

3. Green transformation of network management: Precise monitoring, and management and optimization of energy consumption, serving as a green brain.

3.2 Key Technologies of Green Transformation of Core Network

3.2.1 Infrastructure – Green Foundation

After the CN is fully cloudified, the infrastructure is upgraded from "dedicated equipment + dedicated platform" to "various servers + a cloud platform". The control plane is deployed based on general servers in a centralized manner, and the user plane is deployed to various edge servers as required. Low-carbon & energy-saving focuses on energy control of various servers and the cloud platform.

3.2.1.1 Green Server

Energy efficiency improvement of servers involves power supply and power consumption.

The improvement of the power supply phase focuses on efficiency improvement and transmission loss reduction.

To improve power consumption, the following three aspects must be taken into consideration: On the one hand, the conversion efficiency from power to computing power must be continuously improved to reduce thermal energy consumption. On the other hand, computing power should be matched based on service features and scenarios to improve computing efficiency. At the same time, it is necessary to continue to innovate in new efficient architectures to improve energy efficiency.

Figure 3-2 4-Dimenstion Optimization of Green Servers

3.2.1.1.1 Efficient Power Supply

Power supply includes the power supply link in the equipment room and the power supply system inside the server. The power supply of core network (CN) devices requires high reliability. To cope with the instability of mains supply, the backup power supply system on the power supply link is indispensable. The server supports hybrid power supply of external AC+HVDC, providing more options and possibilities for the backup power system. As a backup power supply system, the HVDC can improve the transmission efficiency comparing with the traditional AC backup power supply system. The battery is directly mounted on the HVDC bus (in the traditional AC mode, the battery needs to pass through the DC inverter to supply power to the servers), which saves the loss of inverters and improves the power supply efficiency.

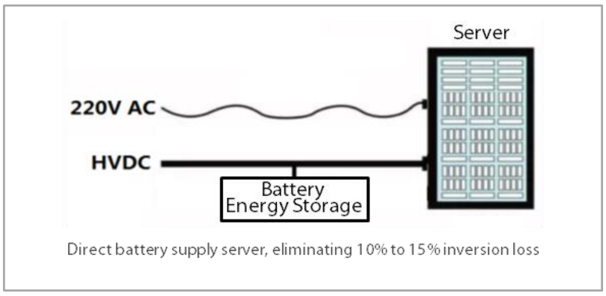


Figure 3-3 AC + HVDC Hybrid Power Supply

The energy efficiency optimization of CPU and memory of the server is critical for the whole power supply system. The major power supply modes of the two elements are multi-phase power supply, and a total quantity of phases is related to maximum power consumption of the CPU and memory. The operating status of the CPU and memory varies with the service situation. In the case of low power consumption, a relatively large quantity of losses will be caused if all phases work, and power conversion efficiency is reduced. The APS (Active Phase Switching) mode is supported in the server power supply design. The number of power supply phases can be automatically adjusted in accordance with the load to improve the conversion efficiency and reduce losses.

Figure 3-4 Active Phase Switching

3.2.1.1.2 Efficient Heat Dissipation

Server performance is continuously improved, and chip power consumption is increasing rapidly. In the future, the power consumption of a single CPU can be increased to 500W. Heat dissipation of chip becomes a huge challenge to the thermal design of products in the future. According to the analysis of ASHRAE, when the power consumption of CPUs exceeds 400W, the current air cooling cannot meet the requirements, and liquid cooling must be used.

Liquid cooling for heat dissipation makes use of the high thermal conductivity and high heat capacity of liquids. It has significant technical advantages in terms of heat dissipation capability, energy saving effect, technical maturity, architecture compatibility, and ecological environment. It has become the focus of governments and enterprises, operators, and Internet enterprises. At present, the entire ICT industry has accelerated the research and deployment of the liquid cooling technology.

Currently, there are two mainstream liquid cooling technologies: cold-plate liquid cooling and immersive liquid cooling. At present, the cold-plate liquid cooling industry is the most mature, the supply chain is the most perfect, which is the most widely used in the market. The cold-plate liquid cooling can continue to adopt the current server architecture, is compatible with air cooling and can be used in both new and legacy equipment rooms. Immersive liquid cooling uses a new architecture, which is more suitable for new equipment rooms, enables lower PUE comparing with cold-plate liquid cooling. It is a development trend of liquid cooling in the future. However, its universality and adaptability still need to be strengthened.

Dimension	Cold Plate Liquid Cooling	Immersive Liquid Cooling
Architecture	Continue the use of the current server architecture, and compatible with air cooling	New architecture, No standard in the industry
	☆ ☆ ☆ ☆ ☆	☆ ☆ ☆
Adaptability	For legacy and new equipment rooms	For new equipment rooms
	☆ ☆ ☆ ☆ ☆	☆ ☆ ☆
Energy Consumption Standard	PUE: 1.2 -1.3	PUE : 1.01-1.09
	☆ ☆ ☆	☆ ☆ ☆ ☆ ☆

Figure 3-5 Comparison Analysis of Two Liquid Cooling Technologies

3.2.1.1.3 Efficient Computing

With the slowdown of Moore's Law and diversified development of computing, especially diversity of application types, general CPUs encounter more and more performance bottlenecks when processing massive computing and massive data/images, for example, low parallelism, insufficient bandwidth, and long latency.

Heterogeneous computing is introduced in more and more scenarios, and hardware such as GPU, FPGA, and NP are used for acceleration. Heterogeneous means that computing units of various architectures such as CPU, GPU, FPGA, and NP are integrated into a hybrid system to perform computing together in a special manner. Each of the different types of compute units can perform the tasks that they do best.

In the 5G era, data traffic will increase exponentially, and it is necessary to focus on bit energy consumption. By using heterogeneous computing (for example, CPU+FPGA) on UPFs of 5GC, the performance and latency bottlenecks of virtualization forwarding can be solved, and the power consumption per bit can be greatly reduced.

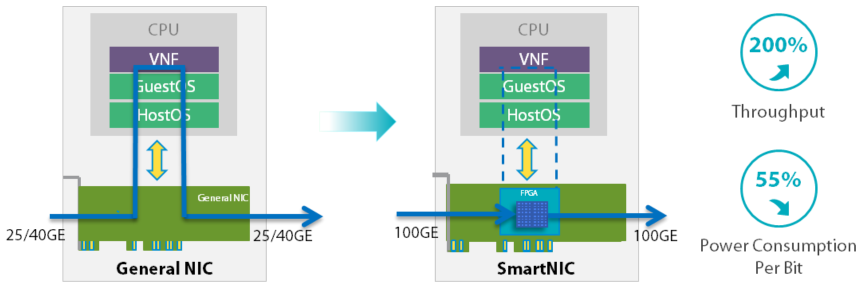


Figure 3-6 Acceleration Solution of SmartNIC

3.2.1.1.4 Efficient Architecture

The classic Von Neumann architecture uses a computing and storage separation architecture, which divides the entire computing architecture into five parts: Computing, control, memory, and input & output. The calculation pipeline includes five stages: instruction fetch (IF), instruction decode (ID), execute (EX), memory access (MEM), and register write back (WB). In many application scenarios, an energy consumption ratio used for calculation is less than 10%, and most energy is used for processing such as memory fetch, cache processing, decoding, and branch prediction. With development of big data and artificial intelligence applications, in a context of a memory wall and a power consumption wall, a conventional computing architecture has increasingly serious impact on emerging data-intensive applications, and a new computing architecture is urgently required to resolve this problem.

The memory and computing integration technology, based on the application requirements, reduces data migration, increases data read and write bandwidth, improves the energy efficiency ratio of computing, and optimizes the joint design of computing and memory to break through the limitations of the existing memory wall and power consumption wall.

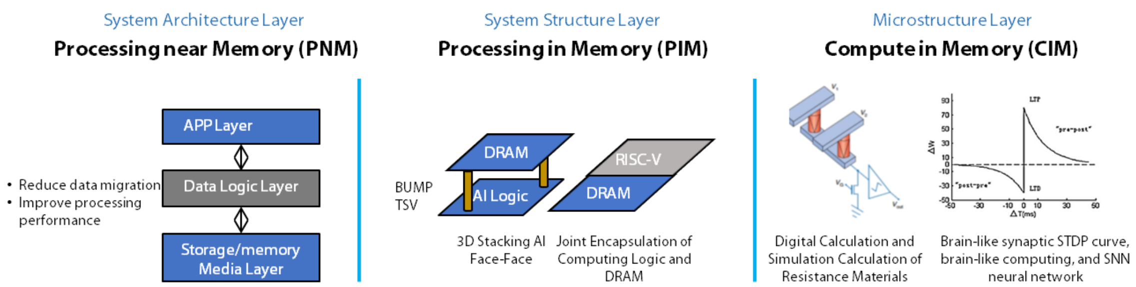
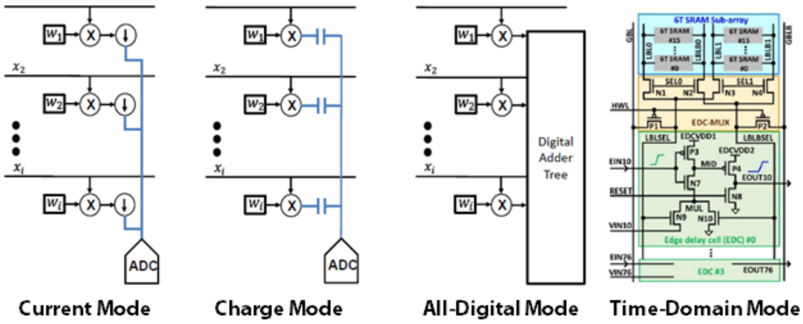


Figure 3-7 Definition of Memory and Computing Integration

Processing Near Memory (PNM) is realized by integrating the computing unit and the memory unit together so that processing can be directly performed in the memory, and a processing result is generated locally and directly returned. This can effectively avoid many problems caused by data migration. For example, reducing data movement and speeding up processing speed can significantly improve performance. Data does not leave the memory for processing, and also helps improve security. The industry is beginning to standardize memory-based PNM. The Storage Networking Industry Association submitted a draft of compute storage/memory system architecture and programming models, including definitions of device architecture, device components, usage models, programming models, and usage examples. One of the NVM Express working groups focused on standardization of storage/memory computing, such as the SNIA draft, will be published in the near future.

Processing in Memory (PIM) is realized mainly by integrating the computing unit inside the memory, and the main storage/memory medium is DRAM. Simple computing, such as conversion between Celsius and Fahrenheit, can be done while the data is being read and written. DRAM is mainly provided by Samsung, Micron, Yangtze Memory, CXMT, and other major manufacturers, so it is difficult for common users to customize it.

There are four main modes of implementation for **Compute in Memory (CIM)**: current, charge, all-digital, and time-domain. Current-based CIM units have fewer transistors; charge-based CIM units can result in high parallelism. The all-digital mode attracts much attention recently, and has the advantage of full precision. The cost per unit area of time-domain mode is lower, but the latency is larger.



Source: Joint Laboratory of Fudan University and ZTE

Figure 3-8 Four Implementation Modes of CIM

Currently, the memory and computing integration technology is still in the exploratory stage. It is expected that it can be put into large-scale commercial use in the near future, reducing the inefficient migration of data inside and outside the data center, and improving the computing energy efficiency at the system level.

3.2.1.2 Green Cloud Platform

In the era of the Internet of Everything (IoE), cloud platforms have been widely deployed and applied in telecommunications, government and enterprise, finance, transportation, energy, and other major industries, facilitating the digital transformation of thousands of industries. Driven by the strategic goal of "carbon neutrality" and "carbon peak", the green cloud platform with energy saving and consumption reduction, efficient deployment, and accurate enabling has become the mainstream underlying infrastructure of modern low-carbon intelligent data centers. How to reduce power consumption has become an important task of the cloud platform.

In terms of deployment and product forms, the green cloud platform can shield the complexity of resources in terms of scale, type, and layout, and provide integrated abstract computing power for multiple applications. In addition, it improves resource utilization, reduces energy consumption, optimizes platform deployment and architecture, precisely empowers applications, and promotes the construction of cloud platforms with energy saving and low consumption in multiple dimensions. It is reflected in the following three aspects:

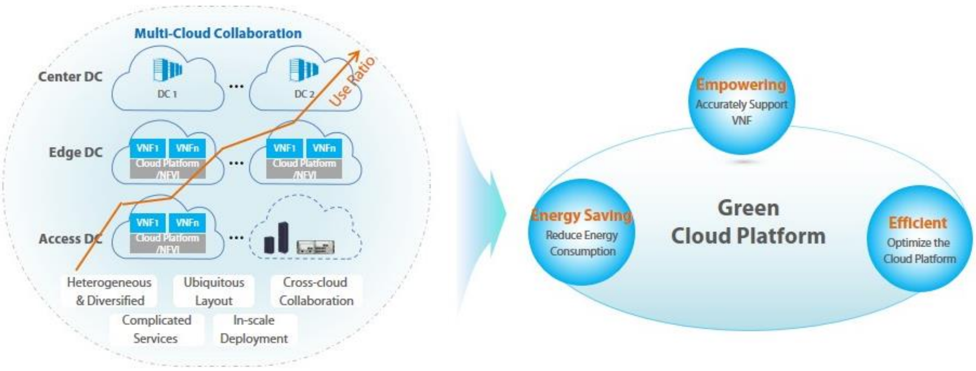


Figure 3-9 Green Cloud Platform

- **Energy saving**: The green cloud platform can predict and adjust resources through intelligent analysis, capacity prediction, and resource pool planning in advance. Afterwards, it uses the dynamic resource scheduling technology and dynamic power management technology to consolidate resource fragments, precisely manage power supplies in a multi-level manner, so as to reduce energy consumption.
- **High efficiency**: The green cloud platform can tailor functions and components in different deployment scenarios to optimize platform resources. And different tidal applications can be deployed together to improve resource utilization.
- **Empowering**: The green cloud platform can tailor and supply resources as required based on application features to achieve optimal resource adaptation. With the features of the cloud native

architecture, core basic capabilities and functional components can be fully shared to empower lightweight applications.

3.2.1.2.1 Dynamical Resource Adjustment to Save Energy and Reduce Consumption

Traditional service platforms are constructed independently, resources cannot be reused or scheduled as required, and the device usage is low. The green cloud platform can be deployed in resource pools and dynamically adjusted on demand, so that all services can be shared and reused to effectively improve resource utilization and save energy. Dynamic resource adjustment refers to the accurate adjustment of resources through dynamic monitoring of resource usage. Key functional technologies involved include intelligent capacity prediction, dynamic resource scheduling, and dynamic power management.

- **Intelligent Capacity Prediction, Planning and Adjusting the Resource Pool Capacity in Advance**

  The intelligent capacity prediction function shows the resource usage by collecting performance indicator data and combining with the AI model, and analyzes the data to obtain the trend prediction and busy/idle period prediction from the global perspective. The trend prediction can guide the expansion and purchase of resource pools, and the busy/idle period prediction can guide the resource pool scaling plan. In accordance with the forecast data, the administrator can precisely define a resource pool scaling plan to manage energy consumption in advance without affecting service QoS, including the setting of power-off resources and low-power-consumption resources. The more accurate the prediction is, the more precise the energy management is, and the more resources are saved.

- **Dynamic Resource Scheduling, Together With Power Management to Reduce Energy Consumption**

  The cloud platform, acting as the infrastructure, uses the dynamic resource scheduling function and the dynamic power management function to implement dynamic optimization of resources and precise control of power consumption. Scattered application instance resources are aggregated to a certain extent, idle nodes are powered off in accordance with the redundancy ratio for energy saving, and power management is performed on powered-on devices in accordance with host load to achieve the purpose of energy saving and efficiency improving.

  The dynamic resource scheduling function supports lossless online migration of service applications to ensure a lower resource fragment rate.

  The dynamic power management function realizes D/G/P/C state management of powered-on hardware based on the policies of services and time periods, the online host redundancy ratio and the limitation of host power.

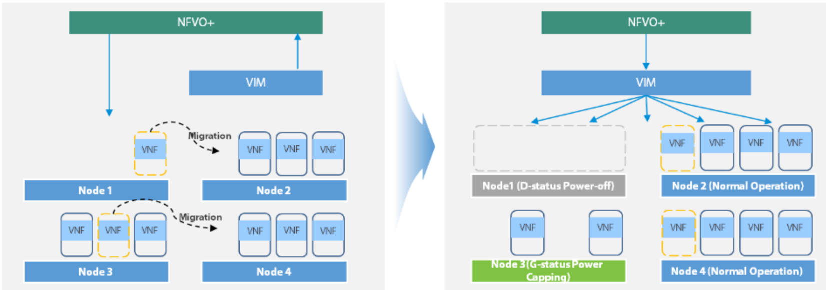


Figure 3-10 Dynamic Resource Adjustment & Dynamic Power Management

According to the practical evaluation of telecom cloud deployment, after dynamic resource adjustment and the precise power management, the cloud platform in the resource pool of 5GC can effectively reduce energy consumption.

3.2.1.2.2 On-demand Platform Optimization to Achieve Efficient Deployment

The cloud platform integrates multiple service types such as IaaS, PaaS and FaaS to enable full-scenario services. However, in some scenarios where the deployment scale is small and hardware resources are limited, such as "edge cloud" and "private cloud", if all cloud platform functions and components are deployed, the system is too “heavy”, and resources are wasted seriously. Therefore, the cloud platform needs to be“tailored”. In accordance with the requirements of service resources in different scenarios, the software is tailored as required, and only necessary components are kept. The redundant components are removed to implement lightweight deployment and improve resource utilization. Based on different tidal effects of services, the "hybrid deployment of applications" can be implemented to improve the comprehensive utilization of resource pools.

- **Software Tailored as Required for Lightweight Deployment of Platforms**

  In accordance with the deployment scenarios and service feature requirements, flexibly combined products and versions can be customized, including component selection, service provisioning, and version tailoring, the lightweight cloud platform is deployed, lowering resource occupation, and reducing energy consumption. The following figure shows an example of tailoring the cloud platform software.

Figure 3-11 Lightweight Cloud Platform

- **Flexible Deployment of Services, Improving Comprehensive Utilization of Resource Pools**

According to the analysis of application scenarios, regular service tidal phenomena exist in network O&M, service provisioning, and other services, with periodic peak and valley traffic ranges. Big data analysis and other services raise high requirements for resources instead of time-insensitive, so the system can operate "at any time".

With service features, the cloud platform can flexibly implement hybrid deployment of applications. Different applications are deployed on the same node. The big data analysis service can operate in the lowest traffic range of the network O&M service, saving costs, and improving the comprehensive utilization of the resource pool.

In "hybrid deployment of applications", the unified intelligent scheduling platform needs to be deployed to dynamically perceive the computing power load of each resource pool, and intelligently predict the characteristics of application tidal waves. In this way, hybrid scheduling of multiple resource pools and services is implemented, the peak value is reduced, and the comprehensive resource utilization is improved.

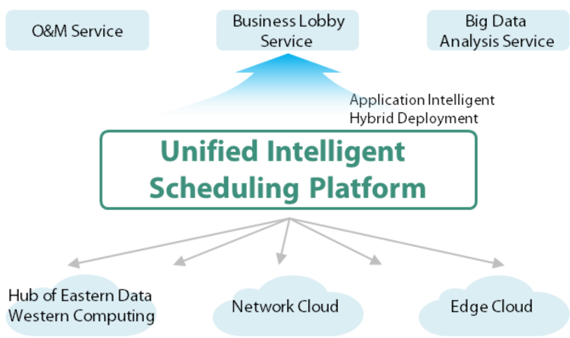


Figure 3-12 Hybrid Deployment of Tide Services

3.2.1.2.3 Accurate applicable to applications, with Agile Empowering

In the digital transformation of different industries, different services in different industries have differentiated requirements for network performance, transmission latency, cost budget, and security protection. However, the cloud platform has limited resources, so it is necessary to focus on application scenarios, precisely orchestrate resources, and provide resource for applications in a differentiated manner to reduce resource consumption.

On the one hand, it is necessary to focus on the key features of applications and differentiated resource requirements, specify resources accurately, and implement fast resource selection and simplified deployment, to build a customized cloud platform for services. On the other hand, focusing on the separated structure of the network, the cloud platform provides precisely adaptive resource in accordance with the service features and environmental conditions in different locations of the network. It achieves the integrated resource supply, O&M, and service operation of "edge", "network", and "cloud" resources in the entire network, so as to achieve the purpose of on-demand resource scheduling, efficient utilization, and quality assurance, and better match the transformation and innovation requirements of different industries. In the 5GC, based on the differentiated requirements of applications and the complexity of cloud platform resource requirements, the enhanced VNF orchestration management center can be used to accurately adapt resources of different clouds and different applications in the telecom network, improving resource utilization.

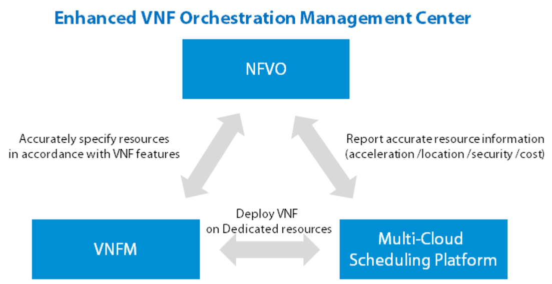


Figure 3-13 Enhanced VNF Orchestration Management Center

In addition, the construction of networks in industrial cloud is moving from the cloud resource phase to the cloud native phase. In the cloud resource phase, resources are regarded as the center, and all applications share a unified resource pool, such as computing, storage, and network resources. However, application components are still in silo-type status. In the cloud native phase, applications are regarded as the center, the core basic capabilities and functional components of service applications are standardized and platform-based to form the common capability layer, so that various types of applications can be empowered agilely to implement agile development, fast delivery, and automated deployment. Therefore, sharing common capabilities and empowering applications with agility will also become the key capabilities of the cloud platform.

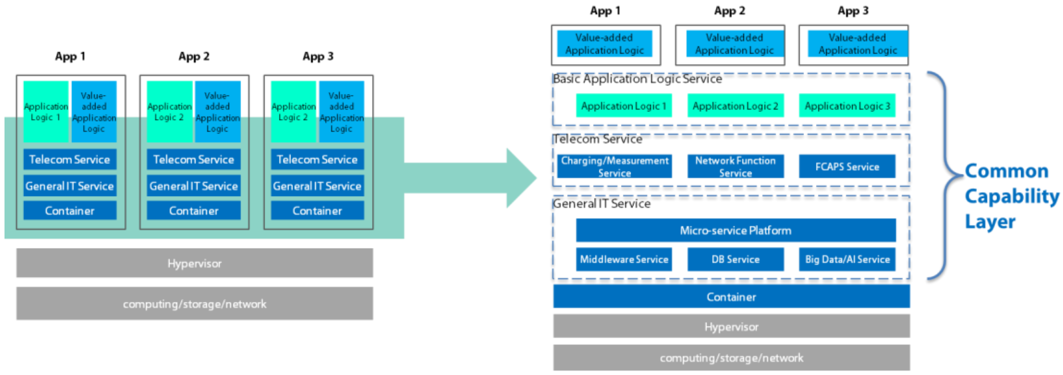


Figure 3-14 Constructing a Capability Sharing Layer

As shown in the above figure, the industry is also promoting the construction of the general capability layer of the 5GC. Based on the container platform, general IT services, telecom services, and basic application logic services are built as the shared PaaS capability layer for unified O&M. These layers can be invoked by various APPs as required to meet the requirements of fast iteration.

3.2.2 NE Optimization – Green Empowering

The 5GC, which is designed based on the Service Based Architecture (SBA), implements green energy saving through the introduction of a new network architecture and cloud technologies. SBA is the most important iconic innovation of 5G, enabling 5GC to modularize functions, unify interfaces, simplify architecture, and decentralize, and makes the 5GC green and energy-saving through on-demand deployment, on-demand reuse, and on-demand interworking. As a key technology of the 5GC, CUPS (Control and User Plane Separation) makes 5GC networking more flexible, services more reliable, the forwarding plane closer to users and applications, and resource more efficient by separating the control-plane and user-plane and interconnecting them through Full-Mesh.

By using cloud native technologies, the 5GC can be orchestrated and assembled in the form of microservices to decouple software from hardware. In addition, cross-resource deployment is supported, so that various hardware resources can be fully utilized to maximize energy saving. In addition, microservices are more lightweight, and the types and specifications can be selected in accordance with different types of services to reduce consumption most effectively.

3.2.2.1 SBA & Cloud Native Architecture, Endogenous Energy Saving

The adoption of SBA & Cloud-Native architecture helps to improve the flexibility and maintainability of service components, achieve continuous iteration and O&M automation through agile methods and

DevOps procedures, and flexibly build an ultra-lightweight 5GC based on fully converged microservice components. Based on the endogenous intelligent algorithms, the status of devices where VNFs are located can be intelligently evaluated based on parameters such as CPU load, user online rate, data throughput, and monitoring results of surrounding NE status, thereby realizing automatic scale in/out, dynamic service scheduling, and automatic adjustment of CPU frequency, to optimize resource utilization. The endogenous energy-saving feature helps operators greatly reduce network energy consumption and realize green 5G.

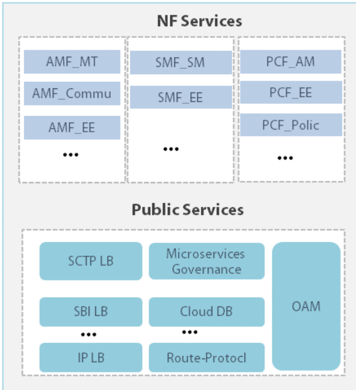


Figure 3-15 Cloud Native & Microservice Architecture

Taking a single DC to meet the traffic of 40 million users as an example, compared with traditional dedicated hardware devices, the annual energy consumption of cloud native hyper-converged 5GC is reduced by 30%.

3.2.2.2 High-Performance Integrated UPF, Intelligent Energy Saving

With the introduction of the C/U separation architecture, centralized, efficient, and intelligent deployment can be implemented on the control plane, and distributed, low-cost, and flexible deployment can be implemented on the user plane through the remote control plane management. Therefore, as the most important data forwarding NE on the user plane, the UPF can focus on high-performance data forwarding, providing users with high-speed bandwidth experience. With the large-scale commercial use of 5G networks, mobile user traffic will grow explosively. The high-performance forwarding of mobile data has become the core requirement of users, and the high-performance and integrated 5G forwarding plane NE UPF becomes an important implementation point for 5G energy saving.

To meet the requirements of high-performance forwarding, the CPUs of traditional UPFs need to run at 100% no matter whether the service is busy or idle, and their power consumption is always the maximum. Because of certain time regularity of work time and rest time, personnel density and motion route, the tidal effect of traffic is obvious. Therefore, how to implement energy saving in accordance with the traffic on the user plane to achieve intelligent energy saving and emission reduction is a problem that needs to be solved for the high-performance integrated UPF.

The high-performance integrated UPF solution has built-in energy-saving algorithms and a multi-thread load balancer to adapt to traffic model changes. According to the network traffic load, hosts are powered off in proportion or CPU cores are slept as required, and a small number of CPU cores are intelligently on-duty observation. The process does not require manual interference. To ensure lossless services upon sleep/wakeup of the core, the traffic of an NE is distributed based on the ring queue. The ring queue is continuously taken over through fast handover of the CPU core switch, so that smooth services are not damaged during the handover.

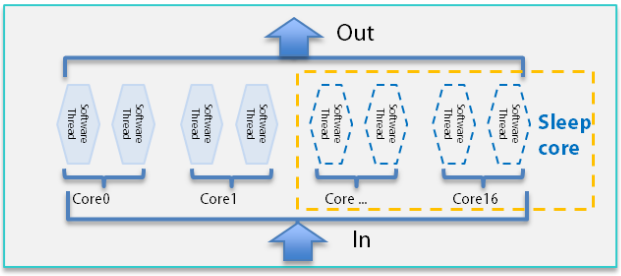


Figure 3-16 High-Performance Integrated UPF Intelligent Energy-saving Solution

For the high-performance integrated dedicated U-plane (UPF/SAE-GW-U/GGSN-U), the 400Gbps forwarding throughput is set as an example. Compared with the SAE-GW/GGSN of traditional hardware, the power consumption can be reduced by 50%.

3.2.2.3 IoT Feature Optimization, Accurate Energy Saving

In the IoT era, different types of IoT users have different service features. For example, features of object-end devices such as a streetlight and a manhole cover include fixed access, small-packet transmission, and timing communication. The features of object-end devices such as security protection and monitoring include fixed access, big data, and real-time transmission. Therefore, for IoT users with differentiated service features, the 5GC can provide specific energy efficiency management solutions:

Service Feature	Energy-Efficiency Management Solution
Fixed access	Low mobility: Current NR/eNB of paging policy
Timing communication	Intelligent power saving: eDRX/PSM
Small-packet transmission	Optimization of small-package transmission: Optimizing the control plane and the user plane
Real-time transmission of massive data	Massive data: Bandwidth control

In addition to the energy efficiency management in the signaling flow, the system can also optimize the storage of user information in accordance with different features of IoT users, and accurately manage and optimize the storage and computing of internal data structures such as user contexts, session management contexts, and mobility management contexts.

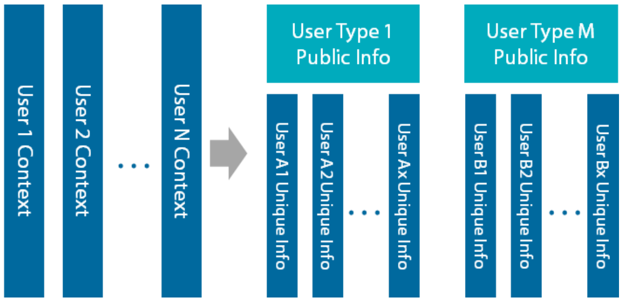


Figure 3-17 Distinguishing User Types for Unified Storage of Common Information

After the above all-round optimization of IoT features, the energy consumption of a single IoT user is about 1/5 of a single ToC user, and the energy consumption is reduced by about 80%.

3.2.2.4 All-Scenario Private Core Network, Green and Empowering

With the continuous maturity of the 5G end-to-end industry chain, digital transformation is urgently required in vertical industries aiming at "5G private network + application" under the three 5G application scenarios. However, there are many kinds of vertical industries, and the refined scenarios are diversified. The 5G ToB market features fragmented requirements. Private network services face the following challenges:

- Convergent access, "multiple" NEs
- Thousands of industries, "multiple" scenarios
- "Multiple" service requirements
- Internet of Industry, "fast" development of technologies

To match the characteristics of industry applications, private networks oriented to ToB industry applications need to support diversified communication service capabilities, and telecom operators and industry customers urgently need all-scenario private 5GCs.

Compared with the large- capacity and full-featured ToC 5GC NEs, the private 5GC needs to be tailored or enhanced in accordance with the requirements of industry customers, and the network is more customized. In addition, the 5GC supports diversified and flexible service deployment capabilities, on-site access, edge-side deployment, and centralized deployment, meeting the requirements of industry customers, and significantly reducing the consumption of physical resources to empower the industry green transformation.

For example, for small/medium-sized industrial applications with 5-10Gbps throughput requirements, the private 5GC is deployed on one server, reducing power consumption by 90% compared with ToC 5GC.

3.2.3 Intelligent O&M – Green Brain

With the large-scale deployment of 5G networks, networks will face the coexistence of multi-system (2G/3G, 4G, and 5G) environments, increasing the difficulty of management coordination and interoperability. To reduce the difficulty in locating faults in a cloud-based network architecture, intelligent O&M has been promoted to an important strategic direction in the current overall network O&M. At the same time, to achieve the goal vision of green transformation of the core network, the O&M management and control layer, which is regarded as the core intelligent brain, becomes an important anchor point for green transformation of the core network. It needs to solve many challenges to the current green operation transformation:

1. Where is low energy consumption? How to achieve optimal design and precise deployment of services and energy consumption in the entire network through eastern data and western computing and dynamic distribution of resources?
2. Which service consumes high energy? Identifying services with high energy consumption is the prerequisite for energy consumption optimization. How to accurately calculate service energy consumption?
3. How to optimize energy consumption? How to save energy at the service level? How do hardware and business optimization collaborate together?

To cope with the above challenges of green O&M, a green brain, which is oriented to intelligent O&M of the core network is built, supports the entire lifecycle of O&M planning, construction, maintenance, optimization, and operation, and achieves accurate and optimal matching of resource utilization, energy consumption, and service quality.

3.2.3.1 Green Network Planning and Deployment

Energy saving should start from the source and meet service requirements for energy saving and emission reduction to achieve a green design. Through resource monitoring, operators can master the real-time operation status and resource utilization of the network. Historical data simulation is used to construct a model that can accurately reflect the actual resource usage, helping to establish a resource simulation planning center. The simulation planning center can predict service and energy consumption development trends, intelligently evaluate resources and energy consumption, and establish resource development trend models. When a new service request occurs and the network needs to be deployed or expanded, the collection and management system sends service indicators and prediction models to the resource simulation planning center. The resource simulation planning center matches the optimal resources that meet the service requirements in accordance with the AI model. Based on service and energy consumption planning simulation, the simulation planning system deploys network services in accordance with the optimal resource indication. The intelligent on-demand deployment of the network not only effectively ensures service QoS, but also achieves the optimal balance of resource consumption, improving the matching accuracy between services and energy consumption.

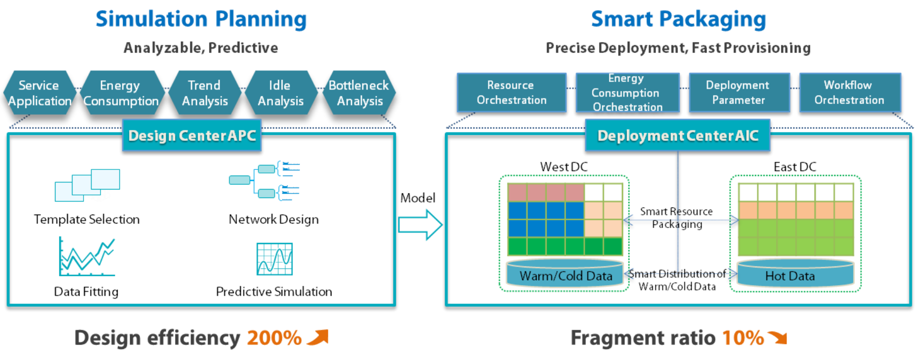


Figure 3-18 Simulation Planning & Smart Packaging

In a cloudified core network, the traditional deployment mode of resources usually adopts basic orchestration and configuration solutions, and resources are sequentially orchestrated and distributed in the existing resource pool system in accordance with requirements. This mode easily causes resource fragments and low overall resource utilization. The smart green packaging algorithm is used to optimize the deployment solution, and match resource specifications with energy consumption requirements to comprehensively evaluate energy consumption and distribute resources. The overall resource scheduling feature is used to maximize the value of each CPU core.

3.2.3.2 Visual Management of Energy Consumption

Traditional network management is only limited to the monitoring of network infrastructure indicators or faults. In addition, traditional energy management is only limited to the hardware energy consumption of equipment rooms and servers, and there is no energy consumption management of software services such as slices and NEs. Because energy consumption management is not associated with service monitoring such as device indicators and faults, the management system cannot accurately locate the root causes of energy consumption.

To solve this problem, comprehensive monitoring algorithms of the resource-energy consumption need to be introduced to detail the overall energy consumption, combine it with the network topology and service features, and form a correlation relationship to build an integrated green intelligent monitoring solution.

First, multi-level monitoring of managed devices needs to be implemented, including resource layer monitoring, network layer monitoring, and service layer monitoring. Second, power consumption of hardware devices such as DCs, racks, and servers needs to be monitored. The energy consumption, resource, network, and service association models are established through manual experience or machine learning, especially the association model at the service layer. The energy consumption ratio is associated through service features, such as traffic, latency, and bandwidth. Through this visual analysis process, it is possible to easily locate the services with high energy consumption and quickly troubleshoot the causes of high energy consumption at the network layer and resource layer. The green intelligent monitoring solution discovers high-energy-consumption devices by analysis, multi-dimensional statistics and intelligent drilling from real devices to virtual devices, from devices to services, to implement accurate energy consumption monitoring.

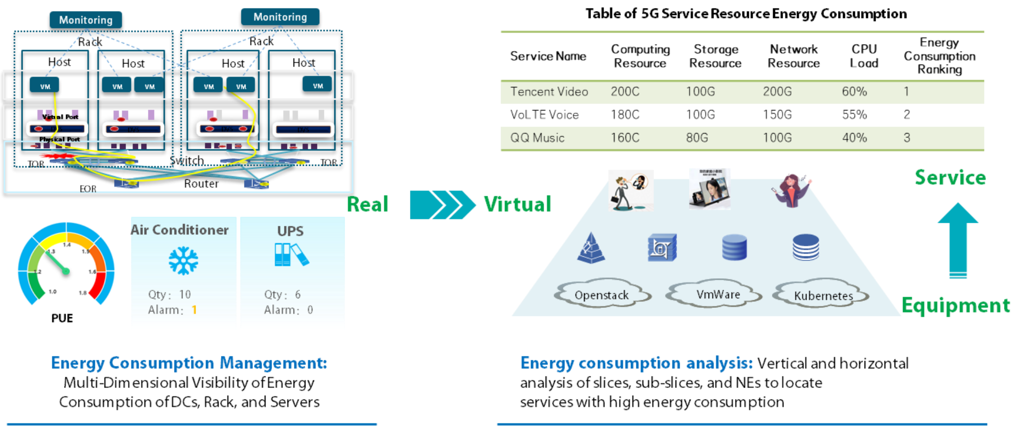


Figure 3-19 Green Intelligent Monitoring Solution

3.2.3.3 Intelligent Energy Consumption Optimization

In traditional network O&M, if a device fault occurs, network fault location or system optimization can be implemented manually or automatically to meet the specified SLA requirements. Continuous system optimization also inevitably causes larger resources to be invested for implementation. For example, if resource energy consumption is not well controlled, a relatively high energy consumption may be used in exchange for a small amount of performance. With the green intelligent monitoring solution, the energy consumption of the DC equipment room can be associated with resources, networks, and services. Based on this relationship, perform joint energy consumption optimization between systems to ensure customer SLA requirements, thus achieving the best match between the service SLA and energy consumption. The energy consumption optimization solution includes three levels:

- **"Cell-level" optimization of green NEs**: The NE-level Microservice components implement system energy saving through self-sleep, intelligent NE scaling, and POOL dynamic migration. Through the perception of NE traffic, the energy-saving mode can be intelligently switched based on the busy/idle status.
- **"Organization-level" optimization of green DCs**: It is possible to make policies for frequency reduction, sleep, power-off, and power control of DC-level devices, and implement energy saving adjustment in different resource statuses. The dynamic cross-DC migration of services are performed without user perception to implement intelligent resource defragmentation, avoiding continuous or spread resource fragments. The intelligent temperature and humidity control policy associated with service or device status can be used to intelligently adjust the environment in the DC equipment room to save energy.
- **"Eco-level" green slices**: Energy consumption optimization at the CN sub-slice level is implemented through policies such as UPF resource energy-saving control. Associated wireless and transmission sub-slice-level energy consumption optimization can achieve end-to-end slice-level optimization. Based on the cross-slice busy/idle scheduling scenario, the energy consumption optimization solution for the entire network with multiple slices can be implemented.

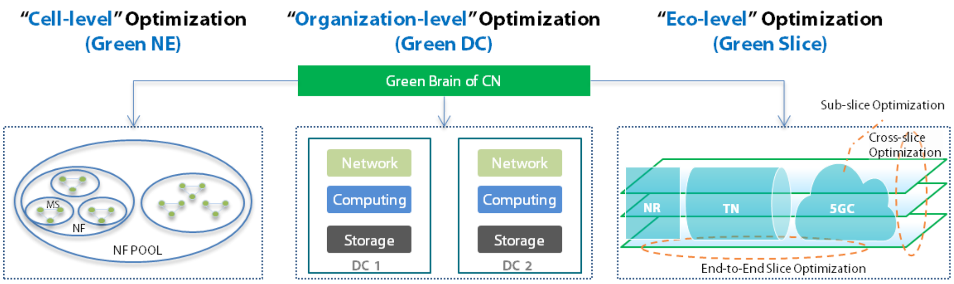


Figure 3-20 Multi-Hierarchical Energy Consumption Optimization

The network-wide multi-dimensional and multi-level optimization solution is used to achieve optimal matching of service quality and energy consumption based on real-time perception analysis of energy consumption, environment, and services.

3.3 Green 5G Core Practice

3.3.1 “4+3+2” Architecture of GreenEngine

With the large-scale construction and wide applications of 5G networks, a large number of base stations, edge sites, and data centers are being deployed, the network traffic continues to increase, the energy consumption increases greatly, and the green transformation of communication networks and applications is urgent. As the brain of the 5G network, the core network needs to continuously adopt innovative technologies and explore energy-saving and consumption-reducing technologies to implement the green transformation, and provide impetus for the green development of other network domains and industrial applications.

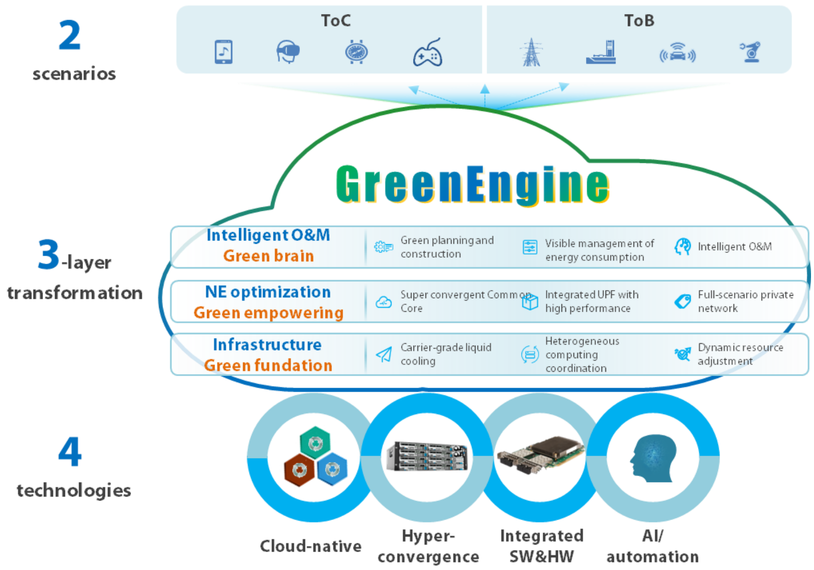


Figure 3-21 GreenEngine “4+3+2” Architecture

Based on 5G Cloud Core, ZTE focuses on "energy saving" and "efficiency improvement", also takes on the mission of "empowerment", to launch the GreenEngine solution. Based on the four major technologies of cloud native, hyper-convergence, integration of software and hardware, and AI/automation, ZTE implements three-layer end-to-end green transformation and upgrade from infrastructure, network elements, to O&M. Facing the ToC and ToB scenarios, ZTE has built the GreenEngine in the computing power era to help operators' networks save energy, empower digital and intelligent production in the industry, and help thousands of industries achieve the "dual-carbon" goal.

- **Infrastructure layer**: ZTE launches cold-plate liquid-cooling servers to dissipate high-energy-consumption components such as CPUs efficiently, safely, and reliably. After cold-plate liquid-cooling servers are large-scale used in the resource pool of the data center, the PUE will be reduced to below 1.2, saving a large amount of power in the core equipment room. The TECS cloud platform supports hierarchical management of power consumption, and performs fine control for hosts in P-state/G-state/C-state in accordance with the power saving policy. The cloud platform supports collaborative scheduling and management of XPU heterogeneous computing power to improve the efficiency of computing resources. For example, the NEO cloud card offloads the virtual layer to the card, reducing the occupation of CPUs of the cloud platform and increasing the computing power density of hosts by more than 20%.

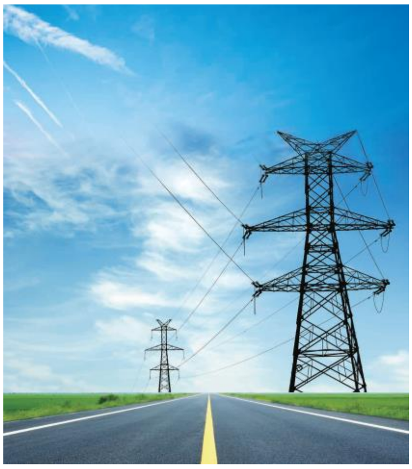
- **NE layer**: The converged core network Common Core is based on cloud native architecture, supports dynamic scaling of VNFs, and implements on-demand resources to save energy endogenously. The high-performance integrated UPF has a built-in RSE inference engine, which triggers intelligent sleep and wakeup of CPUs in accordance with the CPU load to implement intelligent power saving on the user plane. Oriented to diversified ToB scenarios, ZTE provides a full range of hyper-integrated private network products to accurately meet differentiated and customized requirements for network capacity, functions, footprints, and SLAs in the industry, and achieve compact deployment of private networks and optimal resources.

Pocket	Simple	Standard	AIO Cloud-network	Embedded
				
PC version	1 server	3 servers	Cloud-network cabinet	Embedded BBU or OLT chassis
Emergency relief, family	Medium/small industry 5~10 Gbps	Big/medium industry, highly reliable 10~50Gbps	Medium/big industry Cloud/Network/Application Convergence ~20Gbps	Room space is limited or wireless and fixed are deployed together 5~10 Gbps

Figure 3-22 Full-Scenario Product Forms for Private Core Network

- **O&M layer**: The management domain perceives the service load in real time, predicts the busy/idle load based on AI, performs visual management through the energy consumption dashboard, and locks the services with high energy consumption through carbon energy analysis. The O&M layer coordinates with the infrastructure and NE layer to optimize end-to-end energy consumption. Through component-level scaling, resources are allocated on demand. Through dynamic service migration, resource fragments are reduced. In addition, no-load CPUs or hosts are intelligently slept or powered off in proportion, so that the end-to-end intelligent dynamic consumption reduction of the entire system is implemented.

3.3.2 All-Scenario Private Core Network Empowers Smart Grid



In the electric power field, rapid development of new clean energy such as wind energy and solar energy can effectively reduce greenhouse gas emissions caused by thermal power generation. However, wind energy and solar energy are greatly affected by weather and regions, and their output power is unstable. In addition, there are many sites with wide distribution, which brings challenges to the grid-connection of new energy.

To ensure that new energy can be efficiently, stably, and reliably connected into the power grid, ZTE, together with China Mobile and NR Electric, jointly released the industry's first 5G TSN green power grid solution. It uses wireless communication networks including the GreenEngine 5G core network as the technical base and integrates 5G TSN/5G LAN/URLLC/SLA precise control to provide the power grid industry with promised SLAs. This solution helps to implement "one last mile" wireless deterministic access instead of expensive optical fibers, and provide flexible expansibility for large-scale and distributed renewable energy deployment. This solution provides deterministic low latency, and assists the power grid in real-time monitoring and scheduling to avoid the security impact on power grid lines caused by the instability of new energy grid connection, thus ensuring the secure and stable operation of the power grid.

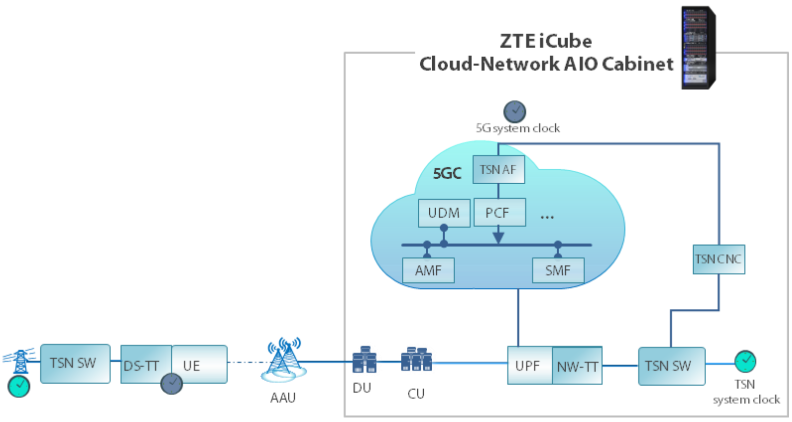


Figure 3-23 5G TSN Green Grid Solution

In addition, this solution can be widely used in power service scenarios such as grid differential protection, power distribution automation, grid synchronous phasor measurement (PMU), and accurate grid load scheduling to accelerate the efficient coordination of different types of power, and help the power industry achieve the strategic goal of "peak carbon and carbon neutrality".

3.3.3 High-performance Integrated UPF Empowers Smart Aluminum Industry

China is a major country of electrolytic aluminum production. How to promote the green and low-carbon transformation of the electrolytic aluminum industry? In addition to introducing clean energy to optimize the energy structure, the digitalization of production process and management has a profound impact on production efficiency and energy conservation.

Yunnan Sunho Aluminum, as one of the largest electrolytic aluminum manufacturers in the southwest China, has cooperated with China Mobile and ZTE to build a new digital infrastructure for the aluminum industry, deployed a private network + MEC for the enterprise based on the GreenEngine solution, and built an industrial Internet innovation platform, to realize refined energy consumption control of production nodes, and collect and analyze real-time data of factory environments, production lines, and personnel to achieve automatic and intelligent production and management.

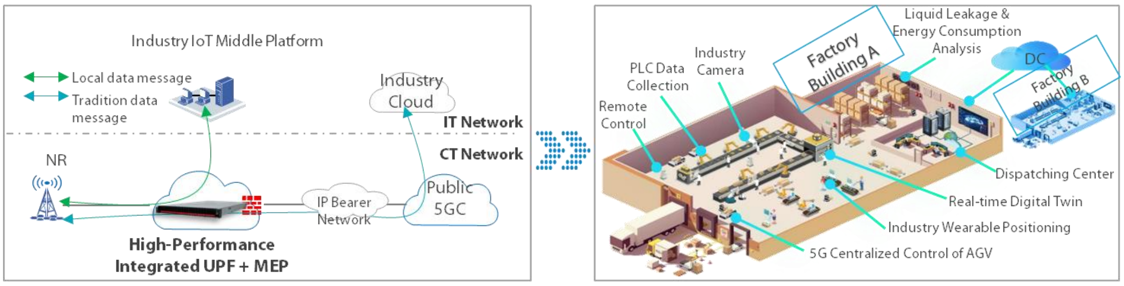


Figure 3-24 Smart Aluminum Digitalized Infrastructure Architecture

With the support of the new digital base and innovative platform, Yunnan Sunho Aluminum has gradually realized automated production, intelligent management and control, and integrated precise operations, greatly reducing the power consumption per unit of aluminum, saving labor costs, and accelerating the transformation to achieve green sustainable development. As the first 5G smart factory in Yunnan, the intelligent green transformation of Yunnan Sunho Aluminum has a great demonstration effect in the region and the industry, and is becoming a future-oriented benchmark for the modern metallurgical industry, promoting the sustainable development of the non-ferrous metal smelting industry.

4 Outlook of Green Evolution of the Future Network

Oriented to 5G-Advanced and 6G, ZTE adheres to the green and low-carbon development concept. By regarding energy saving and green empowering as the target, focusing on intelligent automation, full lifecycle, new architecture, and new materials, ZTE continuously researches innovative energy-saving technologies. On the one hand, ZTE provides green 5GC to reduce the energy consumption of operator network. On the other hand, ZTE acts as a green empowering role to meet the green transformation requirements of industry customers and help China achieve the goal of carbon peak and carbon neutrality. At present, innovative energy-saving and consumption-reducing technologies are emerging for 5G networks. In the process of achieving green transformation through in-depth integration with thousands of industries, 5G networks also require cooperation with industry partners to explore a smooth evolution and work together to solve green problems.

1) Rational Use of AI/ML for Efficient & Intelligent Energy Saving

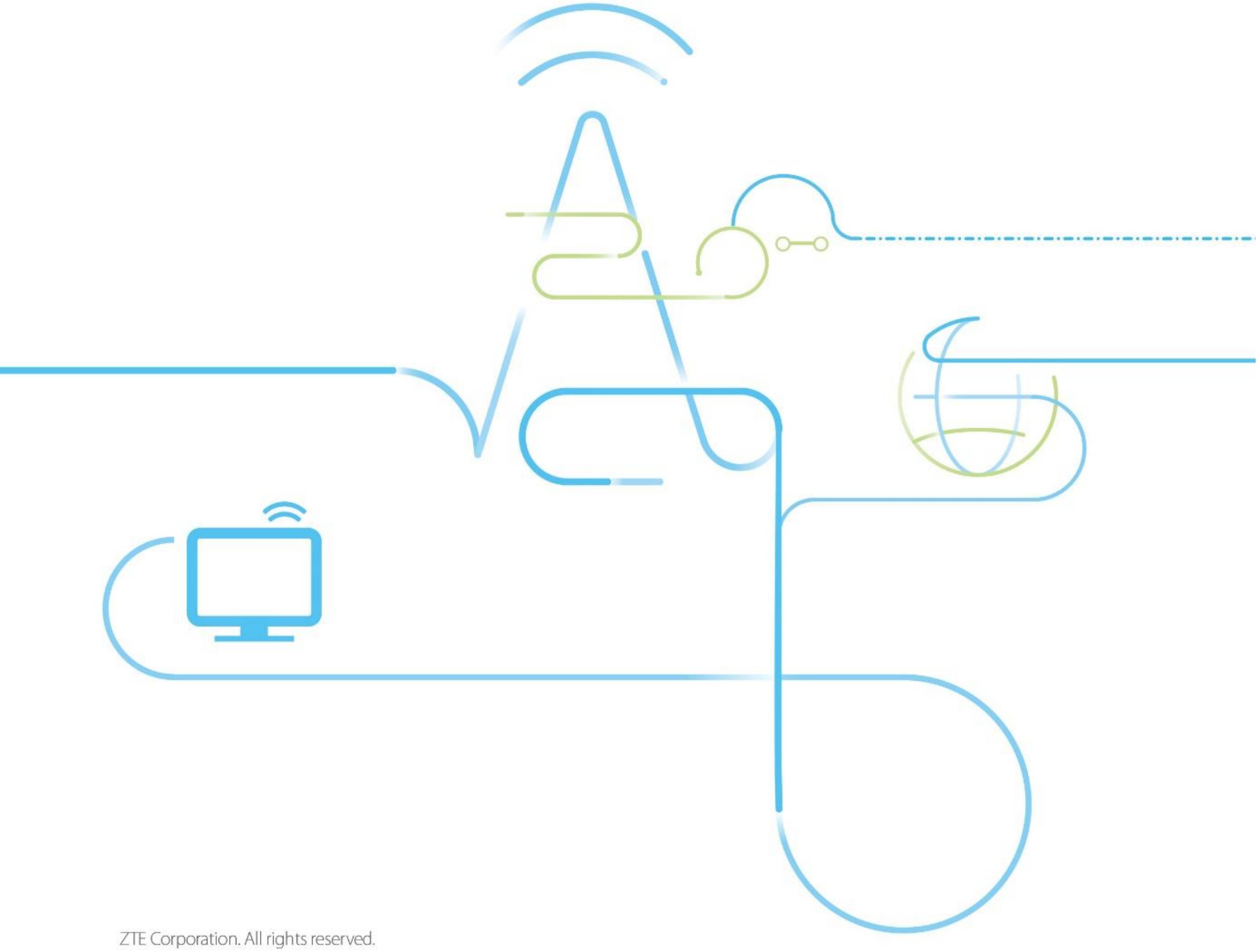
The intelligence and machine learning technologies have inherent advantages in adapting to a constantly changing network. They have a capability of understanding a complex relationship between different network parameters by means of learning and training, so as to help O&M persons implement energy-saving design and optimization. However, model training, verification, and testing on different deep neural networks require a large quantity of data storage and preprocessing devices, which brings high energy consumption. In addition, dedicated hardware used for training and inference also brings high energy consumption. Therefore, the industry needs to jointly promote AI/ML development, improve algorithm efficiency, and co-build an industry-based network energy-saving knowledge base.

2) Single-Domain to Multi-Domain, Online Monitoring and Optimization of End-to-End Network Energy Efficiency

Currently, energy consumption is measured and optimized in offline mode, and the energy consumption of each network domain is independent of each other. Therefore, end-to-end energy efficiency online monitoring, online optimization, and joint energy consumption optimization cannot be implemented, and the best match between end-to-end service SLAs and energy consumption cannot be achieved. Therefore, ZTE needs to promote end-to-end network energy efficiency monitoring and optimization from single-domain online energy monitoring and optimization to multi-domain collaboration.

3) **Design a Green Network Based on the Lifecycle from Network Planning to Deployment Optimization**

The traditional network design is based only on the network services, and power consumption is only regarded as a requirement instead of a network design objective. In the future network design, a green network needs to be designed based on the network lifecycle, aiming to save network energy and reduce consumption. In combination with service requirements, resources and energy consumption are orchestrated and coordinated. Through real-time perception analysis of network-wide energy consumption, environment, and services, the optimal matching between service quality and energy consumption can be achieved.



ZTE 中兴通讯股份有限公司 ZTE CORPORATION

NO.55, Hi-tech Road South, ShenZhen,P.R. China Postcode:518057

Website:www.zte.com.cn Tel:+86-755-26770000 Fax:+86-755-26771999