HUAWEI

Version 2024

Cloud Computing 2030

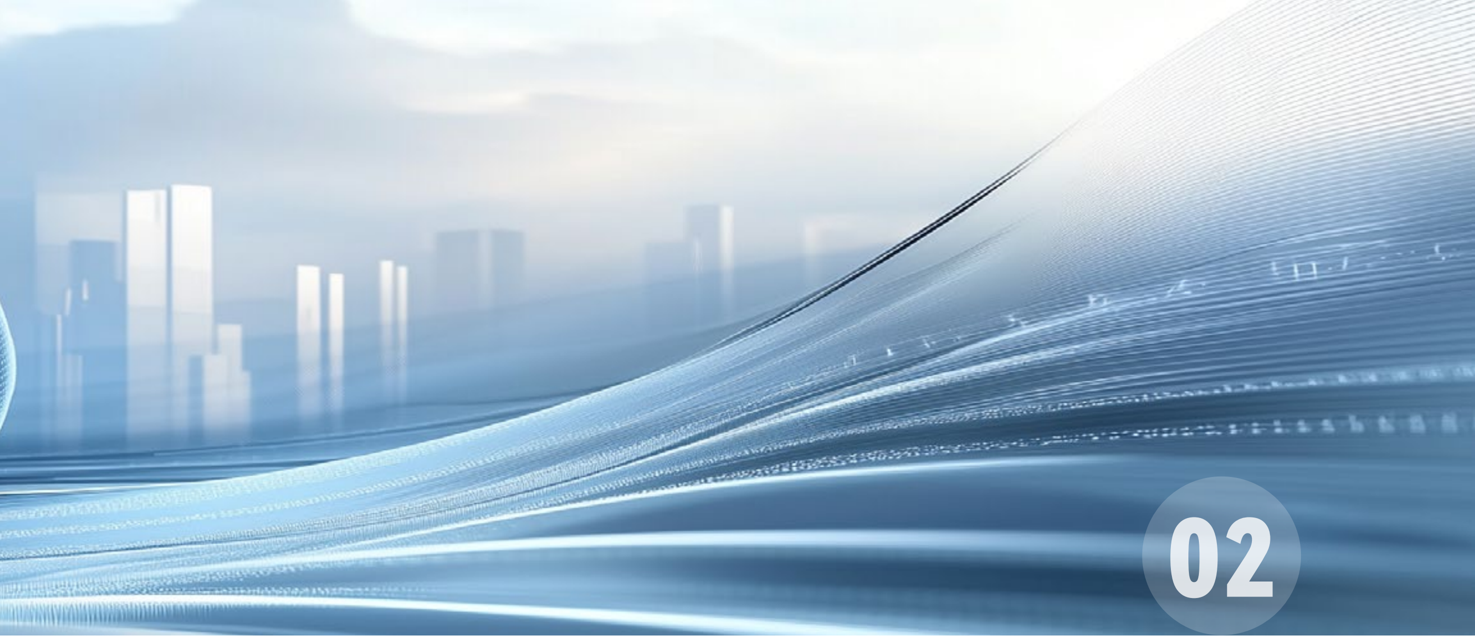
Building a Fully Connected,
Intelligent World

Contents

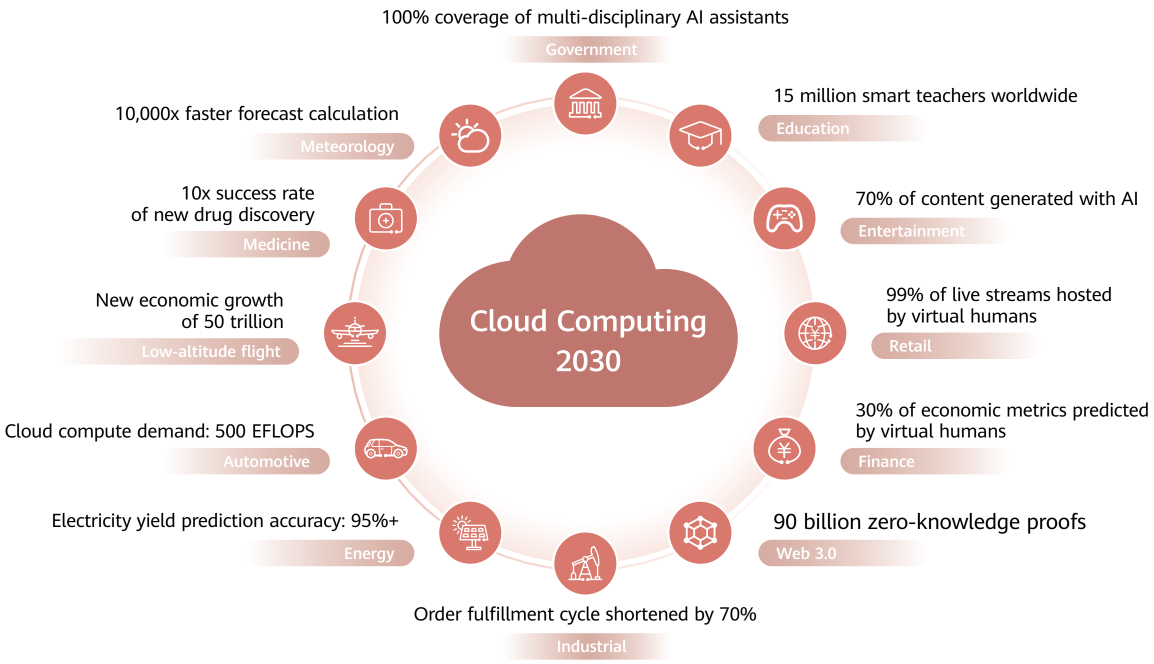
Macrotrends

Since the dawn of the Internet in the late 20th century, we have traversed a remarkable technological landscape. The emergence of cloud computing at the onset of the 21st century and the current surge in artificial intelligence signify not just advancements but a transformation in how we harness human ingenuity. The convergence of cloud and AI has given rise to an intelligent world from smart homes to smart cities, from precision healthcare to personalized education, and from intelligent manufacturing to FinTech innovations. Cloud and AI are the cornerstones of limitless potential for societal development, propelling enterprise innovation and growth to new heights.

By 2030, the ubiquity of cloud computing is anticipated, with an estimated 3 billion smart devices offloading their computational demands to the cloud. Intelligence will be omnipresent, with an expected 1.5 billion enterprise employees benefiting from personalized AI assistants. Furthermore, it is projected that 80% of enterprise applications will be either built anew or rebuilt with AI at their core. The physical world is being reshaped, leading to a data revolution in the 3D space, where the volume of data is expected to exceed current levels by a million-fold. Approximately 500 million individuals will venture into the realm of spatial computing, a fusion of virtual and real worlds.



Future Industry Scenarios


Cloud Computing 2030
100% coverage of multi-disciplinary AI assistants
Government
15 million smart teachers worldwide
Education
70% of content generated with AI
Entertainment
99% of live streams hosted by virtual humans
Retail
30% of economic metrics predicted by virtual humans
Finance
90 billion zero-knowledge proofs
Web 3.0
Order fulfillment cycle shortened by 70%
Industrial
Electricity yield prediction accuracy: 95%+
Energy
Cloud compute demand: 500 EFLOPS
Automotive
New economic growth of 50 trillion
Low-altitude flight
10x success rate of new drug discovery
Medicine
10,000x faster forecast calculation
Meteorology

2.1 Pharmaceutical: AI Boosts Drug Design Success 10-fold and Halves Development Time

The development of a new drug is often a lengthy and costly process, spanning over 10 years and requiring an investment of over USD1 billion. However, despite these substantial resources, the success rate of new drug discovery remains disappointingly low, hovering around 10%. This low success rate can be attributed to various factors, including low clinical efficacy, high toxicity, poor druglikeness, insufficient market demand, and unsuccessful product strategy.

Fortunately, AI is here to help. Throughout the four phases of drug development — target discovery, drug screening, lead optimization, and preclinical testing — AI significantly improves efficiency through tasks like AI molecule generation, ADMET property prediction, and molecular dynamics (MD) simulation. Existing data indicates that AI can accelerate drug design by 70% and boost success rates tenfold. The potential impact is immense.

Today, large scientific computing models play a pivotal role in small molecule drug design, spanning diverse innovative drug R&D tasks. These tasks include the development of new antibiotics, antitumor drugs, and drugs targeting the central nervous system, and the discovery of druglike natural products, all of which have yielded impressive outcomes.

Projections suggest that by 2030, AI technology will be integrated into every phase of new drug discovery, potentially reducing the typical 10-year R&D cycle to just five years or less.

2.2 Meteorology: A Data-Driven Earth Decoder Accelerates Weather Forecasting 10,000 Times

According to a Nature Communications report, between 2000 to 2019, extreme events attributable to climate change resulted in costs estimated at approximately USD2.8 trillion, averaging over USD143 billion per year and USD16.3 million per hour. According to a United Nations Environment (UNEP) report, between now and 2050, the cost of adapting to climate change in developing countries may reach between US$280 billion and USD500 billion annually.

Traditional weather forecasting relies on intricate system modeling, incorporating factors like atmospheric circulation, diverse terrains, ocean dynamics, and their complex interactions. The conventional numerical weather prediction (NWP) methods are highly compute-intensive, employing thousands of processor cores for each computation, often spanning dozens of hours. Despite advancements, overall progress in weather forecasting has been gradual. Furthermore, the global weather forecasting services market is predominantly dominated by European and American agencies.

Since 2023, data-driven AI weather models have emerged as a completely different approach to weather forecasting, attracting unprecedented global attention to the research and application of AI technology in this area. In essence, an AI weather model proposes an efficient 3D earth decoder that can better model and predict problems related to Earth science and meteorology.

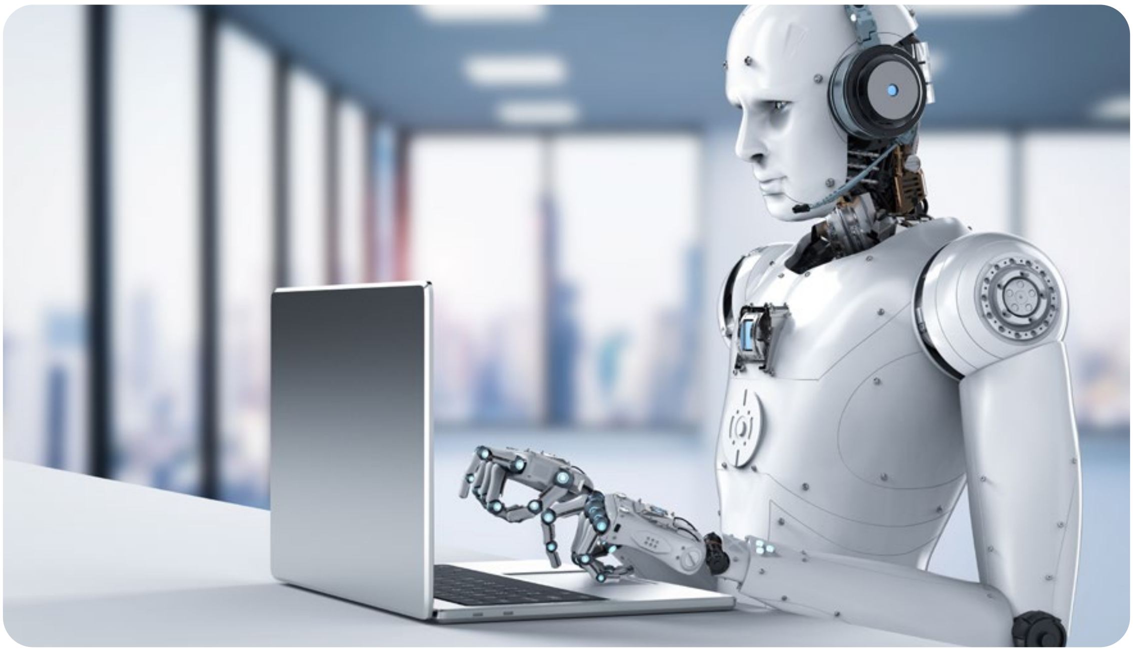
Estimates suggest that by 2030, AI weather forecasting models will be 10,000 faster than traditional NWP systems and 50% more accurate. These advanced models will provide short- and medium-term forecasts for the next 1 to 10 days, mitigating economic losses resulting from extreme weather events, which can reach into the hundreds of billions of dollars.

2.3 Finance: Nowcasting High-Frequency Data for 30% of Economic Indicators

In the financial sector, decisions are made based on external data such as monthly and quarterly economic reports and GDP growth rates. Generally, there is a significant delay before official data like this is published. Nowcasting can use AI models to predict key economic indicators before the official data is published.

Massive amounts of data is collected and processed and AI models are used to improve the sampling frequency from monthly to daily. In this way, economic indicators that you once had to wait over a month to see can now be obtained on the same day or very close to it. By integrating AI algorithms with a real-time, dynamic input-output system, non-linear correlations can be extracted from massive data and factors that always impact the past, present, and future can be discovered. This way, the future states of economic indicators are predicted.

It is estimated that by 2030, nearly 30% of economic indicators and financial benchmark metrics will be nowcasted, with an accuracy of up to 99%.



2.4 Government: Multidisciplinary AI for One-on-One Assistance

Generative AI is fundamentally transforming the way governments operate and serve their public. By 2030, multidisciplinary AI support will be available as both self-developed and externally-provided platforms. These platforms will lighten government workloads by as much as 50% to 70%. With various integrated AI features, such a platform will bring out the best of staff's expertise, responsibility, and availability, working autonomously with people to interact more like a colleague than a mere tool.

Another story worth telling is intelligent assistants, which are built on dedicated government models and constantly evolving their dialog to be more humanlike. **By 2030, each government employee and the enterprises and individuals they serve will have an assistant** for smarter recommendations, easier payments, and better information. These assistants, tailored to each citizen, provide instant, authoritative responses on government affairs and higher administrative efficiency. Citizens stay better informed on policies and regulations, with access to internal guidance and clear explanations, on a single, always-on channel.

By 2030, multidisciplinary AI support will be deployed in more than 95% of governments, and intelligent assistants will be accessible to all served enterprises and individuals.

2.5 Education: Co-Teaching, Co-Learning, and Co-Nurturing with 15 Million Digital-Intelligent Teachers Worldwide

Education is a national priority facing many challenges: In primary education, classroom teaching often involves teachers lecturing to students passively receiving the information. The teaching methods tend to be uniform and untargeted. At home, parents put significant effort into tutoring their children, but they typically lack sufficient guidance and it's difficult for them to focus on interest and innovation. In higher education, university curricula and talent development plans evolve too slowly, so there is often a mismatch between the skills of graduates and industry needs.

The development of artificial intelligence, particularly Artificial Intelligence Generated Content (AIGC), has accelerated the progress of smart education. Large teaching models will reshape education in three main ways:

- Teaching materials can be generated by AI rather than teachers relying on personal experience, and homework can be automatically graded, complete with feedback. These innovations can save teachers over 15 hours per week.
- Adaptive curricula become possible. Study outcomes, including reading and writing abilities and learning attitudes, can improve by up to 90% when students are guided by virtual teachers and supported by regular feedback.
- AIGC can independently conduct certain experiments and generate research papers, reducing the time needed for research by 1/3.

Looking ahead, education will become more open, breaking through the constraints of time, location, and demographics. It will be more sustainable and able to meet the needs of lifelong learning.

By 2030, it is expected that there will be 15 million "digital-intelligent teachers" worldwide working alongside schoolteachers, moving towards a new paradigm of co-teaching, co-learning, and co-nurturing.

2.6 Retail: A Flexible Supply Chain Is Slashing 72% of Inventory Costs. XR and Unmanned Delivery Are Hitting the Mainstream

Global retailers face challenges with perishable goods that can lead to spoilage if they do not sell quickly. With an average daily turnover of USD200,000 per store, just one extra day of unsold inventory across 100 stores can tie up USD20 million in capital.

To solve this, retailers use AI and digital tools to streamline their supply chains, enhancing production, sales, and delivery of goods. AI is really good at guessing how much customers want to buy and figuring out the best prices. This makes their supply chain more flexible and able to work almost on its own. By using AI to look at data, retailers can introduce new products to the market more quickly and handle their stock more effectively. This, in turn, cut down on the costs associated with holding stock. Also, more and more retailers are focusing on making the shopping experience better for customers. They use lots of data and cool tech like XR, unmanned deliveries, and virtual assistants to give customers services that are personalized, fun to use, and understand their needs.

By 2030, it is predicted that AI will be a game-changer for the retailers, affecting up to 95% of customer interactions and having 85% of global retailers investing in it. Full coverage of AI in the retail world could lead to a significant reduction in operating costs for 72% retail enterprises and a revenue increase for 69% retailers. This could potentially add an extra USD611 billion to USD815 billion in overall revenue, a 1.5% to 2% increase.

2.7 Web 3.0: Decentralization in Every Industry, 90 Billion Zero-Knowledge Proofs

Web 3.0 has coexisted with Web 2.0 for a while now, and it is not entirely replacing Web 2.0. Rather, it is evolving and expanding the web's capabilities. Web 3.0 empowers individuals with greater control over their information and reduces reliance on vulnerable centralized servers. It utilizes zero-knowledge proofs to safeguard users against potential privacy breaches and data exploitation. In Web 3.0, there is an enhanced awareness of data and privacy protection. Currently, Web 3.0 remains niche for three main reasons: 1. the poor scalability of existing public blockchains, limiting the scale of upper-layer applications; 2. the lack of industry-changing killer apps; 3. the absence of a development model for Web 3.0 to be integrated with the real economy.

With the approval of BTC/ETH ETFs in Hong Kong (China) and the United States at the beginning of 2024, more real assets will enter the decentralized world. The integration of traditional and decentralized financial systems has been accelerating. Decentralized applications (DApps) will be adopted by all sorts of industries, where they can provide significant automation, transparency, and security for the real economy. Typical use cases include real estate rent collection and property rights contracts, procurement orders, logistics, payments, and settlements throughout the supply chains. DApps have the potential to reduce insurance fraud. In the future, cloud computing performance may be hundreds of times better than today, enabling public blockchains to handle more complex smart contracts and larger transaction volumes. Additionally, cloud computing platforms may integrate quantum computing capabilities to protect data privacy on public blockchains and provide even higher levels of security.

By 2030, it is predicted that the number of global users of digital assets will reach 1 billion, and the Real-World Assets (RWA) tokenization market will reach USD16 trillion. A decentralized finance or game app that impacts our daily lives (with monthly active users exceeding 100 million) may emerge, enhancing the liquidity of these tokenized assets and offering more investment and economic incentive opportunities. By 2030, we expect Web 3.0 applications to generate 90 billion zero-knowledge proofs, creating a computational market space worth tens of billions of dollars.

2.8 Energy: Network-Wide Intelligence of the Energy-based Operating System, Reducing Greenhouse Gas Emissions by 10%

According to the International Renewable Energy Agency, the installed capacity of photovoltaic (PV) energy is expected to reach 5200 GW by 2030, with 68% of the energy coming from renewable sources. The proportion of fluctuating renewable sources, including wind and PV power, is expected to increase to 46%. With the growing power of renewable energy, particularly PV power, the energy mix is undergoing significant changes. This shift represents various challenges for the energy industry in terms of grid connection, operations, and safety. For instance, the intermittent nature of renewable energy affects the stability of the power grid, and the peak load of power demand is becoming increasingly prominent. These challenges significantly impact the safety and stability of the power system.

In the coming years, cloud computing, AI, and 5G will be increasingly used in the energy industry, covering manufacturing, production, and consumption. This will enable the creation of an energy-based operating system that streamlines the entire generation-grid-load-storage-consumption process, leading to network-wide intelligence. By leveraging management systems and AI algorithms to coordinate the running status of each energy node, the overall efficiency of distributed systems can be improved by 5% to 10%. Furthermore, AI prediction models can be applied to new power generation scenarios, such as PV power generation, offshore wind power, and floating PV plants, to improve the accuracy of power generation plans to over 95%. With the help of big data analysis and machine learning algorithms, the power generation prediction error can be reduced to less than 10%, significantly enhancing the grid feed-in capability of renewable energy sources.

By 2030, it is projected that 90% of global PV plants will utilize AI technology, resulting in an annual reduction of 5% to 10% in greenhouse gas emissions.

2.9 Entertainment: AI Creates 70% of Media Content and Unlocks a Personalized Content Market Worth USD500 Billion

Today's media and game sectors face a series of challenges. The speed of content creation and rendering is below expectations, causing huge delays in content distribution. Complex production processes often compromise on reliability, while insufficient collaboration across different areas leaves the potential of teamwork untapped.

However, challenges do come with opportunities. In this case, it's AI, which will transform the production model of media and game content by 2030, with manual video shooting replaced by computer generation. As digitalization is sweeping across industries, AI will dominate some of them, including the media and game sectors. **The maturity of AI supercharges the virtual human industry into a USD38 billion market by 2030.** With AI-generated content (AIGC), reproducing an imaginary world is a breeze, and you can modify your creation whenever you want. Gaming in such a virtual world means unprecedented immersion. In terms of media content, users will no longer just get what they are given, as the popularization of AI makes everyone become content creators. This trend will be accompanied by a revolutionized cloud architecture, industry automation, and new business models.

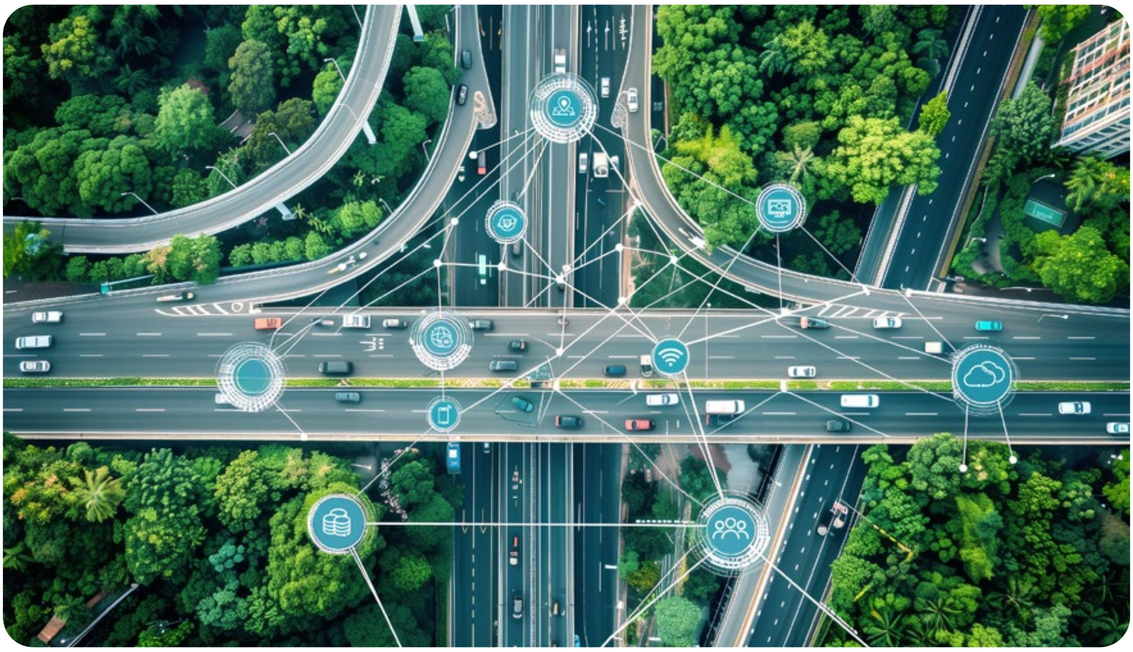
The AI-infused personalized content market is expected to hit USD500 billion by 2030, and contribute to the monthly traffic growth by 85 billion GB.

2.10 Industrial: 50% Supply Chain Cost Slash and 70% Fulfillment Speed Boost Through the Multi-Agent System

The global value chain is swiftly redefining itself as companies are expanding their reach worldwide. Digital and smart supply networks are leading the wave. However, the growing complexity of these networks requires more costs in procurement, production, logistics, inventory, and quality control, and makes order fulfillment less predictable.

To address this, industrial manufacturers are turning to AI for design, production, and quality control, making processes smarter and more efficient. Robots with AI are becoming more agile, and supply chains are getting smarter too, with AI improving order predictions. At the cutting edge, companies are using agent technology to create flexible, collaborative manufacturing processes with group intelligence. Industry leaders are building networks that drive innovation and efficiency in R&D, production, and supply chain management. The key to this advancement is multi-agent systems technology, which is creating intelligent supply chains. This not only greatly reduces costs but also significantly shortens the time it takes to fulfill orders, making the industry more responsive.

By 2030, agent technology in manufacturing is expected to exceed 80% adoption, slashing supply chain costs by half and reducing order fulfillment times by a staggering 70% with the aid of foundation models.



2.11 Smart, Personalized Human-Vehicle Interaction Powered by 500 EFLOP/s Cloud Computing

Currently, the interaction between humans and vehicles is limited to physical controls, voice instructions, navigation systems, and passive safety technology. Communication and collaboration between vehicles and the external environment are still in the early stages, resulting in insufficient traffic safety and efficiency. This cannot satisfy the increasing demand for intelligent vehicles.

In the future, human-vehicle interaction will be smarter and more tailored. Voice assistants will engage in fluent multi-language conversations with 95% accuracy and 90% emotion recognition, minimizing distractions and keeping drivers alert on the road. AR HUD displays will project real-time traffic and weather information, making information acquisition more intuitive and improving security. Touch and gesture control technologies will be popularized, and face and fingerprint recognition will simplify in-vehicle settings and personalized adjustments. AI will personalize the driving experience with 90% accuracy by adapting to drivers' habits, paving the way for a seamless shift to full autonomy. V2X will enable microsecond-level data sharing between vehicles, infrastructure, and pedestrians, optimizing traffic management and reducing congestion and accident risks.

By 2030, it is estimated that the number of global intelligent vehicles will reach 200 million, and cloud computing power will reach 500 EFLOP/s (exaFLOP/s) to support complex AI models, real-time data processing, and autonomous driving.

2.12 Low-Altitude Economy: A USD50 Trillion Growth Opportunity

The low-altitude economy, a nascent industry driven by general aviation, is rapidly expanding globally. This emerging sector encompasses a wide range of activities, including low-altitude flights, aviation tourism, regional airlines, navigation services, scientific research, agriculture, and emergency response.

Beyond traditional general aviation services, drones have introduced a novel service model, further accelerating the growth of the low-altitude economy. The industry is poised for significant expansion, with projections indicating 50,000 drone manufacturers, 5 million registered unmanned aerial vehicles, and 1 million licensed drone operators in China by 2030. The development and adoption of electric vertical take-off and landing (eVTOL) aircraft and intelligent aviation logistics equipment will also contribute to this growth.

Key enabling technologies for the low-altitude economy include AI chips, multimodal foundation models, and big data platforms. AI plays a crucial role in enhancing aircraft intelligence and optimizing airspace utilization and safety through intelligent traffic management. Low-altitude aircrafts will become an integral part of embodied AI, and low-altitude flight management will increasingly rely on AI-powered digital platforms. In the future, autonomous and self-evolving systems may assist or even replace humans in management and service roles.

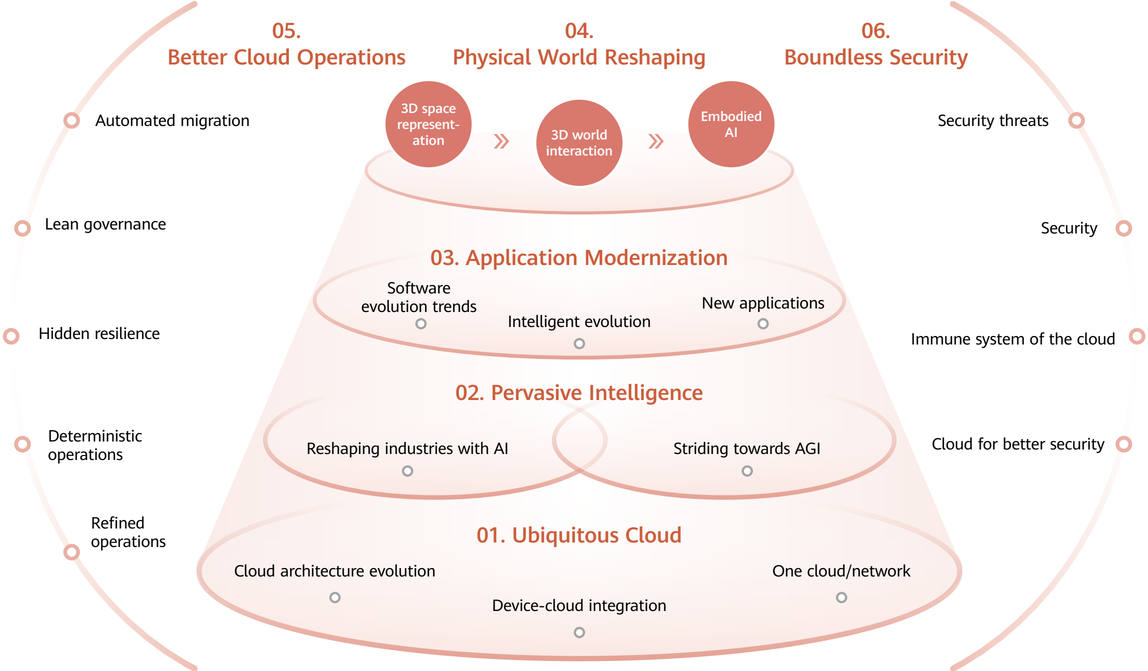
Three primary use cases will drive the growth of the low-altitude economy:

- Green city: Small drones will come into play for sustainable urban development, completing tasks like city inspections.
- Inter-city smart travel: eVTOL aircrafts will be an efficient and convenient option for passengers.
- Large-scale unmanned freight: Large freight drones will address logistics and distribution challenges in remote areas.

The expansion of the low-altitude economy will transform cities from a "2D economy" to a "3D economy", as economic activities extend into the airspace.

By 2030, the global low-altitude economy is projected to contribute USD50 trillion to global economic growth.

Key Technical Features



3.1 Ubiquitous Cloud

3.1.1 Cloud Architecture Evolution: AI Native Architecture

As artificial intelligence (AI) develops rapidly and becomes more widespread, the cloud infrastructure is shifting from a traditional general-purpose computing platform to an AI native architecture. The AI Native architecture should deliver the following technical features to comprehensively improve the performance, efficiency, and environment sustainability of AI applications: diverse compute, peer-to-peer architecture, simplified network, large-scale pooling, proclets, resource flexibility, and low carbon.

1) Proclets: Dynamic Allocation of Process-level Resources Without Compute Unit Segmentation

Cloud resources on VMs or Docker containers require reservation in various ways. However, the granularity of resource allocation is too large. Even serverless applications must run on VMs or containers. Static allocation is inefficient and cannot achieve the best resource utilization.

Proclets categorize resources by memory, CPU, or GPU, and allocate resources at the process level. This eliminates the need for segmentation like function compute units. Proclets use fixed, physical thresholds and combined production, and they are small and fast enough to not be noticed by upper-layer applications.

By 2030, around 20% of compute that is in cloud data centers is expected to be split and assigned as proclets, resulting in significant cost savings of billions of dollars for enterprises.

2) Flexible Compute: The Next Hop of Elastic Compute, Increasing Resource Utilization to 70%

Traditional cloud servers come with fixed specifications and are complex to deploy. This often results in either overprovisioning, which wastes valuable resources, or underprovisioning, which means demand cannot be met during peak hours. Typically, more than 80% of compute resources are allocated, but only a little over 20% are actually utilized.

It is estimated that by 2030, flexible compute will significantly enhance cloud resource utilization, increasing it to 70%.

In the future, flexible compute will incorporate the following key technologies:

Intelligent dynamic overcommitment involves real-time monitoring of resource profiles for individual instances and analyzing their CPU utilization. Resource allocation is then intelligently adjusted to ensure that each instance receives the CPU resources it needs. Dynamic CPU allocation enables zero-latency, user-unaware vertical scaling.

AI can be used to forecast resource requirements by analyzing the resource usage history of applications. It intelligently forecasts service needs and dynamically adjusts compute capacity to ensure that applications always run optimally. The entire process does not require manual intervention. Intelligent horizontal scaling can be automated by recognizing time sequence-based resource needs of large-granularity applications.

Flexible memory enables dynamic memory overcommitment. Unlike traditional methods, that are application-unaware and asynchronous, this technology monitors applications' memory usage and provides synchronous overcommitment. This ensures efficient memory utilization and high application performance, making memory management more intelligent and refined.

3) Diverse Pooling: The Next Step in the Evolution Away from Monolithic Compute

Cloud computing has been evolving from monolithic compute to diverse pooling compute. Traditional servers can only provide the compute of a limited number of CPUs, GPUs, and NPUs, and their ratios are fixed. But AI service requirements are diverse. The same application can have a diverse range of compute requirements, sometimes requiring dozens, hundreds, or even larger number of processors. The cloud architecture needs to be able to flexibly configure diverse compute provided by CPUs, GPUs, and NPUs as required. The selected compute resources can be tightly coupled into a resource pool through an ultra-high-speed network. Resources in the pool are dynamically provisioned

to meet changes in demand. This ushers in a new era of intelligent compute. Tightly coupled resource pools can deliver 10-fold higher computing performance than traditional servers and are suitable for a wider range of scenarios.

4) Peer-to-Peer Architecture: Shifting from Primary/Secondary for Direct Communication

With the rapid growth of the AI and various high-performance computing requirements, the traditional primary/secondary architecture struggles to keep up with to the increasing performance requirements due to its CPU-centric resource bottlenecks and transmission delay. The compute resources in data centers need to be organized and communicate with each other using a decentralized, peer-to-peer architecture. The unleashed compute can reduce the latency as low as microseconds and increase bandwidth sharply. It is estimated that by 2030, 60% of cloud data centers will evolve from a primary/secondary to a peer-to-peer architecture.

5) Simplified Network: An All-in-One Network for Cluster Connectivity Across AZs and Regions

The rapid development of AI applications brings unprecedented challenges.

- As AI applications grow, they require more diverse network capabilities, including scale-out and scale-up, as well as cluster connectivity across AZs and regions. This has led to increasingly complex network architectures.
- Scaling out means retaining redundant resources to handle traffic surges, but over-deployment reduces resource utilization and drives up costs during off-peak hours. In addition, scaling networks out and up, and deploying VPCs independently increases costs and makes O&M harder.
- Traditional network architectures are insufficiently responsive to traffic spikes. Once a cluster's size is determined, it is challenging to expand, limiting flexibility.

To address these challenges, a new simplified network architecture is introduced. The core advantages are as follows:

- Multiple independent networks are integrated so that AI applications, general-purpose compute nodes, and storage resources can share a high-bandwidth network, improving resource utilization, reducing costs, and simplifying management.
- AI applications can run seamlessly across AZs and regions, and the simplified network can scale out during traffic spikes and scale back in when traffic returns to normal.

It is estimated that by 2030, 60% of cloud data centers will adopt the all-in-one simplified network architecture.

6) Core Architecture Transformation: Plug-and-Play, Multi-modal Fusion, Ubiquitous Distribution - Intelligent Cloud-Native Databases

Currently, the core architecture for enterprise services is centered around resources, regions, and loads. Data silos throughout the data lifecycle create problems like inconsistent data across regions, non-uniform security policies, excessive transmission latency, and expensive service development and maintenance.

With the popularization of AI compute power, high-end compute delivered by NPUs and Rack clusters has become more accessible. Against this background, how to use new hardware and new computing power to enable the intelligent transformation of enterprises has become a key issue for data management systems.

In the future, databases will exhibit the following technical characteristics:

Data Access as a Service: Applications do not need to be concerned with the underlying data model that the data is stored in. Serverless, a cloud-native model, empowers databases to manage, query, and retrieve multiple types of data with the key capabilities. It enables query and storage of heterogeneous data, and it supports Hybrid Transactional/Analytical Processing (HTAP) processing. This reduces the cost of managing applications and using data, accelerating the release of data value.

Intelligent Native Data Management: Intelligent SQL optimization and transformation designed for application developers reduces the barrier to entry for developers. Intelligent Q&A and operations and maintenance (O&M) for next-generation DBAs improve O&M efficiency by 80%. Data management systems have been evolving from relational models to various new models integrating Large Language Model (LLM) and SQL execution. This transition supports real-time inferential and knowledge computing.

Cloud-Native Fully Pooled Architecture: Decoupling and pooling of resources (such as CPUs, memory, and storage) enables elastic scaling, transparent to the applications, in seconds. Decoupled resources can allow you to double the performance you get out of the same amount of compute. This provides performance compatibility for enterprise databases with terabytes or more of data, and offers rapid, smooth scalability for distributed databases.

It is expected that by 2030, we expect to see an all-in-one database management system based on new hardware and cloud-native architecture. We expect to see a system with autonomous data management, intelligent optimization of data processing, and intelligent data security protection. We are truly stepping into the era of data intelligence.

7) Cloud Service Reshaping: From Regional to Global

Traditional cloud services allow geographical flexibility and data localization by region. However, distributed applications are limited in performance optimization and management around the world. By 2030, cloud services will evolve from a regional architecture to a global architecture. With a global design (such as regionless), applications will break through the geographic restrictions and be optimally deployed around the world with SLA guaranteed.

Global data services: By 2030, 80% of applications will use unique IDs for cross-region data access, which improves data management efficiency by hiding cross-region statuses and simplifying the data flow process. Nearby data access around the world enables a subsecond latency. The cross-region data computing performance can get a 5-fold increase and the DR time can be shortened as low as 1 minute.

Global storage services: By 2030, the global storage capacity is expected to increase from 12 ZB to 28 ZB, and the cloud storage capacity to increase from 2 ZB to 5 ZB. Global storage services can be quickly accessed, and the data upload speed will be improved by 50% to 70%.

Global network services: Regional network services are transforming to global network services to handle the challenges of geographical isolation. Network services, such as Direct Connect, VPN, Enterprise Router sharing, and Endpoint, will be globalized to provide low-cost, cross-region VPC communications and cross-region access to gPaaS & AI DaaS services for the same tenant. An application-oriented network model will be provided for simplified management and configurations as well as seamless connections between regions or sites.

Global application distribution: By 2030, the global architecture will greatly simply the development of distributed applications. It is estimated that 80% of applications will be automatically distributed based on the SLAs, and 90% will support global distributed architecture throughout their development lifecycle.

8) Cloud Computing Clusters with ZFLOPS of Compute: No Constraints of Compute, Storage, and Networking

It is estimated that the cloud computing clusters will exceed ZFLOPS of compute by 2030. Cloud data center technologies need to make breakthroughs in the following aspects:

Compute: Physical supernodes need to be evolved into logical supernodes in ultra-large AI acceleration clusters. Elastic cluster compute can then provide efficient training and inference support for multi-modal and trillion-parameter models. A peer-to-peer pooled compute architecture can flexibly adapt to the requirements of various AI applications and intelligently optimize resource configuration.

Storage: A system like this will require petabytes of memory, so a high-performance cache pool and a tiered storage solution are used to improve the storage capability and reduce costs by 75%.

Network: The bus bandwidth can be increased by 30 times with ultra-high bandwidth and high-performance interconnection technologies. A unified protocol is used to connect the AI network and data center network (DCN), ensuring high-speed transmission of a large amount of parameters and gradient information. In addition, a new topology architecture and advanced routing technologies will be used to reduce the number of network hops and provide a deterministic transmission solution to solve long tail latency issues such as P95 or P99

9) Cloud Data Centers: Green and Reliable

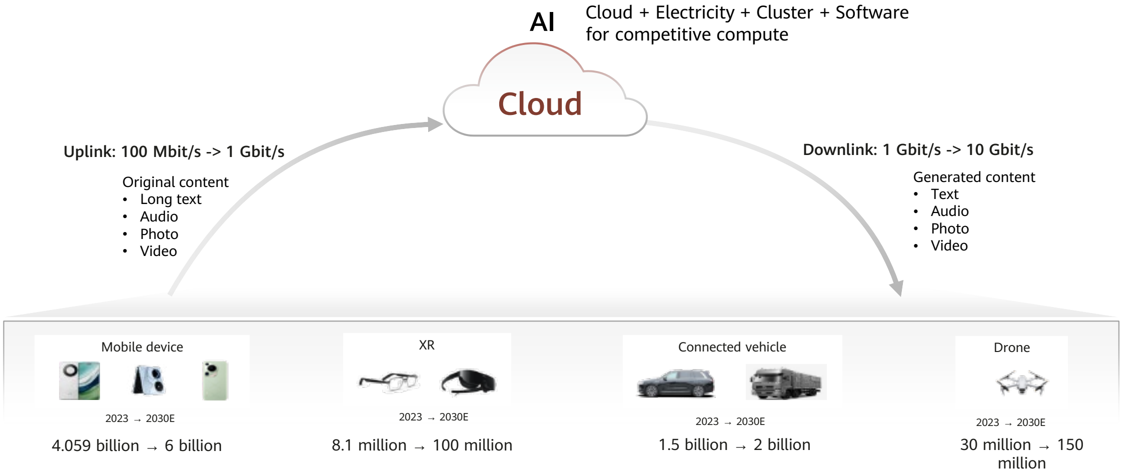
1. Low-carbon power supply: Low carbon propels the innovation of cloud data centers. Tomorrow's cloud data centers will feature advances such as liquid cooling, cooling storage, waste heat utilization, and energy storage. By complementing that tech with renewable energy such as solar, wind, and nuclear, these centers slash their energy consumption and carbon emissions. The combination of new energy and energy storage will improve power supply balance and stability of renewable energy systems. It is estimated that by 2030, more than 70% of data centers will be on the cloud, where 100% of power supply is green.

2. AI-powered O&M: Another prominent feature of cloud data centers is O&M intelligent enough to cover timely identification, analysis, prediction, and handling of risks and issues. Smarter O&M makes data centers smarter and more automated. For example, intelligent inspection is 100 times more efficient than manual inspection in identifying risks.

Robots will significantly improve O&M efficiency and quality while reducing the risk of misoperation, especially in labor-intensive and repetitive tasks. 2030 is predicted to see robots take over half of O&M.

3. Streamlined, reliable architecture: Cooling and power supply systems are still the main culprits of data center service interruptions. Evolving centralized cooling to a distributed structure would prevent the failure of a single device from affecting other devices, meaning less impact and higher reliability. This new architecture prevents a single point of failure in the cooling system from crashing the entire data center. The pooled interconnection of power plants makes it possible to simplify the power supply architecture in 2030, where power utilization is more efficient even while removing diesel generators, UPSs, and batteries.

3.1.2 Device-Cloud Integration: 100-Fold Intelligent Computing Enhancement for Devices and Applications Powered by Device-Cloud Synergy



As devices become smarter, applications like XR and autonomous driving demand higher computation and latency lower than 50 ms. Device-cloud synergy offloads processing, boosting on-device intelligence by 100 times and feeding the next wave of new applications.

1) 5G-A Large Uplink: Stimulating Device-cloud Synergy and Promoting Traffic Growth

5G-A technology is set to revolutionize uplink capabilities, with projections indicating a tenfold increase in uplink peak rates to 10 Gbit/s by 2030, complemented by downlink peak rates of 1 Gbit/s. This advancement ensures near-ubiquitous accessibility for cloud applications, exceeding 99% reliability. It facilitates rapid data upload from devices to the cloud, potentially raising monthly network traffic to over 1000 PB.

Concurrently, device-cloud synergy's end-to-end latency could be sub-50 ms.

2) New Computing Architecture: Distributing Computing Tasks Across Devices, Edge, and Cloud

In the age of AI, the demand for processing power outstrips the capabilities of traditional device-centric models. The strain on devices in terms of computing, memory, and power consumption is growing, posing a bottleneck to on-device intelligence. To address this, a novel computing architecture is emerging that distributes processing tasks across devices, edges, and clouds. By 2030, it is anticipated that 70% of computing tasks will be handled on devices, with the remaining 30% offloaded to the cloud. This architecture will enable the cloud to amplify device computing power by 100 times, leveraging the cloud's virtually limitless power and resources like clusters and software. It will facilitate a computing capacity that could scale up to 40 ZFLOPS. Furthermore, the evolution from device-native to device-cloud collaborative OSs will be pivotal. Future applications will inherently support heterogeneous compute across devices and clouds, ensuring secure and reliable data exchange, adaptive task scheduling, and collaborative training and inference capabilities.

3) Explosive Growth of Intelligent Applications on Devices, Vehicles, and Wearables

AI-powered devices and PCs are leading a technological shift, offering end users advanced assistance in document analysis, meeting summarization, graphic and text generation, and AI-aided home education. These capabilities enable daily activities, learning, and professional tasks to be seamlessly managed through cloud-empowered devices with robust compute. The smart vehicle industry is set to accelerate vehicle-cloud synergy and the development of smart road infrastructures. Smart vehicles will diversify their interactions through the cloud, encompassing smart driving, on-vehicle entertainment, and vehicle-road-cloud smart transportation systems. Moreover, advancements in world model technologies, real-time 3D technologies, and zero-latency interactions are propelling the growth of wearables like smart glasses. By 2030, it is projected that the number of AI devices facilitating device-cloud synergy will soar to 3 billion, and that of intelligent networked vehicles hits 300 million.

4) Personal Privacy Protection: Confidential Computing

Personal privacy extends beyond devices to the cloud. Ensuring privacy and confidential computing becomes paramount. Besides the security for devices themselves, the cloud must safeguard against data breaches, ensuring that the computing environment is trustworthy and that data transmitted by devices is anonymized and encrypted. End-to-end security is crucial, creating a secure link between devices and cloud, preventing unauthorized access and ensuring data in transit remains indecipherable. A new, robust security technology stack is in demand, which includes implementing confidential computing, security sandboxes, no privileged access, and data deletion after use. The cloud itself cannot get its hands on any user data neither.

3.1.3 One Cloud

"One Cloud" is the vision of a globally integrated cloud computing platform, unifying data and services across the world. It transcends geographical boundaries, empowering businesses to deploy applications swiftly and conduct global data analytics. This approach not only fuels digital transformation and innovation but also enhances competitiveness in the market.

1) Customers' Perspective: From Siloed IT to Integrated, Tiered Cloud Architecture

By 2030, the corporate landscape will shift from siloed IT systems to a cohesive, tiered cloud architecture. "One Cloud" will streamline connections between headquarters, branches, and edge facilities, aligning with public cloud services to deliver a seamless customer experience.

One hybrid cloud: Hybrid IT will dominate by 2030, with projections showing 90% of large enterprises and 60% of SMEs adopting this model, a significant rise from 2024's 60% and 30%. This unified platform will harmonize online and offline resources, enabling real-time data sharing and informed decision-making.

One global distributed cloud: The influence of multinational companies will escalate, with an anticipated increase from 100,000 to 140,000 entities by 2030. Global innovation centers are set to more than double, from 800 to 2,000, and the local workforce percentage is expected to grow from 50% to 70%. A unified cloud platform will orchestrate cross-regional services and data, optimizing resource management and service collaboration.

One cloud to connect edge and devices: Edge and on-device computing by 2030 will decentralize data processing further. The global edge device count is expected to surge from 20 billion to 50 billion, with 85% of large enterprises and 50% of SMEs adopting cloud-edge-device architectures. 50% of data will be processed at the edge, and the unified cloud platform will integrate these layers to boost performance, speed, and reliability across the board.

2) Business's Perspective: One Network for High Bandwidth, Low Latency, Massive Connections, and Global Resource Sharing

The evolution of service innovation and network technology has transformed "One Cloud" into a sophisticated global "One Network," enhancing bandwidth, reducing latency, expanding connectivity, and improving security. This advancement empowers businesses to efficiently share global resources and data, driving digital transformation and fostering regional collaboration.

One network for media: By 2030, the online video user base is expected to hit 6 billion. Cloud platforms and real-time streaming will revolutionize sports and entertainment events. With 85% of media companies leveraging cloud and AI for content curation and audience analysis, Huawei integrates high bandwidth and low latency to create one global network for media, offering personalized and interactive experiences.

One network for vehicles: Daily vehicle data generation would range from 0.2 TB to 1 TB by 2030, with a significant rise from 35% to 80% in smart transportation infrastructure in major cities worldwide. The one global cloud, integrated with IoV technologies, autonomous driving, and intelligent transport, utilizes end-to-end encryption to build one global network for vehicles, enhancing mobility intelligence, road safety, and transportation efficiency.

One network for enterprises: The daily data output of global enterprises is set to grow from 125 EB to 500 EB by 2030, with cloud storage capacity expanding from 0.5 ZB to 1.5 ZB and bandwidth demands increasing by 50%. Secure and efficient network technologies would help construct one network for enterprises, enabling efficient global resource sharing and data transmission, and promoting cross-regional collaboration.

One network for cities: The data generated by global cities is anticipated to jump from 0.1 ZB to 1 ZB daily by 2030. 80% of smart cities would employ AI for data analytics, predictive maintenance, and automated management, and 70% of smart city solutions would incorporate edge computing. Integrating intelligent technologies and infrastructures enables one network for cities for efficient digital city management, improves residents' quality of life, and enhances urban operational efficiency.

3.2 Pervasive Intelligence

The vast mountains of knowledge accumulated over the past few decades of the Internet era have been integrated into large language models (LLMs) as tokens, establishing a robust foundation for intelligent services. Looking ahead, these intelligent services are expected to expand from the cloud to the edge and onto end-user devices, permeating every organization and the life of every individual. This evolution will profoundly impact software and applications. By 2030, it is estimated that every enterprise will have at least one custom-developed large AI model, and every employee will be equipped with at least one AI agent. Consequently, every piece of software will be redesigned or refactored using LLMs, and every application will be developed using some form of AI-supported programming tools.

3.2.1 AI Reshapes Industries, Tackling Big Challenges and Driving the Intelligent Economy

Across various industries, AI is being integrated into enterprises' core production systems to address big challenges. It is poised to become a crucial driver of productivity during the current industrial revolution. We anticipate that AI will swiftly transform the workforce and job market, ushering in an intelligent economy.

1) AI Ignites the 4th Industrial Revolution

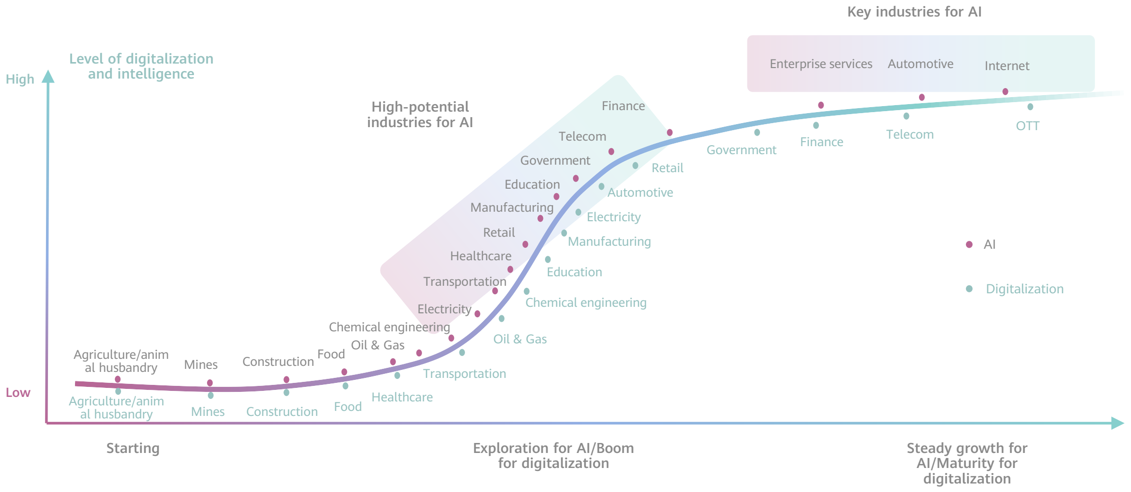
AI is at the heart of the 4th Industrial Revolution, driving innovative business models and fostering an intelligent economy. As the latest general-purpose technology (GPT) following the steam engine, electricity, and information technology, AI has the potential to significantly enhance productivity and reshape the global economy in numerous ways.

Boosting productivity: AI is evolving from AI Copilots, which are intelligent assistants offering information and suggestions, to AI Agents, autonomous agents capable of executing complex tasks independently. The next stage is the AI Workforce, where teams of AI agents collaborate to perform a broader range of complex and creative tasks. This evolution will continuously push the boundaries of automation and intelligence, transforming productivity.

Reshaping the job market: AI is expected to replace some repetitive and routine jobs, such as taxi drivers and graphic designers. However, it will also create new jobs that require skills and creativity, such as AI prompt engineers and AI data scientists. Estimates suggest that by 2030, AI will replace around 400 million jobs, but it will also generate approximately 97 million new jobs.

Increased contribution to economic growth: Statistics indicate that AI currently accounts for approximately 1.6% of the global economy. By 2030, it is projected to contribute an estimated US$2 trillion.

Transforming traditional industries and creating new ones: AI will not only transform traditional industries but also create new ones, such as digital therapeutics, personalized AI companions, and chronic disease management.



2) AI for Industry: Faster Innovation and Higher Efficiency

AI is set to transform industries by fueling innovation and boosting efficiency.

- **Fueling innovation:** AI supports creative jobs by automating the creative process and offering creative ideas. By using AI to drive automation in tasks such as communication, document collaboration, and interpersonal interaction, employees can free up more time and energy to focus on more creative tasks. This goes beyond accelerating innovation in areas such as customer relationship management, marketing, and sales. It will boost innovative power across entire creative industries.

- **Boosting productivity:** Trained on massive amounts of real-world data, AI has reached or even surpassed human-level performance in many domains. This is particularly true in knowledge-intensive tasks, such as software engineering and development, where AI is significantly boosting productivity.

These advancements, however, do not mean that AI will soon replace humans in all fields of work. The speed and likelihood of AI replacing humans depend on the error tolerance of specific jobs. Jobs that are highly error-tolerant are likely to be the first ones to be displaced. Specifically, creative jobs such as writing, graphic design, and creative copywriting are likely to be first ones to be impacted, followed by jobs that require higher precision, such as education and training and financial consultation. Jobs like doctors and lawyers, which require the highest level of precision, are likely to be the last ones to be displaced.

Error Tolerance	Use Cases
10-50%	Content creation assistance: writing/graphic design/social media account/creative images/text/video | text understanding
1-10%	ADAS/assistance algorithm, game NPC/visual effect, marketing/advertising/recommendation systems, language translation
0.1-1%	AI training | Junior AI legal assistant | Financial consultation | GPT generates a semantic summary or serves as a general interface | L3 autonomous driving
0.01-0.1%	Accounting | Doctor's AI assistant | AI teacher | Auto coding after software splitting
0.0001-0.001%	AI doctor | AI lawyer | L4 autonomous driving

By 2030, it is estimated that 75% of AI's contribution to global GDP will originate from creative and knowledge-intensive industries.

3) AI for Science: AI Will Transform Scientific Research, with AI Supporting Over 50% of Scientific Computing Tasks

Traditionally, scientific computing has depended mathematics and numerical methods to solve partial differential equations (PDEs). However, computations can be slow or even fail due to limited computing power or the curse of dimensionality. Today, scientists are increasingly turning to AI for solutions. The key is to use data-driven methods, such as machine learning and deep learning, to discover patterns in vast datasets. Looking forward, we anticipate three major shifts in scientific computing:

- **From traditional methods to intelligent scientific research facilities on the cloud:** Scientific research has long relied on traditional labs and theory-driven methods. In the future, it will come to rely more on intelligent facilities on the cloud, including high-performance computing, large, interdisciplinary scientific computing models, and AI assistants. This will enable more efficient scientific research processes and accelerate new discoveries. With on-demand, theoretically unlimited computing power, the cloud solves long-standing challenges for scientific computing.

- **From manual to autonomous experiments:** With advances in AI4S, self-driving labs will become the new norm. AI-powered robots and intelligent lab environments will enhance efficiency, minimize human errors, and enable scientists to concentrate on more complex issues that demand genuine human ingenuity.

- **From independent research to large-scale interdisciplinary collaboration:** Scientific research is shifting from the independent work of individuals or small teams to large-scale, cross-disciplinary collaboration. The large-scale deployment of technologies such as AI agents and privacy computing allows us to build secure, collaborative environments for scientific research. This helps to eliminate data silos while protecting the data ownership of different parties.

By 2030, it is estimated that AI will be utilized in over 50% of scientific computing tasks, potentially accelerating computational speeds by up to 10,000 times in fields such as meteorology and pharmaceuticals.

3.2.2 Striding Towards AGI

Looking ahead, AI is steadily progressing towards artificial general intelligence (AGI). We anticipate rapid advancements in architecture optimization, enhancement of multimodal capabilities, adoption of next-generation architectures, and manual alignment. Generative AI will continue to revolutionize productivity in the media industry, while discriminative AI will maintain dominance in over 50% of all AI use cases. Additionally, AI agents are expected to evolve into super apps, seamlessly connecting enterprise applications.

1. Intelligence as a Service for All Scenarios

As a new general-purpose technology, AI in all its forms - not just chatbots, AI plug-ins, and AI engines that enhance traditional applications, but also AI-native software and hardware - can be offered as intelligent services to augment human perception, cognition, and decision-making. By 2030, Intelligence as a Service will be ubiquitous.

- **All-scenario applications:** In terms of vertical industries, AI will be more widely adopted by the Internet, government, telecom, finance, manufacturing, energy, education, healthcare, transportation, and retail. In terms of horizontal functional domains, AI will see wider adoption in R&D, production, supply, sales, services, operations, and maintenance. We estimate that by 2030, over 90% of companies all over the world will have adopted AI technology in some shape or form.

- **All-modality:** In AI, modality refers to the type of input and output data. As AI applications shift from making predictions to generating things, AI modalities are extending beyond natural language, speech, images, and videos to even more diverse forms, for example, the sense of touch, taste, and smell for humans, and infrared, inertial navigation, and remote sensing data for machine perception. Enhanced data modality support will significantly boost AI's ability to create digital worlds and transform the physical world.

- **Cloud-edge-device synergy:** AI models will be deployed across the cloud, edge, and end-user devices. Smaller models distilled from larger models will be deployed to personal computers in the hundreds of millions, mobile phones in the billions, and IoT devices in the hundreds of billions. Along with new, AI-native devices, they will make Intelligence as a Service available everywhere, facilitating people's life and work everywhere.

- **Multi-size deployment:** 1B to 3B models will be deployed on end-user devices such as mobile phones and tablets, 6B to 7B models on personal computers, and 10B, 100B, or even larger ones on large servers and clouds. Models of various sizes can meet diverse AI needs. Estimates show that by 2030, models smaller than 10B will account for over 95% of all models developed and deployed, instead of the 38% today.

- **Inclusive services:** The marginal cost of Intelligence as a Service is near zero. Today, AI models' inference cost has dropped to less than CNY10 per million tokens. Comparing this with the cost of a typical free search, which is approximately US$ 2 cents, it is safe to say that Intelligence as a Service is becoming more and more inclusive.

- **Data Intelligence:** From Data-Centric to Knowledge-Centric Enterprises currently use data primarily for business intelligence, process optimization, and control - mainly analyzing structured data. However, in the era of large AI models, current data platforms fall short in data cleansing and knowledge extraction. Building a knowledge-centric data foundation is now crucial. Extracting knowledge from vast datasets helps enterprises consolidate business logic. When combined with industry expertise, this knowledge can be integrated into IT systems, providing companies with quick, relevant, and actionable insights. For instance, in human-computer interaction, it can support efficient, accurate, and intelligent business decision-making. Additionally, it supplies high-quality datasets for AI model training, fine-tuning, retrieval-augmented generation (RAG), and prompt engineering, enhancing both training and inference processes. By 2030, the digital economy is projected to constitute over 60% of global GDP. As a key production element, data will drive enterprise innovation and growth. Intelligent data operations will form the cornerstone of industrial intelligence.

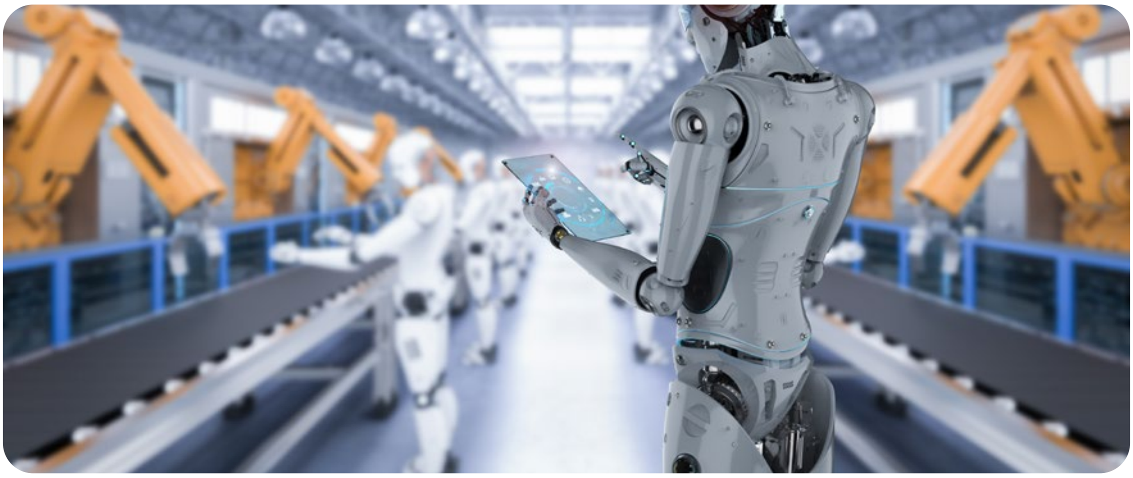
2. More Than Transformer: The Rise of Hybrid Architectures in AI

Convolutional neural network (CNN) and recurrent neural network (RNN) were once the mainstream architectures for deep learning. Later, the Transformer architecture, which supports highly parallel processing by leveraging the attention mechanism and abandoning the recurrence mechanism, emerged as the leading choice for implementing the Scaling Law. We identify four key technologies and trends that are crucial in our pursuit of AGI.

- **More than Transformer:** Transformer models have advantages in handling tasks that rely on context learning and complex reasoning, while pure Mamba models have better performance in long-sequence training and inference. A hybrid architecture that combines the strengths of both models has already been applied to various tasks, including HD image generation, point cloud analysis, and time series forecasting. This makes it a promising candidate for developing future foundation models. Additionally, these models must continuously increase intelligence per FLOPS as well as intelligence per bit.

- **The shift from multimodal to all-modal, and from language-centric to native multimodal:** The results from existing multimodal models have proven that native multimodal training methods can effectively improve model performance. Currently, data processing predominantly uses "Any to Text" and "Text to Any" approaches. In the future, we may see the rise of native multimodal data processing methods in the form of "Any to Any." By 2030, it is anticipated that unified methods for tokenizing multimodal inputs for models will have emerged, facilitating a comprehensive understanding of the world. Concurrently, unified methods for tokenizing multimodal outputs will also be developed, ensuring AI-generated content more closely mirrors reality.

- **Next-gen neural networks improve model adaptability:** Mimicking natural neural networks, spiking neural networks (SNNs) have demonstrated unique advantages in processing time series data such as voice and video. Liquid neural networks (LNNs) feature adaptive weights, unlike the fixed weights typical of traditional models. LNNs can dynamically adjust their weights based on input data, resulting in a smaller and more interpretable neural network architecture. By 2030, it is estimated that new network architectures, such as SNNs and LNNs, will surpass current mainstream models in terms of performance and cost-efficiency in specific domains.



- **Addressing hallucination and explainability issues through manual alignment:** In tasks that prioritize high precision, AI hallucination can be a significant problem. However, in creative tasks, it may be viewed as imaginativeness that inspires human creativity. AI alignment is about aligning AI with human preferences, goals, values, and ethical principles. Future advancements in AI alignment techniques and capabilities will make AI more explainable and reliable. Specific AI alignment methods include learning from feedback, learning under distribution shifts, alignment assurance, and AI governance. As engineering practices improve, models' ability to handle data contamination and misleading prompts will also significantly improve.

Cloud services have a significant role to play in fueling the development of AI. During model training, the cloud can provide a reliable, large-scale network and essential SRE capabilities, which are crucial for improving model floating-point operations utilization (MFU). Currently, the average MFU is between 30% and 50%. Cloud technologies and optimizations could increase this to 60% to 70%, significantly cutting the training compute cost. Cloud services are also equally important during the inference phase. By increasing the batch size, cloud services can control the overall latency and ensure the efficiency of the inference process. Also, since most of the latency during inference is due to computation rather than data transmission, the powerful compute capacity of the cloud is expected to reduce inference latency to near-zero. Additionally, in a multi-device application scenario, cloud services can consolidate data from different devices to provide rich context, ensuring a consistent user experience.

By 2030, we anticipate that the number of parameters of large AI models will match the synaptic connections in the human brain, that is, between 100 trillion and 1,000 trillion. Models of this scale will require much larger training clusters on the cloud, scaling from 100,000 xPUs to millions of xPUs. In the meantime, the energy requirements for data centers to support the training and inference of these massive models

will surge, increasing from tens of megawatts to hundreds of megawatts.

3. Discriminative AI Continues to Create Value for Enterprises

Discriminative AI models have advanced significantly over the past two decades. They are predominantly used to make predictions based on input data and their training. In the field of computer vision, this means to carefully analyze specific images and determine their categories (image classification), or to identify and locate objects in images (object detection). When applied to structured data, discriminative AI focuses on parsing input data and predicting the target values (regression on structured data).

Although the emergence of generative AI has opened up endless new possibilities, discriminative AI models remain valuable. Their potential lies in the following key areas:

- **Unified discriminative AI models:** Most of today's discriminative AI models are small, task-specific models. For each downstream task (such as image classification, object detection, and structured regression), often a dedicated model needs to be developed from scratch, or highly customized development is required. In the future, large, pre-trained foundation models may be developed for discriminative AI. Such models are expected to achieve the desired performance level without fine-tuning. For more performance-demanding tasks, they will only need minimum supervised fine-tuning (SFT).

- **All-modal pre-trained model:** Compared with a pre-trained, single-modal discriminative AI model, a future all-modal model will support multiple data modalities, such as images, videos, structured data, point clouds, remote sensing, and audio, and integrate different modalities to enrich information. This allows it to significantly improve performance in downstream tasks.

- **Collaboration with generative models:** We see huge opportunities in combining discriminative and generative models. This synergy not only bridges the gap between generation and discrimination but also propels advancements in multimodal fusion, reinforcement learning, and adaptive systems. Such progress will drive AI technology towards more intelligent phases, ultimately leading to AGI.

Due to their reliable performance, small footprints, and high computational efficiency, discriminative AI models are widely adopted in enterprise scenarios. They excel in making accurate, highly generalized predictions and classifications based on multimodal inputs, such as camera-captured images, audio signals, and specialized sensor data. By 2030, it is estimated that discriminative AI will still account for over 50% of all AI use cases and nearly 70% of all enterprise applications.

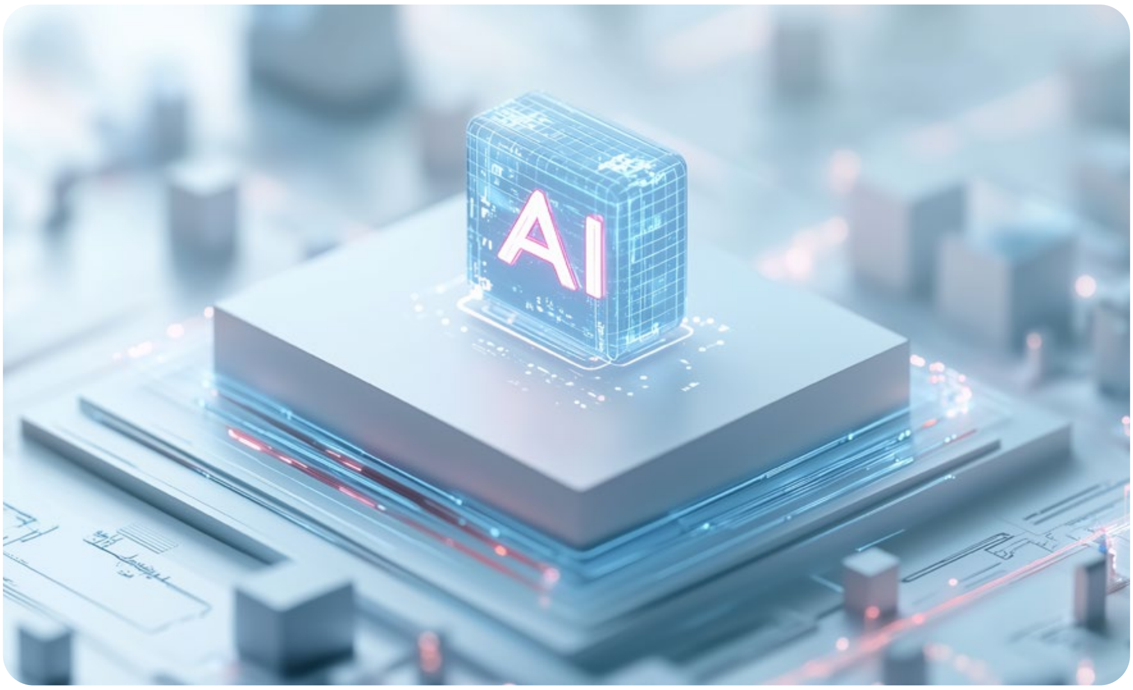
4. AIGC Transforms Media, Offering New Digital Experiences

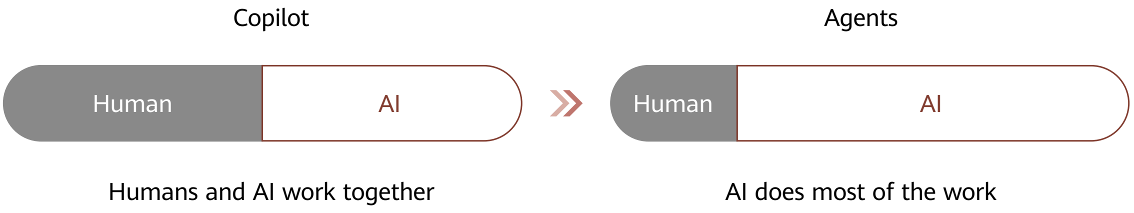
- Traditional media content is primarily generated by professionals with cameras. Media platforms support 1-to-N unidirectional content distribution only. It is estimated that by 2030, more than 70% of media content will be AI-generated or generated with the support of AI. This will enable more personalized content creation and dissemination. Key features include:

- **Personalized content creation:** AIGC will enable faster creation of personalized content, which will be delivered to the audience by matching the user profile, ushering an era of N-to-M personalized content distribution.

- **Real-time interaction between tens of thousands or even millions of viewers:** Traditionally, online spaces can only serve a limited number of users, often in the hundreds, due to server performance limitations. Now, high-speed cloud interconnects and distributed platforms allow for real-time interaction between tens of thousands or even millions of viewers.

- **Compute-network converged media platform:** Traditional media networks rely solely on a network of caches for content distribution, while generative media platforms integrate caching,





computing, and distribution. Through a synergy between the cloud, edge, network, devices, and chips, along with OS optimization both in the cloud and on end-user devices, we can build a network with zero latency and no stalls, delivering the ultimate viewer experience.

- **Fully digitalized generative media engine:** Deep integration of CG and AI technologies, used along with 3D content segmentation and controllable generation, enable large-scale, distributed, real-time content rendering. AI technology and digitalized devices together enable all-digital content collection, editing, and dissemination.

5. AI Agents Become the New Paradigm for Application Interaction, and the New Enterprise Super App

AI agents integrate four key elements of productivity: expert knowledge, data, models, and computing power. They enable intelligent interaction between users and AI-powered applications and systems. Traditionally, users interact with enterprise applications through a software interface. Now, they interact with AI agents, which autonomously execute complex tasks without being given specific instructions on how to navigate these tasks. AI agents give each employee a dedicated, intelligent assistant with enhanced capabilities, potentially making them super apps within enterprises. By 2030, AI agents will impact two-thirds of the world's jobs, and have the potential to replace 30% of the work hours across the globe. By then, over 40 million people may need to be reskilled or change jobs, and 1.5 billion company employees worldwide will have their own intelligent assistants.

Key features of future AI agents include:

- **Autonomous actions based on the preset goals:** Different from today's LLMs, which lack the ability to translate specific goals into actions, AI agents evaluate their goals, develop plans, and take appropriate actions to accomplish those goals. All of these are done autonomously.

- **Persistent memory and intelligent state tracking:** AI agents can be equipped with long-term memory or the ability to track past states. Their knowledge can be accumulated and used as the basis for subsequent decision-making and actions. All these enable more intelligent AI systems.

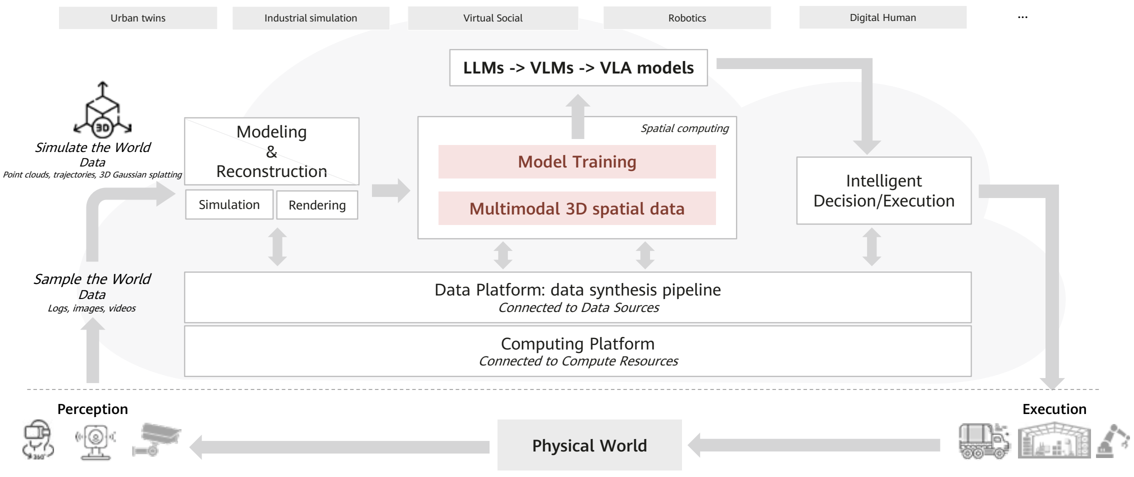
- **Interaction with the environment:** AI agents can perceive and understand their environment, regardless of whether it is a digital world or robotic system. In the future, AI agents will be able to interact with the physical world through sensors or other physical components.

- **Long-term learning and accumulation:** AI agents can autonomously learn from their interactions with new environments and in dealing with new situations. They can continuously optimize their knowledge systems and update their skills.

- **Multi-domain task handling:** AI agents have the potential to become general-purpose, multi-tasking AI systems that seamlessly integrate multiple skills, such as language processing, logical reasoning, perception and understanding, control and manipulation. They will assist humans in tackling a wide range of complex problems.

3.3 Transforming the Physical World

By 2030, intelligent technologies will have profoundly transformed our physical world. This fundamental transformation encompasses the entire process, from perception to computation, and ultimately to action.



New technologies such as XR devices, eye tracking, gesture recognition, and voice interaction will enable more natural and more efficient interaction (visual, voice, and action) between humans and the 3D digital world, ushering in the new era of spatial computing. Estimates show that by 2030, 60 million XR devices will be shipped annually, and around 500 million people will spend an average of 5 hours per day in a world of spatial computing that joins the physical and virtual worlds.

The transformation in perception is underpinned by wide interoperability between assets and models, enabling us to build global-scale digital twins. Digital twin and generative AI, combined together, will enable people to use these new spatial devices more effectively, generating more diverse, higher-precision 3D spaces by accurately extracting the features of the real world. The data of the 3D digital world, combined with data synthesis technology, will lay the foundation for spatial intelligence and embodied intelligence. The 3D digitization of the real world and the integration of embodied AI models into robots will enable seamless interaction and in-depth integration between the digital and physical worlds.

3.3.1 Representation of 3D Spaces: Integrating AI and CG to Accelerate Information Exchange in a 3D Digital World

The digital representation of 3D spaces has evolved from manual processing of mesh geometry, materials, and illumination to new methods that combine photo-shooting and AI generation. With convenient collection of spatial information using various cameras and innovative representation techniques, key details about real-world spaces—such as illumination, colors, materials, and spatial depths—can now be quickly and accurately gathered at low costs. Additionally, spatial-temporal data can be seamlessly overlaid onto 3D models, accelerating the comprehensive, standardized representation of complex scenes by hundreds of times.

We estimate that by 2030, city-level 3D reconstruction solutions will cover areas up to 10,000 square kilometers, incorporating real-time city events across all seasons. Such solutions will enable real-time city-level simulations, serving as digital training grounds for L4 autonomous vehicles, drones, and robots.

3.3.2 3D World Interaction: New Paradigm of Spatial Computing, Million-Time Increase in 3D Training Data

New, AI-powered interaction devices are becoming increasingly popular. Unlike traditional cameras and LiDARs, these devices are connected to cloud-based computing power and multimodal AI models, enabling them to capture and collect information about the physical world on a much larger scale. This means computer vision technology is now shifting from "sampling the world" to "simulating the world." Training data is changing from text, images, and videos to fine-grained 3D spatial simulation data across all modalities. Data from the physical world is preprocessed using data engineering pipelines and then synthesized, providing 3D spatial data for training vision-language-action (VLA) models that power spatial intelligence and embodied intelligence. The size of this training data is 106 times larger than the currently available training data for LLMs, which is approximately 13 trillion tokens.

VLA models, pre-trained using real-world data preprocessed using AI, rendering, and simulation techniques, offer enhanced spatial perception and AI capabilities. They are driving new computing paradigms and empowering more end-user devices, creating a data flywheel that will keep reinforcing itself well into the future.

We estimate that by 2030, under the new paradigm of spatial computing, human interaction will demand 1 zettaFLOPS computing power, while data preprocessing and training of VLA models that power spatial intelligence will demand 100 zettaFLOPS computing power.

3.3.3 Embodied Intelligence: Human-like Robots Are Seeing Wider Adoption, Super-human Robots Are Taking Shape

Embodied intelligence, or embodied AI, refers to the integration of AI and robots that enables robots to understand their surroundings and themselves and interact with the physical world to perform designated tasks. It adds the element of embodiment on top of spatial intelligence, enabling the ability to act and interact with the real world.

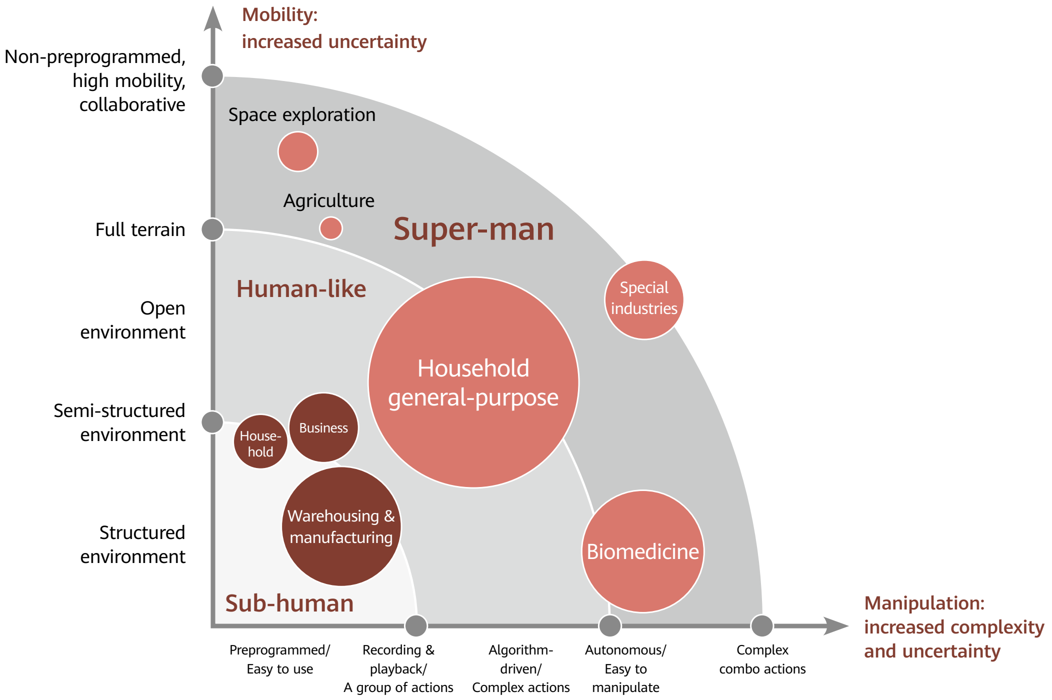
With embodied AI, robots are expected to evolve from sub-human, to human-like, and on to super-human. Sub-human robots primarily handle routine and repetitive tasks. Human-like robots are more adaptable and capable of performing more complex tasks, while super-human robots are supposed to surpass humans in all aspects, working in extreme environments and managing very complex tasks. Rapid advances in LLMs and VLAs are driving exploration of general-purpose, human-like robots. Intelligence lies at the core of robots, while data is key to intelligence. Key technologies include:

A big brain on the cloud and smaller brains on the edge: Together, they power super intelligent robots that may eventually surpass human-level intelligence.

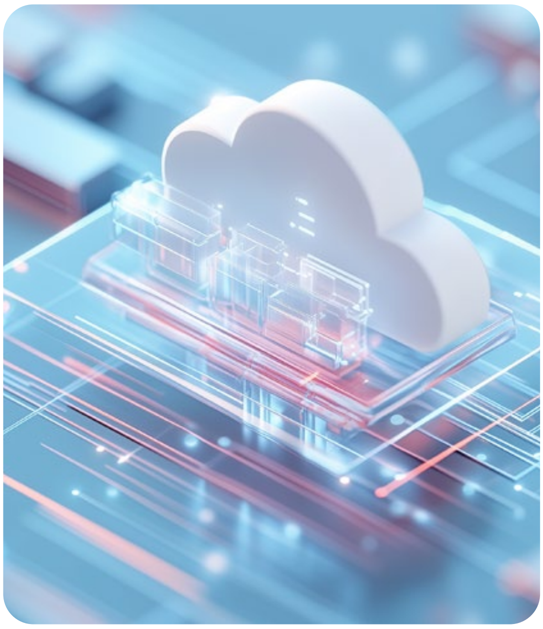
Cloud-based simulation and data synthesis: Data for embodied intelligence is scarce. This means synthetic data is key to training the brains that power embodied intelligence.

Online learning and a closed loop of data: Knowledge is not intelligence. True intelligence can only be learned through interaction with the real world.

We estimate that by 2030, general-purpose service robots, general-purpose factory robots, and general-purpose household robots will see widespread adoption. Powered by embodied AI, their shipments are expected to reach 30 to 50 million units annually. By then, early forms of super-human robots may have emerged, capable of performing tasks that are beyond human capabilities, such as space exploration.



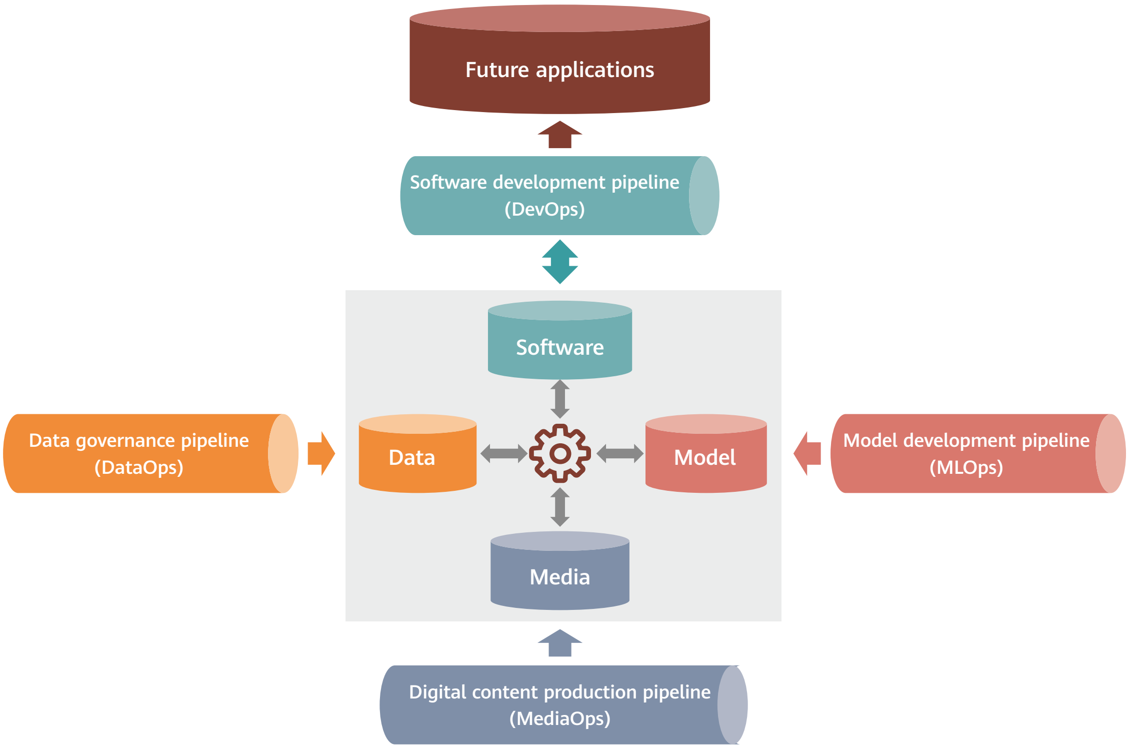
3.4 Application Modernization



According to Gartner, the digital economy will be made up of 500 million new applications in 2025, equivalent to all of the applications of the past 40 years combined. Traditional applications do not have the development, scalability, resource utilization, or O&M efficiency that are needed to adapt to this economy. In response, applications need to modernize, as can be seen by the growing trends in microservices, serverless, DevOps, and low-code development.

Future applications need an intelligent platform that integrates data governance, model development, digital content production, and software development, as envisioned in Software 3.0. They will combine data, models, digital media content, knowledge, code, and services to evolve from Cloud Native to AI Native for human-machine collaboration, continuous learning, growth, high autonomy, and swarm intelligence.

Such applications have an increasingly complex structure and will foster an ecosystem that dynamically combines and superposes various software elements in both the social (use and management) and physical (generation) spaces, which will be integrated with the information space for a new business model by 2030.



Future applications
Software development pipeline (DevOps)
Software
Data governance pipeline (DataOps)
Data
Model
Model development pipeline (MLOps)
Media
Digital content production pipeline (MediaOps)

3.4.1 Trends

1) From Code to Integrated Elements, AI Reconstructs 80% of Applications

Future applications are not just about code. They blend code, data, AI models, and digital content for more diversified and intelligent experience.

These composite intelligent systems will comprise system engineering mindsets, foundation model services, hybrid models and architectures, as well as personalized storage, retrievers, generators, and external tools. Yet their development remains economical, thanks to intelligent workflows, job and service orchestration, and component assembly in a converged cloud and AI native framework.

Future applications will integrate traditional code with data for processing, with AI and large language models for intelligence, and with digital content (such as text, images, videos, and interactive elements) for experience.

Their quality will not be just measured quantitatively, but by trustworthiness. These metrics include data privacy, output randomness, result explainability, and legal compliance.

All applications are worth rebuilding with AI foundation models, and 80% of them are expected to achieve this by 2030.

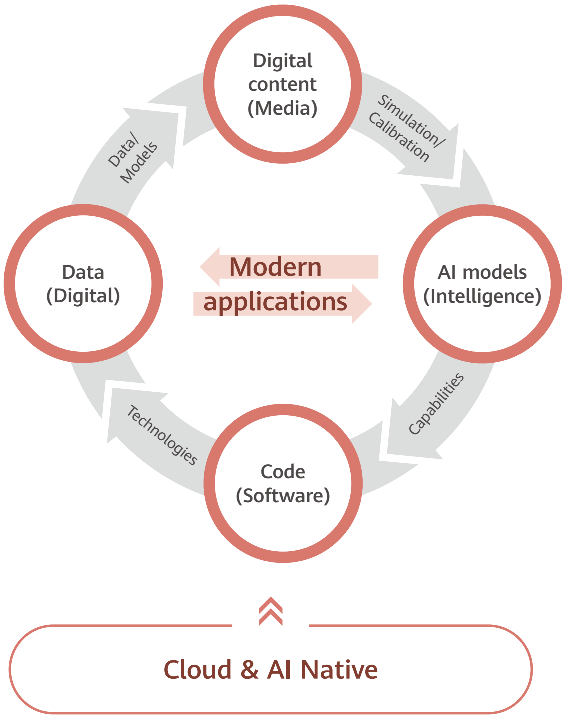
2) From Conventional to AI-powered Convergence Pipeline for 100x Efficiency

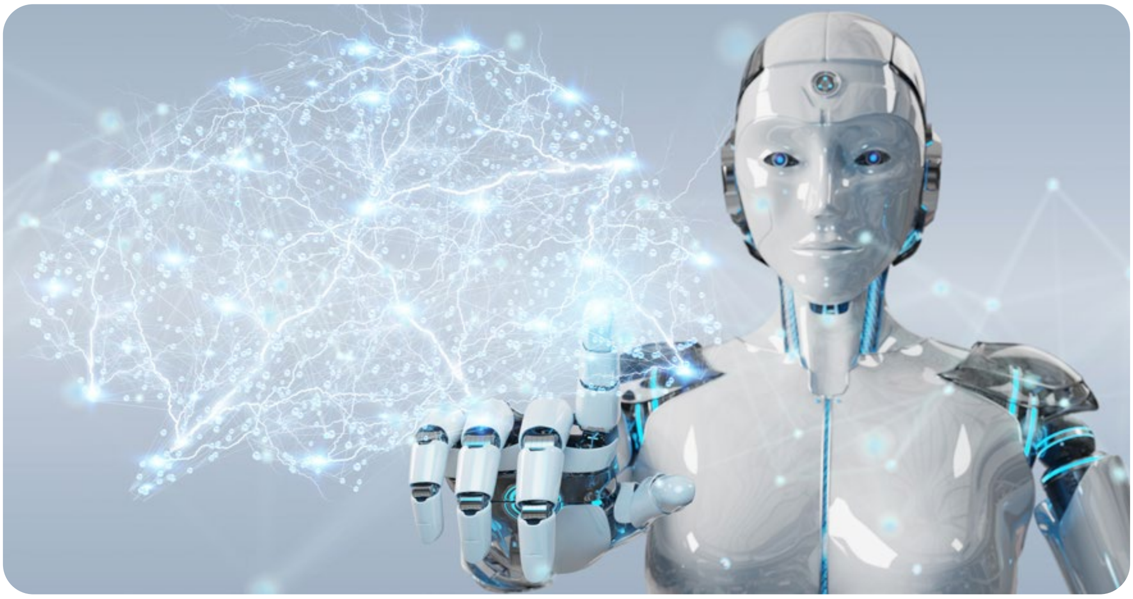
AI drives software engineering into a new era characterized by:

- **Directed programming and development:** Software is now independently developed on the loop instead of in the loop. This intelligent shift involves automated project management, risk control, team collaboration, and analysis. This journey of collaboration between people and highly autonomous agents is not only a technological rejuvenation, but also a fundamental reshaping of R&D roles.

- Intelligent pipelines: DevOps, DataOps, MediaOps, and MLOps remold services and tools to embrace new software engineering concepts, methodologies, and practices.

1. DevOps: This evolution has three phases.





Intelligent assistance (2023–2025): AI enhancement technologies, including R&D models, model fine-tuning, retrieval-augmented generation (RAG), and prompt engineering, are incorporated into existing processes and tools to boost efficiency.

Intelligent collaboration (2026–2028): Professional R&D assistants empowered by AI agents, task decision-making, and tool ecosystems will work with humans on complex tasks to unlock productivity.

Intelligent autonomy (2029–2030): "Intelligent independent developers", fueled by artificial general intelligence (AGI), will complete R&D from end to end with little human intervention, improving productivity by 10 to 100 times.

Today, DevSecOps is essential for the secure use and governance of open-source software. The agile and automated philosophy of DevOps will merge with industries to form "DevIndustryOps", enhancing competitiveness and market responsiveness.

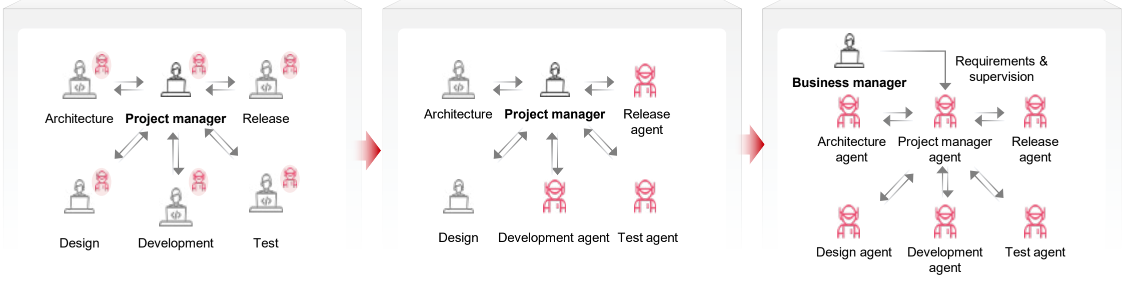
2. DataOps: Data collection, processing, and analysis are boosted by in-depth AI convergence. Augmented analytics combines GenAI with business intelligence, data science, machine learning, anomaly detection, and action assistance to provide new human-machine interactions in natural languages, automating decision-making with insights, code, and data.

3. MLOps: This key practice streamlines the development, deployment, and monitoring

Phase 1: Intelligent assistance

Phase 2: Intelligent collaboration

Phase 3: Intelligent autonomy

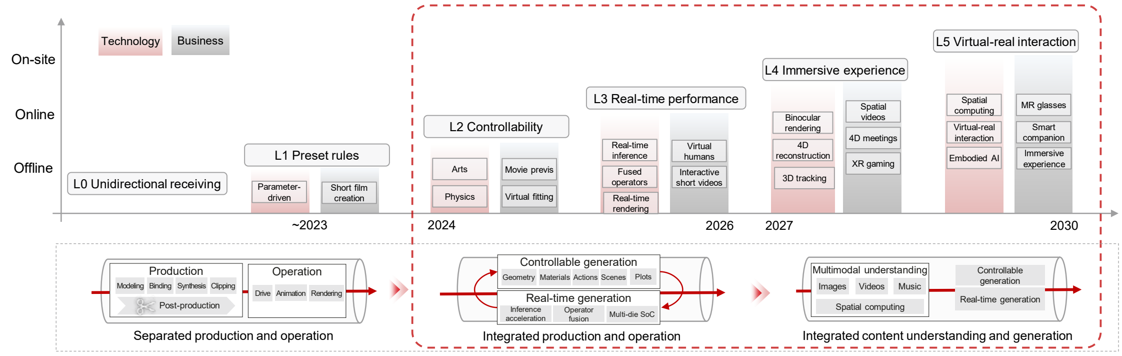


AI-assisted coding

AIGC architecture

Human-supervised AI development

Intelligent interactive media L0-L5



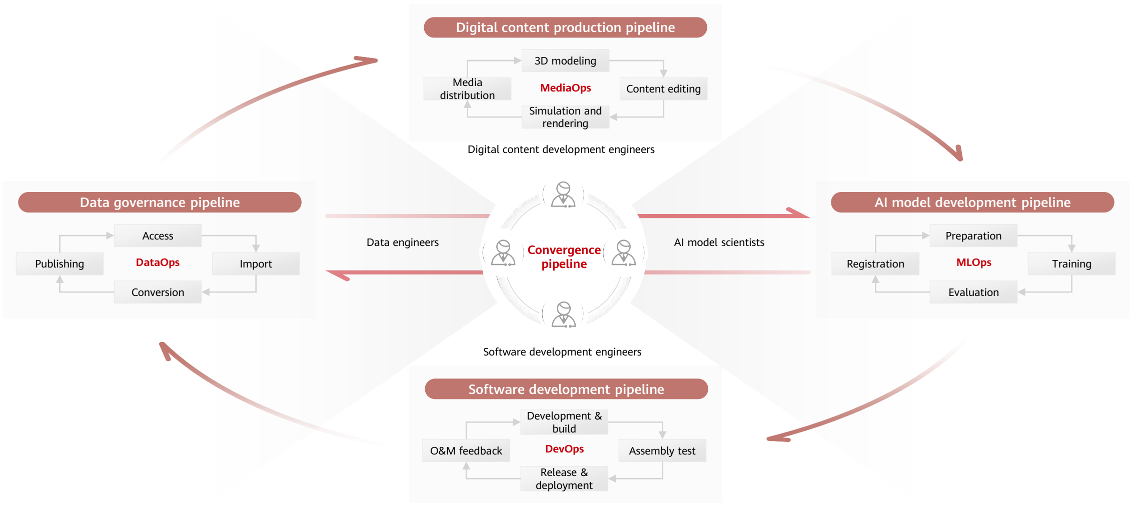
of machine learning models for iteration and processing. It will evolve and deploy AI models more quickly and widely.

4. MediaOps: The media industry is redefined with interactive short videos, scenario-aware ads, six degrees of freedom (6DOF) e-commerce, and immersive movies. These intelligent technologies profoundly change the paradigm of short videos, movies, and games from production and distribution to real-time generation and operation. It is estimated that by 2030, "AI+CG" will help 100 million digital content producers create up to 1 trillion hours of interactive media content.

- **Multi-modal software pipeline**

Each pipeline converges with human wisdom and machine efficiency to make software development more efficient, flexible, and innovative. Combined, this automated, omniscient platform integrates DevOps, DataOps, MediaOps, and MLOps to eliminate silos and bolster efficiency.

By 2030, the convergence pipeline will be widely used with any skill level for software development and innovation, increasing efficiency by 10 or even 100 times.



3) From Led by Humans to Driven by Humans + Compute + Data

Although traditional development has been simplified by emerging IDE tools and high-level programming languages, it still relies on human effort for tasks like requirement analysis, coding, project management, and team collaboration and must be measured by manpower.

Future software will involve more elements such as data, models, and digital content, which complicates the cost structure. Integrating, using, and supervising these elements still require human participation, but compute and data resources play an increasingly important role. By 2030, software costs will not be measured with just manpower, but with a mixture of elements including compute, data sets, and digital copyright.

3.4.2 Intelligent Evolution

Applications utilize cutting-edge AI to optimize, predict, and make decisions autonomously, mimicking human-like thinking and learning. In today's highly competitive market, intelligent applications not only personalize services at low cost, but also promote high-quality decision-making.

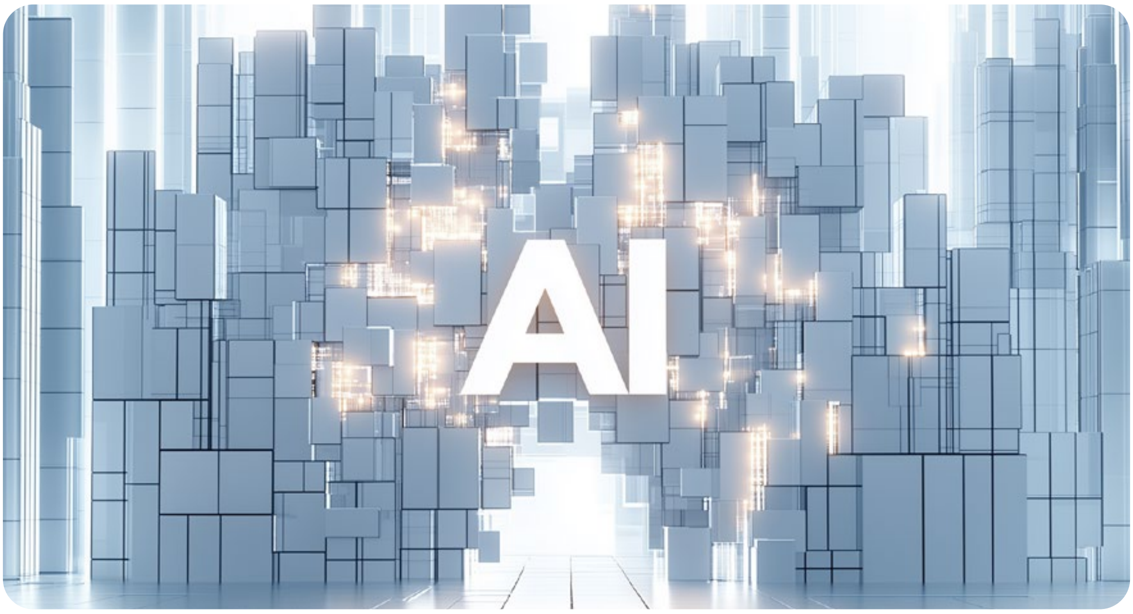
1) From Man-Machine Dialogue to Real Communication

Natural language processing (NLP) takes man-machine dialogue to the next level, where intelligent applications can comprehend complex languages and contexts and provide accurate information and services. Advancements in sentiment analysis allow these applications to recognize and react well to user emotions, interactively comforting or encouraging users based on their tone and language.

By 2030, NLP technology will humanize applications with a speech recognition accuracy of up to 98%. Customer service with sentiment analysis will reduce complaints by 30% to 50%.

2) From Traditional Operations to Intelligent Processes

Intelligent service processes are a key innovation in enterprise operations and management. By 2030, predictive analytics and pattern recognition



will be necessary for efficient operations and quality decision-making. Enterprises will use machine learning and big data analytics for insights into market trends, consumer behavior, and potential risks.

In this transformation, personalized AI assistants will analyze staff workloads and recommend custom improvements. These assistants will take on cumbersome tasks to free employees to focus on more strategic work.

In the future, every employee will have their own AI partner that will always be by their side to help them work more efficiently. This transformation will not only boost efficiency but reduce human error by over 50%. Significant decision-making in enterprises will take only half the time it used to, enabling quick responses to market changes.

3.4.3 New Applications

1) Web3 Applications: Digital Attestations and Decentralized Architecture for Trustworthiness

Web3 establishes a trustworthy foundation that is unavailable in Web2 for the digital world. Its decentralized architecture records all transactions and data in immutable ledgers, eliminating the risks of manipulation and fraud. Here, users own and control data of Web3 applications and profit by participating in the ecosystem. They are no longer passive content consumers, but creators and beneficiaries. Web3 applications can effectively supplement traditional ones for:

- Content creation and copyright protection: Non-fungible tokens (NFTs) and smart contracts help creators own and distribute earnings.

- Supply chain management: A decentralized record system improves transparency and traceability and reduces fraud.

- Financial services: Decentralized finance (DeFi) allows users to manage assets and gain profits without traditional bank middlemen.

Smart contracts and zero-knowledge proofs are crucial for Web3 applications.

Smart contracts provide code functions and the conditions for verifying and executing contracts, helping Web3 decentralized applications (DApps) execute and manage various functions and services. Smart contracts will be deeply integrated with AI to process more complex logic and decisions via autonomous learning and optimization. Bloomberg predicts that the smart contract platform market will soar to USD15–25 trillion by 2030.

Zero-knowledge proofs enable off-chain data and accounts to get verified and obtain trust in the Web3 ecosystem. This can be used for diverse scenarios such as asset proofs, anonymous voting and payment, and transaction privacy protection. With the approval of BTC/ETH ETFs in Hong Kong (China) and the United States at the beginning of 2024, more real assets will enter into the decentralized world, and more traditional and decentralized financial systems will be integrated. The current 400 million people using digital assets around the world are projected to reach 1 billion by

2030. By then, 16 trillion real-world assets will be traded and circulated. All mortgages and pledges that are now physically proven will be electronically verified and evaluated for their actual asset value. By 2030, Web3 applications will execute 90 billion zero-knowledge proofs with USD10 billion compute resources.

2) Quantum Applications: Better for Finance, Physics, Energy, and Biology

A historic "quantum advantage" is expected to happen by 2030 with practical algorithms and powerful computers. Although quantum algorithms like Grover and Shor have shown theoretical potential, they still rely on quantum computers with enough qubits and robust error correction. By 2030, qubit error correction will be precise to 99.999% to 99.9999%, and the number of qubits will grow from 1,000+ to 10,000+. These technological advancements will make it possible for quantum computing to be applied in various industries, including finance, physics, energy, and biology.

By 2030, quantum computing as a service (QCaaS) will become a new industry standard that empowers enterprises and research institutes to easily access quantum computing resources through a cloud platform without investing in or maintaining expensive hardware. Quantum computing infrastructure, algorithm/operator platforms, and application development platforms will mature on the cloud. Presently, quantum computing hardware is developing towards superconducting, ion trap, optical quantum, cold atoms, and spin quantum computing, technologies that have the potential to emerge as the major directors of wide application and commercialization by 2030. At that point, quantum computing will not only be present in scientific research but in industry innovation and transformation.

By 2030, the global application modernization market will be fueled by enterprise digitalization, legacy system reconstruction, cost optimization, and experience enhancement to reach USD380 billion, with a CAGR of 17%.

3.5 Better Cloud Operations

Cloud vendors are continuously refining their enterprise cloud concepts and frameworks based on real-world experience. They are constantly improving their understanding of cloud products and their abilities to combine them in different ways to create new product offerings. Moving forward, we can expect to see AI integrated to deliver advanced capabilities like automated migration, lean governance, hidden resilience, deterministic operations, and refined FinOps. This allows enterprises to fully leverage cloud computing and achieve their business goals faster.

3.5.1 Automated Migration: E2E Automation, 10 Times Faster Than Before

Enterprises are rapidly migrating their services to the cloud, driving the development of migration tools. These tools streamline the migration process by tracking status updates, analyzing services, designing architectures, implementing the migration itself, and verifying the data afterwards. By doing so, they help mitigate the risks and costs throughout the cloud migration. However, enterprises are still facing many migration challenges, such as identifying application dependencies, evaluating technical feasibility, selecting the most appropriate cloud resource specifications.

In the future, cloud vendors will incorporate AI into migration tools to analyze the code, configuration files, and runtime logs of the source application system. They will make efforts to identify application architectures and component dependencies.
The target end's technical architecture will be automatically designed based on design principles and best practices. Cloud vendors will use AI to analyze historical performance data of the application system and forecast workloads. They will use it for automated performance testing and capacity planning. AI can help more accurately forecast resource requirements. The most cost-effective cloud resource specifications will be automatically recommended based on the cloud service pricing model. As migration pipeline technologies continue to improve, it is possible that cloud migration will be fully automated. By 2030, cloud migration is expected to be fully automated and unattended. The migration efficiency will be over 10 times higher. Complex application systems will be smoothly migrated from data centers to clouds and between multiple clouds.

3.5.2 Lean Governance: Eliminating 90% of Compliance Risks Early On

A change in quantity can entail a change in quality. Large-scale cloud migration inevitably brings many

challenges in cloud governance. Managing resources and ensuring compliance become exponentially challenging due to the complex dependencies among hundreds of service systems and tens of thousands of cloud resources. Additionally, hundreds of thousands of employees from various business lines and partners require access to these resources. If responsibilities are unclear, you can end up with disorganized permissions configurations. This can result in an exponential increase in data leak channels, making risk control much challenging. To address these challenges, cloud vendors launched the Landing Zone or Governance, Risk, and Compliance (GRC) solution, but there is still significant room for improvement. A survey shows that 76% of enterprises consider cloud governance a significant challenge. As multi-cloud and hybrid cloud environments become more popular, enterprises encounter even more difficulties in cloud governance.

Moving forward, the Landing Zone solution will use zero trust systems to establish comprehensive data perimeters. This ensures that sensitive data can only be accessed when environments, identities, networks, and resources are all trusted. AI will be used to analyze users' historical access and automatically generate and configure strategies to help enforce the principle of least privilege (PoLP) based on user roles and responsibilities. Users will be continuously authenticated and dynamically authorized based on their attributes, the resources they need, and the environments they are in. AI will be deeply integrated into the Landing Zone solution, making it easier to govern people, finances, resources, permissions, and security compliance in a well-architectured, centralized, fined-grained, and intelligent manner. It will help identify most potential risks in advance, based on complex correlation analysis. It can then automatically eliminate these risks early on. Furthermore, a unified platform will probably be used to manage multi-cloud and hybrid cloud environments. AI and automation technologies will be utilized to integrate management interfaces from different cloud vendors to establish consistent cloud governance frameworks, strategies, and processes, simplifying management. By 2030, cloud governance is expected to become more intelligent and streamlined, and over 90% of compliance risks will be eliminated in the early stages.

3.5.3 Hidden Resilience: Zero-Burden, Fully Managed, and Highly Self-Healing

Cloud application resilience refers to the capability of an application system to keep running and quickly recover from faults caused by cloud infrastructure issues, external attacks, external dependency problems, or even regional disasters. Many enterprises enhance application resilience using microservice architectures, containerization, and distributed systems. They implement automated monitoring, fault detection, and recovery systems so that applications can be automatically adjusted and rectified when faults occur. Cloud vendors also provide various tools and services to help enterprises establish more resilient application systems, which further drives the rapid advancement of resilience technologies.

By 2030, application resilience is expected to enter a new era, where it is zero-burden, fully managed, highly self-healing, and completely hidden from view. Nano-level intelligent self-healing technologies will be used to provide micro self-healing for different processes. A global unified sensing system will be used to detect faults and disasters in real time. There is no need to create numerous redundant instances in advance. New instances can be created in seconds when a fault occurs, ensuring seamless fault handling without service interruptions. Tenants no longer need to rely on O&M for resilience. Instead, technologies will be leveraged to implement unattended resilience and O&M. Enterprises will shift their focus from improving application resilience and reliability to using the resilience bastion technology to harden applications through containerization. Traffic control and perception will be incorporated. Steady intelligent monitoring will be used to implement evolutionary traffic consumption and balance. Chaos engineering will no longer be limited to a specific script. Unpredictable chaos engineering will be used to cultivate more creative blue teams.

3.5.4 Deterministic Operations: 80% of Cloud Faults Fixed in Just 10 Minutes

Digital transformation poses severe challenges on O&M. Multi-cloud environments and multiple technology stacks often have to coexist for a long time, which makes O&M more complicated and availability assurance more difficult. It is hard to balance service agility and live network stability, which increases production risks. Various incidents occur all too frequently. Ensuring system stability and reliability has become the lifeline of service development. It is increasingly difficult to establish a highly available service system without breaking the bank. The skills of O&M personnel do not align with the way the future has been shaping up. Traditional O&M capabilities can no longer meet the demanding requirements for stability and reliability of digital services. Enterprises need a more efficient, secure, and reliable O&M quality assurance system.

Deterministic operations continuously develop full-stack technologies and O&M processes for all scenarios. This includes designing and evaluating high-availability architectures, forecasting and preventing risks through proactive O&M, conducting chaos drills, and analyzing end-to-end observations. It also involves recovering from faults in a deterministic manner, dynamically governing risks for large-scale resources, and controlling change risks. These measures greatly improve enterprise cloud O&M, reduce system faults and downtime, and reduce labor costs and operational risks.

In the future, AIOps will be used to create a digital twin world of hundreds of millions of O&M objects on the cloud. All changes to cloud resources will be perceived in real time, and global O&M statuses will be visualized. Foundation models and SRE expertise will be utilized to provide intelligent O&M diagnosis and decision-making for automatic fault decision-making and recoveries. By 2030, enterprises are expected to fully implement deterministic operations for their management systems. This will help them detect 80% of cloud faults within 1 minute, respond within 5 minutes, and recover within 10 minutes. It will automatically track, make decisions for, and handle 80% of O&M tasks.

3.5.5 Refined FinOps: Saving Over $200 Billion USD per Year for Cloud Users

A survey shows that managing cloud expenditures is a major challenge for many enterprises. On average, 27% of public cloud expenditures are unnecessary. 51% of enterprises have established dedicated FinOps teams to cope with this issue. FinOps has become a critical component of many enterprise cloud strategies. By managing and optimizing costs effectively, enterprises can better control cloud expenditures, enhance their ROI, and promote sustainable business growth.

In the future, AI will be deeply integrated into FinOps. AI can accurately forecast costs, identify expenditure spikes and resource waste, and trigger automated repairs and optimization. The entire process, from cost visualization to optimization, will be fully automated. FinOps and AIOps will work closely to combine cloud cost data with operations data. This will provide more comprehensive insights and enable more intelligent decision-making for operations. AI will also make it easier for enterprises, including small- and medium-sized ones, to benefit from FinOps. FinOps will also use GreenOps to help enterprises reduce their carbon footprint and energy consumption while optimizing costs. By 2030, intelligent and refined FinOps is expected to be a must for enterprises to use the cloud. It will save global users over $200 billion USD per year by reducing cloud resource waste.

3.6 Boundless Security

3.6.1 Threats: The Most Frequent and Complex Cyber Attacks Ever

By 2030, cybersecurity will face unprecedented changes resulting from advances in AI and quantum computing, changes in the geopolitical landscape, and the emergence of new services and scenarios. Quantum computing will probably be able to make quick work of the cryptographic methods that are widely used now. There will be an increase in the generative adversarial attacks (such as malicious content, deepfakes, model architecture attacks, and various forms of malware) driven by AI large models. National-level advanced persistent threats (APTs) and professional commercial ransomware will greatly increase the intensity and impacts of network attacks. More countries and companies will need to pay special attention to the security and trustworthiness of supply chain hardware and software. Networks will be more vulnerable with the increase of sensors and controllers of smart devices (such as smart vehicles, smart home devices, and robots). The wide use of digital assets and smart contracts will create severe security challenges for Web 3.0.

It is estimated that by 2030, there will be more professional hacker organizations, highly intelligent attack tools, and new commercialized cyber attack services. The intensity and complexity of cyber attacks will increase by several times, maybe even by as much as 10 fold.

3.6.2 Defense: A Complete, Cloud-oriented, In-depth, Zero-trust System

Traditional security systems only provide reactive or static defenses, depend on border defense, and lack collaboration between products. A future-oriented, intelligent defense system needs to feature proactive defense, data sharing, internal and external consistency, and cloud-network-edge-device-chip collaboration. The architecture needs to be tightly coupled to the services being protected (native security that is deeply integrated into the services). A future-oriented system needs to offer a holistic security view. It needs to be a zero-trust system focused on cloud security, a system able to withstand the fierce network attacks of the future.

1) Zero Trust: Protecting Clouds, Networks, Edges, Devices, and Chips; Changing from Add-on Security to Native, Distributed Security

In the future, a zero trust architecture will protect corporate networks, cloud computing environments, mobile devices, remote offices, remote access, supply chains, IoT devices, applications, data, and more. Companies will deploy more zero trust products than they do today. They will fully integrate different zero-trust solutions, the original cybersecurity architecture, and business applications. It is estimated that more than 95% of companies will implement zero trust policies by 2030. Separate zero trust solutions will be integrated into a zero trust architecture with cloud security at its core.

The zero trust architecture will expand from network border (Zero Trust Network Access [ZTNA] and microsegmentation) to cover chips, devices, identities, data, applications, networks, and infrastructure. It will reshape the network architecture. Zero trust will also expand from external security to cloud native security, which means that security will become an inherent part of the system architecture, not just an additional layer for protection. Security will be shifted left in application development. In the future, security measures will be integrated right from the design stage to ensure synchronous development of security and other functions. In the future, security will be distributed and dynamically scalable based on service and traffic requirement. Security will be an integral part of the services, just like the immune system is an integral part of the human body.

2) Security System: Evolving to Intelligence for Proactive and Automatic Defense

Traditional rule-based security solutions struggle to process massive amounts of data and identify new attack patterns. By 2030, it is expected that 80% of organizations will adopt AI-powered cybersecurity products to address the evolving threat landscape and implement more proactive and intelligent security defenses. The security industry will fully leverage AI's technical advantages to achieve intelligent upgrades, particularly in threat detection, prediction, automatic response, and security decision-making optimization.

Intelligent Threat Detection: AI can identify unknown threats through behavior analysis and extensive threat intelligence, significantly enhancing the detection capabilities of traditional products. AI is primarily utilized in EDR, firewall, and APT products.

Event Prediction: AI introduces new technologies and paradigms for predicting imminent threats. Using techniques such as graph neural networks, malicious behaviors are modeled, and a security foundation model is trained with both benign and malicious samples. This enables the foundation model to deeply understand the internal relationships and rules between attack events, allowing it to predict the types, targets, and methods of future attacks.

Automatic Response and Security Decision-Making: AI-generated threat intelligence helps security operation systems proactively and accurately formulate defense policies for both current and impending attacks, and automatically respond to them. This proactive approach minimizes security risks by implementing targeted protection measures in advance. Additionally, AI assists security experts in tracking attackers comprehensively, enhancing their situational awareness, and supporting proactive defense actions such as deterrence, trapping, and source tracing.

3) Security Ecosystem: Intelligent Cloud-Based Innovation, Industry Integration, and Win-Win Cooperation

The security industry faces challenges such as fragmentation, market segmentation, lack of systematic architecture for vendors, and insufficient ecosystem channels. In the new wave of security transformation driven by intelligence and cloud technologies, the industry will undergo intelligent upgrades and cloud-based innovations on cloud platforms. This will promote industry integration (resources, technologies, and markets) and foster a win-win, open, innovative, and future-oriented security ecosystem.

In the future, the new security defense system will leverage cloud platforms for technology, intelligence, and market sharing, enabling joint defense and coordinated intelligent operations against security threats. The traditional security ecosystem, which includes security software, hardware, service providers, consulting firms, and training institutions, will adapt to technological trends through cloud-based and intelligent transformations. By 2030, it is estimated that 90% of security vendors will transition to the cloud, embracing it proactively and integrating deeply to address more advanced cyber threats.

3.6.3 Security: The Immune System of the Cloud

1) Large Models: Comprehensive Protection Covering Four Aspects

By 2030, 50% of enterprises are expected to have their own large models. Large model security covers the following aspects: data security, confidential computing, model hardening, and content moderation.

Data security: Protect sensitive data and assets involved when training, using, and storing large models, including data property protection (data source validity, copyright authorization, and open source license), data lineage (upstream and downstream relationships and associations formed during data processing, transfer, and convergence), data encryption, and data cleansing.

Confidential computing: Ensure that enterprises can use sensitive data to train large models, protect data and models, and prevent attacks and damages caused by privileged and internal personnel. The combination of the following ensures security: First, hardware-based security and isolation. Second, zero-trust architecture (ZTA). The authentication service is used to verify the credibility of compute assets. Third, data owners are allowed to collaborate in model training while protecting the confidentiality of dedicated data.

Model hardening: Ensure the security of large models throughout their lifecycle, including development, training, testing, deployment, and management. Model leakage prevention and model encryption are widely used for security hardening. Preventing model leakage aims to protect models, including its parameters and functions, from being replicated or migrated through model technologies such as model theft and model compression. Encrypting models aims to protect model confidentiality, integrity, and availability. The taken measures are as follows: model leakage prevention, that is, replicating or migrating large model capabilities, including parameters and functions, through; model encryption, that is, ensuring with encryption. There are still other measures commonly used, such as model tampering prevention, model backdoor detection, and model copyright protection.

Content moderation: User inputs and large model outputs are checked to ensure that there is no immoderate content, such as the one against social values and privacy laws, or the one including bias, extremism, and discriminatory remarks. Technologies such as text review, image review, video review, AIGC privacy desensitization, AIGC watermarking, and AIGC authentication are used.

2) Cloud-Network-Edge-Device Synergy: A Full-Stack Defense System

It is estimated that by 2030 more than 90% of enterprises will seek for the comprehensive integration of security suppliers. The cloud-network-edge-device synergy will be implemented for the cloud platform by then to create a unified and comprehensive defense system.

Device security: ZTA will be extended to the edge to implement cloud-assisted device, device-cloud synergy, and situational awareness. It is the shift from endpoint detection and response (EDR) to extended detection and response (XDR). XDR driven by AI large models will implement real-time analysis and automatic protection of enterprise terminal devices and IT assets to prevent virus software and network attacks, enabling end-to-end data protection.

Edge security: As edge nodes are distributed, to implement decentralized edge security, there are various challenges, such as increasing border defense risks and physical hardware risks. In the future, applications that are deployed at edges will be more advanced and intelligent. It is vital to create a cloud-centered global security scheduling system based on the synergy of cloud and cloud-assisted edges.

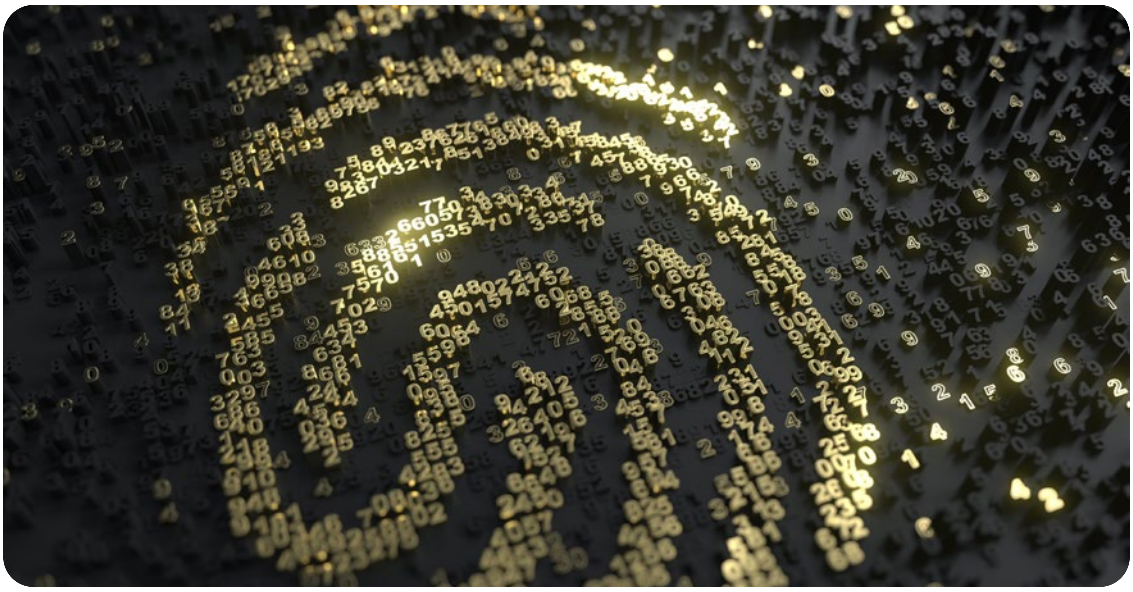
Network security: This is all about comprehensive monitoring and management of network traffic. AI-powered real-time analysis can identify and detect potential threats in the network traffic more efficiently. Based on the cloud and AI algorithms (network traffic models), the analysis precision will be improved greatly, making it possible to quickly identify and defend against more devastating DDoS attacks in the future.

Cloud load security: Apart from baseline defense, the traffic security is combined to monitor and protect the data traffic in the cloud environment, identify potential threats and malicious activities, and ensure data in transit security. As infrastructure as code (IaC) and policy as code (PaC) approaches are used, automated moving target defense (AMTD) for cloud workload security will soon be put into commercial use.

3) Data: Ensuring Encryption, Status Visualization, Transparency, Traceability, and Auditability

By 2030, over 95% of enterprises will adopt data-centric security protection policies.. A confidential computing platform (confidential environment, remote attestation, trusted nodes, and multi-node architecture) is built through automatic data discovery, intelligent classification, full-link encryption, and status visualization. Ensure the security and accuracy of data asset transmission between different systems and platforms. Protection for enterprises can be provided to data assets.

In the future, many new technologies will emerge to ensure the security of core data assets. Hardware-based confidential computing will be widely promoted. Memory encryption will be enhanced based on the traditional data isolation. Zero-knowledge proof will be widely used in financial services, Web3, and blockchain aspects. Secure multi-party computation and homomorphic encryption allows operations on encrypted data, and the operation result is the same as that of the original data after decryption. On the premise of ensuring data privacy, federated learning, zero-knowledge proof, and differential privacy support data interaction and processing. These technologies are expected to mature in the next few years and will be commercially available in 2030.



3.6.4 Cloud for Better Security

1) Network Security: Distributed and Intelligent Network Attack Defense System

With the high bandwidth, abundant computing resources, and hundreds of data centers in more than 200 countries, the cloud platform provides a solid foundation for meeting the increasing computing power requirements. It is estimated that by 2030, the anti-DDoS capability of the cloud platform will reach over 10 Tbit/s, far exceeding the bandwidth of common enterprises' self-built defense systems and several or even dozens of times the defense bandwidth of enterprises. This provides enterprises with more powerful and reliable network security assurance.

The cloud platform uses elastic resources, advanced protection technologies, professional security teams, global data center networks, and automatic response mechanisms to build a defense system that can defend against large-scale network attacks.

2) Data Security: From Enterprise-built to On- and Off-Cloud integrated Protection Policies

Data has become the most valuable asset of enterprises. The security of data is highly valued by countries and enterprises. It is difficult for enterprises to invest in long-term and continuous data security resources, including technologies, talent strategies, and data asset risk management systems.

The cloud platform has a more complete data security defense system (end-to-end encryption, access control, and identity authentication) and more advanced technologies (confidential computing and encryption technologies). In addition, the cloud platform meets the data compliance requirements of governments and administrative organizations and has a more comprehensive data asset governance mechanism, helping enterprises comply with data protection regulations and standards.

Cloud data security technologies will continue to innovate and improve. Confidential computing and cryptographic algorithms will significantly improve the performance of cloud platforms. By 2030, the cost-effectiveness of cloud data security solutions will be more than 50% higher than that of enterprise-deployed solutions. This will further promote enterprises to use cloud-based data security solutions.

3) Security Operations: Embracing AI-powered Cloud Native SecOps

Cyber security depends 30% on systematic construction and 70% on constant operations. Security operations is a big challenge facing traditional enterprises because of expertise and resource shortages. While a cloud platform makes security operations simple and efficient based on its advanced security tools, professional services, and extensive expertise.

A cloud platform has a wide range of data sources, including run logs of security and cloud services, experience in defense against trillions of attacks, hundreds of millions of pieces of professional security knowledge, tens of billions of malware detection parameters, and more comprehensive, accurate, real-time, and in-depth threat intelligence and indicators. This makes it easier for a cloud platform than a traditional security platform to build threat identification models, provide automated response playbooks, and generate more powerful security models. In the future, the "security brain" powered by large models will help significantly improve security operations, including alarm noise reduction, attack analysis, automatic response, and automatic report generation. By 2030, 70% of enterprises are expected to adopt the AI-powered cloud native SecOps model.

4) Security Innovation: Technological Innovation Is Gradually Migrated to the Cloud

In the future, root technologies such as cryptography, security chips, and certificates will reshape the traditional cyber security industry. The cloud will implement the atomic root of trust based on cryptography and security chips. The cloud will facilitate the innovation, commercial use, and implementation of security technologies.

Cryptographic technologies, as the basis of security, will face new challenges brought by quantum computing threats, international situation changes, and emerging technologies. The cloud can provide a solid foundation for building a comprehensive, efficient, and reliable password protection system based on cloud native encryption services, centralized key management services, more advanced identity authentication mechanism, and more real-time and intelligent intrusion detection and defense capabilities. This will vigorously promote the rapid development of technologies such as post-quantum cryptography, SM series cryptographic algorithms, and hardware-based confidential computing.

Security chips are basic components for trusted computing. The implementation of security chips requires an end-to-end reconstruction of operating systems and applications, which is a complex process that involves in-depth integration and optimization of hardware and software. A cloud platform features software and hardware integration, capability integration, process standardization and automation, as well as good flexibility, security, and economic benefits. These make it an ideal platform for end-to-end reconstruction, making it easier to put security chips into commercial use. By 2030, the availability and cost-effectiveness of security chips will be greatly improved. It is estimated that the market penetration rate will be over 20%. Security chips will be popular in the market by then.

Call to Action

The best way to predict the future is to create it. Intelligence is not a mere vision; it is the defining trend of our times. As we stand on the cusp of the intelligent world of 2030, let us join forces to catalyze this transformation. Together, we will navigate the dynamic cloud landscape and harness the full potential of intelligence to reshape industries.

Appendix: Abbreviations and Acronyms

Abbreviation/Acronym	Full Spelling
ADMET	Absorption Distribution Metabolism Excretion Toxicity
AGI	Artificial General Intelligence
AI4S	Artificial Intelligence for Science
AIGC	Artificial Intelligence Generated Content
AMTD	Automated Moving Target Defense
APT	Advanced Persistent Threat
CNN	Convolutional Neural Network
CPU	Central Processing Unit
Dapp	Decentralized Applications
DataOps	Data Operations
DBA	Database Administrator
DCN	Data Center Network
DDOS	Distributed Denial of Service
DeFi	Decentralized Finance
DevIndustryOps	Industry Development Operations
DevOps	Development and Operations
DevSecOps	Development, Security, and Operations
EB	Exabyte
EDR	Endpoint Detection and Response
ER/EP	Enterprise Router/Endpoint
ETH ETF	Ethereum Exchange-Traded Fund
eVTOL	electric Vertical Takeoff and Landing

Abbreviation/Acronym	Full Spelling
FinOps	Finance Operations
GPU	Graphics Processing Unit
GRC	Governance, Risk, Compliance
GreenOps	Green Operations
GW	Gigawatt
HTAP	Hybrid Transactional/Analytical Processing
IaC	Infrastructure as Code
IDC	Internet Data Center
IDE	Integrated Development Environment
LLM	Large Language Model
MediaOps	Media Operations
MFU	Model Floating-point Operations Utilization
MLOps	Machine Learning Operations
MLOps	Machine Learning Operations
NFT	Non-Fungible Token
NLP	Natural Language Processing
NPU	Neural Processing Unit
OS	Operating System
PaC	Policy as Code
PDE	Partial Differential Equation
Proclet	Processlet
RNN	Recurrent Neural Network

Abbreviation/Acronym	Full Spelling
RWA	Real World Assets-tokenization
SecOps	Security Operations
SFT	Supervised Fine-Tuning
SLA	Service Level Agreement
SQL	Structured Query Language
SQL	Structured Query Language
SRE	Site Reliability Engineering
UPS	Uninterruptible Power Supply
V2X	Vehicle-to-Everything
VLA	Vision-Language-Action
VPC	Virtual Private Cloud
VPC	Virtual Private Cloud
VPN	Virtual Private Network
XDR	Extended Detection and Response
XR	Extended Reality
XR	Extended Reality
ZB	Zettabyte
ZTNA	Zero-Trust Network Access

HUAWEI TECHNOLOGIES CO., LTD.
Huawei Industrial Base
Bantian Longgang
Shenzhen 518129, P. R. China
Tel: +86-755-28780808
www.huawei.com

Trademark Notice

HUAWEI , HUAWEI, are trademarks or registered trademarks of Huawei Technologies Co., Ltd.
Other Trademarks, product, service and company names mentioned are the property of their respective owners.

General Disclaimer

The information in this document may contain predictive statement including, without limitation, statements regarding the future financial and operating results, future product portfolios, new technologies, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.