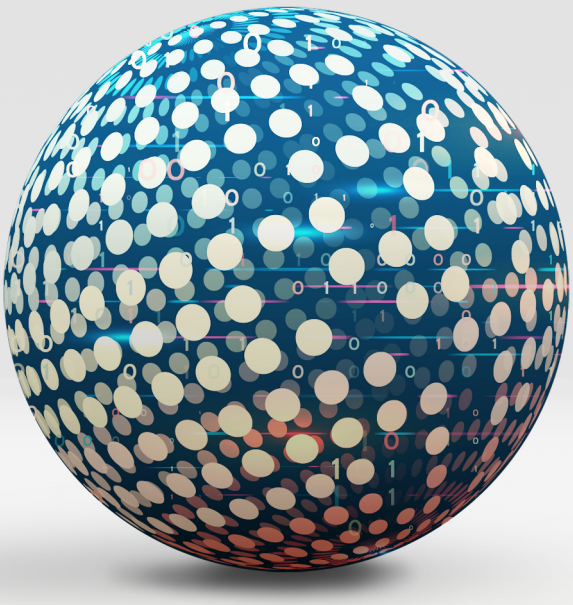
HUAWEI

Computing 2030



Building a Fully Connected, Intelligent World

Foreword

A decade ago, humanity generated just a few zettabytes[1] of data every year, and mobile Internet, cloud computing, and big data were still in their infancy. Today, these technologies are profoundly changing our world, and computing is playing an unprecedented role.

By 2030, we will be producing yottabytes[1] of data every year. The amount of general computing power in use will increase tenfold, and AI computing power will increase by a factor of 500[2]. The digital and physical worlds will be seamlessly converged, allowing people and machines to interact perceptually and emotionally. AI will become ubiquitous and help people to transcend human limitations. It will serve as scientists' microscopes and telescopes, enhancing our understanding of everything from the tiniest quarks to vast cosmological phenomena. Industries already making extensive use of digital technology will now use AI to become more intelligent. Computing energy efficiency will increase, bringing us closer to low-carbon computing, so that digital technologies can become a tool for achieving the global goal of carbon neutrality.

In the next decade, computing will help us move into an intelligent world – a process of the same epochal significance as the age of discovery, the industrial revolution, and the space age.

Contents

Macro trends

After half a century of development, computing has become deeply integrated into every aspect of our work and lives. In the next decade, computing will become the cornerstone of the intelligent world and continue to support economic development and scientific advances.

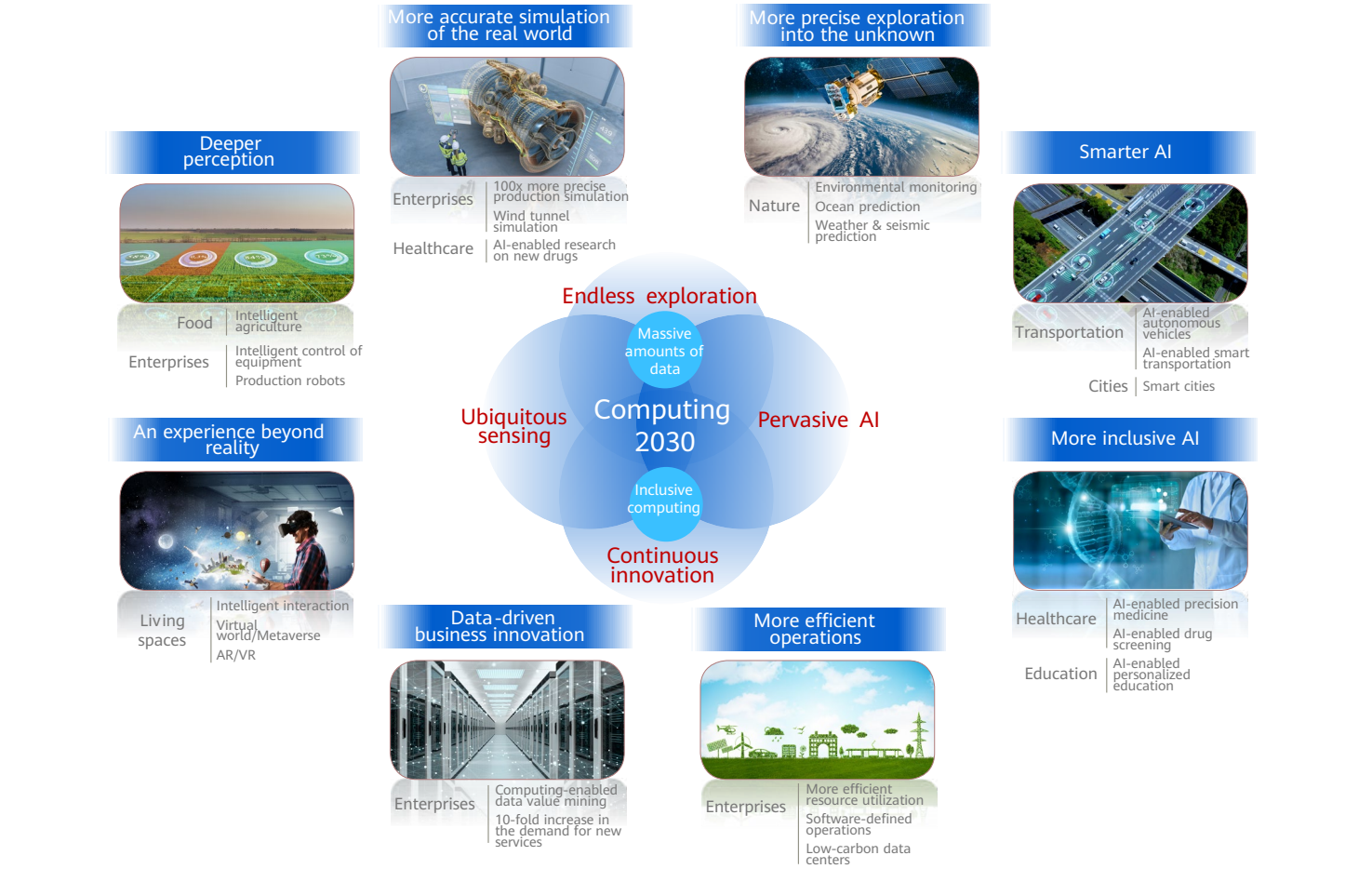
Looking ahead to 2030, many countries and regions, including China, the EU, and the US, will prioritize computing in their national strategies. China's 14th Five-year Plan and Vision 2035 define high-end chips, artificial intelligence (AI), quantum computing, and DNA storage as technologies of strategic importance for the country. The EU's 2030 Digital Compass: the European Way for the Digital Decade lays out a plan whereby, by 2030, 75% of EU companies will be making full use of cloud, AI, and big data, and the EU will have its first homegrown quantum computer. The US has reintroduced the Endless Frontier Act, which authorizes the government to legislate and make grants to promote US research in areas such as AI, high-performance computing, semiconductors, quantum computing, data storage, and data management technologies.

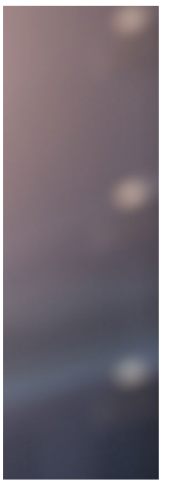
In 2030, the digital and physical worlds will be seamlessly converged. People and machines will interact with each other perceptually and emotionally. Computing will be able to simulate, enhance, and

recreate the physical world. Hyper-real experiences will drive computing to the edge, and necessitate multi-dimensional collaborative computing between cloud and edge, between edge and edge, and between the digital and physical worlds. AI will evolve from perceptual intelligence to cognitive intelligence and develop the capacity for creativity. It will become more inclusive and make everything intelligent. As the boundaries of scientific exploration continue to expand, the demand for computing power will increase rapidly. Supercomputers that can perform 100 EFLOPS[2] and a new, intelligent paradigm for scientific research will emerge. In the push toward global carbon neutrality, computing of the future will be greener, and service experience will get better.

The semiconductor technologies that computing relies on are approaching their physical limits, and this will spark a golden decade of innovation in computing. Innovation in software, algorithms, architecture, and materials will make computing greener, more secure, and more intelligent. It is estimated that by 2030, global data will be growing by one yottabyte every year. Total general computing power will see a tenfold increase and reach 3.3 ZFLOPS, and AI computing power will increase by a factor of 500, to more than 100 ZFLOPS[2].

Future computing scenarios

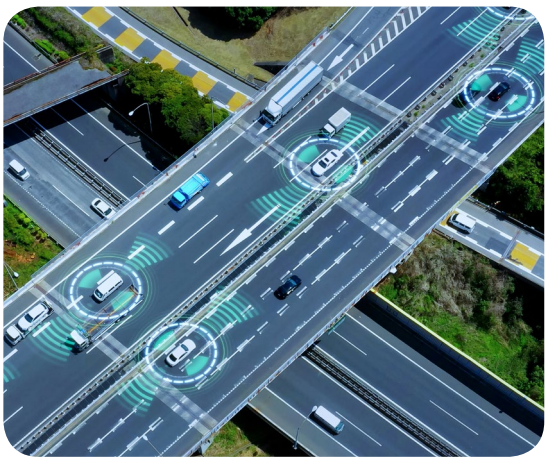




Smarter AI

AI-enabled smart transportation

By 2030, the number of electric vehicles, vans, heavy trucks, and buses on the road worldwide is expected to reach 145 million. Today, all these means of transportation run up against the limited capacity of our road networks. Intelligent transportation is the key to solving this problem.



There will be a wide range of intelligent transportation use cases that use cameras, radars, and weather sensors to collect various types of data. At the edge, data will be read to identify vehicles, traffic accidents, road conditions, and more, and to generate a multidimensional representation of a stretch of road. In the cloud, a digital twin of roads across the city will be produced, constituting a multidimensional representation of real-time and historical road conditions. Policy-based computing on the cloud can help generate different commands for every vehicle and every road, and manage vehicles and traffic signals.

The sheer volume of data involved means that the bottleneck to be addressed is not the capacity of our roads, but the capacity of our computing networks. Suppose a vehicle runs two hours a day on average. For each running vehicle, the compressed data uploaded per second will increase from 10 KB today to 1 MB in 2030, meaning that for every 100,000 intelligent connected vehicles, about 720 TB of data will need to be transmitted every day. The data generated by each running vehicle will need to

be frequently exchanged between the vehicle itself and the city.

With the help of intelligent transportation infrastructure that can store and analyze such massive amounts of data, urbanites can look forward to quicker daily commutes (15–30 minutes shorter, on average), less frequent traffic accidents, and vehicles with lower carbon footprints. Increased computing power will boost transportation safety, efficiency, and user experience, facilitating socioeconomic development.

AI-enabled autonomous vehicles

L4 autonomous vehicles will be commercially available on a large scale, and data will be continuously sent to the digital twin. AI learning and training will continue in the digital world, so that AI models will become smarter and eventually outperform humans in coping with complex road conditions and extreme weather. In time, AI will even make L5 autonomous vehicles a reality. The computing power required for intelligent driving will far outstrip what Moore's Law can provide. The corner case library will continue to expand and the demand for computing power will increase. By 2030, an L4 or higher-level autonomous vehicle will require computing power of 5,000 TOPS.

The training of AI models will involve introducing unsupervised or weakly supervised learning into closed-loop data, and using images and visual information obtained from vehicle snapshots to support automatic, unsupervised, video-level AI machine learning and training. Autonomous vehicles demand device-cloud computing. In the future, a vehicle manufacturer will need at least 10 EFLOPS of computing power on the cloud.

Smart cities

Urban areas make up 2% of the world's land surface, and are home to more than 50% of the world's population. Cities consume two thirds of the world's energy and are responsible for 70% of global greenhouse gas emissions (including over 25 billion tons of carbon dioxide). Smart city governance will be the way forward for cities that want to achieve sustainable development. IoT sensors will collect the environmental data that is needed to support the operations of smart cities. In the future, every physical object will have a digital twin. Digital cities made up of digital buildings, digital water pipes, and other infrastructure will be a powerful tool for intelligent urban management. Smart city governance will aggregate 100x more data than conventional city governance and make our cities more efficient.

The data storage and analysis capabilities of smart energy infrastructure will make it possible to manage urban energy supply and demand in one system, and to schedule urban energy more efficiently through real-time data processing. For example, a real-time energy efficiency map can be drawn based on urban energy consumption data. This will help dynamically monitor energy usage and ensure targeted energy scheduling, which will cut average electricity consumption in peak hours by more than 15%.

The quality of public services like meteorology, oceanography, and earthquake prediction can deeply affect the life of each resident in a city, and these services rely on massive data computing and processing. With a greater

volume and diversity of urban and natural environment data, smart public services will help better predict the impact of weather, oceans, and earthquakes on urban life, making cities more resilient to extreme events. With these smart public services, residents can gauge the impact of climate or emergency events on themselves and their communities using the push messages tailored to their geographic locations.

Data will be at the core of efficient operations of smart cities. How can we effectively manage and use the massive data generated? This is a question we must answer if we want to promote the development of smart cities.

More inclusive AI

AI-enabled precision medicine

In the healthcare sector, AI is already able to automatically identify tiny lung nodules, saving doctors a lot of time compared to conventional identification with the naked eye and manual tagging. AI will play a bigger role in more complex consultations. It will be deeply integrated into the diagnosis process, providing explainable diagnoses and predicting outcomes. The future will bring a new model of healthcare in which AI will provide solutions, and the role of doctors will be to check and approve them. The World Health Organization estimates there will be a shortage of 18 million healthcare professionals by 2030, and AI offers a viable solution to this problem.

AI-enabled drug screening

AI will become more transparent. It will not make judgments inside a black box. Instead, it will show the reasoning behind its conclusions so that we can understand its thinking process. Greater transparency will allow AI to play a greater role in more domains and help us perform more complex tasks, such as screening antiviral drugs. AI will be able to tell us why the drugs are selected, instead of just giving us a list of drugs selected. Results on their own, without the decision-making processes, cannot help us make informed decisions.

AI-enabled personalized education

The process of AI training is also a process of better understanding ourselves. AI makes it more important to understand human intelligence and how the human brain works. This will in turn push humans to rethink and reform education[3]. AI of the future will change our learning and cognition processes. For example, AI instructors will analyze students' behavior, habits, and abilities in detail and then develop personalized teaching content and plans. This will help students acquire knowledge more easily and realize their full

potential.

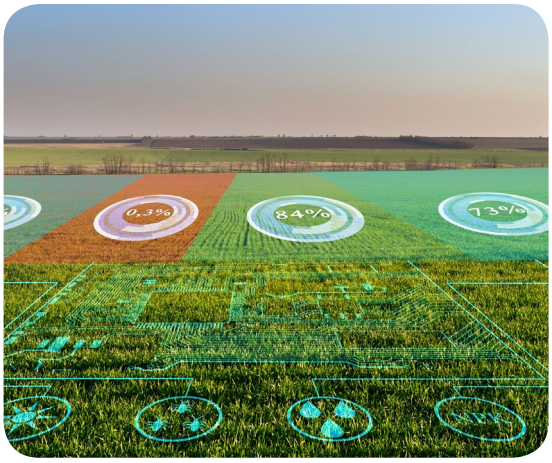
AI will be integrated into every aspect of our lives. It will allow us to analyze, create, and study more efficiently, and open up high-quality resources to many more people. AI will make services like precision healthcare, creative design, cultural education, elderly care, community services, and autonomous driving more inclusive.

Deeper perception

By 2030, there will be 200 billion connections. Hundreds of trillions of sensors will be collecting information about the physical world, including temperature, pressure, speed, brightness, humidity, and chemical concentration. Turning this basic data into sensory information to give robots vision, hearing, taste, smell, and a sense of touch will require deeper perceptual capacities. Issues of data quantity and latency mean that the process of computing for generating sensory information must be completed at the edge. The edge will therefore need to be able to intelligently process data, which would include simulating how the human brain processes information. In the future, a large amount of perceptual computing will be completed at the edge, where about 80% of data will be handled.

Perceptual intelligence makes the gathering and analysis of vast flows of data possible. It will enable more industries to perceive their work, and to build digital twins in the cloud. Digital twins remain in constant balance with their physical models, and support digital innovation.

Intelligent agriculture



In the future, an intelligent space-air-ground integrated network will be built and continuously optimized for remote sensing and monitoring of agricultural information. Advanced information technologies, such as the Internet, the Internet of Things (IoT), big data, cloud computing, and AI, will be deeply integrated with agriculture. This will create a brand-new model of agricultural production that features agricultural information sensing, quantitative decision making, intelligent control, targeted investment, and personalized services. Applications like smart fields, smart greenhouses, smart farming, smart planting, and spraying drones will have increased demand for edge AI computing. Intelligent agricultural sensing and control systems, intelligent agricultural machinery, and autonomous field operation systems will be deployed. These will promote the development of e-commerce, food source tracing and anti-counterfeiting, tourism, and digitalization in the agriculture sector. Agriculture will become more digital, connected, and intelligent.

Intelligent control of equipment

AI will be increasingly adopted in enterprises' production systems. It will support every aspect of company operations, improving workflows, staffing, and coordination across different departments and different sites. Over the next decade, AI will bring massive improvements in quality and cost savings in critical production processes. With the support of AI, manufacturers can achieve intelligent operations and management, massive data analysis and mining, and lower-latency diagnosis and warning.

The Made in China 2025 plan has a target of universal adoption of AI in key manufacturing sectors, with 50% reductions in operating costs, production time, and defects in showcase projects. In deep learning use cases, such as bearing fault diagnosis, steel furnace thermal anomaly detection, and power device overhaul, factories can use AI to diagnose problems and send warnings faster, detect production problems more efficiently, and shorten order delivery cycles.

Production robots

Workers who once operated machines in harsh environments will be able to operate them remotely. More non-operational tasks of enterprises will involve AI. Humans and machines will seamlessly collaborate with each other. AI will reshape enterprises' business operations at every level, from product design, production, and sales to enterprise architecture, employee hiring, and training. For example, enterprises will use AI to analyze factors such as economic development and current events and assess their own growth and trends across the industry. They will then optimize their production plans and create new solutions as input for decisions on new product concepts. AI will play an especially important role in flexible manufacturing that meets personalized needs. It can design customized products and even generate new product designs based on demand changes and product usage. We project that by 2030, there will be 390 robots per 10,000 workers. These robots will be able to accurately understand people's instructions, sense the environment, and provide recommendations.

Lights-out factories, with no human workers at all, are already in widespread use. AI robots are busy on production lines and in logistics, freeing humans from repetitive, boring tasks. In the future, machines will help humans handle dangerous jobs in harsh environments, even in highly variable scenarios. People will no longer need to operate machines onsite. Instead, they will be able to command the machines remotely from the safety and comfort of a control room.

In the mining industry, for example, China has set the goal of achieving intelligent decision-making and automatic collaborative operations by 2025 in large coal mines and mines where severe disasters have occurred in the past. Key roles down in the mines will be assumed by robots, and few, if any, actual workers will have to work underground. The longer-term goal is to build an intelligent coal mine system featuring intelligent sensing, intelligent decision-making, and automatic execution by 2035.[4]

AI will make enterprises more intelligent as it will play a greater role in creative work, rather than just operational work. AI will be more deeply involved in our thinking process and will better interact with people while showing the reasoning behind its conclusions. It will become more reliable and take on a bigger role in complex fields that require high-quality decisions, such as finance, healthcare, and law. In the next decade, AI will continue learning about the physical world and will become smarter. AI will move beyond well-understood scenarios and play a bigger role in empowering humans to do better in more complex tasks. AI will help people transcend human limitations.

An experience beyond reality

Intelligent interaction in living spaces

The AI of today has already helped people complete tasks that were impossible in the past. For example, we can use the cameras on our phones to identify plants and obtain information about their habits and how to grow them. Robots are helping humans perform better. For example, exoskeleton robots can help patients recovering from accidents. Home robots can perform intelligent work like keeping the elderly company and doing household chores. It is estimated that more than 18% of homes will use intelligent robots by 2030.

When AI participates in human thinking and creation, it must be able to explain its thought processes in terms that people can understand. This means that AI needs to be able to use natural language to articulate the logic behind its recommendations. AI will make a leap from perception to cognition, and from weak AI to strong AI.

AI has already made initial attempts at poetry writing and painting. The AI of the future will be able to perform more complex creative work, like film making, art, and industrial design. AI will provide highly customized content services, so that people can get a tailor-made painting or movie at any time. When watching a movie, the audience will be able to decide how the story goes. Based on audience choices, AI will analyze potential storylines and develop the video in response. Each viewer will experience the movie differently, making the content richer. It will also be possible for people to supply a theme and let a creative AI fill in the blanks. This will inspire our creativity and add another layer of richness to our lives.

AR/VR in living spaces

Data will create many digital spaces, such as virtual tourist attractions, holographic conferences, and virtual exhibitions. These digital spaces, together with the physical world, will form a hybrid world. Virtual tours can give us a true-to-life experience of scenery on the other side of the world. They will also allow us to sit side by side and talk with friends thousands of miles away, or have wide-ranging conversations with luminaries of the ancient world. The way people communicate with other people, communities, nature, and machines will be revolutionized, and our ways of living, work, and study will be redefined. It is estimated that by 2030, more than 30% of businesses will operate and innovate digitally, and there will be one billion augmented reality (AR) and virtual reality (VR) users.

Virtual world / Metaverse in living spaces

The seamless convergence of the digital and physical worlds requires the ability to accurately perceive and recreate the physical world, and

the capacity to understand user intentions in the hybrid world. The demand for a hyper-real experience means that computing will be brought closer to the edge. Multi-dimensional collaborative computing is required between cloud and device, device to device, and between the virtual and physical words. The physical world will be modeled and mirrored on the cloud, and following a process of computing and the addition of virtual elements, will be recreated digitally. Edge devices will be able to hear, see, touch, smell, and taste, and real-time interaction between people and devices will be possible. Multi-dimensional collaborative computing will change a user's environment into a supercomputer that is able to compute environment information, identify user intentions, and display a virtual world using technologies such as holography, AR/VR, digital smell, and digital touch.

More precise exploration into the unknown

The "high-performance computing (HPC) + physical models" approach has been widely applied in many scientific domains. As humans continue to study quantum mechanics, life sciences, the Earth's atmosphere, and the origins of the universe, our cognitive boundaries will continue expanding to embrace phenomena at both the subatomic and cosmological scales, in which the distances can be as short as 10^{-21} m, or as vast as 10^{28} m. The amount of data and computing that scientists have to process will grow exponentially. The amount of computing power available in the digital world determines how deep and how broad we can explore in the physical world.

CERN, the European Organization for Nuclear Research[5], built a computing pool by connecting supercomputers located worldwide. Scientists used this pool to analyze nearly 100 petabytes of data generated by its Large Hadron Collider (LHC), and ultimately proved the existence of the Higgs boson in 2012. The CERN plans to use the High-Luminosity LHC (HL-LHC), a major upgrade of the LHC, by the end of 2027, which will be able to produce more than 1 billion proton-proton collisions per second. The amount of data to be computed will be 50–100 times greater than that used to prove the existence of the Higgs boson, and zettabytes of data will need to be stored. By 2030, computing will help scientists solve basic problems in more domains.

Environmental monitoring

Environmental protection is a top priority for humanity. New technology and equipment will be powered by AI to ease environmental problems such as the greenhouse effect, soil desertification and salinization. Models built based on big data will help predict the results of different management measures, which can be fed back to algorithmic models to come up with better governance models, like accurately locating pollution sources and predicting pollution diffusion.

Weather forecasting

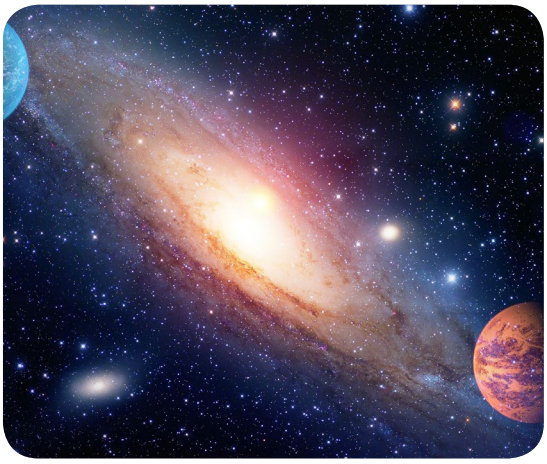
Future weather forecasts will use more complex dynamic numerical models to predict the weather more accurately. Potential applications include weather radar quality control, satellite data inversion and assimilation, as well as weather and climate analyses (e.g. short-range and imminent weather forecasts, probability forecasts, typhoon forecasts, extreme or catastrophic weather warnings, storm environment feature classification, and environmental forecasts). Take short-range local weather forecasts as an example. Torrential rainfall in a short period of time is an extremely destructive phenomenon, but it is difficult to forecast when it will happen, because it requires massive amounts of data and huge computing power. If we were to increase the granularity of weather forecasts from the current 10 x 10 km to 1 x 1 km, that would increase the amount of data and computing power needed by two or three orders of magnitude. It is expected that by 2030, with the emergence of supercomputers that can perform 100 EFLOPS, more accurate climate simulations and weather

forecasts will be possible.

Seismic and ocean prediction

In the future, AI will be used to monitor earthquakes and estimate the focus of earthquakes in real time, which will make prediction much more accurate. It is very time-consuming to calculate the focal mechanism (also called a fault-plane solution) based on seismic records. Ever since seismologists began calculating fault plane solutions in 1938, focal mechanism parameters have been a huge challenge. AI can effectively solve this complex computing problem. Seismic data can be used to train AI neural networks, which can make prediction systems more accurate and reliable. This will further drive breakthroughs in earthquake prediction.

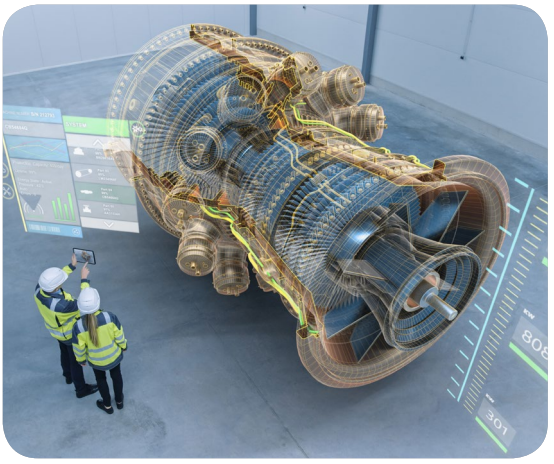
Exploring the structure of the universe



The large-scale structure of the universe is one of the most important current fields of science. Scientists are studying the formation and evolution of cosmic structures over time, to find answers to questions about the composition of the universe, the process of cosmic evolution, dark matter, and dark energy. The conventional method is to use a supercomputer to calculate the evolution of various large-scale structures in the universe based on our current physical theories, and then compare the results with observed data. This, however, requires accurate calculations for hundreds of thousands—or even millions—of cosmological objects. As of today, there are two trillion galaxies and countless planets in our observable universe. Even if we were to pool all of the world's computing resources together, it would still be impossible to complete the calculations.

More accurate simulation of the real world

More precise wind tunnel simulation



Computer wind tunnel simulation is now an important test method for high-speed vehicles such as aircraft, high-speed trains, and automobiles. However, due to the huge amount of computing required for these simulations, the testing system needs to be broken down into sub-systems like tire and engine, and then further divided into even smaller systems to get precise simulation results. This will pose new challenges in verifying whether system design meets requirements. As computing power will increase by 2 to 3 orders of magnitude in the future, wind tunnel simulation is expected to be used in larger sub-systems, or even for the entire system.

AI-enabled research on new drugs

When it awarded the 2013 Nobel Prize in Chemistry to three scientists "for the development of multiscale models for complex chemical systems", the Nobel Committee stated that, "Today the computer is just as important a tool for chemists as the test tube. Computer models mirroring real life have become crucial for most advances made in chemistry today."

Quantum mechanics/molecular mechanics (QM/MM)[6] modeling is one of the most reliable methods for simulating the catalytic mechanisms of enzymes. The high-precision QM model is used in core regions of the enzyme, and the low-precision MM model is used in peripheral regions. This approach combines the accuracy of QM and the fast speed of MM. To use this model to simulate the growth and reproduction of 0.2-micron Mycoplasma genitalium cells over a period of two hours would take the supercomputer Summit[7] one billion years. For more complex studies of thinking, memory, and behavior in the human brain, vastly more computing power would be needed. To predict the response of the human brain to a particular stimulus, it would take Summit 10^{24} years to simulate one hour of brain activity[8].

Turing Award winner Jim Gray divided scientific research into four paradigms: experimental, theoretical, computational simulations, and data-intensive scientific discovery. As we continue with research in dynamically complex fields such as biology, material science, chemistry, and astronomy, it will be increasingly difficult to make progress relying on traditional computation methods. The curse of dimensionality may occur as the number of variables and degrees of freedom increase, and this means that the demand for computing power will increase exponentially.

AI will provide a new solution to the curse of dimensionality and a new path for scientific research. Using conventional methods, it would take scientists several years to analyze the folding structure of a single protein, but with the help of AI, scientists are able to learn the 18,000 known protein structures and produce simulations with atomic levels of precision for unknown protein structures within just a few days. This kind of research is giving us new ways to discover therapies that could prevent and treat cancer, dementia, and other diseases caused by changes in the structure of proteins in cells. The winners of the 2020 Association for Computing Machinery (ACM) Gordon Bell Award[9] simulated a system of more than 100 million atoms using AI. The system was more than 100 times larger than current models and the time-to-solution was 1,000 times faster. This project has brought accurate physical modeling to larger-size material simulation[10].

The scientific computing of the future will rely on a combination of data, computing, and intelligence, which will give rise to new paradigms for scientific research. AI will study existing knowledge, analyze, and draw new conclusions. Online iteration, combined with traditional modeling methods, will speed up scientific exploration and further expand people's cognitive boundaries.

Data-driven business innovation

Computing-enabled data value mining

Cloud computing and big data are now the foundation for digitalization in any industry. They are driving the digitization processes that are making many industries more efficient. A key feature of digitization is that it improves the matching of producers to consumers. Examples include e-commerce platforms and online-to-offline (O2O) models.

10-fold increase in the demand for new services

Full-stack, serverless device-edge-cloud computing will become a key technology for enterprises to modernize and go digital and intelligent. Programming languages, language runtime, as well as application scheduling, operations, and O&M based on the cloud-native computing model will be the foundation for building modern full-stack serverless software. This will allow all applications to be migrated to the cloud, and will result in tenfold gains in performance, efficiency, and cost reduction.

More efficient operations

More efficient resource utilization

The wide adoption of cloud allows companies to use computing resources more easily and quickly. New computing technologies will give companies access to these resources in smaller packages, available more quickly. This will reduce waste in the way companies use these resources. For example, before the cloud, central processing units (CPUs) were used only 10% of the time. Containerization raised this indicator up to 40% or higher. In the future, the wide adoption of new resource management technologies will reduce waste by 50% or more.

Software-defined operations

IT is now one of the core components of any operational system. Internet companies use a DevOps[11] model and are becoming more agile and efficient. By 2030, companies in the manufacturing sector will achieve highly efficient software-defined operations in their more complex value chains.

The industrial Internet will connect the supply chain, manufacturing, maintenance, delivery, and customer service processes. All companies will form a value network that spans the globe. The digital transformation inside a company will expand into an improvement of entire industries, which will translate into greater synergies. And the dependence on data will change: from a company being highly dependent on its own data to being dependent on data from up and down the value chain, or even from other industries.

Companies of the future will use software to manage complex cross-organizational coordination and to define their own operations. For example, they can use technologies like robotic process automation, no-code/low-code development, and AI-supported natural language programming to invoke the capabilities of robotic automation software, obtain required services, and orchestrate business processes. This will mean that even personnel without much expertise can improve processes and fix problems on their own.

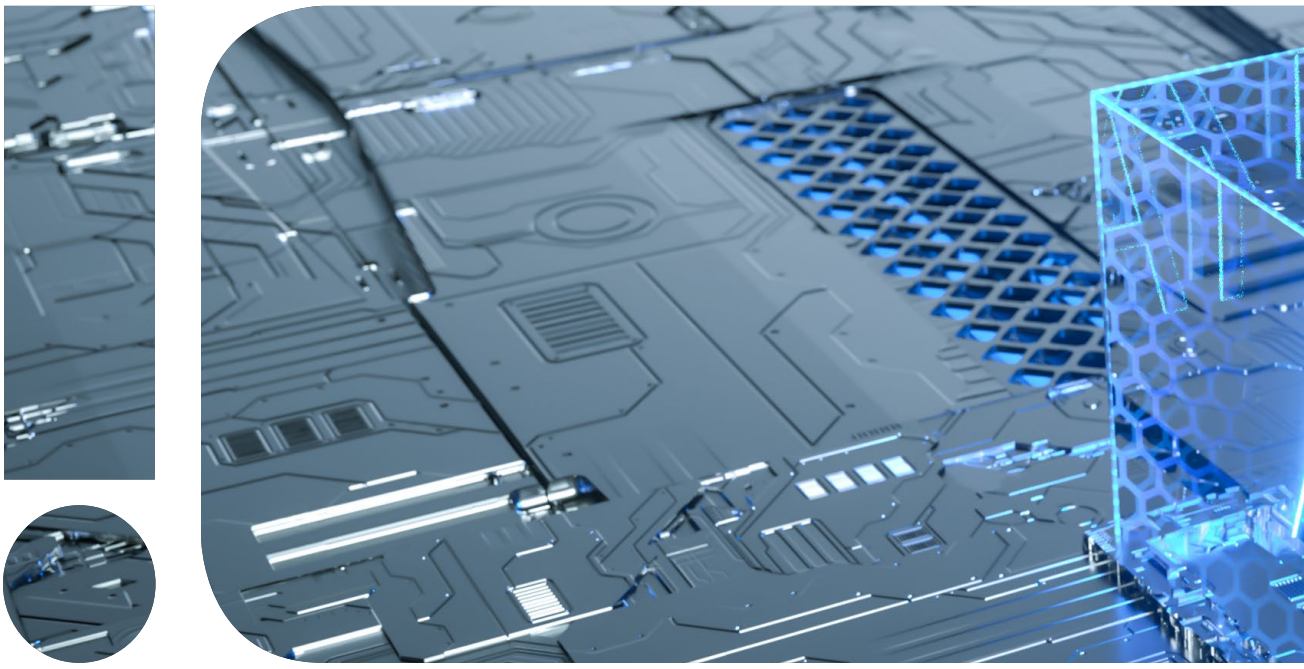
Low-carbon data centers

By 2030, data centers (DCs) will deliver a 100-fold increase in computing power while achieving low-carbon operations, giving companies access to green computing resources.

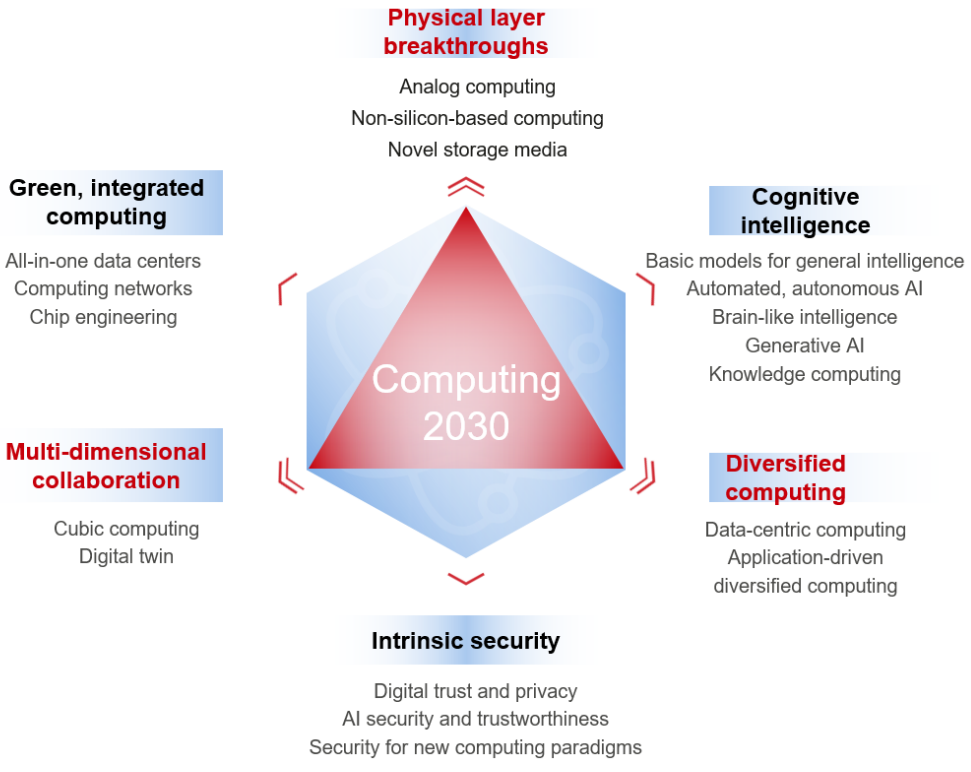
New computing architectures will massively boost energy efficiency. In a conventional computing

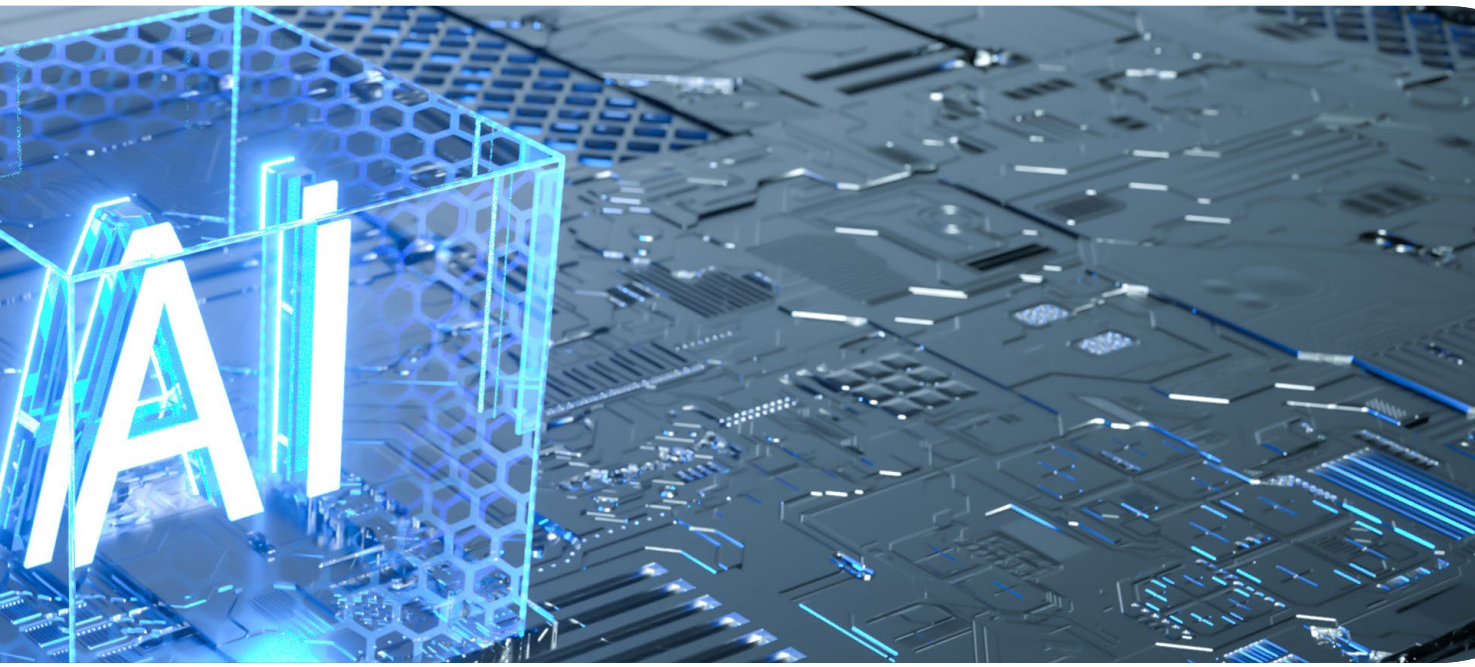
process, more than 60% of the energy is used shuffling data. The data centers of the future will make computing tens of times more energy efficient. Analog computing such as quantum computing and analog optical computing will be important sources of computing power, driving energy efficiency indicators up exponentially.

In the push toward carbon neutrality, data centers will be positioned near energy resources and near areas with high computing demand. This will change the computing architecture on a larger geographic scale. Computing networks can balance the needs of green energy, latency, and costs and achieve optimal global power usage effectiveness (PUE) and cut carbon emissions. Tasks like AI model training or gene sequencing can be done in places with abundant green energy sources and low temperature while tasks like industrial control and VR/AR can be performed in places that are closer to customers' production environments.



Vision and key features of Computing 2030





Cognitive intelligence

AI is evolving from perceptual intelligence to cognitive intelligence. Cognitive intelligence is an advanced stage of AI evolution, at which machines are given the capabilities of data understanding, knowledge representation, logical reasoning, and autonomous learning. It will make machines powerful tools for humans to become more capable and change the world. In the evolution toward cognitive intelligence, semantic and knowledge representation and logical reasoning are important means of cognition, and multimodal learning is an important way to realize information fusion and collaboration. By building large-scale multimodal basic models, AI systems can learn converged representation of multiple types of information to establish multimodal transfer and concordance. This improves an AI system's ability to perceive and understand complex environments, thereby enabling AI applications to work in different environments and on a wide range of different tasks.

Basic models for general intelligence

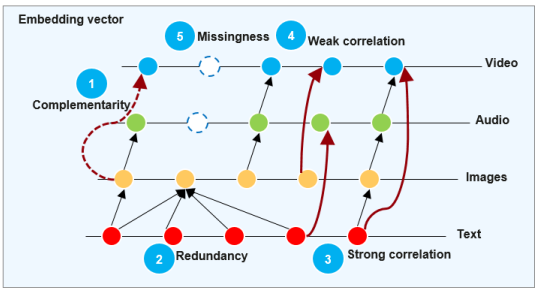
AI's ongoing evolution from perceptual intelligence to cognitive intelligence: AI has delivered computational intelligence and perceptual intelligence; it is now on the way to developing cognitive intelligence. Machines have strengths in computing speed and storage. Today, deep learning and big data analytics are enabling machines to perform certain tasks through vision, hearing, and touch, similar to how a human being would. Cognitive intelligence will allow machines to understand and reason like humans. When machines have these abilities, they will become powerful tools that help humans to understand and change the world.

Improving the ability of machines to generalize in the process of solving problems is an important evolutionary path from weak AI to strong AI. AI systems will be given the ability to

solve multiple different problems using large-scale, domain-general basic models that can generalize from one situation to another; from one modality to another; and from one task to another.

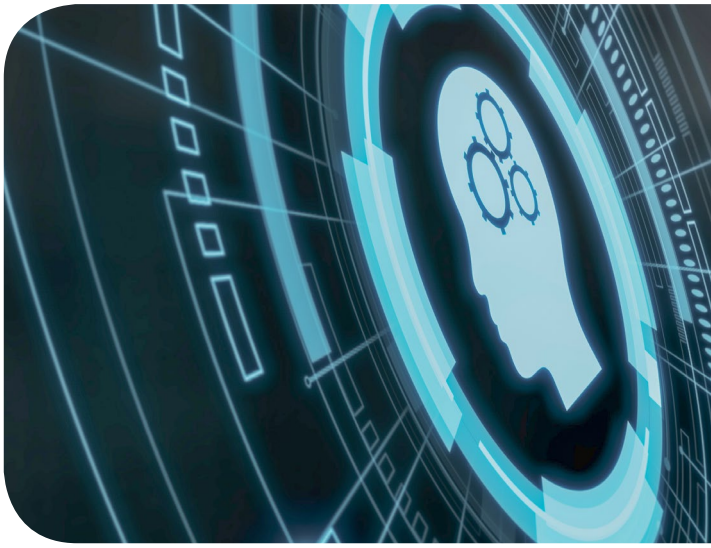
Multimodal learning is an important approach to building basic models: In multimodal learning, data heterogeneity is the first problem that needs to be solved, which creates a number of challenges: (1) How can complementarity and redundancy in multimodal data be used for representation learning? (2) How can the strong and weak correlations between these representations be processed to produce relational vector maps between modalities? (3) During adaptive learning and multimodal transfer for model training, how can we keep model accuracy within an acceptable range when one piece or a type of data is missing in a certain modality? (4) During inference, when one piece or a type of data is missing in a certain modality, how should model topology and routing adapt for maximum inference gains?

Based on progress to date, we expect that multimodal models will become capable of multimodal, self-supervised learning and the transfer of generally-applicable knowledge. This means that tasks in different domains can be approached using the same multimodal framework.



- **Efficient representation learning** 1 2
- **Precise mapping** 3 4
- **Self-adaptive learning (training)** 5
- **Model topology and routing (inference)** 5

Breakthroughs in multimodal learning will require advances in the following key areas:



First, the technology to tag training data to associate captions, audio, video frames, etc. Second, multi-stream codecs from single-modal pre-training models to multimodal association coding, which enables multimodal learning with weak information association, with the decoder providing support for cross-modal transfer and generation. Third, self-supervised learning technology, involving semantic alignment and inter-modal predictions between text, speech, vision and other modalities. Fourth, technologies for downstream task fine-tuning that support multimodal semantic understanding and multimodal generation. Fifth, multimodal models that are smaller.

Automated, autonomous AI

Deep learning has not yet successfully developed beyond the stage of supervised learning. Data cleansing and tagging, and the design, development, training, and deployment of models, all require extensive manpower. Development in domains such as transfer learning, few-shot/zero-shot learning, self-supervised/weakly-supervised/semi-supervised/unsupervised learning, and active learning, will eventually drive AI to reach autonomy, eliminating our dependence on manual training, design, and iteration of models. AI autonomy will make models more homogenized, with the same model serving multiple purposes. The

amount of data learned will increase without manual intervention. Models will learn to pick up and train on new data as they operate, improving their capabilities in the process. The scaling of data and the prevalence of online learning will lead to more centralized model production. Industry applications in multiple domains will converge to a handful of or even a single ultra-large model.

However, there are still some major challenges that the developers of autonomous AI must overcome:

1) Training signals can be incorporated online in a self-supervised fashion, so that feedback is available during inference, not just during the training phase.

2) At present, a model's learned representations are formed without constraints. The representations that result from different training sessions may be radically different even if they are of the same model structure. Models need to overcome the problem of catastrophic forgetting, so that learning can be carried out continuously, and training and inference can converge into a single process.

3) Models manually designed for different tasks need to be replaced by models that can learn to encode for different tasks and switch between different modalities in context and on demand.

Brain-like intelligence

Current deep learning technology is largely data-driven and relies heavily on large quantities of labeled data and powerful computing. Backpropagation training algorithms need continuous enhancement in terms of model robustness, capacity to generalize, and interpretability. Drawing on and imitating the way biological neurons work, brain-like intelligence creates digital neurons with richer functionality and promises to enable learning that is event-triggered, uses pulse encoding, and is coordinated both temporally and spatially. Using neurodynamic principles, brain-like computing can deliver both short-term plasticity and long-term memory, and is capable of adaptive adjustment and learning in open environments. Inspired by the sparse connectivity and recursive form of the biological brain, no computation will be performed without pulses, which greatly reduces energy consumption. In the next five to ten years, if breakthroughs in related technologies are made, brain-like computing is likely to begin to outperform other models, as well as consuming less power, in many computing tasks, and be applied in smart devices, wearables, and autonomous vehicles.

At present, brain-like systems are still inferior to the deep learning systems in terms of learning efficiency and computing accuracy, because our understanding of the human brain's learning mechanisms is too shallow. Research in this field will need to advance in two major areas. First, from the bottom up, the systems can simulate pulses in the biological brain, and use neuromorphic chips to recreate neurons and synapses at scale, which should support low power and low latency in time-dependent applications. Second, from the top down, more comprehensive theories of neurodynamics and cognition are needed from a functional perspective, which can then be applied in combination with AI to achieve more robust and general intelligence.

Generative AI

Generative AI powers automated content production: It allows computers to abstract the underlying patterns related to a certain input (such as text, audio files, and images) and use it to generate expected content. Generative AI is used in identity protection, image restoration, audio synthesis, and antimicrobial peptide (AMP) drug research, among other fields.

Generative AI generates data that is similar to training data, rather than simply replicating it, so it can incorporate human creativity into processes of design and creation. For example, a game generation engine can generate 3D games to test the vision, control, route planning, and overall gaming capabilities of an intelligent agent, in order to accelerate the training of the agent. In the development of generative AI applications, the key objective is generation models that are capable of evolving and dynamically improving over time.

The field of generative AI is facing the following challenges:

1) Some generative models (such as generative adversarial networks, or GANs) are unstable, and it is difficult to control their behavior. For example, generated images may not be sufficiently accurate; they may not produce the desired output; and the cause cannot be located.

2) Current generative AI algorithms still require a large amount of training data and cannot create new things. To address this, algorithms capable of self-updating and evolving are needed.

3) Malicious actors can use generative AI for spoofing identities and can exploit vulnerabilities in AI tools to conduct remote attacks, resulting in serious threats to online information security such as data breaches, model tampering, and spam.

Knowledge computing

The industrial application of AI needs the ability to make high-quality decisions based on expert domain knowledge across multiple disciplines. A complete technical system is needed for knowledge extraction, modeling, management, and application. In the next decade, knowledge computing will make a leap forward: In knowledge extraction, the data source will not only include text and structured features, but also complex and multi-level knowledge, which includes several areas of research such as multi-modal knowledge alignment, extraction and fusion, complex-task knowledge extraction, and cross-domain knowledge extraction.

Knowledge modeling will move from developing scenario-specific, atomized, automated, and large-scale knowledge graphs to integrating these scenario-specific graphs into general knowledge graphs. The applications of knowledge will develop, from simple query and predictions to high-order cognitive tasks such as causal reasoning, long-distance reasoning, and knowledge transfer.

The application of knowledge computing will require breakthroughs in algorithms for massive retrieval of sparse information, capture of dynamic-length knowledge, knowledge attention, and large-scale graph computing. The training schema for cognitive intelligence will

require advances in high-frequency knowledge retrieval during training and inference and feature enhancement based on knowledge combination. In terms of computing, it will be necessary to solve a number of problems such as training and inference for high-frequency random retrieval, high-speed data communication, and some graph computing puzzles such as random walk and structural sampling.

Intrinsic security

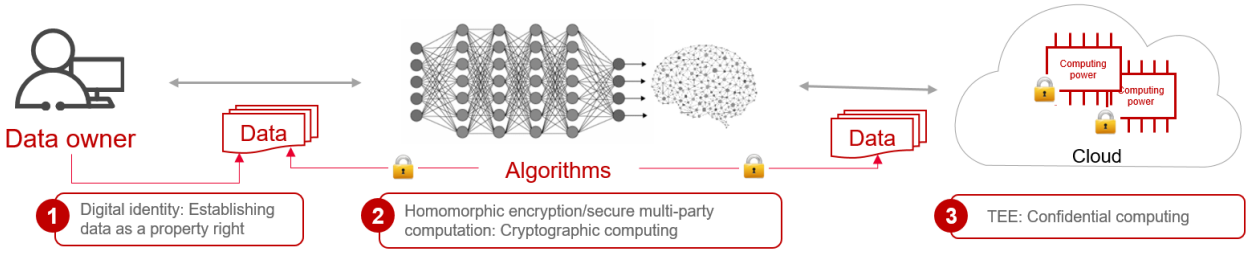
The migration of computing resources to the cloud has gone beyond traditional security boundaries. Traditional add-on security based on the division of trust and untrust zones cannot withstand new types of attack. In order to protect users in an evolving threat landscape, security must become intrinsic. Specifically, that means:

- Security must be an intrinsic capability of a system and a basic feature of chips, firmware, and software.
- Security should be ensured throughout the entire data processing lifecycle (including storage, computing, and transmission), to defend against all kinds of attack.
- A hardware-based root of trust is essential. Due to the system access control model, security functions must be implemented based on the highest hardware privilege in order to provide reliable security services on the operating system and applications. In addition, hardware acceleration can improve the performance of security services.
- The principle of open design should be adopted, which means the security of a mechanism should not depend on the secrecy of its design or implementation. Security services should be made open source. This way, service software can embed security into itself based on its own architecture pattern to ensure service security.

Digital trust and privacy

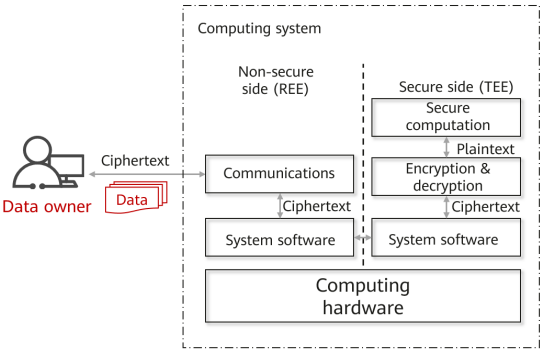
Data processing, in essence, is the process of computing data using algorithms. If all the three elements – computing power, data, and algorithms – are controlled by the data owner, data security and privacy are not really an issue. However, during cloud computing, these elements are often separate. Algorithms and computing power are provided by computing service providers, while users (i.e., data owners) need to upload data to the cloud for processing. Even if users trust the computing service providers, they don't trust the computing service provider administrators who have access privileges. Therefore, the major security challenge of cloud computing lies in protecting user data and privacy. To address this challenge, digital trust systems need to be rebuilt.

Governments worldwide have enacted data protection laws, providing a legal basis for rebuilding digital trust systems. Digital identity and privacy computing are key technologies in this rebuilding process. Digital identity is the



basis for establishing data as a property right, while privacy computing can be used for data analysis and processing without compromising data.

1) Hardware isolation technology that is based on trusted execution environments (TEEs) can be used to process sensitive data. However, the completeness of hardware security isolation mechanisms cannot be mathematically proven, so it may be hard for the mechanisms to prove their own innocence, and security vulnerabilities may exist. On the other hand, TEEs have a smaller impact on performance than cryptographic technology. In the future, privacy computing based on TEE technology will be widely adopted in public cloud, Internet, and major enterprise services. It's expected that TEE technology will be used in more than half of all computing scenarios by 2030.



2) Homomorphic encryption and secure multi-party computation are considered to be the most ideal privacy computing technologies because it is possible to verify their security level mathematically. However, both of these technologies come with a significant performance cost (their processing is over 10,000 times slower than conventional computing). Significant performance improvements must be made if these technologies are to be applied in real-world scenarios. Approximation algorithms are maturing, and homomorphic encryption and secure multi-party computation technologies have already been applied in face authentication, the sharing of health data, and other specific domains. In the future, further breakthroughs based on hardware will be made in these technologies, which are expected to be commercially used in scenarios that require high security, such as in finance, healthcare, and other security-conscious sectors.

3) Multi-party computation is built on the sharing of secret slices between multiple parties. Cryptographic methods like zero-knowledge proofs come with a high performance overhead. However, TEE technology can greatly improve the performance of multi-party computation, while being used to enable the sharing of secret slices between multiple parties. In addition to that, security can be proved mathematically based on TEEs. So this technology is expected to be used in various scenarios in the future.

AI security and trustworthiness

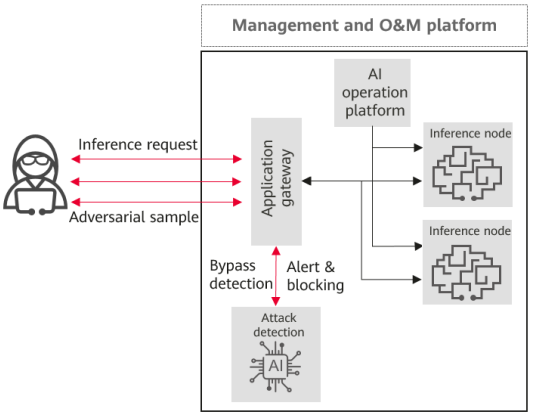
As AI applications become more popular, especially in fields like healthcare and autonomous vehicles, AI-related security challenges are increasing. AI models and training data are core assets of AI application providers. If not properly protected, they may be maliciously recovered and can be used to trace back to the data subjects. In addition, AI models are vulnerable themselves, resulting in more and more evasion and poisoning attacks on AI models. Attacks on AI models in key fields can have serious consequences. As concern about AI increases, there are challenges regarding AI ethics and forensics that will have to be overcome.

To address these challenges, all participants in the AI ecosystem must work together to ensure AI regulatory compliance and governance. They also need to adopt innovative technologies to trace the responsibilities of multiple participants, so as to support responsible AI.

1) Protection of AI models and training data: Encryption, mandatory access control, security isolation, and other mechanisms must be implemented to ensure security of AI models and training data throughout the data lifecycle, from

collection and training, to inference. The major challenge lies in encrypting the high-bandwidth memory data of neural network processing units (NPUs) in real time while ensuring no performance loss. In the future, breakthroughs need to be made in high-performance and low-latency memory encryption algorithms and architecture design for a hardware memory encryption engine on NPUs to provide full-lifecycle protection.

2) AI attack detection and defense: Adversarial sample detection models should be implemented to better identify physical and digital evasions and other attacks on AI models, block attack paths, and prevent misjudgment when AI models are attacked. The main challenge lies in continuous adversarial training against new types of attacks. In the future, independent security products and services to defend against AI attacks will emerge.

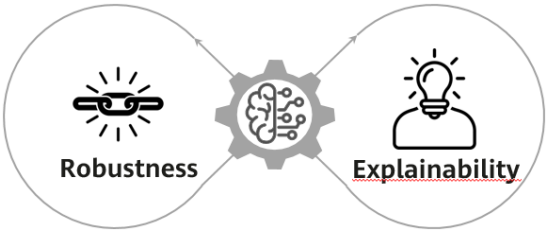


3) In addition to defense against known attacks, the security of an AI model itself must be enhanced to avoid the damage caused by unknown attacks. This can be achieved by enhancing model robustness, verifiability, and explainability.

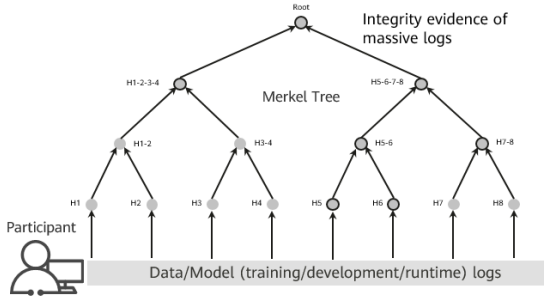
Adversarial training is one of the key technologies for improving the security of AI models. Regularization of models and generalization of adversarial samples are key technologies that need to be improved. Adversarial robustness is expected to increase from current low levels to 80%.

Effective formal verification methods will be available to prove the security of small AI models. However, the formal verification of large models still faces huge challenges.

The ability of AI models to justify their decisions will be vital to minimizing legal or logical risks. Moving forward, an explainable model can be built through explainable data before modeling. Currently, linear models are basically explainable, but there are still huge challenges to be overcome in making non-linear ones explainable. It's still hard to make AI models explainable globally, which means that making some layers of network models visible and explainable may remain the most technically feasible approach for a long time to come.



4) AI models should also be continuously monitored and audited to comply with AI regulations, and blockchain and other related technologies can be used to ensure reliable audit results and real-time tracking of issues.



Security for new computing paradigms

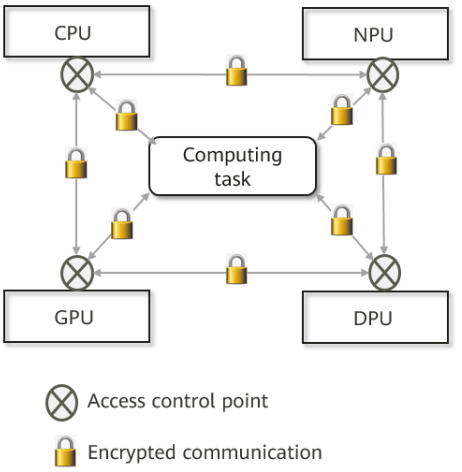
In data-centric computing scenarios, computing power extends beyond CPUs, and in particular

computing power is moved to memory, using the processing in memory (PIM) technology. This causes the failure of traditional memory encryption mechanisms, making it impossible to deploy hardware-based privacy computing technologies. Even if data is encrypted at the application layer, data is still processed as plaintext, which means privileged users and processes cannot be prevented from data breaches. The only solution for this scenario is to deploy cryptography-based privacy computing technologies (e.g., homomorphic and multi-party computation) to build users' trust in computing service providers.

In data center scenarios where diversified computing power is provided, the migration to the cloud is blurring the boundaries of security, leaving traditional security approaches that were based on security boundaries out of date. That's where the Zero Trust Architecture [12] model comes into play. This architecture addresses the security challenges of untrusted environments by enhancing access policies, proactive monitoring, and encryption. The Zero Trust Architecture model and diversified computing power together plot out the path of security technologies for diversified computing.

1) Security + in-network computing architecture: The Zero Trust Architecture model erases the old boundaries of security, so it employs a finer-grained access control mechanism to support dynamic authentication and resource access policies. That means software implementation consumes a large amount of CPU resources. However, an in-network computing architecture that uses hardware acceleration mechanisms for regular expressions can make policy execution 10–15 times more efficient.

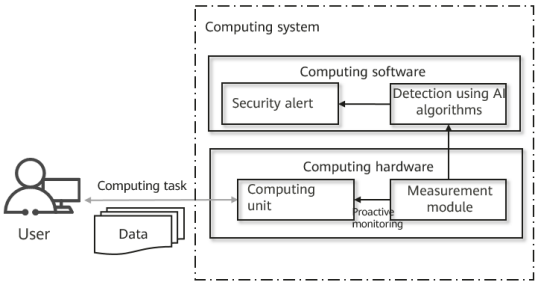
2) Security + diversified computing architecture: A Zero Trust architecture assumes that the network environment is untrusted. It requires encrypted communication throughout the network, including between compute nodes and data centers. Therefore, each xPU in a diversified computing architecture is required to implement the high-performance hardware encryption engine that supports post-quantum encryption algorithms to withstand potential quantum attacks.



3) Security + data-centric peer-to-peer computing architecture: In a data-centric peer-to-peer computing architecture, high-performance SCM will connect with the memory bus in the system. There are increasing risks of data and privacy leakage, as no mechanisms are in place to encrypt residual data in the memory after a power off. Ensuring data security in a data-centric peer-to-peer computing architecture will be a new challenge. For example, in a distributed cluster system where memory is shared across hundreds of compute nodes, it's challenging to protect data without greatly impacting bandwidth performance (keeping the impact close to a theoretical limit that is less than 3%).

4) DC-level dynamic measurement and proactive monitoring: Current computing platforms are generally unaware of the computing tasks running within systems. Even if the systems are attacked, the platforms cannot effectively distinguish malicious behaviors from normal computing tasks. In data centers, we are still facing many challenges in terms of detecting behavior of computing tasks in the system, so that they can measure system status proactively and monitor computing tasks, to detect and defend against potential malicious behaviors adaptively, thereby assuring computing power

security.



Green, integrated computing

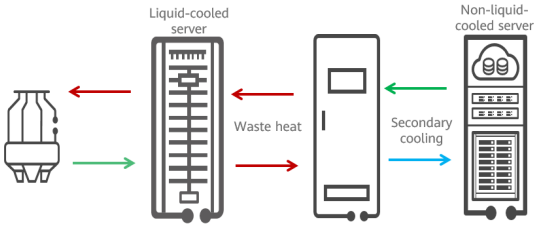
Data centers currently account for about 1% of global electricity consumption. The total energy consumption of general computing has been doubling every three years. The push toward global carbon neutrality will drive a 100-fold increase in computing power while increasing energy efficiency. Ongoing improvements in chip packaging and chip architectures are increasing computing power density and energy efficiency. Co-packaged optics can reduce losses in high-frequency data exchanges. All-in-one data centers will use AI to coordinate power supply, servers, and workloads to achieve an optimal PUE. The ultimate goal is to reduce the PUE to less than 1. Computing networks will connect distributed data centers that provide equivalent services while respecting differences in latency, cost, and green power use, achieving a globally optimal PUE and lowering carbon emissions.

All-in-one data centers

1) DC-level full-stack, converged architecture

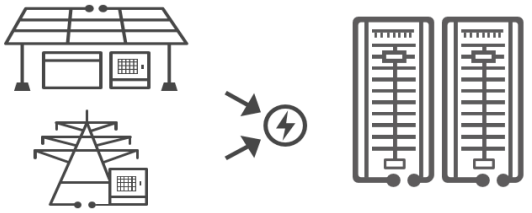
Rapid development of compute-intensive technologies such as AI, supercomputing, and cloud computing will enable large data centers to accommodate millions of servers. This will create challenges such as end-to-end heat dissipation, hardware configuration and resource utilization, and unified O&M for millions of central nodes and massive numbers of edge devices.

All-in-one data centers will consume megawatts of power, so we will need to continuously increase their energy efficiency in order to deploy them at scale. Air conditioner-free and chiller-free data centers are now common and liquid cooling technologies are seeing wide adoption. Reuse of waste heat from liquid cooling for heating, secondary cooling, and power generation has become a new growth opportunity in the industry. New technologies are being improved and put into commercial use. As a result, the PUE of some data centers is approaching 1.0, and some are expected to achieve PUE below 1.0 in the foreseeable future. As chip manufacturing and packaging technologies continue to advance, the heat flux density of chips for compute-intensive tasks such as AI and high-performance computing will exceed 150 W/cm^2, and may even go beyond 200 W/cm^2. Native liquid-cooled chips are emerging. With wider adoption of AI, we will see full-stack, automatic, coordinated optimization at the DC level, from power supply and cooling to chip working modes, based on service scheduling and workload features.



Power for data centers needs to be delivered on shorter and more efficient supply paths. New packaging technologies such as 2.5D, 3D, and wafer-level chip (WLC) will enable kiloampere-level chip power supply, which will require new processes, components, and topologies. Power fluctuation due to overclocking and heavy, dynamic loads will require us to rethink server power supply design. Liquid cooling is more complex than air cooling, meaning more difficulty during the construction of equipment rooms, server production, installation, and O&M. It also demands higher skills in data center

personnel. Core components such as cold plates and coolants need to be improved in terms of processes and reliability if they are to be deployed at massive scale.



The temperature rises inside 3D chip packages are higher than existing packages. 3D packaging is responsible for nearly 50% of the temperature rise along the heat dissipation path. Therefore, 3D packaging will present new heat dissipation challenges. The thermal resistance of thermal interface materials (TIMs) and cold plates will need to be reduced by 50%, and achieving this will require innovation in materials and processes. Large chip packages like WLC will also require advances in cold plate assembly, coplanarity, and reliability. One viable heat dissipation solution is integrating the chip packaging technology and the liquid cooling technology. With the TIM layer removed, the coolant comes in direct contact with the die inside the chip package. However, this will give rise to reliability issues such as long-term erosion and corrosion, and challenges related to heat dissipation on the surface of the die, jet uniformity, and package sealing.

Waste heat can be reused much more efficiently when water temperatures are high, but for efficient cooling and high chip performance, coolant water temperature must not be too high (not higher than 65°C). Low water temperature presents challenges for data center heat reuse systems. Waste heat reuse in secondary cooling is expected to be in large-scale use by 2025. However, the current efficiency of power generation from waste heat is less than 5%. Large-scale adoption will require breakthroughs in key technologies, such as new power generation materials with high ZT values. In addition, stable heat sources are required for waste heat reuse. The temperature of the liquid-cooled return water depends on chip workloads. Therefore, service scheduling, workload control, and coolant flow control will be needed to help provide stable heat sources for the waste heat reuse system.

Data center-level full-stack energy efficiency optimization will require open interfaces to monitor and control cooling towers, water pumps, coolant distribution units (CDUs), uninterruptible power supply (UPS), electricity meters, and servers. Developing the specifications of these interfaces will be another challenge.

Flexible hardware configuration: As service types and processor platforms become increasingly diversified, IT resources in cloud computing and 100 EFLOPS supercomputing data centers will see a dramatic rise in both scale and complexity. There will be a gradual evolution from the current server-based delivery model to a component-based one. As a result, resource utilization will increase from the current 30% to 70%. To support automated O&M and component-based supply, specifications must be developed for hardware form factors and software and hardware interfaces.

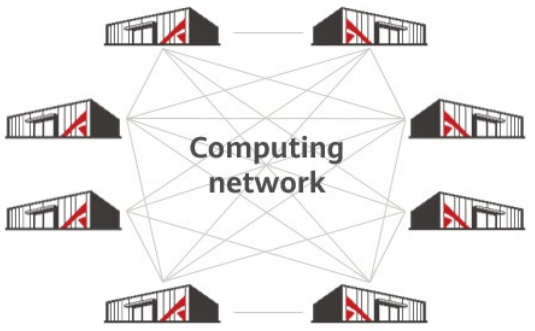
Automated, intelligent equipment O&M: With millions of servers deployed in data centers, automation can improve the efficiency and accuracy of construction and O&M by orders of magnitude. Large numbers of nodes are being deployed at the edge, and automating their integration will spare us corresponding increases in labor and operation costs that edge deployment would otherwise bring. Automation will also improve our ability to troubleshoot edge systems. AI and big data will help make better informed decisions; learning algorithms and dynamic adjustment of hardware and software configurations will increase IT resource efficiency and energy efficiency. Incidents like the COVID-19 pandemic will require data centers to support contactless delivery and O&M.

As Industry 4.0 and AI continue to develop, automation technologies are rapidly maturing. Intelligent and unmanned adaptive data centers (ADCs) will be deployed widely, making automatic and dynamic matching between data centers and service workloads a reality.

Computing networks

1) Cross-region distributed super data centers

The central idea of a computing network is to use new network technologies to connect computing center nodes distributed across different geographical locations. The purpose of such a network is to achieve real-time awareness of the status of computing resources, to coordinate the allocation and scheduling of computing tasks, and to transmit data, so that the system as a whole forms a comprehensive network that senses, allocates, and schedules computing resources across the board. Through this network, computing power, data, and applications will be aggregated and shared.



Computing centers have multiple layers and management domains. Different computing centers differ greatly. The types of applications deployed, datasets stored, and computing architectures may vary from site to site. Management policies, billing standards, and carbon emissions standards may also vary. If we are to build computing networks, there are several things that need to be sorted out first: coordination between different computing centers; a trusted transaction and management mechanism for computing power, data, and applications; and unified standards. The ultimate goal is to build computing architecture that is open, energy efficient, and delivers high resource utilization.

2) Converged applications will form a digital continuum

Hyperscale AI models, the explosive growth in the volume of data, and the increasing requirements for precision and speed in scientific computing will require massive computing power and new applications. The distributed applications of the future will integrate real-time and non-real-time data processing, model training and inference, simulation and modeling, IoT, and information physics to form a "digital continuum". This will solve the problems that individual computing centers find hard to solve. For example, a digital meteorological model, which combines neural networks and real-time data, can provide short-term and imminent weather forecasts at high frequency and high resolution, bringing tangible benefits to our everyday lives. Distributed large-scale models can use the resources of multiple computing centers to speed up model training. New applications will support the connectivity between different computing centers and between computing centers and edge computing facilities. Computing centers will no longer be independent systems; instead, each center will be a node in an interconnected computing network. In order to meet the computing and data processing requirements of complex applications, users from multiple organizations can share computing power and data distributed across multiple computing centers.

3) Collaborative scheduling for cross-domain computing centers

Multiple computing centers distributed at different geographical locations will be connected to support new distributed converged applications. Training hyperscale models will require the resources of multiple computing

centers, and complex converged applications may also rely on the computing power and datasets of different computing centers. Application diversity, resource heterogeneity, and inconsistent management strategies will all pose new challenges to the scheduling system. The scheduling system needs to be aware of the computing power and storage resources required by applications; it will need to know the locations of data, to reduce data movement overheads; and it will need to understand how applications communicate to reduce communication overheads. The scheduling system also needs to be aware of the availability and heterogeneity of resources in different computing centers in real time, and the network status of different computing centers. In addition, the system needs to make the optimal decisions while taking into account the required cost-effectiveness and energy efficiency, in order to adapt to differences in resource pricing and carbon emission standards that apply to different computing centers. That is, the scheduling system must be capable of discovering resources, aware of the characteristics of applications, aware of software and hardware heterogeneity at computing centers, and aware of local management policies. This will make it possible for the scheduling system to deliver globally optimal efficiency in computing, data movement, and energy use.

Chip engineering

1) 2.5D chiplet packaging and integration technology will continue to improve chip computing power and product competitiveness

The hard dimensional limits on wafer exposure (25 mm x 32 mm for one reticle) present huge technical barriers to increasing total die size and die yield. This issue is impeding efforts to improve chip performance and cut chip costs.

2.5D silicon/fan-out (FO) interposer + chiplet technology can increase die yield and reduce chip costs. Stacking and integration help achieve greater chip performance, and provide better adaptability to different product specifications. In addition, the energy consumption per bit in 2.5D packaging is just half that of the board-level interconnection solution used in conventional packaging.

As the industry continues to advance and the demand for chips grows, it is estimated that by 2025, the size of a 2.5D silicon/FO interposer will be more than four times that of a reticle, and the substrate is expected to be larger than 110 mm x 110 mm. Larger 2.5D and substrate processes pose engineering challenges in terms of yield, lead time, and reliability. To address these challenges, converged, innovative substrate architectures will be needed.

2) 3D chip technology is expected to outperform conventional architectures by dozens of times

3D chip technologies present significant advantages over advanced 2D/2.5D packaging and heterogeneous integration: better interconnection density, bandwidth, chip size, power consumption, and overall performance. 3D chip technologies will be critical to chip and system integration in key scenarios such as high-performance computing and AI.

3D chip technology will evolve from die-to-

wafer (D2W) to wafer-to-wafer (W2W) and µbumps, and then to hybrid bonding, and finally to monolithic 3D technology. This technology will be widely used in different types of stacking, including 3D memory on logic, logic on logic, and optical on logic, and will gradually extend to multi-layer heterogeneous stacking.

3D chip stacking requires the use of ultra-high-density bonding technology with pitches smaller than 10 µm. 3D chips have significant advantages over 2.5D packaging in terms of bandwidth and power consumption, so power consumption per bit is expected to fall by 90%. Ongoing research is required into technologies for working with smaller through-silicon vias (TSVs), both in materials and processes. One drawback of 3D stacking is that it multiplies local power density and current density, with implications for the system's power supply and heat dissipation paths.

3) Co-packaged optics for Tbit/s-level high-bandwidth ports

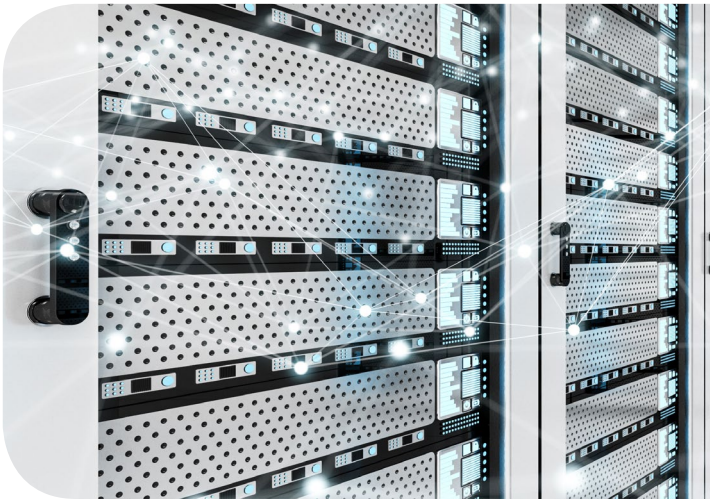
Compute-intensive chips (e.g. xPUs, switches, and FPGAs) will deliver increasingly higher I/O bandwidth. It is expected that the port rate will reach terabits per second or higher by 2030. As the speed per channel increases, serial communications at speeds of 100/200 Gbit/s or higher will create challenges in power consumption, crosstalk, and heat dissipation. Conventional optical-to-electrical conversion interfaces will no longer meet the demands of increasing computing power. Co-packaged optics are expected to cut end-to-end power consumption by 2/3. Co-packaged optics can replace pluggable optics and on-board optics, and will become a key technology for higher port bandwidth. If the technology is to be widely adopted, challenges in engineering technologies will need to be addressed, including 3D packaging of photonic integrated chips (PICs) and electronic integrated chips (EICs), ultra-large substrate and optical engine (OE) integration, and chip power density.

4) Power supply for power-intensive chips

The demand for increasing computing power and the development of chiplet technology continue to drive up chip power consumption. The power supply for kW-level chips will no longer be a problem, but more innovative and efficient power supply strategies will be required for 10kW-level wafer-level chips. New power supply architectures such as high-voltage single-stage conversion and switched-capacitor hybrid conversion, combined with engineering

technologies such as low-voltage gallium nitride (GaN) power devices and high-frequency integrated magnets, can further improve the end-to-end energy efficiency and power density of board power supply.

High-voltage (48V) direct power supply is key to addressing the problem of chip power supply. To implement this technology, it will be necessary to first develop new materials for substrates and packaging, along with the processes necessary to apply them, which can accommodate the high voltages. Efficient on-chip voltage conversion and core-based power supply are the way forward for research.



5) Chip-level heat dissipation technology

The power consumption of computing chips has risen sharply, and heat dissipation has become a major barrier to further chip performance improvement. There is an urgent need for new heat dissipation technologies and materials. Lidless chips, advanced package- and chip-level liquid cooling, and high-conductivity TIM1 materials that reduce path thermal resistance are expected to provide the heat dissipation capacity necessary for kW-level chips, and even 10 kW-level chips. They will open the way for major advances in chip performance. Dynamic chip thermal management and system-level coordinated heat dissipation will also be key technologies for ultra-power-intensive chips.

Diversified computing

In the future, data will be processed in the right place, using the right kind of computing. For example, network data will be processed on data processing units (DPUs) and neural network models will be trained on NPUs. Computing power will be everywhere. Peripherals such as hard disks, network adapters, and memory will gradually become capable of data analysis and processing. Converged applications call for a unified architecture for diversified computing. Currently, tools from different vendors are siloed from each other, greatly hindering the development of diversified computing.

Data-centric computing

1) Symmetric computing architecture (in-memory data processing)

In Von Neumann architecture, data needs to be moved from storage to the CPU for processing, and this movement of data consumes a large amount of computing power and energy in the system. In addition, numerous memory, storage, and transport formats need to be converted back and forth during data processing and exchange, which consumes a lot of CPU time and leads to low energy efficiency. At the same time, data volumes are mushrooming, and hardware deployment cannot keep pace. This will exacerbate existing issues related to input/output (I/O), computing power, and networks. Such issues slow down data migration, hinder processing efficiency, and affect a system's overall energy efficiency.

These issues can be properly addressed with a symmetric computing architecture that supports memory pooling. Under this architecture, unified memory semantics will be used to process and exchange data throughout the data lifecycle, and even ensure that all data is processed in the memory. This architecture can eliminate

the need for format conversions, improve data migration speeds, expand the memory available for applications, and ultimately enhance the entire system's data processing capability. This will be one of the major approaches to faster computing. Building this architecture will require breakthroughs in multi-level memory architecture, large-capacity non-volatile memory, and other key technologies.

2) Ubiquitous computing (intelligent peripherals)

In the future, a diverse range of xPUs will provide different types of computing power. In addition, we believe that an architecture with ubiquitous near-data computing will be a way forward. Under this architecture, data will be processed in the right place with the right kind of computing power, which will help reduce data migration and boost overall system performance.

Ubiquitous near-data computing may involve the following directions:

(1) Near-memory computing. In current systems, the effective bandwidth available for data migration is limited by the bandwidth of the external bus. In the future, multiple concurrent programmable computing units will be added to the dynamic random access memory (DRAM) control circuit, and the DRAM array structure will be optimized to improve concurrent internal data access. This will multiply effective bandwidth for data computing in the DRAM, and help overcome the bandwidth bottleneck caused by the memory wall.

(2) Near-storage computing. Currently, a fixed data acceleration unit (such as a compression engine) can be added to a solid state drive (SSD) controller specifically to process data. In the future, multiple operator engines in the SSD controller could be invoked on demand through application programming interfaces (APIs). Coupled with compilers, this approach can support more flexible offloading of compute workloads, and improve the energy efficiency of data computations in general scenarios.

(3) Computing using memory based on SmartNIC, which will evolve to a DPU-based, data-centric computing architecture. In the future, in-network computing power will be flexible and programmable, existing within open, heterogeneous programming frameworks, for a service-driven in-network computing paradigm. This will support acceleration across the board, including storage, security, and virtualization, and will greatly improve the performance of distributed applications, such as HPC and AI convergence, big data, and databases. Fine-grained dynamic scheduling and efficient

interaction of all computing resources in data centers will become possible.

3) Computing using memory

Computing using memory is a tight coupling between processing and storage units, which allows storage media to function as both a storage unit and processing unit. This erases the boundary between computing power and storage, effectively overcoming the power wall and the memory wall. This technology is expected to be at least 10 times more energy efficient than traditional Von Neumann architecture.

Computing using memory based on mature memory technologies like static random-access memory (SRAM) and NOR flash is expected to be in commercial use on a large scale within two to three years. This technology will make AI inference and operation on devices and the edge 10 times more energy efficient. Computing using memory, powered by new non-volatile memories like resistive random-access memory (ReRAM), phase change memory (PCM), and magnetoresistive random-access memory (MRAM), is still in the experimental phase, but given their high performance and low energy consumption, they have the potential to be used in data centers in the next decade.

Breakthroughs in the following areas will also be required before computing using memory can become commercially available on a large scale.

Computational precision: Computational noise and issues of component consistency and stability can cause computational errors, so computing using memory is less precise than conventional computing systems. Therefore, algorithms will need to be optimized to account for the kind of compute circuit on which they are running.

Software ecosystem: Computing using memory is a type of data-driven computing. Neural network models need to be deployed on the right storage units, and the entire computational process will be controlled through data flow scheduling. This necessitates the development of more intelligent, efficient, and convenient data mapping tools.

System architecture: Computing using memory, powered by new non-volatile memory, uses a calculation method that multiplies matrices by vectors. Today, these systems are often used in specific machine learning applications (e.g., neural network inference and training), and it is difficult to extend them to other use cases. In addition, they cannot cooperate with existing storage systems to efficiently process data. To overcome these challenges, a full-stack design that facilitates synergy between storage devices, programming models, system architecture, and applications will be essential to ensure that the architecture of computing using memory works for general purposes.

4) Buses: From board-level buses to DC-level buses

With the exponential growth of computing power and data, large, centralized data centers that focus on AI, HPC, and big data will become more important. Compared with intra-node buses, the networks connecting entire data centers has a huge latency, bandwidth gap, and heavy network software stack overheads. All of these features degrade computing power. Lightweight software stacks, with high bandwidth, low latency, and memory semantics, exist at the board level, and will be extended to the entire data center through the memory-semantic bus. This will enable optimal performance and energy efficiency for the entire data center.

For memory-semantic buses, the biggest challenge lies in building open, equal, interoperable buses, interfaces, and protocol standards. This helps prevent the fragmentation of standards for computing system buses, which would only hinder advances in computing performance and large-scale computing.

Application-driven diversified computing

The next generation of computing systems will bring a new paradigm, characterized by domain-specific hardware, domain-specific programming languages, open architectures, and native security architectures.

1) New paradigm for scientific computing

With breakthroughs in AI computing methods and AI computing architectures, a new paradigm is emerging in scientific research, in which machine learning is combined with first-principles-based physical modeling. In the next decade, intelligent scientific computing will be involved in every aspect of scientific research and technological innovation. The effort to efficiently integrate AI algorithms with scientific computing presents unprecedented challenges and opportunities.

- In terms of the fundamentals, there are challenges regarding the computational frameworks and mathematical methods of the new computation approach. There is a need for new frameworks and approaches that ensure a given problem can be effectively solved using AI. That is, the mathematical methods and frameworks must ensure computability, learnability, and interpretability. Therefore, over the next decade, hardware and software infrastructure must be built based on mathematics and AI research and provide appropriate implementation, assessment, and testing systems.

- In terms of data, a large number of different data sources are required to boost scientific research, engineering, and manufacturing using AI. First, different fields of scientific research rely on different sources for their data. These data sources may include instruments, simulations, sensor networks, satellites, scientific literature, and research findings. Currently, there are still great challenges to overcome regarding the availability and shareability of this data. Second, there are challenges in using AI to generate effective data that is based on physical principles and complies with basic laws of physics (such as symmetry and conservation laws). To address these challenges, scientists from different domains, AI experts, mathematicians, and computer scientists need to work together.

2) AI enabling intelligent storage

Storage systems are now expected to address loads of increasingly diverse and complex service requirements and to offer simplified system management and O&M.

Storage systems of the future will be able to use AI to proactively manage and respond to their internal and external environments,

to learn continuously, to be workload-aware and adaptive, and to automatically optimize themselves to deliver gains in resource allocation, cost, performance, reliability, usability, etc. In addition, manual O&M will need to evolve to automated intelligent O&M using AI.

Progress has already been made in the application of AI in indexing, automatic optimization, and resource allocation in storage systems. However, breakthroughs in the following four areas are still needed:

- **Workloads:** The impact of I/O workloads on system performance needs to be modeled to identify the key indicators and factors affecting module performance, to accurately assess system performance, and to simulate real-world service scenarios.

- **Data:** Data distribution, data lifecycle, and data context need to be perceived so that systems can improve data access performance, reduce the consumption of resources by back-end garbage collection, and improve data reduction ratios.

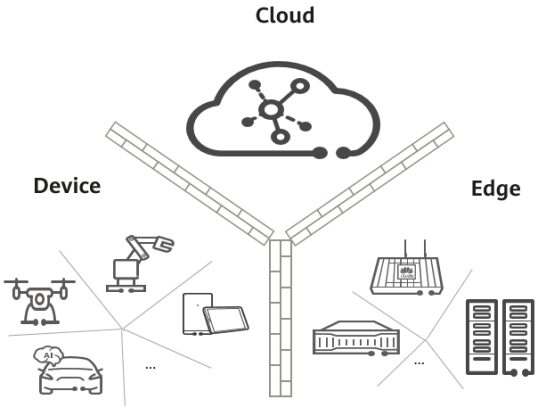
- **Systems:** Rules and patterns need to be identified based on past data, computing tasks need to be arranged and scheduled efficiently, and systems need to be optimized during runtime to improve system parameters and resource allocation, reduce system power consumption, and ensure that fluctuations in system performance are controllable and do not undermine reliability.

- **Operations:** Automated O&M is needed to eliminate the need for manual work; faults need to be automatically analyzed to identify root cause; and any system suboptimality needs to be detected, prevented, and rectified automatically.

Integrating top-down load modeling and bottom-up adaptive learning to support intelligent storage has become an area of interest. A great deal of current research is aimed at developing intelligent storage systems featuring automatic performance optimization, automatic QoS control, intelligent data awareness, self-learning of rules and policies, intelligent scheduling, low-power controls, simplified planning and configuration, prediction of system issues, and automatic root cause analysis.

Multi-dimensional collaboration

Computing and storage infrastructure are distributed in different locations on the cloud, edge, and devices. Such infrastructure can be horizontally and vertically coordinated to complement each other and enable cubic computing. This addresses problems such as poor service experience, uneven distribution of computing, low utilization of computing resources, and information silos.



Multi-dimensional sensing and data modeling enable the physical world to be mirrored, computed, and enhanced to form digital twins. With light field holographic rendering and AI-assisted content generation, the digital world is precisely mapped to the physical world. Multi-dimensional collaboration between time and space, as well as between virtuality and reality, enables seamless integration of the physical and digital world.

Cubic computing

1) Edge computing

The world of the future will be an intelligent one in which everything is connected. As 5G technologies mature and see increased application, edge computing will be widely deployed in the ICT industry. It is expected that the global edge computing market will be worth hundreds of billions of US dollars by 2030, but at present the value of this market is US$10 billion. To apply edge computing on a large scale, we must first confront challenges in areas like edge intelligence, edge computing network, edge security, edge standards, and open ecosystems.

Edge intelligence: Intelligent upgrades of vertical industries like manufacturing, power grids, city administration, transportation, and finance are important drivers of the exponential growth of edge computing. Development kits for basic AI capabilities, such as incremental learning, transfer learning, device optimized model compression, and inference scheduling and deployment, are needed to solve common issues encountered by many industries currently undergoing intelligent transformation. A development kit is needed to address common issues unique to intelligent manufacturing. This industry is characterized by samples or images with complex backgrounds and low contrast, small size training samples, and weak supervision. Development kits should also be developed for other industries, to form a comprehensive set of software development kits (SDKs) for application enablement.

Edge computing network: Future service demands will drive edge devices to support a greater range of services. As such, these devices will need to be mobile, low-power, and smaller, but computing, storage, bandwidth, and latency will become bottlenecks. Holographic and multi-dimensional sensing services require 100 times more computing power than is currently available, storage capacity will need to expand by 100 or even 1,000 times, and network bandwidth will need to increase to tens of terabits per second. Industries such as intelligent manufacturing, intelligent power grids, and intelligent transportation require millisecond-level deterministic latency. To meet the demands of edge acceleration, offloading, and performance breakthroughs, we need hyper convergence of computing, storage, and networking, with efficient use of diversified computing. This will pose new challenges to edge software and hardware architecture.

Edge security: Edge devices are physically closer to attackers. Being located in complex environments, edge devices are more vulnerable to attacks from physical hardware interfaces, southbound and northbound service interfaces, and northbound management interfaces. Data is often a core asset of users, so data loss or theft may cause significant losses to users. It is estimated that 80% of data will be processed at the edge by 2030. It is thus paramount to strengthen security and privacy protection during data collection, storage, processing, and transmission at the edge. In addition, the security and privacy of core assets such as edge applications and models must be strictly protected. Data silos caused by data privacy protection must be prevented as this would make it difficult to fully unleash the potential value of data and AI algorithms in sectors such as healthcare, finance, and industry.

Edge standards and open ecosystems: Edge devices for different industry applications differ greatly in computing power, functions, software and hardware formats, and interfaces. Proprietary software and hardware solutions and interface protocols from different vendors are often not interoperable, which greatly hinders the adoption of edge computing. The edge computing system, software and hardware frameworks, and related interfaces and protocols need to be standardized, and corresponding test and acceptance standards need to be established for better interoperability between edge devices, software, and protocols. In addition, open ecosystems need to be built for each industry to attract investment from more vendors and partners.

2) Multi-device collaboration

Animals like ants and bees create swarm intelligence through collaboration. The multi-device collaboration technology aims to achieve similar breakthroughs to improve the problem-solving capabilities, overall performance, and robustness of multi-device systems.

Multi-device collaboration takes various forms, such as task sharing, result sharing and intelligent agents. In task sharing, devices collaborate by performing subtasks of a particular task. In result sharing, devices collaborate by sharing parts of the results. The processing capability of each device at any given moment depends on the data and knowledge that the device owns or receives from other devices. In the form of intelligent agents, devices collaborate on the basis of independence and autonomy.

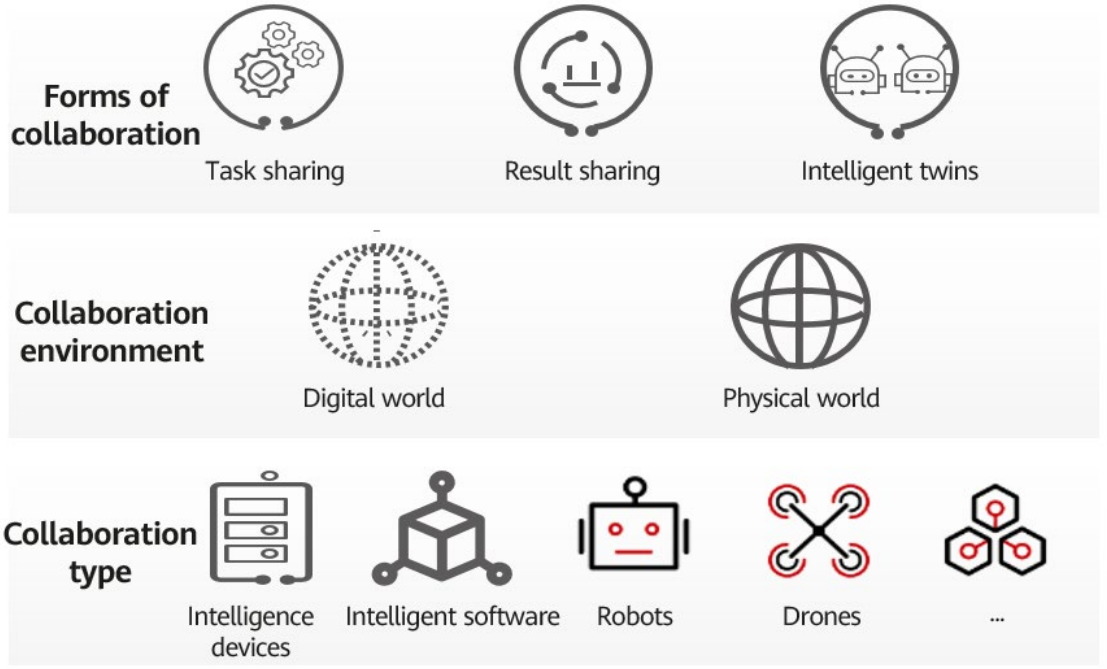
Effective multi-device collaboration requires solving problems related to cooperation and conflict resolution, global optimization, and interaction and collaboration consistency.

Cooperation and conflict resolution: A deadlock or livelock may occur during multi-device collaboration. Deadlocks make devices unable to perform their respective next-steps, and livelocks make devices work continuously without making any progress. Coordination mechanisms and algorithms are critical for preventing deadlocks and livelocks in interactive processes.

Global optimization: It is difficult to achieve global optimization when multiple devices are collaborating based on local information. However, collaboration based on a global view often means large communication traffic, which can overburden the system. Efficient and secure acquisition of high quality and reliable global situation estimations determines the efficiency and effectiveness of multi-device collaboration.

Interaction and collaboration consistency: Each device obtains information from other devices through network communication and adjusts its own state. In practice, because the connectivity between multiple devices is unreliable or there are barriers to communication, collaboration consistency issues may arise. Therefore, the ability to address such issues determines the robustness of a multi-device collaboration system.

Multi-device collaboration systems will gradually



evolve from simple cooperation and connection to autonomous swarm intelligence.

3) Device-edge-cloud computing

AI and emerging data-intensive applications, such as intelligent manufacturing, intelligent cities, smart inspection, and intelligent transportation, are developing rapidly. The need to improve application experience, such as by reducing latency, reducing bandwidth costs, and enhancing data privacy protection, drives the development of device-edge-cloud computing. To develop an integrated computing architecture, the following challenges need to be addressed.

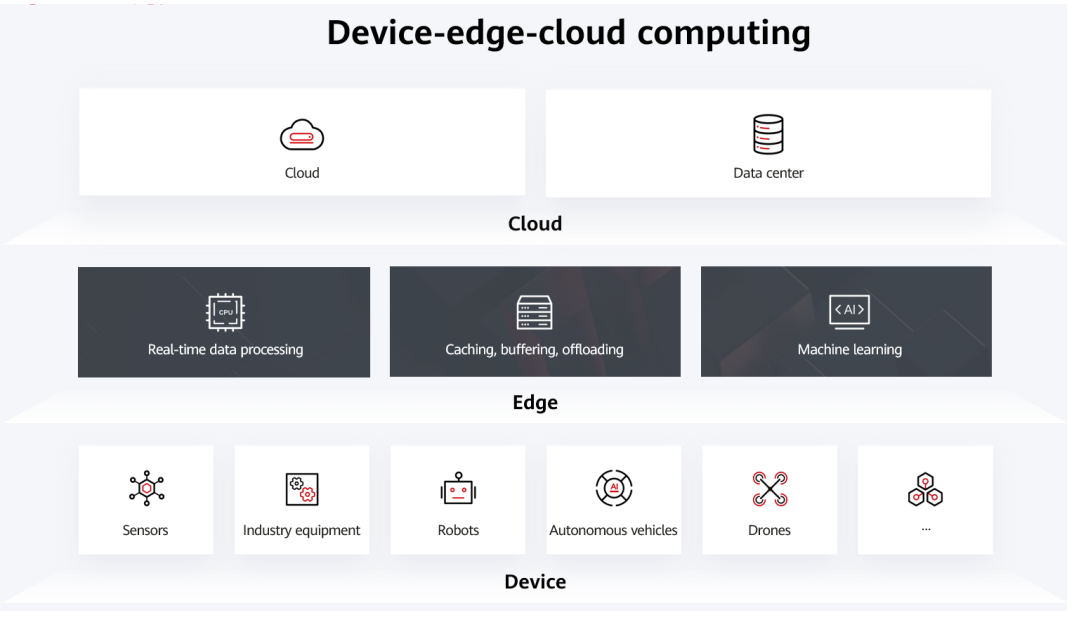
Task collaboration: How should a computing task be divided into multiple subtasks? How should subtasks be deployed and scheduled on the device, edge, and cloud? Where should a subtask be performed (on the device, edge, or cloud) and when? The migration of computing subtasks across clouds, clusters, and nodes is also a challenge.

Intelligent collaboration: The model of training on the cloud and inference at the edge is moving toward device-edge-cloud collaborative training and inference. Challenges in the following areas need to be addressed to achieve device-edge-cloud synergy: precision and rate of convergence of collaborative training; latency and accuracy of collaborative inference; and data silos, small sample sizes, data heterogeneity, security and privacy, communication cost, and limited device/edge resources.

Data collaboration: Data is the basis of intelligence. Diversification and heterogeneity pose challenges for data access, aggregation, interaction, and processing.

Network collaboration: As the scale of the device-edge-cloud computing network grows, access by a large number of devices and subnets brings great challenges to device, network, and service management. We need solutions for the challenge of ensuring reliable real-time connectivity.

Security and trustworthiness: How can security and privacy be ensured when edge devices and their data are connected to the cloud? How can the cloud protect itself from edge-side attacks? How can the data sent from the cloud to the edge be protected?

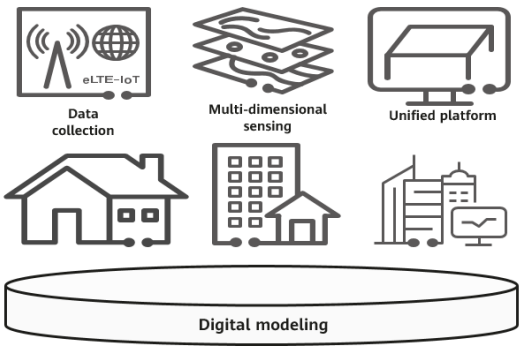


Digital twin

1) A unified digital twin platform is the way forward

Under the digital tide of various industries, such as smart factories, smart cities, and virtual social media, there is no unified platform for creating personalized digital twin systems. This platform needs to focus on the unification of data formats and development tools of 3D models, and provide diversified computing power and storage space required for modeling large amounts of data.

2) Multi-dimensional sensing and digital modeling technology



The physical world of the future will have a digital twin. These two worlds will be seamlessly converged and work in tandem to improve the efficiency of product design, product manufacturing, medical analysis, and engineering construction. The process of mapping the physical world to its digital twin will face numerous challenges, such as multi-dimensional sensing, 3D modeling, and light field data collection and storage.

Multi-dimensional sensing: Massive amounts of data on the physical world, including images, videos, sounds, and temperature, humidity, and mechanical records are collected and stored. The acquisition, processing, and convergence of data with more dimensions requires high-resolution sensing, object location, imaging, and environment reconstruction, and the amount of data generated in this process is even larger. The process of screening, preprocessing, modeling, and simulation of such massive amounts of data relies on powerful computing and the deep integration of multiple disciplines, such as artificial intelligence, cognitive science, control science, and materials science.

3D modeling will require 100 times more computing power. 3D modeling, which is based on images and video streams of different angles and massive amounts of data collected by array cameras and depth cameras, requires huge computing power. The volume of high-precision data collected by a 100 plus-channel camera array is 100 times higher than that of 2D images. The resolution will increase to 8K and the required computing power per channel will see a 4-fold increase. The required computing power for modeling is 100 times higher. Managing this massive amount of multi-dimensional data and transforming it into a 3D model is a big challenge. In addition, in the consumer market, depth information of images can be obtained using the 3D camera on a phone, and medium- and low-precision modeling based on the depth information can be performed on the phone. The 3D camera of a phone is usually a binocular camera, structured light camera, or time-of-flight (ToF) camera. A unified, efficient, and economical software and hardware system for 3D modeling is required for high-level and consumer-level modeling, the digital transformation of various industries, and the flourishing of the digital twin industry.

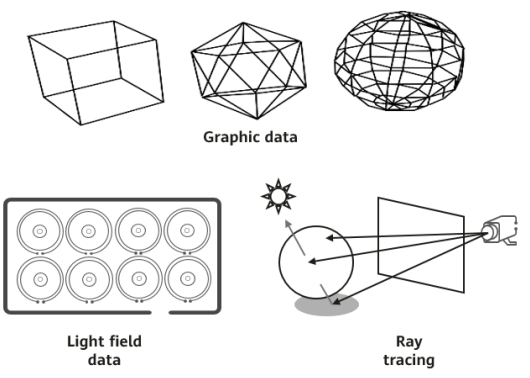
AI-enabled material generation in digital modeling: Powered by AI image recognition technology, intelligent generative algorithms, and strong AI computing power, digital models can automatically recognize the physical properties of images, such as metalness, roughness, reflectivity, refractivity, and surface normal vector. This would then generate materials like we see in the real world in the form of a 3D model. To support this process, a unified and open material

description language is needed to exchange 3D graphic data between different industries.

100 times more light field data will make compression a key technology: Light field camera arrays will collect 100 times more image and video stream data, which will then be used for synthesizing 3D video streams and light shading in rendering. Such massive amounts of data mean that data storage and processing will be a huge challenge. Breakthroughs in fast compression and storage of light field data are therefore essential, as these are the key to subsequent rendering and imaging.

3) Light field holographic rendering technology

Breakthroughs in visual and interactive technologies need to be made for a digital twin display system to provide users with the same experience as they have in the physical world. Most products currently on the market have deficiencies in rendering quality, fidelity, and rendering delay. Real-time ray tracing and zero-delay transmission can directly improve user experience and are key technologies for photo-realistic authentic rendering. Advanced rendering such as ray tracing requires 10 times more computing power than traditional rendering. Utilizing storage to replace computing can meet part of the demand for computing power while reducing latency, but this would necessitate greater storage space. Moving forward, cloud-based holographic rendering of light fields will be an important area of research.



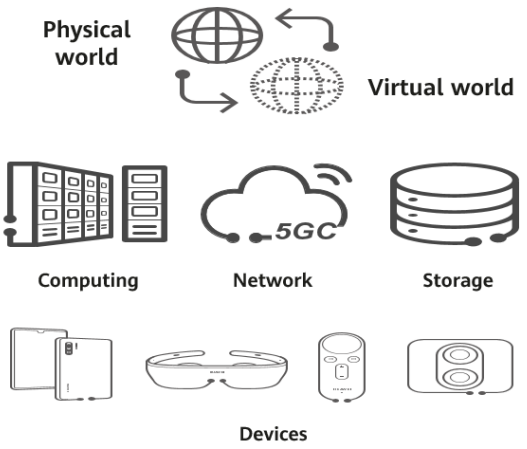
Advanced rendering technology will deliver a 64-fold increase in resolution: The mainstream technology of holographic rendering of light fields has evolved from rasterization rendering to much more advanced rendering technologies such as ray tracing. In scenarios such as gaming and extended reality (XR), a near-real experience can be made possible with 16K binocular resolution, 120 frames per second, and a latency of no more than 8 ms. Strong interaction services use 64 times more computing power and require a latency of 5 ms. These services need breakthroughs in key technologies such as 3D modeling, material generation, and ray storage. Device-edge-cloud computing clusters can provide converged computing power for rendering, AI, and video streaming. When these compute resources are combined with content creation software for advanced rendering, near real-time and high-performance rendering solutions can be created.

AI-based content generation: AI can enable 3D modeling, automatic material generation, super resolution, and noise reduction. AI technologies such as generative adversarial network (GAN), natural language processing (NLP), and natural language generation (NLG) will generate 3D images of avatars and allow them to have vivid expressions and engage in natural language conversations. This will greatly aid communication between people in different parts of the world. AI content generation will also be used in industrial design, XR content creation, and special visual effects.

4) Interaction between the physical and digital worlds for hundreds of millions of users

Allowing hundreds of millions of users in the physical world to interact with digital twins places high demands on computing, storage, and network bandwidth. This is because it requires a large amount of state queries and message transmission. When people and things can interact with each other at latencies less

than 5–10 milliseconds, the bandwidth reaches hundreds of Mbit/s per user, and the required computing power increases to tens of TFLOPS per user, network-edge-cloud collaboration and real-time data processing and transmission for hundreds of millions of users will be possible, but this is a very challenging goal.



Physical layer breakthroughs

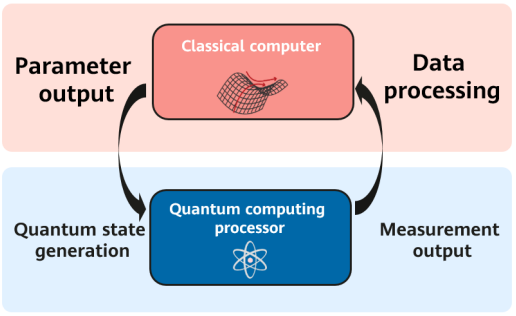
Both academia and the industry are exploring potential breakthroughs at the physical layer, including analog computing, non-silicon-based computing, novel storage media, and optimized chip engineering, to keep improving the energy efficiency of computing and storage density. For example, quantum computing offers exponential advantages over traditional computing in data representation and parallel computing. Analog optical computing consumes little power yet achieves high performance for certain computing tasks. 2D materials and carbon nanotubes have high carrier mobility and shorter channels, and are expected to replace silicon. Significant breakthroughs have been made in ferroelectrics, phase change materials, and device structures, resulting in significant improvement of storage density and read/write performance. Multi-layer and multi-dimensional optical storage has huge potential for long-term storage of cold data. Breakthroughs in DNA storage will need to be made. These breakthroughs in key technologies at the physical layer will revolutionize computing and storage.

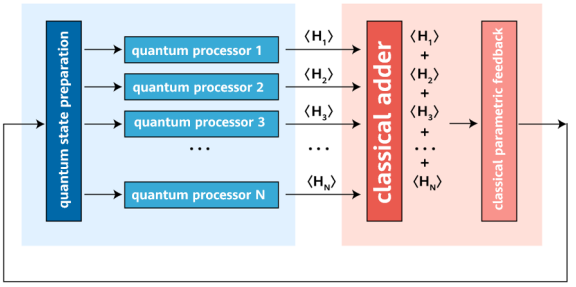
Analog computing

1) Quantum computing: A technology of strategic importance for the future of high-performance computing

Quantum computing is undergoing rapid progress in engineering, and a chip with more than 1,000 qubits is expected to appear within the next five years. Quantum computing is now in an era of noisy intermediate-scale quantum (NISQ). The most feasible path forward is building a hybrid computing architecture that combines the accuracy of classical computers and the performance of quantum computing. This hybrid computing architecture will be used in quantum chemical simulation, quantum combinatorial optimization, and quantum machine learning, as those are the three scenarios that have the greatest commercial potential. Quantum chemical simulation can provide new computing power for research and development of pharmaceuticals and new materials. Quantum combinatorial optimization, where combinatorial optimization problems are encoded as quantum dynamics, can be used to optimize logistics scheduling, route planning, and network traffic distribution. Quantum machine learning will provide a new path for accelerating AI computing.

The focus of the next decade should be on developing a dedicated NISQ-based quantum computer, while continuing to increase the number of qubits in a single quantum chip, prolong coherence time, and enhance fidelity. More efforts should be made to optimize the interconnection of quantum chips to enhance system scalability, so that sufficient computing power will be available to solve those complex problems. At the same time, we also need to make quantum computing more fault tolerant, improve system reliability, optimize quantum algorithms for different application scenarios, and improve the quantum software stack, while reducing circuit depth and complexity. These





are part of the broader efforts to bring NISQ-based quantum computing to commercial use. However, building a universal quantum computer will be a long, challenging process.

2) Analog optical computing: Competitive in certain complex computing tasks

Light propagates at a high speed with negligible power consumption. In certain optical systems, mathematical models are used to describe their associated physical phenomena, such as interference, scattering, and reflection. Certain computing tasks can be accomplished by utilizing the physical characteristics of light, such as amplitude and phase, and the interactions between light and optical devices. In addition, as a boson, a photon allows parallelism in degree of freedom, such as wavelength division multiplexing, mode division multiplexing, and orbital angular momentum (OAM) multiplexing. Multi-dimensional parallelism is an important direction forward for optical computing.
Early breakthroughs of optical computing are expected to appear in convolution computing, Ising model solving, and reservoir computing, followed by application in signal processing, combinatorial optimization, sequence alignment, and AI acceleration.

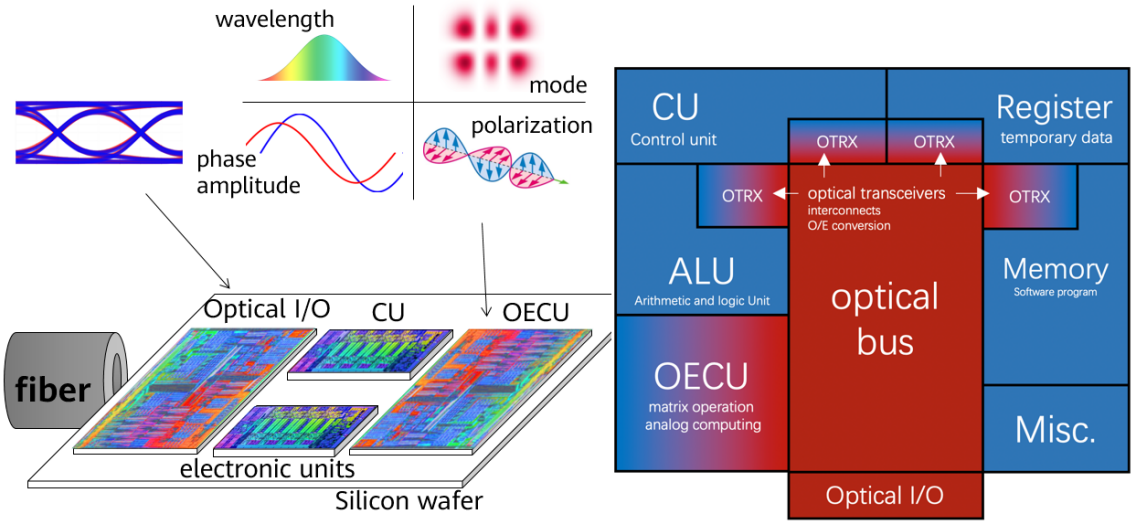
There are still formidable challenges for the commercial application of optical computing, such as insertion loss, noise control, heterogeneous integration, and co-packaging of electronic and optical devices. The drive circuits used in optical computing also need to be further integrated with optical chips to reduce power consumption and area. As optical computing and electrical computing each have their own advantages, optoelectronic hybrid computing architecture is a promising direction for future development.

Non-silicon-based computing

1) 2D materials: A potential material to extend Moore's law

2D materials offer several advantages, including shorter channel length, high mobility, and the possibility of heterogeneous integration, and are expected to be used as transistor channel materials to extend Moore's law as far as 1 nm technology node. In addition, 2D materials with ultra-low dielectric constants can be used as the interconnect isolation materials of integrated circuits. 2D materials are expected to be first adopted in domains such as optoelectronics and sensors, and eventually in large-scale integrated circuits and systems.

At present, 2D materials and relevant devices are still in the basic research stage, and many of the necessary breakthroughs in materials, devices, and processes have yet to be made. Over the next five years, we need to realize industrial-grade wafers made of 2D materials and constantly improve their yield. In addition, we need to keep optimizing the electrode contacts and device structures to improve the comprehensive performance of 2D transistors. Once these improvements are made, 2D materials are expected to be applied in large-



scale integrated circuits within ten years.

2) Carbon transistors: The most promising technology to extend Moore's law

Carbon nanotubes have great potential in both high performance and low power consumption because of their ultra-high carrier mobility and atomic-level thickness. In cases of extreme scaling, carbon nanotube transistors are about 10 times more energy-efficient than silicon-based transistors. Carbon nanotubes are expected to be commercially used in biosensors and radio frequency circuits in 3 to 5 years.

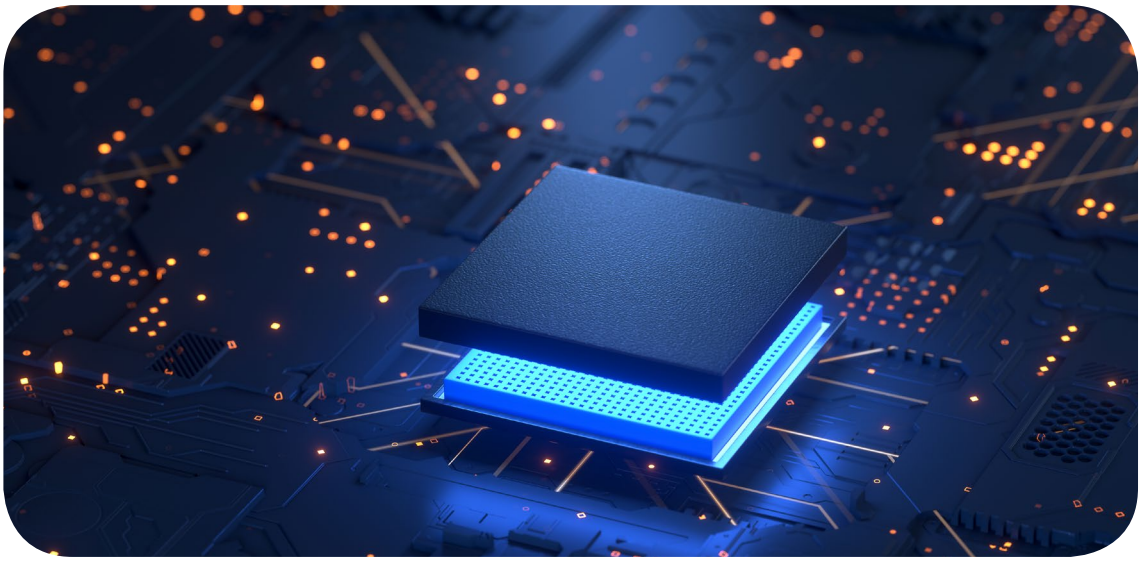
The next five years will see more efforts invested to improve the fabrication process of carbon nanotube materials, reduce surface pollution and impurities, and improve material purity and carbon nanotube alignment. In addition, the contact resistance and interface state of these devices need to be optimized to improve injection efficiency. Supporting electronic design automation (EDA) tools also need to be developed. Small-scale integrated circuits can be used to verify end-to-end maturity of carbon-based semiconductors, which are expected to be initially applied to flexible circuits. Looking ahead to the next decade, when carbon-based semiconductors are scaled down to the level of advanced silicon-based processes, there will be opportunities for large-scale application of this technology in high performance and high integration scenarios.

Novel storage media

While traditional storage mainly uses magnetic media, it is predicted that by 2030, 72% of enterprise storage, including both primary and secondary storage, will be based on all-flash. Furthermore, 82% of enterprise service data will require backup. Because of the differences between hot, warm, and cold data throughout the lifecycle, the evolution of storage media will diverge in two directions: higher speed with better performance, and massive scale at lower cost.

1) Novel media for memory

Currently, hot data is stored in SSDs and transmitted to DRAM for processing, because the latency of DRAM can be up to 1,000 times lower than that of SSDs. However, physical conditions limit DRAM from further density or voltage expansion. Therefore, neither SSDs nor DRAM are the best options for hot data storage. There are now many novel media technologies for memory, such as PCM, MRAM, ferroelectric RAM (FeRAM), and ReRAM, and those media



outperform DRAM in performance, capacity, cost, lifespan, energy consumption, and scalability. They also support byte-level access and persistence, making data migration unnecessary. Eventually, they will become commonly used media for hot data storage, but for now they face two major technical challenges:

Capacity: The total amount of hot data in 2030 will be equal to the total amount of the data stored on SSDs today. The capacity density of hot data media needs to be increased by at least ten times to reach the current level of SSDs, which is 1 TB/die. Such media should also support on-demand expansion unrestricted by processors, memory interfaces, network latency, and bandwidth. Media such as FeRAM, ReRAM, and MRAM face structural and material challenges.

Energy consumption: In the global push toward carbon neutrality, there is considerable pressure to reduce the power consumption of storage media for massive amounts of hot data. Resistor-based data storage technologies such as PCM and ReRAM require high data write voltages and therefore consume more power. The operating voltage of FeRAM, however, is relatively low, and its power consumption per bit is just one tenth that of ReRAM and MRAM, and a mere hundredth of that of PCM, making FeRAM the most promising candidate.

2) High-density NAND flash media

In the future, most hot data will be generated from warm data, which means warm data will become the largest reservoir of hot data. Therefore, warm data media must balance performance, capacity, and cost. NANDs will replace hard disk drives (HDDs) as the primary storage medium for warm data and are evolving towards multi-level cells and 3D stacking. The biggest challenge is to expand the capacity and reduce the cost of NANDs while achieving the same level of performance and lifespan as quad-level cells (QLCs).

Performance and lifespan of multi-level cells: For every additional bit a cell stores, the voltage needed for the data doubles, reducing read/write performance and lifespan by several folds.

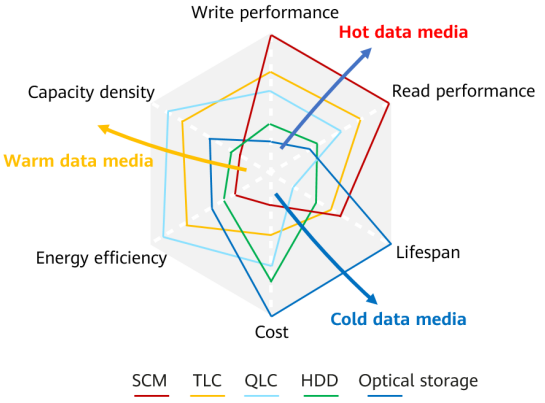
3D stacking process: At present, no mainstream 3D NAND SSDs contain more than 200 layers, but by 2030, we are likely to see products that stack close to 1,000 layers, and the aspect ratio of dielectric through-silicon vias will reach 120:1 (more than double the current level), making processing much more difficult.

3) Optical storage

In the future, the amount of cold data requiring long term storage will increase from 1.2 ZB to 26.5 ZB, and their retention time will grow by 5–10 times. At the National Archives Administration of China, for example, the retention time of key file data has been extended from 100 years to 500 years, and the amount of cold data that needs to be stored is expected to grow from 100 PB to 450 PB. Traditional hard disks and tapes can no longer meet such requirements. With the ongoing research on codec algorithms as well as the read/write mechanisms of transparent materials such as quartz glass and organic glass, optical storage will become the leading storage medium for massive cold data. Before that, however, two challenges must be overcome:

1. The service life of optical storage media needs to be extended tenfold and adapted for use in various complex and harsh environments.

2. Compared with Blu-ray, future optical storage media are expected to have ten times the capacity, perform ten times better, and be available at 1/5 the current cost.



Call to action

Over the past half-century, computing has accelerated scientific advances and economic development, and has been deeply integrated into all aspects of our society. Computing is a resource shared by everyone and will be the cornerstone of the future intelligent world.

Looking ahead to 2030, computing will become both more open and more secure. Every person and every organization will be given equal opportunity to build a more innovative computing industry and share in its value.

Let's work together to usher in a new era of computing.

Appendixes

References

[1] Zettabyte (ZB) and yottabyte (YB) are units of data storage capacity. 1 ZB = 10^{21} bytes, 1 YB = 10^{24} bytes
[2] Huawei predicts that by 2030, there will be 3.3 ZFLOPS of general computing power (FP32) available, a 10-fold increase over 2020; and 105 ZFLOPS of AI computing power (FP16), a 500-fold increase over 2020. FLOPS is short for floating point operations per second. 1 exaFLOPS (EFLOPS) = 10^{18} FLOPS. 1 zettaFLOPS (ZFLOPS) = 10^{21} FLOPS
[3] Speech by Li Deyi, academician of the Chinese Academy of Engineering, at the 1st China Intelligent Education Conference, 2018
[4] China's Guiding Opinions on Accelerating the Development of Intelligent Coal Mines, March 2020
[5] CERN, the European Organization for Nuclear Research, https://home.cern/science/computing
[6] In quantum mechanics/molecular mechanics (QM/MM) modeling, some systems use the QM model for processing, which is very time-consuming, while some use the MM model.
[7] Summit, a supercomputer at Oak Ridge National Laboratory that can perform 148.6 PFLOPS, making it the world's second fastest computer in 2021.
[8] Roland R. Netz, William A. Eaton, Estimating computational limits on theoretical descriptions of biological cells, PNAS 2021
[9] The Gordon Bell Award is presented by the Association for Computing Machinery (ACM). The prize tracks the progress over time of parallel computing and recognizes outstanding achievements in high-performance computing applications.
[10] Weile Jia, Han Wang, Mohan Chen, Denghui Lu, Lin Lin, Roberto Car, Weinan E, Linfeng Zhang, Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning, 2020
[11] DevOps: An approach to agile development that brings development and O&M teams together
[12] The Zero Trust Architecture model was created by John Kindervag in 2010 during his time as an analyst for Forrester Research.

Acronyms

Acronym	Full name
3D	3 Dimensions
AI	Artificial Intelligence
API	Application Programming Interface
AR	Augmented Reality
BP	Back Propagation
CDU	Coolant Distribution Unit
CERN	European Organization for Nuclear Research
CPU	Central Processing Unit
CSP	Cloud computing Service Provider
D2W	Die-to-Wafer
DC	Data Center
DNA	Deoxyribonucleic Acid
DPU	Data Processing Unit
DRAM	Dynamic Random Access Memory
EDA	Electronic Design Automation
EFLOPS	exa Floating-Point Operations Per Second
EIC	Electronic Integrated Circuit
FeRAM	Ferroelectric Random-Access Memory
FPGA	Field Programmable Gate Array
GAN	Generative Adversarial Network
HDD	Hard Disk Drive
HL-LHC	High Luminosity - Large Hadron Collider
HPC	High-Performance Computing
ICT	Information and Communications Technology
IO	Input/Output
KA	Kiloampere
MM	Molecular Mechanics
MR	Mixed Reality
MRAM	Magnetoresistive Random-Access Memory
NISQ	Noisy Intermediate-Scale Quantum

NLG	Natural Language Generation
NLP	Natural Language Processing
O2O	Online to Offline
OAM	Orbital Angular Momentum
OE	Optical Engine
PCM	Phase Change Memory
PB	Petabyte
PIC	Photonic Integrated Circuit
PIM	Processing-In-Memory
PUE	Power Usage Effectiveness
QLC	Quad-Level Cell
QM	Quantum Mechanic
REE	Rich Execution Environment
ReRAM	Resistive Random-Access Memory
SDK	Software Development Kit
SRAM	Static Random-Access Memory
SSD	Solid State Drives
TEE	Trusted Execution Environment
TIM	Thermal Interface Material
ToF	Time of Flight
TSV	Through Silicon Via
UPS	Uninterruptible Power Supply
VR	Virtual Reality
W2W	Wafer to Wafer
WLC	Wafer Level Chip
xPU	x Processing Unit
XR	Extended Reality
YB	Yottabyte
ZB	Zettabyte
ZT	Thermoelectric Figure of Merit

Acknowledgments

During the drafting of this Computing 2030 report, we received invaluable support from Huawei's own team and external consultants. More than 300 experts and professors participated in the discussions that led to this report, contributing ideas and sharing their vision of Computing 2030. We would like to extend our special thanks to them.

(Contributors listed in alphabetical order)

André Brinkmann (Professor, Johannes Gutenberg University Mainz)
Bill McColl (Former professor at the University of Oxford)
Chen Wenguang (Professor, Tsinghua University)
Feng Dan (Changjiang Distinguished Professor, Huazhong University of Science and Technology)
Feng Xiaobing (Professor, Institute of Computing Technology, Chinese Academy of Sciences)
Gan Lin (Assistant Professor, Tsinghua University)
Guan Haibing (Changjiang Distinguished Professor, Shanghai Jiao Tong University)
Guo Minyi (Professor, IEEE Fellow, member of the Academia Europaea, Shanghai Jiao Tong University)
Jarosław Duda (Assistant professor, inventor of asymmetric numeral systems-based compression algorithms, Jagiellonian University)
Jia Weile (Associate professor, Institute of Computing Technology, Chinese Academy of Sciences)
Jin Hai (Changjiang Distinguished Professor, IEEE Fellow, Huazhong University of Science and Technology)
Jin Zhong (Professor, Computer Network Information Center, Chinese Academy of Sciences)
Miu Xiangshui (Changjiang Distinguished Professor, Huazhong University of Science and Technology)
Onur Mutlu (Professor, ACM Fellow, IEEE Fellow, ETH Zurich)
Pan Yi (Professor, Fellow of the American Institute for Medical and Biological Engineering, Foreign Fellow of the Academy of Engineering Sciences of Ukraine, member of the UK's Royal Society for Public Health, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences)
Shu Jiwu (Changjiang Distinguished Professor, IEEE Fellow, Tsinghua University)
Sun Jiachang (Professor, Institute of Software, Chinese Academy of Sciences)
Tian Chen (Associate professor, Nanjing University)
Tian Yonghong (Professor, Peking University)
Wang Jinqiao (Professor, Institute of Automation, Chinese Academy of Sciences)
Wu Fei (Professor, Zhejiang University)
Xie Changsheng (Professor, Huazhong University of Science and Technology)
Xue Wei (Associate professor, Tsinghua University)
Yang Guangwen (Professor, Tsinghua University)
Zheng Weimin (Professor, academician of the Chinese Academy of Engineering, Tsinghua University)

global industry vision

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