2023

Top Ten Technology Trends

of DAMO Academy

DAMO
ALIBABA DAMO ACADEMY

Preface

Since its founding, the driving force behind Alibaba DAMO Academy has always been our desire to play an integral part in shaping the future of humanity. We devoted ourselves to the research of basic sciences, innovative technologies, and the application of technologies in everyday life. Each and every day, our scientists are pushing the boundaries of technology as they venture further into the unknown. In a world where entropy reigns supreme, establishing a clear direction is akin to having a guiding light that helps us stay the course towards our destination. As we innovate technologies through the constant clashing of ideas, we create value that benefits the masses and brings us one step closer to the future we envision.

Our theme this year revolves around our Back to Basics strategy. For us, this meant putting ourselves in the shoes of various industry verticals as we populated the list for DAMO Academy's Top 10 Technology Trends for 2023. We identified the typical technologies that have been engineered and are expected to be available for large-scale commercial use in the near future; the innovative products that have been proven in practice; and technological applications that already have robust value chain ecosystems. We hope that these technology trends will inspire and resonate with scientists, entrepreneurs, engineers, and technophiles. We seek to collaborate with all these people to promote technological innovation and achieve shared success in globalization.

Looking towards 2023, the advancement of various technologies will drive software/hardware co-design and the integration of computing and communications technologies. The wide application of technologies will facilitate the rollout of AI and other digital technologies in vertical markets and promote the collaboration of public and private sectors and individuals in security technology and security management. The innovation driven by the advancement of technologies and their industry-specific application has become an irreversible trend.

As our vision "Tech for the Future" indicates, new-generation ICT technology will create a better future where enterprises can enjoy high-quality development and people can enjoy better lives. We hope that the insights and ideas of DAMO Academy in scientific research will inspire everyone and contribute to scientific and technological development.

Jeff Zhang
Head of Alibaba DAMO Academy

Contents

Redefining Paradigms

Revolutionizing Industries

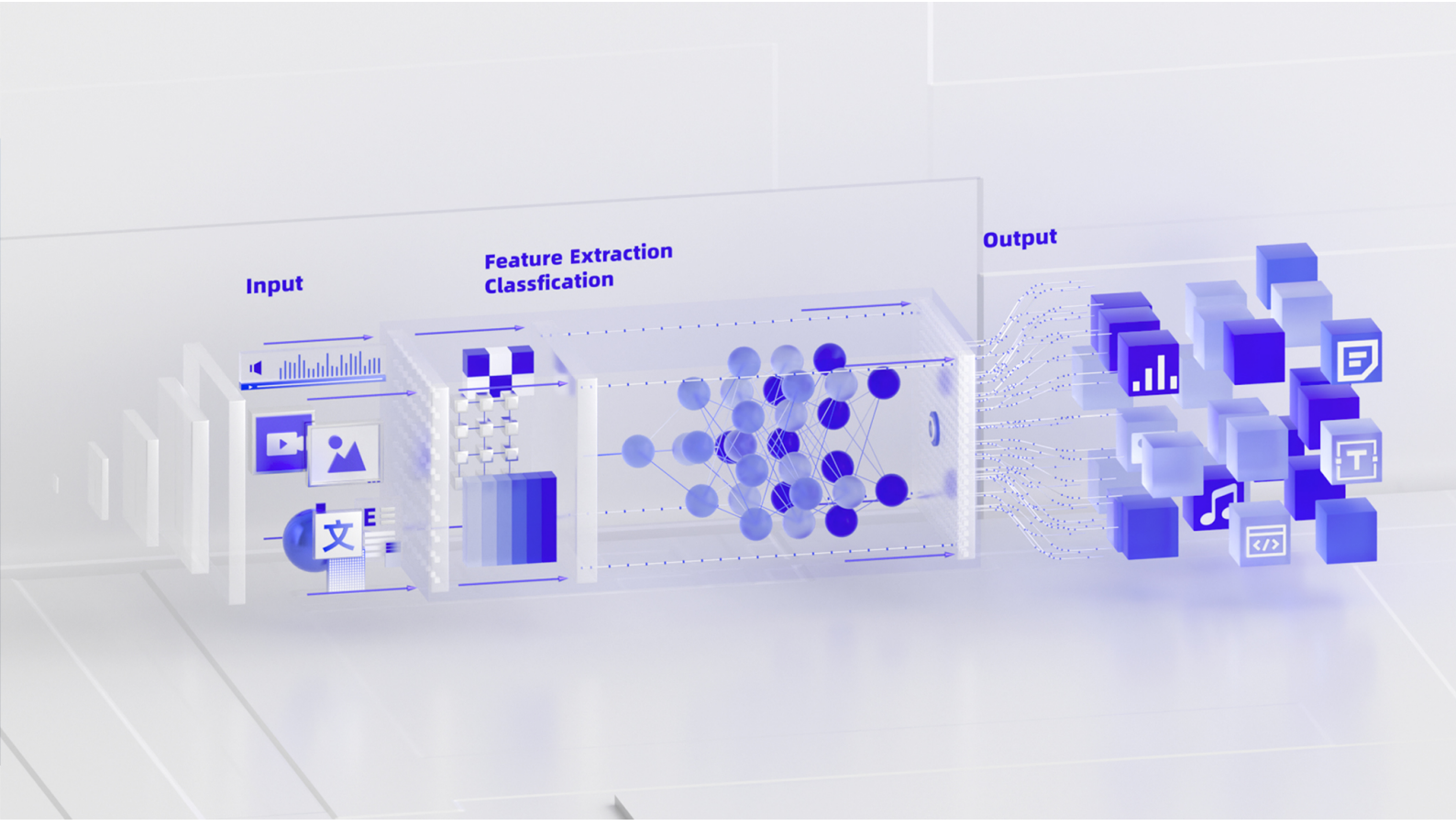
Incubating New Applications

Pre-trained Multimodal Foundation Models

Pre-trained multimodal foundation models have become a new paradigm and infrastructure for building artificial intelligence (AI) systems. These models can acquire knowledge from different modalities and present the knowledge based on a unified representation learning framework.

Introduction

AI systems are evolving from mono-modal systems that focus on either text, speech, or visuals, towards general-purpose systems. At the core of these general-purpose systems is multimodality. Multimodality helps endow AI systems with cross-modal semantic enhancements, combine different modalities, and standardize AI models. The emergence of cutting-edge technologies, such as Contrastive Language-Image Pre-Training (CLIP) and the general-purpose multimodal foundation model BEiT-3, breathes new life into the development of AI. The future of AI is to develop multimodal foundation models that can transfer knowledge across tasks or scenarios based on a wide variety of intelligence accumulated from all industries. In the future, foundation models are set to serve as the basic infrastructure of AI systems across tasks of images, text, and audio, empowering AI systems with cognitive intelligence capabilities to reason, answer questions, summarize, and create.



Trend Analysis

Multimodal model pre-training based on deep learning is the momentum that will propel AI towards cognitive intelligence. Pre-trained foundation models, which only need to be trained once and subsequently fine-tuned to handle different tasks or scenarios, will speed up the standardization of AI models and be used as the basic infrastructure for building more advanced AI systems. Combining sophisticated deep learning models, vast troves of Internet data, and the wide adoption of generative pre-training, we have seen remarkable progress in the implementation of AI models in the natural language understanding, speech processing, and computer vision sectors.

The release of the BEiT-3 multimodal foundation model in 2022 marks a breakthrough in the AI model technology. BEiT-3 achieves state-of-the-art transfer performance on vision-language tasks, primarily across visual question answering, image description generation, and cross-modal retrieval. BEiT-3 advances a large convergence in model framework and backbone architecture, which makes modality-specific encoding and the processing of vision-language tasks much easier. Meanwhile, CLIP continues to be widely used in multimodal model development. CLIP is a pre-trained model developed based on contrastive learning. It pairs image and text features to guide image generation by using Generative Adversarial Networks (GANs) or diffusion models. Stable Diffusion in the text-to-image sector also adopts CLIP to help fine-tune the model based on text-based hints and improve the quality of images with the help of diffusion models. Open source greatly facilitates the convergence of multimodality and the development of pre-trained models by narrowing the technical gap for beginners. Open source transforms foundation models from a new AI technology to a more sophisticated infrastructure that is widely adopted by developers.

The development of pre-trained multimodal models has not only reshaped the business model around AI, but also offers a rich variety of benefits towards daily life. For individual users, the emergence of multimodal models such as CLIP encourages more enthusiasts to put their innovative ideas into action without relying on tools and technical expertise. For enterprises, pre-trained multimodal models are the key to high productivity. Technology enterprises that are capable of providing big data computing capabilities (alongside model development capabilities and compute resources) will become model service providers, helping other enterprises integrate the capabilities of foundation models into production to improve the overall efficiency and reduce costs.

The development of cognitive intelligence will not be confined to mono-modalities such as text and images. The greatest challenge is building an efficient modeling framework and a unified backbone network to allow foundation models to better adapt to various tasks. Developers will also find new and exciting challenges as they explore the correlation between various modalities (such as image-to-text, text-to-natural-language, and video-to-text) and design pre-training jobs to train high-precision models.

The adoption of pre-trained speech, vision, and multimodal models is expected to accelerate the transition of AI models towards general-purpose foundation models. During this stage, deep learning and reinforcement learning can continuously benefit each other and integrate a large amount of industry knowledge at the same time. New models that can adapt to the ever-changing environment more easily will emerge. Multimodal foundation models with the capability to handle a wide variety of tasks and scenarios will become a paradigm of development across the field of AI. Continuous development of foundation models allows for the technology to outperform alternatives in terms of development costs, ease of use, development lifecycle, and performance, and also helps develop more business opportunities.

EXPERT OPINIONS | Pre-trained multimodal models can perform presentation learning by connecting text and images, and then transfer the knowledge to other modalities, such as speech and video. Pre-trained multimodal models outperform mono-modal models in terms of understanding, retrieval, generation, and question answering. Pre-trained multimodal models can acquire knowledge from a wide array of industries to perform presentation learning. Pre-trained multimodal models have become the most widely adopted foundation models across all industries.

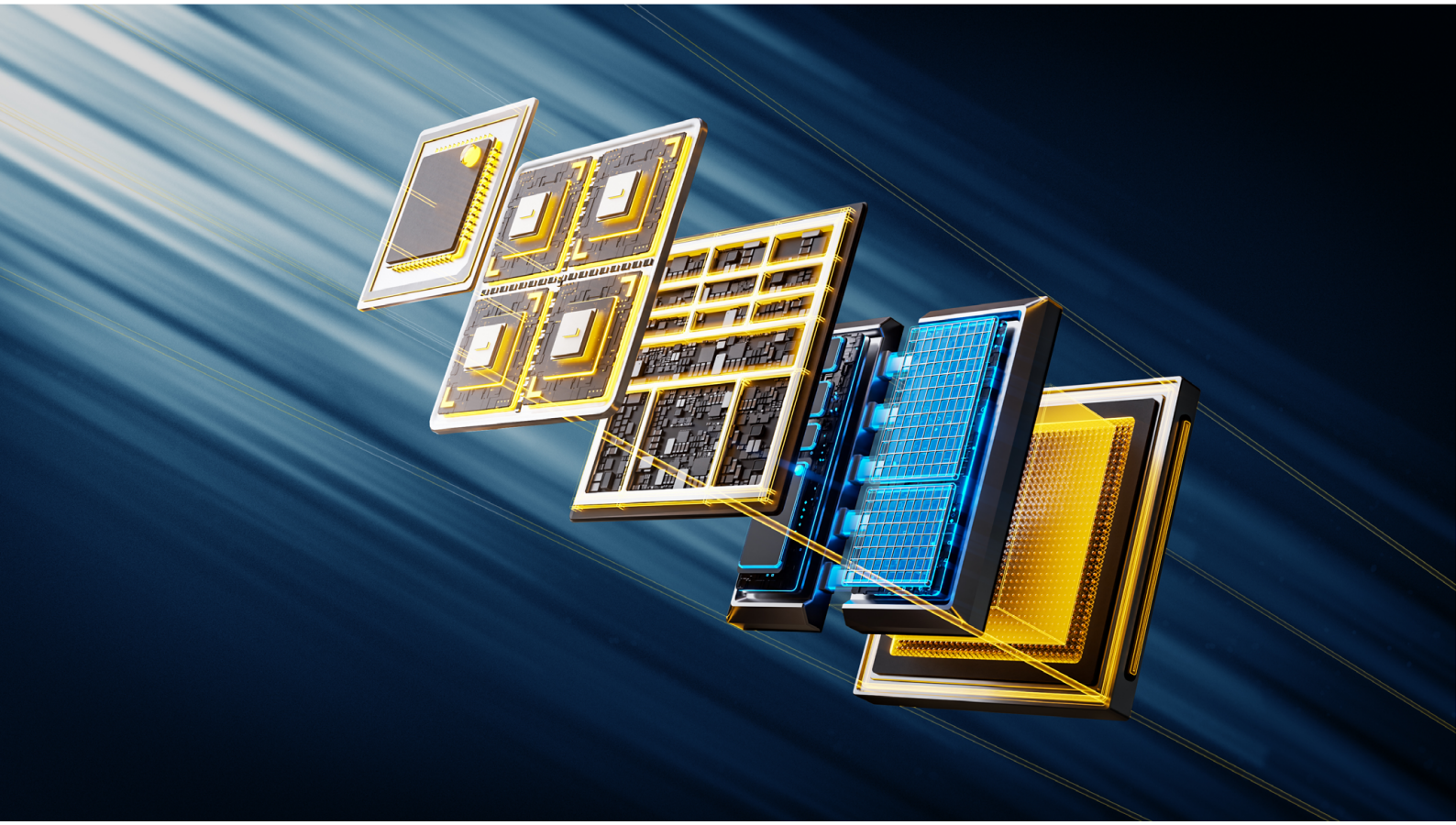
Fei Huang
Head of Language Technology Lab of DAMO Academy

Chiplet

The interconnect standards of chiplets will gradually converge into a unified standard, bringing in a new wave of change to the R&D process of integrated circuits (ICs).

Introduction

Chiplet-based design revolves around three core processes: deconstruction, reconstruction, and reuse. This method allows manufacturers to break down a System on a Chip (SoC) into multiple chiplets, manufacture the chiplets separately by using different processes, and finally integrate them into an SoC through interconnects and packaging. This method can significantly reduce costs and give rise to a new form of reusing intellectual property. As the integration and computing power of SoCs approach the physical limits, chiplets provide an important way to continuously improve the integration and computing power of SoCs. With the establishment of the Universal Chiplet Interconnect Express (UCIe) consortium in March 2022, the interconnect standards of chiplets are being unified into a single standard, accelerating the industrialization process of chiplets. Powered by advanced packaging technologies, chiplets may bring in a new wave of change to the R&D process of integrated circuits, from electronic design automation (EDA), design, and manufacturing to packaging and testing. This will reshape the landscape of the chip industry.



Trend Analysis

Since the advent of Moore's Law in 1965, the IC industry has been driven forward by the Moore's Law. Until recent years, as the size of transistors approaches the physical limits of miniaturization, further advancing manufacturing processes has yielded low returns and minimal performance gains. Against this backdrop, the chiplet technology is expected to become the second growth curve of the IC industry to further improve the performance and reduce the costs of chips.

Chiplets provide an innovative packaging method. This method allows manufacturers to break down an SoC into multiple chiplets, manufacture the chiplets separately by using different processes, and finally integrate them into an SoC through interconnects and packaging. For example, modules that stand the most to gain from advanced processes – such as CPUs and GPUs – are produced by using expensive, advanced processes, while modules that do not require advanced processes are produced by using cost-effective, mature processes. Chiplets are smaller than SoCs and can greatly improve the yield rate of chips and the utilization of wafers, which reduces the manufacturing and R&D costs of chips. In addition, modular chiplets can reduce repeated design and verification, reduce chip design complexity, and accelerate chip iteration. As a proven method to reduce the manufacturing costs of chips, chiplets have become a key area of focus among leading manufacturers and investors.

The core element of the chiplet technology lies in the high-speed interconnection between chiplets. The decomposition of SoCs into chiplets increases the difficulty in packaging. Therefore, the key to the chiplet technology is ensuring the reliability and universality of the chiplet interconnect and packaging processes. This helps increase the bandwidth and reduce the latency when data is transmitted between chiplets. However, advanced 2.5D and 3D packaging technologies have brought many complex physical problems such as electromagnetic interference, signal interference, heat dissipation and stress. These factors need to be considered during chiplet design and have since given birth to new requirements for EDA tools.

In recent years, advanced packaging technologies have developed rapidly. Through Silicon Via (TSV) is the key enabling technology for 2.5D and 3D packaging solutions and the maximum number of TSVs that can be used in a die is 1 million. It has proven that the advancement of packaging technologies can promote the application of chiplets in large chips such as CPUs and GPUs. In March 2022, several leading semiconductor companies jointly established UCIe to unify the chiplet interconnect standards and build an open ecosystem.

In the post-Moore era, the chiplet technology may be the most feasible technical route to address current challenges facing the IC industry. As a key area of focus of the IC industry, chiplets can reduce the industry's dependence on advanced processes and deliver high performance on par with advanced processes. From the perspective of yield on cost, 2D, 2.5D, and 3D packaging will exist side-by-side for a long period of time. Homogeneous and heterogeneous multi-chiplet packaging will also coexist for a long time. Different advanced packaging technologies and processes will be used together. Chiplets are expected to restructure the chip R&D process, from EDA, design, and manufacturing to packaging and testing, which will reshape the landscape of the chip industry.

EXPERT OPINIONS | The chiplet technology is an important technical means to improve chip integration, reduce chip costs, and make dies reusable. In the future, the chiplet technology will play an important role in numerous domains, such as high-performance computing and high-density computing. Advanced chiplet technologies will continue to be dominated by semiconductor manufacturers, and the combination of advanced packaging technologies such as 2D, 2.5D, and 3D will further improve the cost-effectiveness and competitiveness of chips.

Haiyang Wang
Vice President of Xiangdixian Computing Technology Co., Ltd.

Processing in Memory

The large-scale commercial employment of compute-in-memory chips in vertical markets is mainly driven by the growing industrial demand and capital investment.

Introduction

Processing in Memory (PIM) technology is the integration of a CPU and memory on a single chip, which allows data to be directly processed in memory. This not only reduces data transmission costs while improving computing performance, but also requires less energy to operate. PIM is an ideal solution for AI computing scenarios that require frequent, concurrent access to data. Driven by the growing industrial demand and capital investment, compute-in-memory chips are being rolled out from the production line and tested in real-world applications. These chips make use of static random-access memory (SRAM), dynamic RAM (DRAM), and Flash storage, and are targeted towards products that require low power consumption and low computing power, such as smart home appliances, wearable devices, robots, and smart security products. In the future, compute-in-memory chips are projected to be used in more powerful applications such as cloud-based inference. This will shift the traditional computing-centric architecture towards the data-centric architecture, which will have a positive impact on industries such as cloud computing, AI, and Internet of Things (IoT).

Trend Analysis

With the widespread application of AI in various sectors, neural network algorithms are becoming commonplace in today's world. Applications like deep learning models are required to be able to efficiently process large amounts of unstructured data, such as text, video, image, and audio. Traditionally, computers are built on the von Neumann architecture and consist of compute units and storage units. When computing tasks are performed, data is frequently migrated between the processor and memory. Computing performance is not only dependent on CPU performance, but also on the speed at which data can be accessed. This leads to "memory wall" and "power wall" challenges, which gives rise to demands for more powerful chips in terms of parallel computing, latency, and bandwidth.

The leading companies in the industry have been dedicated to advancing the research of cutting-edge technologies for PIM in recent years. Samsung published the world's first in-memory computing paper based on magnetoresistive RAM (MRAM) on *Nature*. TSMC published six papers on in-memory computing storage over IP (SoIP) on *ISSCC*, which greatly advanced the in-memory computing solutions based on resistive RAM (ReRAM). SK Hynix published its research on DRAM in-memory computing based on the graphics double data rate (GDDR) interface. The academic and industrial communities both believe that PIM is one of the effective methods to address performance and power consumption bottlenecks. Especially in large-scale parallel computing scenarios, such as VR/AR, autonomous driving, astronomy data computing, and analysis of remote sensing data, compute-in-memory chips hold distinct advantages in delivering high-performance bandwidth with low power consumption. From the standpoint of advancing computing power, factors such as latency, accuracy, and cost-effectiveness all play important roles. To optimize the computing architecture at the micro level, PIM requires support from the entire industry to overcome challenges in memory chip design and the optimization of production techniques.

PIM is mainly implemented by the following technical routes:

- Near-memory computing with high technical maturity. Near-memory computing uses advanced packaging technologies to package the computing chip with the memory. This shortens the path between memory and compute units, and improves I/O throughput to ensure high memory/bandwidth and low access costs. Near-memory computing is widely used in various CPUs and GPUs by using 2.5D packaging and 3D packaging.
- In-memory computing favored by the recent investor enthusiasm. In-memory computing is implemented by using traditional storage media, such as DRAM, SRAM, NOR Flash, and NAND Flash. Computing is done by independent compute units inside the memory chip/area, which is more suitable for scenarios that use fixed algorithms.
- New storage devices based on non-volatile memory (NVM). Many new data storage solutions that use NVM such as ReRAM are in the exploration stage. Other devices, such as phase-change memory (PCM) and MRAM, are also possible routes to achieve PIM. PIM makes use of both digital and analog computing technologies. Digital computing mainly uses SRAM as storage media, which has high performance and high accuracy and is more suitable for scenarios that require high computing power and low power consumption. Analog computing usually uses NVM such as Flash and ReRAM as storage media, which has high storage density and high parallelism and is more suitable for scenarios that require low computing power and low accuracy.

PIM has set off a wave of entrepreneurship in vertical markets and has garnered attention and investment from the capital market and industry. PIM is evolving to be more accurate, efficient, and cost-effective. Driven by the growing industrial demand and capital investment, in-memory computing that uses sophisticated storage media such as SRAM and NOR Flash will enter the stage of large-scale commercial use in vertical markets. Product and ecosystem upgrades may first take place for scenarios with low computing power and low power consumption. In-memory computing scenarios that require large computing power may enter the initial stage. New storage devices based on NVM depend on the process and yield improvements and require approximately 5 to 10 years to be available for commercial use.

EXPERT OPINIONS | PIM is a crucial technology in our pursuit of high-performance computing. The development of Internet of Everything (IoE) and AI in recent years has put PIM onto an accelerated track for commercialization, as the industry is still trying to explore and advance PIM technology. In the coming future, in-memory computing products will coexist in the form of System on a Chip (SoC) and chiplet. The application scenarios will continue to be expanded from IoT edge devices to general-purpose computing that requires high computing power. PIM is expected to become the mainstream computing architecture in the AI era.

Keyi Li
Founder & CEO of General Processor Technologies

Cloud-native Security

Security technologies and cloud computing become fully integrated, which boosts the development of new platform-oriented, intelligent security systems.

Introduction

The approach of cloud-native security represents a shift from perimeter-focused defense to defense-in-depth, and from add-on deployment to built-in deployment. Cloud-native security is implemented to not only deliver security capabilities that are native to cloud infrastructure, but also improve security services by leveraging cloud-native technologies. As a result, security technologies and cloud computing are becoming more integrated than ever before. We have witnessed applied technologies evolve from containerized deployment to microservices and then to the serverless model, and security services embrace the shift to become native, fine-grained, platform-oriented, and intelligent.

- Native security services mean embracing the shift-left security approach to build a product security system. The system integrates product R&D, security, and O&M, boosting the collaboration among the R&D, security, and O&M teams.
- Fine-grained security services support precise access control and dynamic policy configuration while providing unified authentication and configuration management capabilities.
- Platform-oriented security services use a hierarchical in-depth defense system and a platform integrated with security products to implement precise proactive defense, eliminating the need to deploy a number of isolated security products.
- Intelligent security services are driven by security operations, and provide the following features to deliver end-to-end protection for applications, cloud products, and networks: real-time monitoring, precise response, quick attack tracing, and threat hunting.

Trend Analysis

Cloud computing is widely adopted across the industrial landscape, and is used to some extent by almost all industries. As such, these industries urgently need to upgrade their security systems to become more effective and versatile to respond to the challenges in the era of cloud computing, such as fast iteration, auto scaling, and big data processing, and to protect dynamic complicated runtime environments.

During the implementation of cloud-native security, security systems are optimized and restructured based on the cloud-native approach and technologies, and security services become lightweight, agile, fine-grained, and intelligent. Meanwhile, security capabilities that are native to cloud infrastructure are developed and further enhanced. Cloud-native security features end-to-end DevSecOps, unified identity and permission management, a platform-oriented in-depth defense system, and a comprehensive security system that supports real-time visualization, management, and control.

The development of cloud-native security includes multiple stages. At the beginning, security services protect cloud-native products. Gradually, these security services become more powerful as they leverage cloud-native technologies. Eventually, a protection system that covers infrastructure, applications, data, R&D, testing, and security operations is born. New cloud-native protection measures, such as the cloud-native application protection platform, digital forensics and incident tracing, extended threat detection and response, and attack surface management, have emerged, and are rapidly evolving. These measures are now accepted and recognized across the industry.

Cloud-native security provides the following benefits to service management, service operations, and users:

- End-to-end observability and risk control. Security and compliance requirements are met across the stages of software development and provisioning, and scans and checks are performed for all important processes. This way, potential risks are avoided, and overall risk control costs are minimized.
- Closed-loop, efficient management of infrastructure. Security features are converged, which helps achieve closed-loop management of issue response and handling. The automated application of policies reduces dependency on security operations personnel and lowers the risk of human error. In addition, automated attack blocking allows for more time to respond to the attacks and fix issues.
- Comprehensive asset protection on the cloud. All types of data assets are monitored in real time. Flexible diversified security services are available for authentication, configuration management, application runtime monitoring, and data security.

In practice, the development of cloud-native security is filled with challenges and obstacles. For example, how can we quickly and efficiently collect monitoring data of data assets from complex heterogeneous environments and aggregate the data? How can we clarify rights and responsibilities of different parties on the cloud and build an open collaborative ecosystem?

In the next three to five years, cloud-native security will become more versatile and can adapt more easily to multi-cloud architectures. It will also become more conducive to building security systems that are dynamic, end-to-end, precise, and applicable to hybrid environments. New governance systems and professionals will be in place. In terms of protection effectiveness, smooth user experience will be delivered and ensured based on fine-grained access control, authentication, data security management, and automatic risk identification and handling that are provided on top of intelligent technologies. Innovative cloud security services will be provided. Cloud native approaches such as security hosting and improving defense capabilities through attack and defense drills will become mature and an integral part of the native security system.

EXPERT OPINIONS | Making accurate predictions about technology trends is quite difficult. However, making predictions about network security technology, which is considered reactive technology, is not as difficult as you think. The candidates for the top technology trends in network security technology include the changes in cryptographic technology in a post-quantum era, the future of trusted privacy computing, and the changes in network attack-defense technology. After careful consideration and research, we have selected cloud-native security as one of top ten technology trends. Network security technology is reactive and future-facing, and these characteristics are manifested in the decision-making process. For example, traditional technology such as cryptographic technology must be advanced to prepare for the challenges brought by quantum computing in the post-quantum era, although the widespread adoption of quantum computing seems light years away. Similarly, cloud-native security will attract a lot of attention in the network security field in 2023 although it is quite new when compared with cryptographic technology.

Qibin Zhai

Professor at State Key Laboratory of Information Security, Chinese Academy of Sciences

The scope of cloud-native security is not limited to the security of cloud-native products. Cloud native security refers to the elastic, unified, and intelligent security capabilities that are native to cloud infrastructure and are provided by leveraging the native capabilities of the cloud.

If you are a cloud service provider, you need to ensure the security of both your infrastructure and your cloud products. You also need to clarify the responsibilities that you share with your customers based on your service model. If you are a security service provider, you need to make full use of the advantages provided by the cloud, and provide your customers with integrated and comprehensive products and services. Moreover, all relevant parties need to be more open and collaborative because cloud-native security involves a wider range of fields and complicated technological stacks. Openness and collaboration will aid in the development of more comprehensive and efficient security products and services that are observable, manageable, and controllable. I believe that in the next one to two years, the focus of the industry will shift towards identity security on the cloud and intelligent security operations centers.

Xin Ouyang

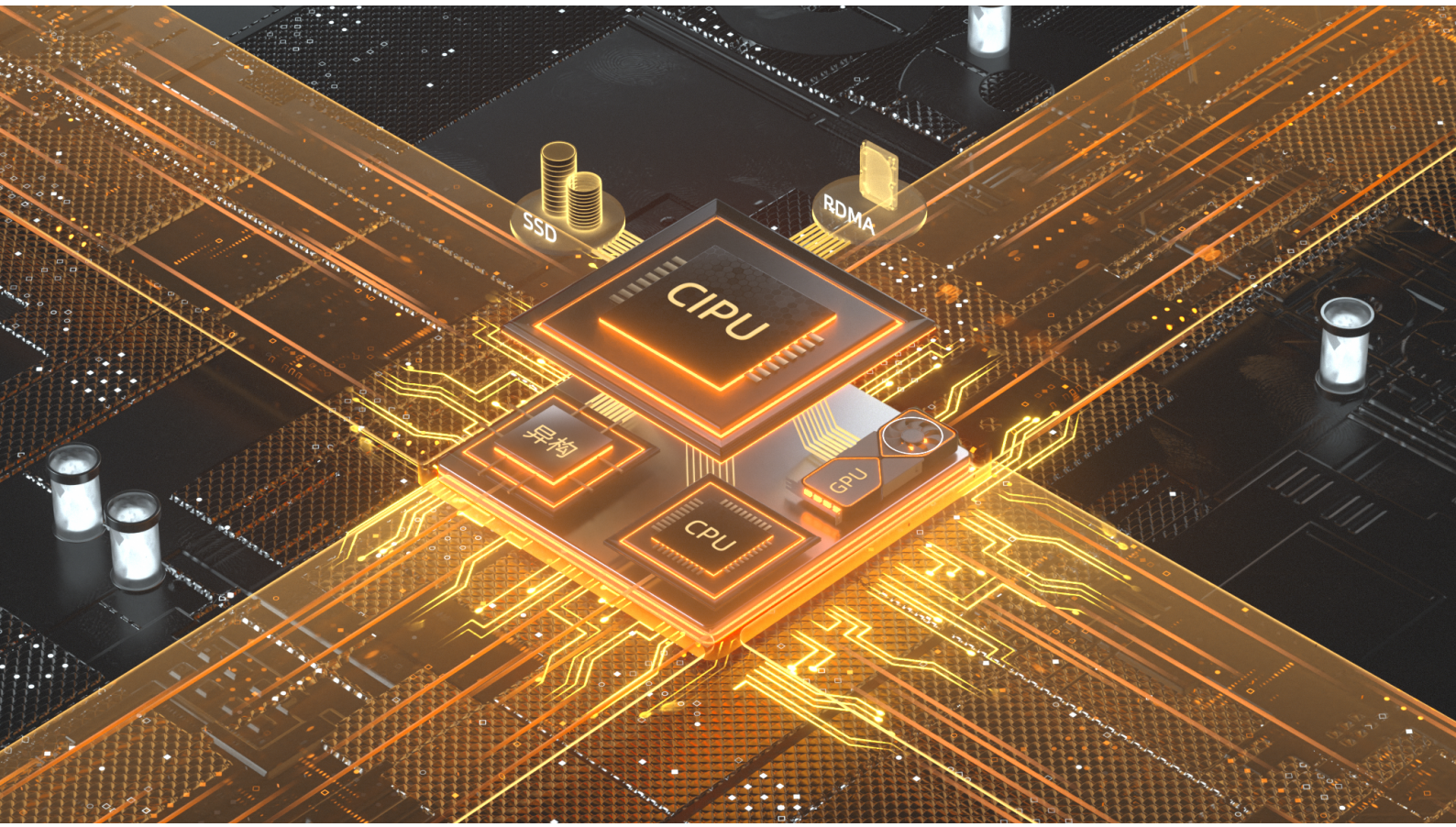
Chief Risk Officer of Alibaba Cloud, General Manager of Alibaba Cloud Security Services Business Unit

Hardware-Software Integrated Cloud Computing Architecture

Cloud computing is evolving towards a new architecture centered around CIPU. This software-defined, hardware-accelerated architecture helps accelerate cloud applications while maintaining high elasticity and agility for cloud application development.

Introduction

Cloud computing has evolved from the CPU-centric computing architecture to a new architecture centered around Cloud Infrastructure Processor (CIPU). This software-defined, hardware-accelerated architecture helps accelerate applications while maintaining high elasticity and agility for cloud application development. The new architecture incorporates both hardware infrastructure and software systems. As a processing unit, CIPU enables acceleration of hardware resources on the cloud, including computing, storage, and network resources. CIPU-accelerated hardware resources can connect to the distributed operating system through the cloud resource controller for flexible management, scheduling, and orchestration. CIPU will become the de facto standard of next-generation cloud computing and bring new development opportunities for core software R&D and dedicated chip design.



Trend Analysis

In the post-Moore era, CPU performance struggles to keep up with the exponential growth of data needs and deal with data-intensive computing scenarios brought on by big data and AI applications. Against this backdrop, the CPU-centric cloud computing architecture becomes insufficiently equipped to tackle increased latency and bandwidth requirements. The shift from the traditional CPU-centric cloud computing architecture to an integrated software and hardware infrastructure is imperative.

Cloud computing architecture has gone through three stages of development, and has managed to withstand ultra-high concurrency and deliver enormous computing power. Around 2010, we underwent the first stage of development. Distributed architecture built on x86 servers and Internet middleware dominated the cloud computing world. The second stage came around 2015. At this stage, cloud vendors built software-defined virtual private clouds (VPCs) and resource pool architecture in which computing and storage resources are decoupled so that computing, storage, and network resources can be separately scaled on demand.

At present, cloud computing has entered the third stage where virtualized workloads are handled by dedicated hardware to enable full hardware acceleration. The challenge at this stage is to offer more cost-efficient computing and infrastructure management in response to the growing volumes of east-west traffic in cloud data centers. Ultimately, improvement in computing efficiency comes down to better chipsets at the hardware level.

The CIPU-centric cloud computing architecture has made the following breakthroughs in engineering implementation:

First, this architecture centrally manages the underlying hardware infrastructure in a fine-grained manner to enable full hardware acceleration. Coupled with the cloud operating system, CIPU delivers improved networking, storage, security, and computing performance, turning a data center into a high-speed bus-connected supercomputer. CIPU provides lower latency and higher throughput for network and storage access.

Secondly, eRDMA is implemented on top of this end-to-end hardware acceleration stack. This groundbreaking innovation enables large-scale networking and does not require modification of users' application code, making high-performance computing on the cloud much more accessible. Finally, under the new hardware architecture, CIPUs and servers can be mixed-and-matched to suit almost any requirement, efficiently addressing the requirements for flexible distribution of east-west traffic in different computing scenarios.

While retaining software-defined hardware, the hardware-software integrated cloud computing architecture offers the agility and flexibility of the distributed architecture as well as the elasticity, reliability, and availability of the resource pool architecture. It also significantly improves computing performance, enabling comprehensive acceleration of cloud applications.

In the next three years, cloud computing will evolve toward a new architecture centered around CIPU. Cloud services such as Function-as-a-Service (FaaS), container services, database services, big data services, and AI-inspired services will also be fully accelerated by CIPU. CIPU's hardware acceleration capability will benefit users throughout their cloud computing experience, from purchasing cloud resources to using the provided cloud services. From resources to services, the core value of cloud computing services will largely depend on the underlying computing power and computing efficiency that cloud vendors can provide.

EXPERT OPINIONS | Hardware-software co-design is a major strategy for creating next-generation computing architectures. In cloud computing scenarios, which bear higher complexity, collaborative optimization and iterative upgrade of software and hardware are especially important. The CIPU developed by Alibaba Cloud is a representative result of the hardware-software co-design approach. It can be used together with the company's Apsara OS to efficiently manage data center infrastructure resources, including computing, storage, and network resources, and to collaboratively accelerate applications on the cloud. Alibaba Cloud's CIPU is a testament to the feasibility of a virtualized hardware-software-integrated computing architecture and will lead the technological advancement in the cloud computing industry.

Ju Ren

Associate Professor, Department of Computer Science and Technology, Tsinghua University

CIPU is a culmination of numerous innovations. Implementing RDMA on the cloud is a crucial capability of CIPU. CIPU can provide significant hardware acceleration across inclusive, distributed, large-scale, and high-performance VPCs, offering 20% to 80% performance improvements for cache, database, big data, and AI services with no code modifications required.

Thanks to the continuous architectural innovations, cloud computing is quickly overtaking on-premises servers in terms of the amounts of computing power provided. Simply by migrating to the cloud, enterprises can receive the ever-growing benefits of cloud computing resources or cloud services at low costs.

Linquan Jiang

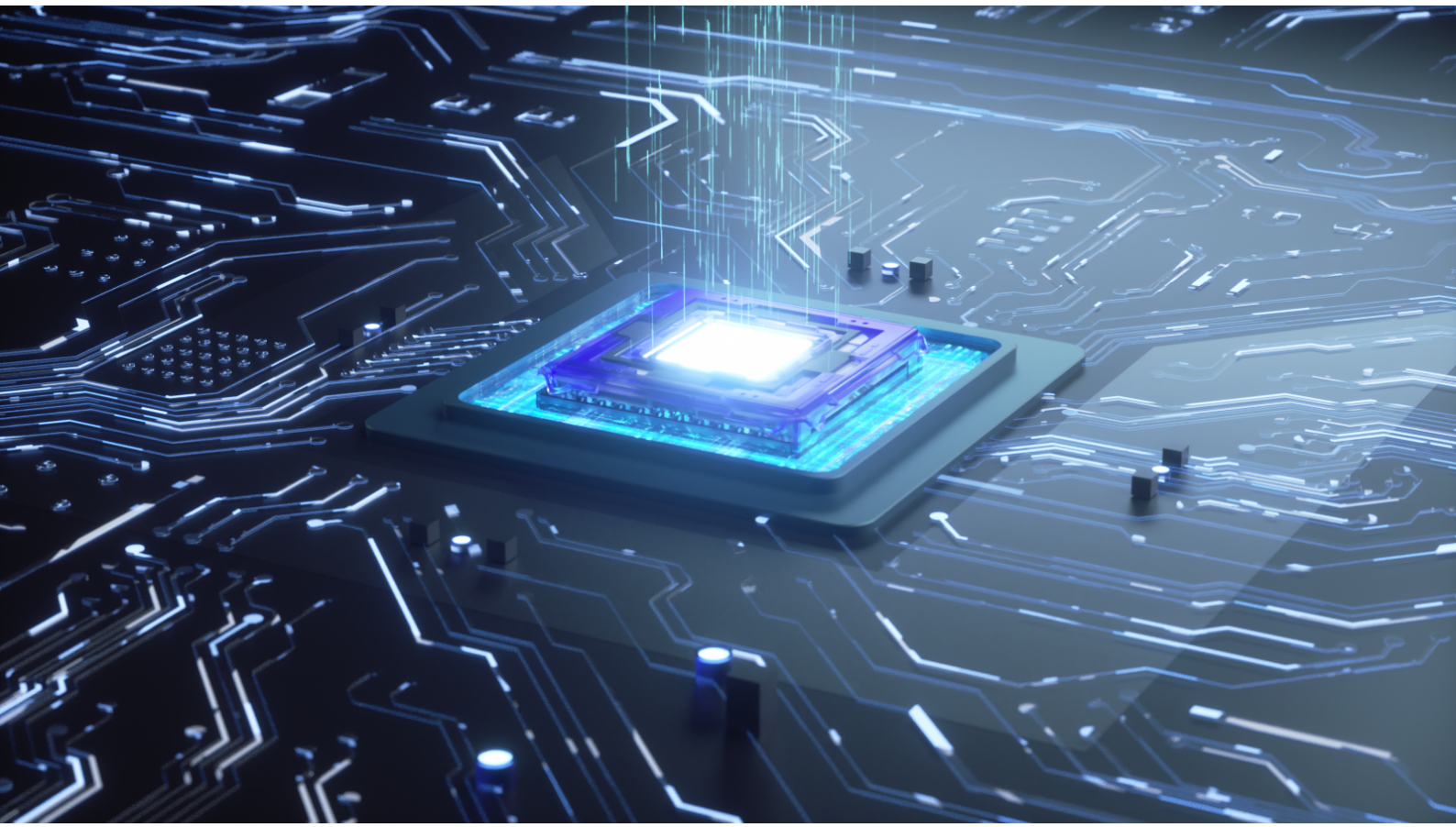
Alibaba Cloud Researcher, Head of Alibaba Cloud Shenlong Compute Platform

Predictable Fabric

Cloud-defined predictable fabric featuring host-network co-design is gradually being adopted from data center networks to wide-area cloud backbone networks.

Introduction

Predictable fabric is a host-network co-design networking system driven by advances in cloud computing, and aims to offer high-performance network services. It is also an inevitable trend as today's computing and networking capabilities gradually converge on each other. On the cloud service provider's side, predictable fabric paves the way for computing clusters to expand at scale through high-performance network capabilities, allowing the formation of large resource pools with immense computing power. On the consumer's side, predictable fabric unlocks and delivers the full potential of computing power in large-scale industrial applications. Predictable fabric not only supports emerging large computing power and high-performance computing scenarios, but also applies to general computing scenarios. It represents the industrial trend of traditional network and future network integration. Through the full-stack innovation of cloud-defined protocols, software, chips, hardware, architecture, and platforms, predictable fabric is expected to subvert the traditional TCP-based network architecture and becomes part of the core network in next-generation data centers. Advances in this area are also driving the adoption of predictable fabric from data center networks to wide-area cloud backbone networks.



Trend Analysis

Predictable fabric is a host-network co-design networking system driven by advances in cloud computing, and aims to offer high-performance network services. It is also an inevitable trend as today's computing and networking capabilities gradually converge on each other. On the cloud service provider's side, predictable fabric paves the way for computing clusters to expand at scale through high-performance network capabilities, allowing the formation of large resource pools with immense computing power. On the consumer's side, predictable fabric unlocks and delivers the full potential of computing power in large-scale industrial applications. Predictable fabric not only supports emerging large computing power and high-performance computing scenarios, but also applies to general computing scenarios. It represents the industrial trend of traditional network and future network integration.

The essence of networking is connectivity. High bandwidth, low latency, high stability, and minimal jitter have always been the holy grail of networking. Although the traditional TCP protocol stack has been widely deployed in the Internet, the network bandwidth and service quality cannot catch up to the increasingly rigorous bandwidth and connection quality demands of today's data center networks.

Traditional networks are based on two ideas: 1) decoupled host and network design; 2) opaque network device-based "best effort" architecture. This design is unable to meet the high-performance bandwidth requirement of today's large computing power pools. This requirement becomes the driving force behind the "predictable" high-performance network architecture, and presents new challenges to the traditional "best effort" network architecture.

The core of predictable fabric is to deliver immense amounts of computing power. To this end, predictable fabric embraces the host-network co-design architecture instead of host-storage-network decoupling architecture adopted by traditional networks. Specifically, the host-network co-design architecture innovatively adopts the idea of synergistic component design and deep integration on both the host side and the network side. Alibaba Cloud has proposed and is building a new-generation computing-network architecture based on a series of novel technologies such as the transport protocols specific to host-network co-design, congestion control algorithms, multi-path intelligent scheduling, deeply customized chips, and programmable hardware. Predictable fabric technology is able to significantly improve the network communication efficiency of distributed parallel computing, thus allowing for the formation of efficient computing resource pools and achieving the flexible supply of large computing power on the cloud. The cloud-defined predictable fabric will have a profound effect on the upstream and downstream of the industrial chain and the evolution of chip technology, becoming a new paradigm in the delivery of computing power.

The computing power network is a concept based on the delivery of computing power as a utility. It is currently taking shape as Internet service providers and cloud service providers define their needs and required capabilities. When the computing power network evolves to become a key part of the nation's infrastructure, there is no doubt that the demand for predictability will become the number one priority. Even as we develop as a digital society, our basic needs will drive data center network technology to deliver even greater performance. This in turn will place demands on resource-pooled cloud computing, and drive advancements in predictable fabric technology. In the next two to three years, we can predict with confidence that predictable fabric will become a mainstream trend in technology.

EXPERT OPINIONS | The most significant change in the network industry over the past decade is how the leading web companies have redefined the network system in response to requirements on large-scale computing power. The protocol-centric classic network was replaced by the software-defined network with the white box solution. Driven by the demand for large-scale computing power, cloud computing will once again redefine the high-performance network system for the next decade. In the network ecology, data processing unit (DPU) chips that integrate computing and networks are being adopted across the board, kicking off a new round of technological evolution. The network is bound to evolve toward predictability via host-network co-design.

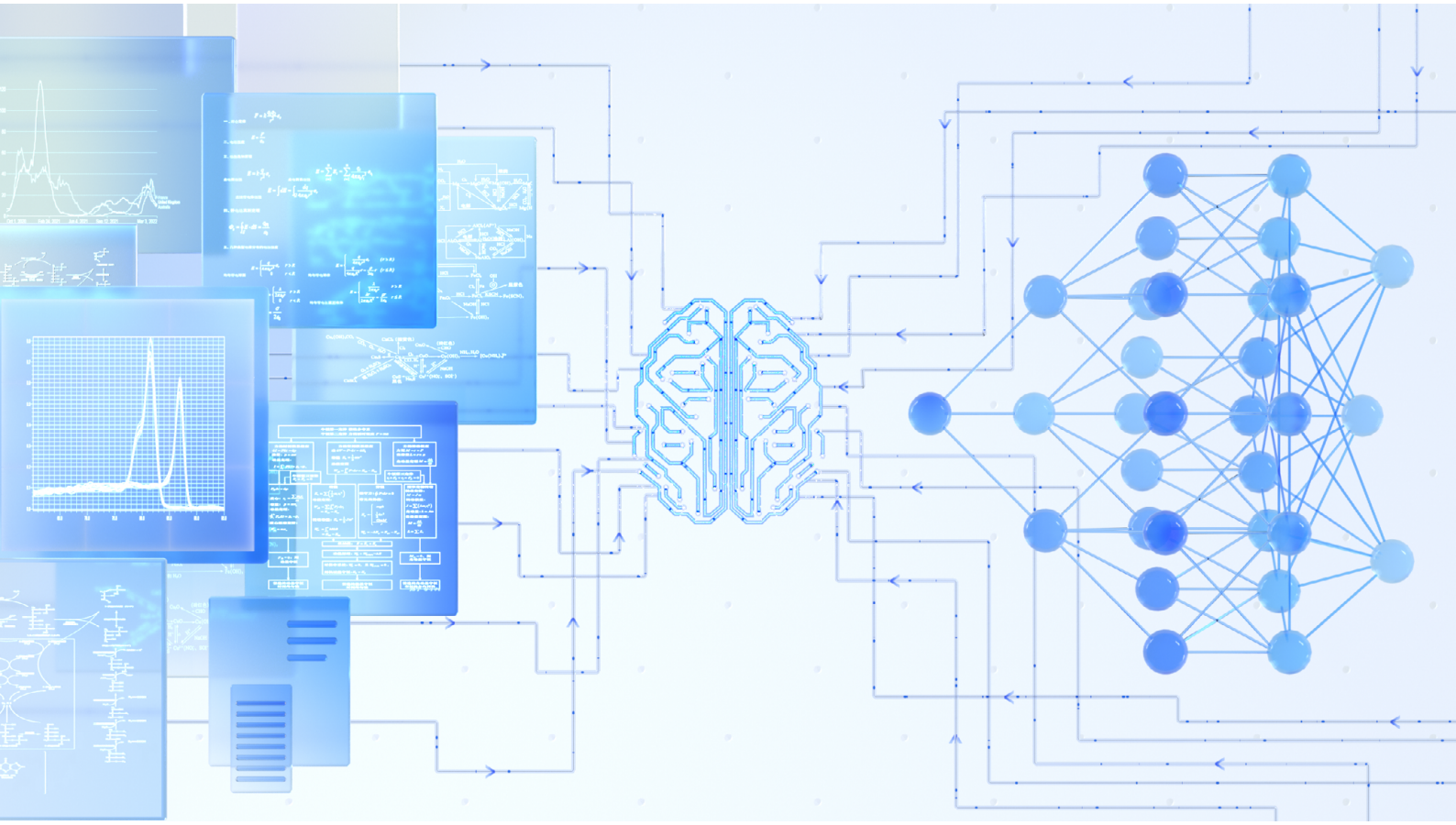
Dennis Cai
Vice President of Alibaba Cloud Intelligence and Head of the Network Infrastructure Department

Dual-engine Decision Intelligence

Decision intelligence supported by operations optimization and machine learning will facilitate dynamic and comprehensive resource allocation.

Introduction

In today's fast-changing world, enterprises must be able to take quick actions and make informed business decisions. The traditional decision-making method is based on Operations Research, which combines mathematical models built for real-life situations and algorithms for Operations Research optimization to find the best solutions under multiple constraints. However, as the world around us becomes more complex, this method is becoming less and less useful, due to its limitations in handling problems with great uncertainty and its slow response to large-scale problems. Therefore, academia and industry began to introduce machine learning into decision optimization, building dual-engine decision intelligence systems that utilize both mathematical and data-driven models. The two engines are perfect complements to each other. When used in tandem, they can improve the speed and quality of decision making. This technology is expected to be widely used in a variety of scenarios to support dynamic, comprehensive, and real-time resource allocation, such as real-time electricity dispatching, optimization of port throughput, assignment of airport stands, and improvement of manufacturing processes.



Trend Analysis

In recent years, events like pandemics, wars, and supply chain disruptions are on the rise, adding to the complexity and uncertainty on the global stage. Amid the turbulence, markets are evolving faster, posing greater challenges to enterprises. To adapt to the fast-changing world, enterprises must take quick actions to make optimal business decisions.

Decision intelligence is a process where intelligent tools and technologies are used to model and analyze data and distill the results into guidance on making optimal decisions towards an objective. This process takes into account the constraints, strategies, preferences, and objectives, and runs them through mathematical models to arrive at an optimal decision. Decision intelligence is designed to solve complex problems that may change on the fly, such as how ride-hailing platforms dispatch drivers, how to select sites for charging stations, and how factories can schedule production for minimal downtime.

Traditionally, decision intelligence equals decision optimization, which is a branch of Operations Research that can be traced back to optimizing air combat tactics during World War II. It combines mathematical models built for real-life situations, and algorithms for Operations Research optimization to find the best solutions under multiple constraints. This process requires only a relatively small amount of data but produces high-quality results that are highly interpretable and actionable. Therefore, it has been widely used to aid decision making.

However, as real-life situations become more complex, this method is becoming less and less useful, due to its limitations in handling problems with great uncertainty and its slow response to large-scale problems. Thus, academia and industry began to introduce machine learning into the decision optimization process, building dual-engine decision intelligence systems that utilize both mathematical and data-driven models. The machine learning method is based on data-driven models, which can narrow down the candidate solutions and boost the efficiency of decision making. This method is ideal for scenarios that have a high degree of uncertainty and require quick responses. However, it also has its limitations. It requires a considerable amount of investment in its learning phase, and the quality of the decisions it delivers may not be up to par when compared to the traditional method. Therefore, the operations research optimization and machine learning methods are used in tandem as perfect complements for each other, enabling responsive and high-quality decision making.

Dual-engine decision intelligence is still in its infancy, but is quickly gaining influence in various scenarios, such as optimization of port throughput, assignment of airport stands, improvement of manufacturing processes, and sales and operation planning. Among these scenarios, the most typical and challenging one is electricity dispatching.

- Objective: While ensuring stability of the power grid system, lower the cost of electricity, maximize social welfare, and boost the utilization efficiency of renewable energy.
- Constraints: 1) safety constraints, including the limits on nodal voltage and on the thermal stability of lines and sections; 2) constraints on load balancing between power generation and consumption; 3) physical characteristics, such as load ramping, start-stop curve, and cascading hydropower plants.
- Challenges: 1) Electricity dispatching is a highly complex operation that involves massive amounts of data. Even at the provincial level, tens of millions of variables and constraints are involved. By the time China's carbon peaking and carbon neutrality goals are achieved, the number is expected to reach over a billion, due to the rapid expansion of clean energy and the introduction of dispatching on the load side; 2) Renewable energy is taking a larger share of all electricity produced, thus its volatility and randomness will bring great challenges to decision optimization based on mathematical models; 3) Machine learning may not be able to ensure that all safety constraints are met.

Dual-engine decision intelligence couples machine learning and the underlying optimization technologies. While meeting all safety constraints, it can deliver ten times higher efficiency and is expected to shrink the response time required for dispatch to seconds, thereby breaking the performance bottleneck for dispatching electricity in the new power system that involves large numbers of clean energy sources.

In the future, dual-engine decision intelligence will be applied in more scenarios. It will serve to increase the number of entities and expand the scale in regional resource allocation scenarios, and eventually achieve dynamic, comprehensive, and real-time resource allocation.

EXPERT OPINIONS | In recent years, carbon reduction programs across the world are gaining popularity, and clean energy devices, such as solar panels, wind turbines, batteries, microgrids, charging stations, and electric vehicles, are being put into widespread use. They have resulted in new requirements for energy management during generation, allocation, and consumption, such as multi-energy complementarity, peak shaving and valley filling, prediction-based optimization, and flexible charging and discharging.

The new decision intelligence system for energy management will integrate information from all sides, including sources, grids, loads, and storage. This allows the system to quickly analyze the massive amounts of energy-related data. It will implement automatic optimization and real-time response to improve resource allocation both at the global and regional scale. It will transform the traditional extensive energy management systems into intensive and intelligent ones.

Intelligent energy management systems will lead to remarkable technological advancements, bringing fresh application possibilities and business opportunities that can influence people's everyday life. Dual-engine decision intelligence systems will continuously improve the efficiency of energy management, laying a solid foundation for the fulfilment of China's carbon peaking and carbon neutrality goals.

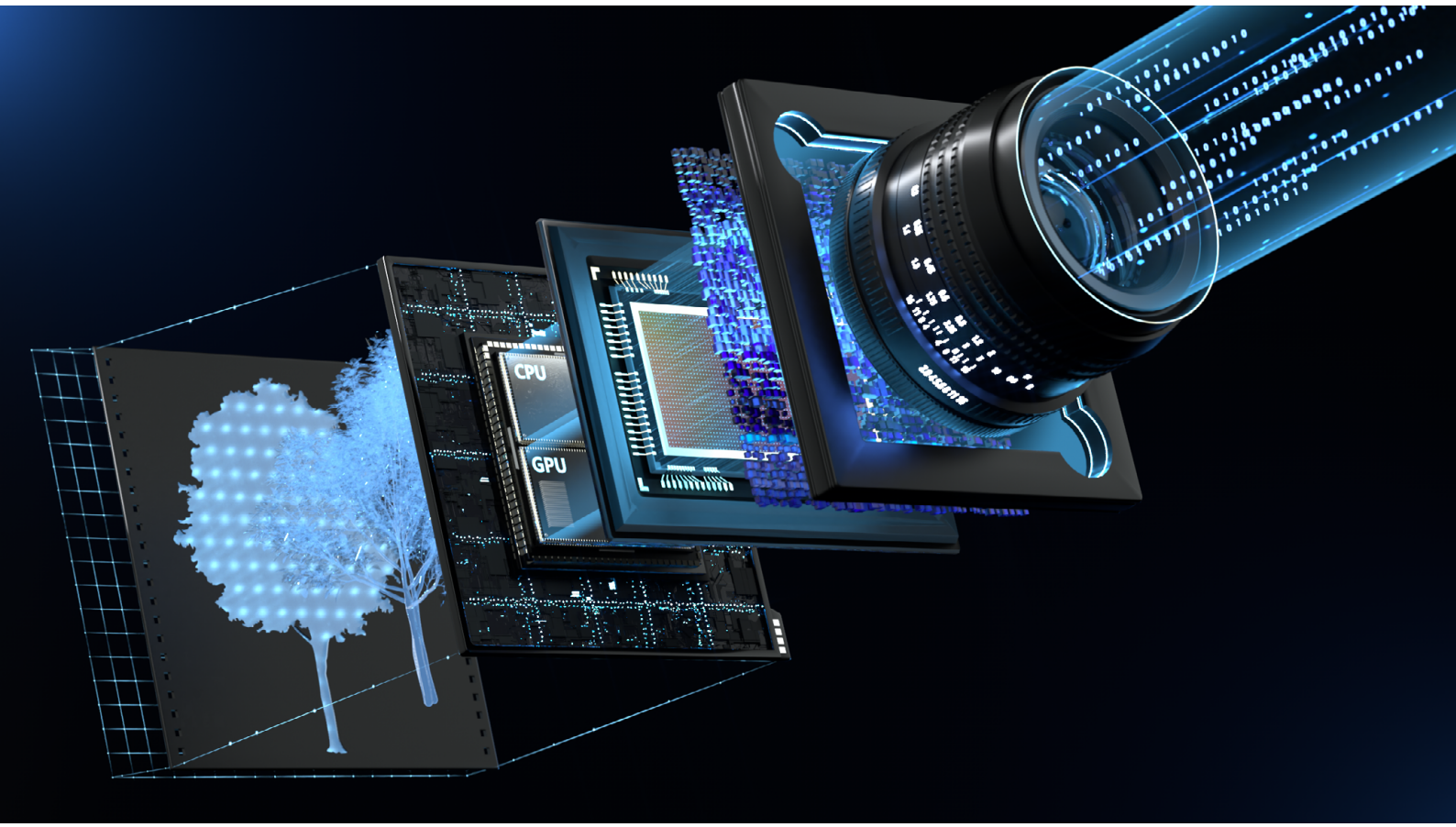
Yuxiang Luo
Partner of ESG - Sustainable Value Chains, PwC

Computational Imaging

Computational imaging goes beyond the limits of traditional imaging, and brings about more innovative and imaginative applications in the future.

Introduction

Computational imaging is an emerging interdisciplinary technology. This application-oriented technology collects and processes multidimensional optical information that cannot be detected by human eyes, such as lighting angles, polarization, and phases. It is a complete rethinking of the optical-based imaging system, and a new paradigm in the sensor technology. In contrast with traditional imaging techniques, computational imaging makes use of mathematical models and signal processing capabilities, and thus can perform unprecedented in-depth analysis on light field information. Computational imaging has been developing rapidly, with many promising research results. This technology has also been used on a large scale in areas such as mobile phone photography, health care, and autonomous driving. In the future, computational imaging will continue to revolutionize traditional imaging technologies, and bring about more innovative and imaginative applications such as lensless imaging and polarization imaging.



Trend Analysis

Traditional imaging is reliant on the principles of geometrical optics, and captures optical information in a similar way as human eyes do to produce an isomorphic image. Due to physical limitations on hardware functions and imaging performance, traditional imaging techniques are unable to obtain multidimensional optical information and deliver high-quality images that are required by latest application scenarios. Examples of this can be seen in everyday life. Mobile phones rely on bulky optical components to ensure imaging performance, resulting in the much-criticized design of notches and protruding lenses. Optical microscopes that use traditional imaging can only either focus on a single point or frame a larger field at lower resolution. In remote sensing or monitoring, images delivered by traditional imaging tend to be blurry in environments that have inferior lighting conditions or restricted visibility.

As next-generation information technologies such as sensors, cloud computing, and artificial intelligence evolve side-by-side, computational imaging becomes an attractive replacement for traditional imaging. This application-oriented technology collects and processes multidimensional optical information that cannot be detected by human eyes, such as lighting angles, polarization, and phases. It is a complete rethinking of the optical-based imaging system, and a new paradigm in the sensor technology. In contrast with traditional imaging techniques, computational imaging makes use of mathematical models and signal processing capabilities, and thus can perform unprecedented in-depth analysis on light field information. Figure 1 shows the different working mechanisms of traditional imaging and computational imaging.

Computational imaging is an emerging interdisciplinary technology that was proposed and defined in mid-1970s. The rapid advancement in information technologies brings computational imaging under the spotlight. No consensus has been reached on the classification of extensive computational imaging researches. However, computational imaging technologies can be categorized by objectives into the following types:

- Function enhancement: Computational imaging captures or measures optical information that traditional imaging cannot obtain, such as the light field, polarization, and coherence.
- Performance improvement: Computational imaging improves the performance of traditional imaging in metrics including spatial resolution, temporal resolution, depth of field, and robustness in complex environments.
- Simplification and intelligentialization: Computational imaging applies single-pixel imaging, lensless imaging, and some other technologies to simplify the imaging system. Computational imaging can also be used to complete specific AI tasks by leveraging the speed of light.

Figure 2 shows the categories of computational imaging technologies.

Computational imaging has been evolving rapidly, and facing many challenges. Firstly, the optical system must be redesigned to place sensors at its core. Secondly, new optical components and a light field control mechanism are required to capture multidimensional optical information. This in turn leads to higher hardware costs, and longer cycles of R&D and

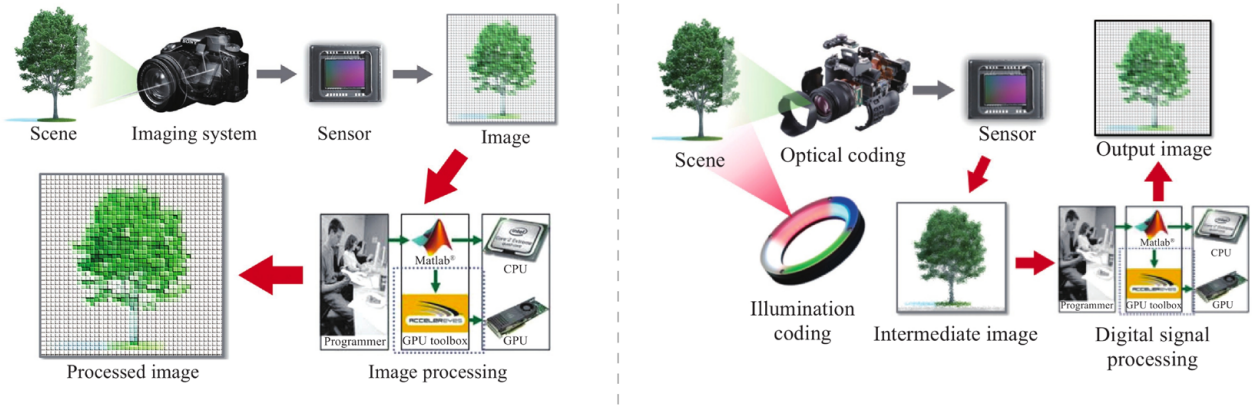


Figure 1: Working mechanisms of traditional imaging (left) and computational imaging (right)

debugging. In addition, new algorithms must be developed to better coordinate hardware and software. Last but not least, computational imaging depends heavily on computing power, which poses high requirements on the performance and adaptability of device chips.

Despite being an emerging technology, computational imaging has become a star in scientific researches and even contributed to Nobel-winning techniques such as super-resolved fluorescence microscopy (2014 Nobel Prize in Chemistry) and cryo–electron microscopy (2017 Nobel Prize in Chemistry). Computational imaging has been used on a large scale in areas such as mobile phone photography, health care, monitoring, industrial inspection, and autonomous driving. For example, in mobile phone photography, major mobile phone manufacturers have introduced the idea of computational imaging and combined compact hardware and algorithms. Many of today's mobile phones can deliver high-quality images that are comparable to and even exceed those taken by single-lens reflex cameras (SLRs).

In the future, computational imaging will continue to revolutionize traditional imaging technologies, and bring about more innovative and imaginative applications. The meta-imaging chip can be used to perceive three-dimensional optical information in a wide field of view with no optical aberrations, which may one day make the protruding rear lenses of mobile phones a thing of the past. Lensless imaging such as FlatCam simplifies and minifies the lens-based imaging system of cameras, which allows imaging systems to be integrated into various wearables. Polarization imaging can be used to construct clear images of scenes with restricted visibility.

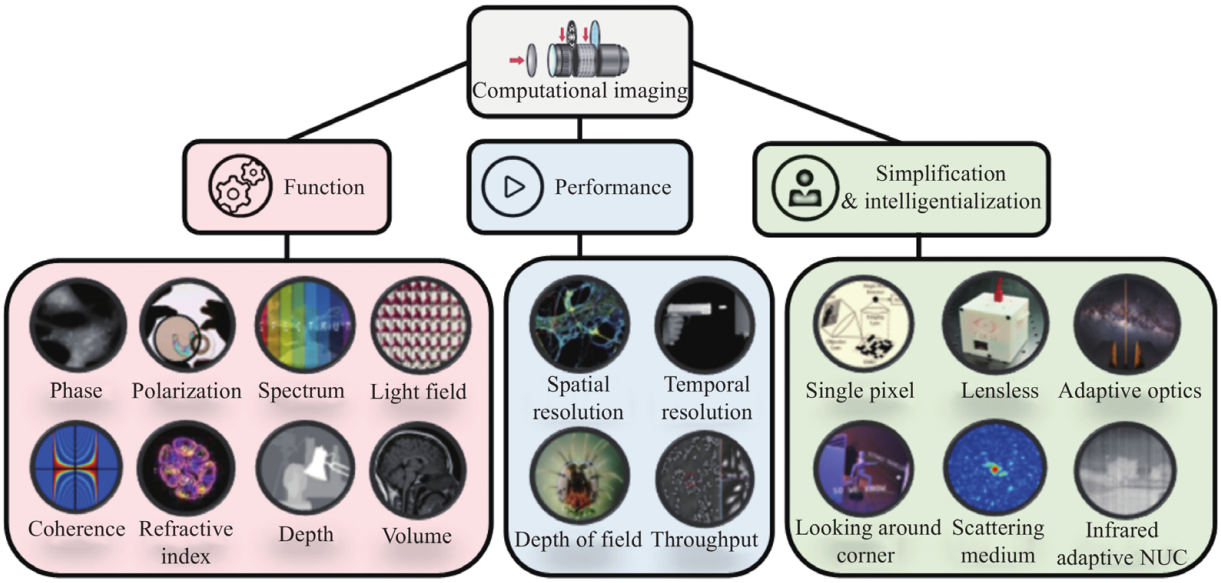


Figure 2: Categories of computational imaging technologies

EXPERT OPINIONS | Over the past decade, the evolution of information technologies has continually reshaped imaging technologies. The emergence of computational imaging is set to transform the way human beings and machines perceive the world. While traditional imaging only reveals what our eyes can see, computational imaging goes beyond the limits of line of sight. Computational imaging leverages high computing power to obtain and process multidimensional light field information. This allows us to trace light propagation, reconstruct sounds from a distant place, explore the myth about life, and observe details through fog and clouds. I believe that computational imaging plays a crucial role in expanding the observable universe, from the microscopic world of microorganisms to the vast expanse of space. Computational imaging also fits into many aspects of everyday life, such as autonomous systems, mobile photography, and industrial inspection. In closing, computational imaging not only expands our horizons, but is also a key technology that will advance the digital economy.

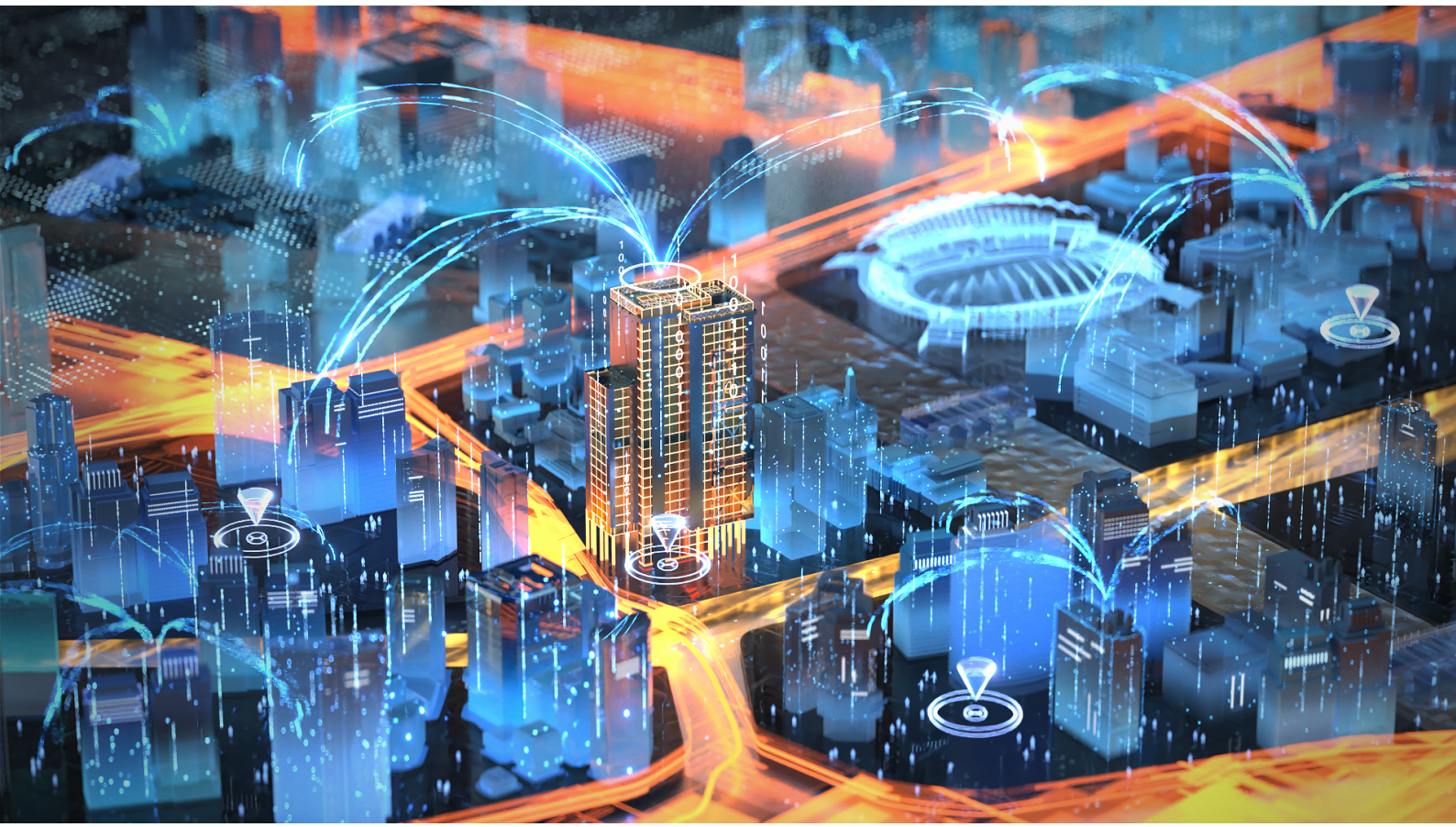
Jiamin Wu
Assistant Professor of Department of Automation, Tsinghua University

Large-scale Urban Digital Twin

Large-scale urban digital twin technology is evolving towards becoming more autonomous and multidimensional.

Introduction

The concept of the urban digital twin has been widely promoted and accepted since its debut in 2017 and has become a new approach to refined city governance. In recent years, we have witnessed significant technical breakthroughs in this field. Several technologies are enhanced for large-scale application. For example, large-scale dynamic perceptual mapping lowers modeling costs, large-scale online real-time rendering shortens the response time, and large-scale joint simulation-enabled inference achieves higher precision. So far, large-scale urban digital twins have made major progress in scenarios such as traffic governance, natural disaster prevention, and carbon peaking and neutrality. In the future, large-scale urban digital twins will become more autonomous and multidimensional.



Trend Analysis

The concept of the urban digital twin has been widely promoted and accepted since its debut in 2017. Since then, Alibaba DAMO Academy has been paying close attention to the development of urban digital twins, and elaborated the industry dynamics in its *Top Ten Technology Trends of DAMO Academy* reports released in 2019 and 2021.

Urban digital twins are virtual cities in the digital world and act as one-to-one digital mappings of real cities in the physical world. Through digital mapping, data can be mutually synchronized in real time between virtual cities and physical cities based on multidisciplinary mechanisms and simulation-enabled inference. The last two years have witnessed major breakthroughs of key technologies in urban digital twins, such as precision mapping, rendering, and simulation-enabled inference. As a result, dynamic perceptual mapping, real-time rendering, and joint simulation-enabled inference have been successfully applied on a large scale, which indicates a qualitative transformation rather than merely quantitative changes.

In terms of precision mapping, we have achieved a cost-effective method of perceiving the position, status, and other properties of static elements and dynamic objects such as vehicles in real time. This is done through combining data from a variety of sensors such as remote sensors, radar sensors, visual sensors, and positioning sensors with the existing plotting data. In the future, all kinds of sensors deployed in space, in the air, and on the ground will be integrated with the perceptive capabilities of AI to converge heterogeneous data about each entity, and build relationships among different entities in cities. By then, we will be able to create an accurate, real-time reflection of the physical world in the digital world at a relatively low cost.

In terms of rendering, 3D city models of different layers, sizes, and resolutions can be automatically generated based on precision mapping data and technologies such as AI-generated content (AIGC) and professionally-generated content (PGC). Urban digital twins on the cloud will be powerful enough to allow multiple users to simultaneously perform multiple real-time simulations of different scenarios with complex model interactions.

In terms of simulation-enabled inference, both multidisciplinary, large-scale mechanisms and simulation models have been applied to the same digital world to construct a "simu-

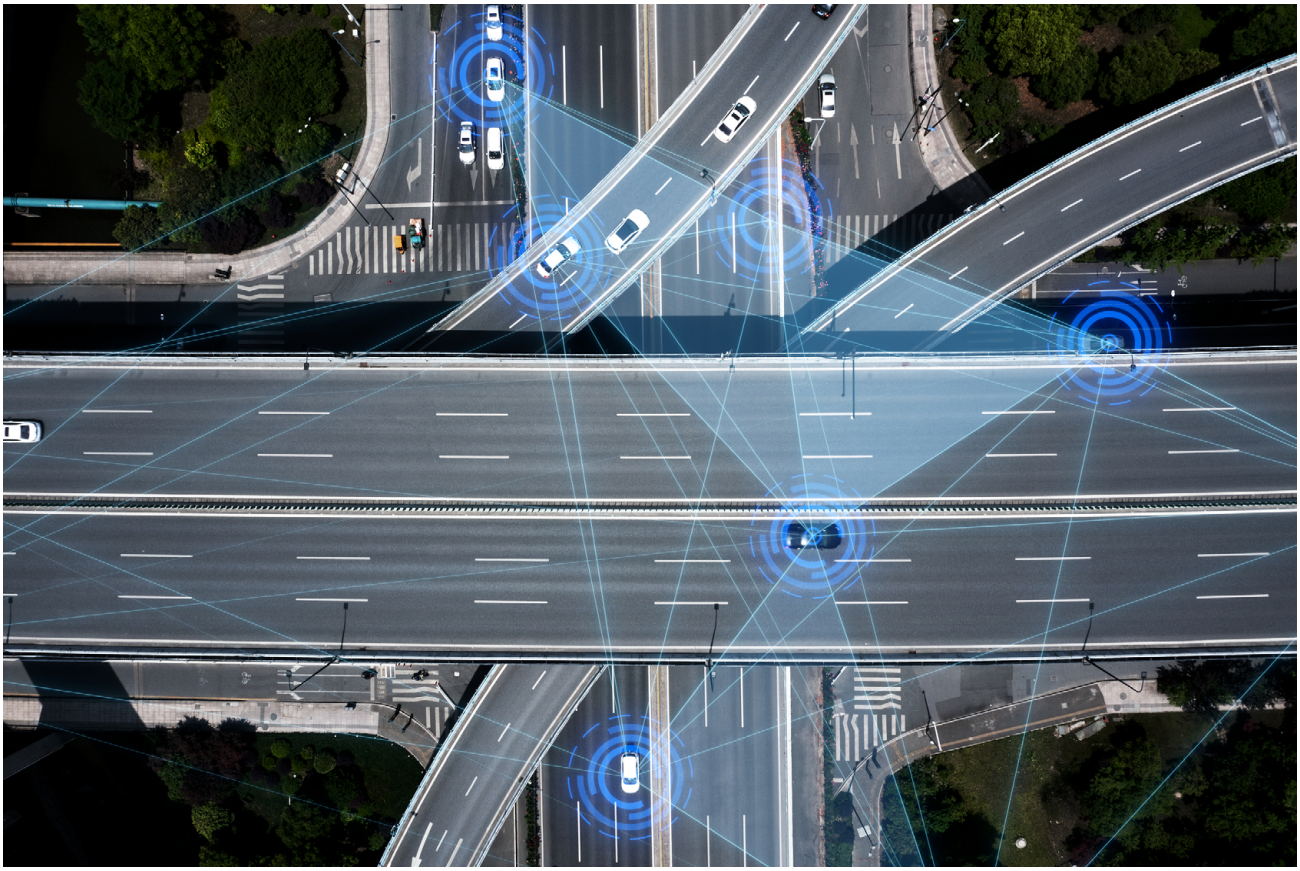


Figure 3 Urban Traffic Analysis

lation mechanism-enabled metaverse". The implementation of the two key technologies will enable humans to interact with objects in the virtual world and vice versa. One of these technologies is the cloud-native simulation system. Leveraging cloud-native scheduling capabilities and high-performance solvers, computing resources are delivered to where they are needed most, greatly accelerating simulation for time-sensitive, city-level scenarios with millions of entities and delivering real-time response plans. The other technology is integrated computing based on the unified interface. It supports real-time integrated computing of multiple mechanism and simulation models to provide simulation capabilities that can solve complex, multi-model interactions.

Driven by technologies and demands, large-scale urban digital twins have made major progress in scenarios such as traffic governance, natural disaster prevention and management, and carbon peaking and neutrality. For example, in terms of traffic governance, 3D modeling and real-time rendering have been performed on urban entities to build virtual high-precision road networks, water networks, rivers, and vehicles. As a result, the modeling cost is reduced by more than 90% and the modeling and rendering period is shortened from several months to only a few days. Furthermore, by integrating joint-simulation models, like road traffic, urban waterlogging, and autonomous driving in the digital twin, comprehensive response plans can be easily developed and evaluated in different scenarios such as traffic control, severe weather preparedness, and public transport management. In actual practice, such plans can reduce the response time to within a minute, and have emergency personnel arrive on-site within 5 minutes.

The urban digital twin technology is the main driving force behind the broad market prospects of smart cities. IDC predicted that more than 100 billion US dollars will have been invested into the smart cities market by 2025, with a compound growth rate of 30% in the next 5 years. Currently, the biggest bottleneck for urban digital twins is that we have not constructed full-fledged digital twin entities for cities, which consist of large-scale object twins and business process twins. Large-scale urban digital twins are evolving towards becoming more autonomous and multidimensional. In the future, urban digital twins will function as not only a development and test environment for the multidimensional and integrated autonomous systems such as drones and robots, but also a system that supports comprehensive perception and resource allocation.

EXPERT OPINIONS | Large-scale urban digital twin technology becomes possible when the perceptive capabilities and computing capabilities such as restoration, modeling, rendering, and simulation-enabled inference of physical cities reach their critical points. Urban digital twins represent a new computing paradigm for urban governance. They are not only the results of visualized rendering, but also a new carrier of diversified urban business to support business innovation. With urban digital twins, we can capture the present and record the history of real cities.

More importantly, as we migrate the urban digital twin technology to a cloud-native environment and integrate simulation capabilities from different areas, we can build a more powerful simulation platform. Large-scale, multi-model complex interactions can be simulated, and we can infer how cities will continue to develop based on their histories.

Zhenyu Zeng
Vice President of Alibaba Cloud Intelligence, Head of Industry Solutions Research and Development Department

Generative AI

The widespread application of Generative AI is transforming how digital content is produced.

Introduction

Generative AI (or AI-generated content, AIGC) generates new content based on a given set of text, images, or audio files. In 2022, its progress is most prominent in the following fields:

- Image generation, driven by diffusion models (DALL·E-2 and Stable Diffusion)
- Natural language processing (ChatGPT based on GPT-3.5)
- Code generation (GitHub Copilot based on OpenAI Codex).

Currently, Generative AI is mainly used to produce prototypes and drafts and is applied in scenarios like gaming, advertising, and graphic design. Along with future technological advancement and cost reduction, Generative AI will become an inclusive technology that can greatly enhance the variety, creativity, and efficiency of content creation.

Trend Analysis

Generative AI learns the patterns of a given set of data by using machine learning algorithms, and then generates content like video, image, text, audio, or code. The content that it generates originates from the given data but is not a direct copy. Its advancement in 2022 was driven by the breakthrough of applying foundation models in fundamental research and deep learning, as well as the accumulation of real-life data and the decline of computing costs. In 2022, AI's ability to create has been greatly boosted by the progress in Generative AI.

In 2022, Generative AI has made major progress in the following fields:

- In image generation, the progress is mainly driven by the application of diffusion models, especially in models such as DALL·E-2 and Stable Diffusion. Diffusion models are used to generate image from noise. They are underpinned by Contrastive Language-Image Pretraining (CLIP) and pre-trained models with higher accuracy for understanding the semantics of natural language. They are conducive to better creativity of the generated images.
- In natural language processing (NLP), the progress is most prominent in ChatGPT based on Generative Pre-trained Transformer 3.5 (GPT-3.5). ChatGPT is a text generation deep learning model that is trained on data available on the Internet. It can answer questions, summarize text, translate text, perform categorization, and generate code. Reinforcement Learning from Human Feedback (RLHF) is used in the training and optimization of ChatGPT, thanks to the development of pre-trained foundation models that combine the use of text and code. RLHF allows ChatGPT to create an iterative closed loop where ChatGPT understands the intentions of a human prompt, determines the quality of the answers it provides based on human feedback, further interprets the previous answers, and smartly handles inappropriate questions.
- In code generation, AlphaCode and GitHub Copilot are two of the most advanced models. AlphaCode, the latest AI coding system from DeepMind, was released in February 2022. It surpassed about 47% of the human programmers in the programming competitions on Codeforces, becoming the first AI code generator able to solve competitive programming problems. GitHub Copilot is an AI coding assistant that is provided as a subscription-based service. It uses OpenAI Codex, a large language model, and is trained on public code. Though the code it generates requires manual adjustments in most cases, it can relieve developers of creating repetitive code patterns and is drastically changing the IDE industry.

Generative AI is not free from challenges. Along with the exponential growth in the quantity of content, controlling the quality and semantics has become a serious challenge for Generative AI. Also, for the marketization of this technology, cost is a major consideration. Models like ChatGPT can be marketized only after their costs for training and inference are significantly lowered. In addition, data security and controllability, IPR, and trust in AI are all issues that require careful handling.

In the next three years, we will see business models emerging and ecosystems maturing as Generative AI becomes widely marketized. By then, Generative AI will have gained content creation abilities on par with those of human beings, and tech giants, with their advantages in data, computing capabilities, and productization experience, will be the major players in bringing Generative AI to market. Computing infrastructure and platforms designed for relevant models will be produced, which will make the models accessible to customers without professional knowledge. Generative AI models will be more interactive, secure, and intelligent, assisting human beings to complete various creative work.

EXPERT OPINIONS | The year of 2022 witnessed the breakthrough of Generative AI, with significant improvements in the quality, reasonability, and security of the generated image, code, and open-set text. Generative AI will be used in much more scenarios in years to come. However, we still need to work on the security, controllability, morality, and accountability of this technology and take special heed of the social hazards posed by fake AI-generated content.

Fei Huang
Head of Language Technology Lab, DAMO Academy

Bibliography

[1] Gautam Kumar, et al. Swift: Delay is Simple and Effective for Congestion Control in the Datacenter. SIGCOMM 2020.

[2] Wang, Shuai, et al. Predictable vFabric on Informative Data Plane. SIGCOMM 2022.

[3] Gibson, Dan, et al. Aquila: A Unified, Low-latency Fabric for Datacenter Networks. NSDI 2022.

[4] V. Olteanu, et al. An Edge-queued Datagram Service for all Datacenter Traffic. NSDI 2022.

[5] H. Bao, L. Dong, S. Piao, F. Wei. BEiT: BERT Pre-Training of Image Transformers, Microsoft Research. https: //aka.ms/beit. arXiv:2106.08254v2,2022.

[6] X. Pan, T. Ye, D. Han, et al. Contrastive Language-Image Pre-Training with Knowledge Graphs. arXiv:2210.08901,2022.

[7] C. Saharia, W. Chan, S. Saxena, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487, 2022.

[8] R. Zhang, B.Li, W. Zhang, et al. Collaboration of Pre-trained Models Makes Better Few-shot Learner. arXiv:2209.12255,2022.

[9] W. Ma, M. Zhao, X. Xie, et al. Is Self-Attention Powerful to Learn Code Syntax and Semantics?. arXiv:2212.10017.2022.

[10] G. Gao et al. Die to Wafer Hybrid Bonding for Chiplet and Heterogeneous Integration: Die Size Effects Evaluation-Small Die Applications. 2022 IEEE 72nd Electronic Components and Technology Conference (ECTC), 2022, pp. 1975-1981, doi: 10.1109/ECTC51906.2022.00310.

[11] John H. Lau. Recent Advances and Trends in Advanced Packaging. IEEE Transactions on Components,w Packaging and Manufacturing Technology (Volume: 12, Issue: 2, February 2022).

[12] T. Tang, Y. Xi. Cost-Aware Exploration for Chiplet-Based Architecture with Advanced Packaging Technologies. arXiv:2206.07308.2022.

[13] A. Mullen et al., Gartner Top Strategic Technology Trends for 2022 [EB/OL].2021. https://www.gartner.com/en/newsroom/press-releases/2021-10-18-gartner-identifies-the-top-strategic-technology-trends-for-2022.

[14] Gartner. 中国云安全市场概览 [EB/OL].2022.

[15] 中国信息通信研究院 . 云原生架构安全白皮书 .[R]，2021.

[16] 何宝宏 . 云与安全深度融合推动原生云安全发展 .[J]. 中国信息安全，2022.

[17] 袁曙光 . 云安全的未来是云原生安全 .[J]. 中国信息安全，2022.

[18] 宋胜攀等 . 零信任在云原生安全中的应用研究 .[J]. 保密科学技术，2021.

[19] 世界经济论坛 . 数字孪生城市：框架与全球实践洞察力报告 .[R]，2022.

[20] 中国信息通信研究院 , 中国互联网协会 , 中国通信标准化协会 . 数字孪生城市白皮书 .[R]，2021.

[21] IDC. 中国数字孪生城市市场分析，2021.[R]，2022.

[22] 左超，陈钱 . 计算光学成像：何来，何处，何去，何从？ .[J]. 红外与激光工程，2022，51(2).

[23] 戴琼海，赵建林，司徒国海，方璐 . 计算光学成像专题 .[J]. 光学学报，2020，40(1).

Visit our Website for More