CONTENTS

01

BACK TO BASIC WITH CIPU: THE FUTURE OF CLOUD INFRASTRUCTURE MANAGEMENT

By **Jeff Zhang**,
President of Alibaba Cloud Intelligence,
Head of Alibaba DAMO Academy

The digital economy has become the key engine that accelerates global economic growth. Cloud computing plays a pivotal role in the new digital economy by offering necessary computing resources to power various businesses on the cloud. However, with the rapid proliferation of AI, big data, and IoT solutions in the digital era, the demand for more reliable computing power and high network bandwidth has exponentially increased across industries. Cloud vendors must shift from the traditional cloud computing architecture centered on CPU to effectively deal with data-intensive scenarios represented by big data applications.

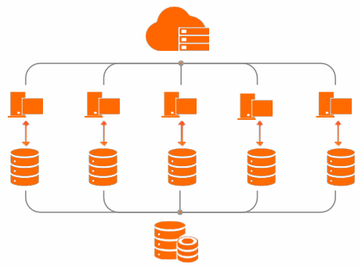
At Alibaba Cloud, we realized that the best way to overcome the challenges brought by the increase in data volume and higher requirements for low latency requires us to "**go back to basic**" - innovate our services by reshaping our cloud computing architecture. Therefore, we have introduced **Cloud Infrastructure Processing Unit (CIPU)** and placed it at the core of our new architecture to improve cloud infrastructure operations. As a processing unit, CIPU shifts the compute, storage, and network resources of data centers to the cloud and enables hardware acceleration. With the CIPU-centric architecture, Alibaba Cloud aims to deliver enormous computing power, high performance, and agile deployment to meet the demands of new-age digital enterprises.

To give you a better understanding of CIPU and its importance in modern cloud data centers, let's first explore the development history of cloud computing.

EVOLUTION OF CLOUD COMPUTING

EARLY STAGE ARCHITECTURE OF CLOUD COMPUTING

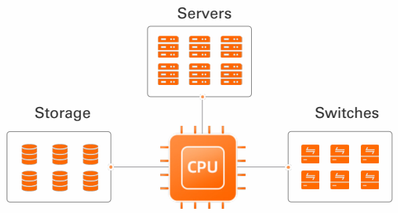
Distributed Architecture



Scalability

Virtualization

Resource Pool Architecture



Decoupled storage and compute resources

Highly available and reliable computing power

The development history of cloud computing can be mainly divided into two stages. In the early stage of cloud computing, distributed architecture was key to success for cloud vendors. At this stage, cloud computing largely replaced mainframes with distributed systems and virtualized technologies to meet the computing and storage demands of enterprises of that time. In the next phase of cloud computing, Internet companies increased their demand for elastic computing power. To meet this demand, cloud vendors adopted CPU-centric resource pool architecture. Under this architecture, cloud vendors decoupled computing and storage resources and then rearranged them to set up large-scale computing and storage resource pools to improve the reliability and availability of their services.

In recent years, with the rapid growth of the digital economy, the digital transformation of enterprises requires more intensive computing and high-speed network transmission capabilities, which are difficult to achieve with the traditional or CPU-centric resource pool architecture.

IN SHORT, CLOUD VENDORS WITH CPU-CENTRIC ARCHITECTURE FACE THREE MAJOR OBSTACLES WHILE DELIVERING CLOUD SERVICES IN TODAY'S DIGITAL LANDSCAPE

Network Latency

The cloud vendors experience significant delays during computing and network transmission due to CPU-centric architecture.

Network Capacity

The rapid proliferation of big data applications has increased the amount of data exchange between servers within data centers, and the CIPU-centric architecture cannot provide high bandwidth to support data intelligence applications.

System Complexity

Data center systems are getting larger and more complex in today's data-driven digital economy. Cloud vendors with CPU-centric architecture find it challenging to manage the super-large-scale infrastructure and the super-large applications on the cloud.

THE RISE OF CLOUD INFRASTRUCTURE PROCESSING UNIT (CIPU)

Alibaba Cloud has introduced a new cloud computing architecture centered on CIPU to resolve the network latency, bandwidth, and infrastructure management-related issues faced by today's data centers. Under this new architecture, CIPU connects to Alibaba Cloud's Apsara Distributed Operating System to efficiently manage and schedule the computing, storage, and network resources on the cloud.

CIPU also serves as the processing powerhouse of next-generation of cloud data centers. As a hardware accelerator, CIPU enables Alibaba Cloud to deliver improved networking, storage, security, and computing performance. For instance, CIPU can enhance computing power performance by up to 30 percent in big data and AI-intensive scenarios.

Let's understand the different functions of CIPU, the core of Alibaba Cloud's new infrastructure architecture for cloud data centers

Network Resource Acceleration

CIPU can perform hardware acceleration on high-bandwidth physical networks to **reduce network latency to as low as 5 microseconds.** It also plays a crucial role in setting up a distributed, large-scale, and high-performance eRDMA network. Remote Direct Memory Access (RDMA) is a next-gen technology to improve network performance; however, providing RDMA as a Service on the cloud is challenging. CIPU resolves this challenge through the cloudification of RDMA.

Storage Resource Acceleration

CIPU can perform hardware acceleration on block storage devices in a compute-storage separated architecture. It helps address latency and bandwidth-related issues during storage resource pooling. With CIPU-driven hardware acceleration, the **cloud disk IOPS can reach up to 3 million**, whereas the **long-tail latency can be reduced by 50 percent.**

Computing Resource Acceleration

CIPU can provide quick access to Alibaba Cloud ECS instances while **ensuring zero loss of computing power.** It also supports security hardening and isolation in the cloud, which is equally good as traditional hardware methods. Further, cloud servers require additional overhead to implement virtualization. CIPUs help **reduces this server overhead to zero** and delivers high-performance hardware isolation capabilities.

Cloud Resource Controller

CIPU ensures flexible management, orchestration, and scheduling of CIPU-accelerated resources in the Apsara Distributed Operating System. This enables Alibaba Cloud's customers to rapidly provision, manage, and scale cloud servers.

ADVANCED CLOUD COMPUTING ARCHITECTURE

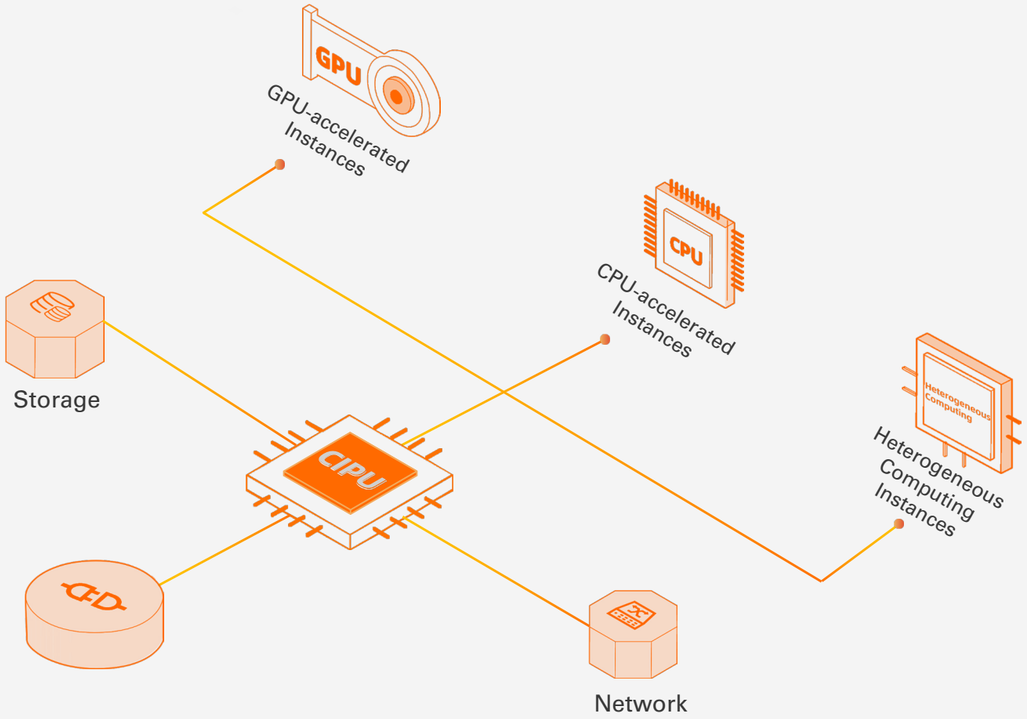
Cloud Infrastructure Processing Unit (CIPU)

Storage Resource Acceleration

- Hardware virtualized storage devices
- Hardware acceleration for storage IO and forwarding
- Hardware based encryption and decryption

Cloud Resource Controller

- More flexible management, scheduling, and orchestration of CIPU-accelerated resources in the Apsara Cloud Operating System

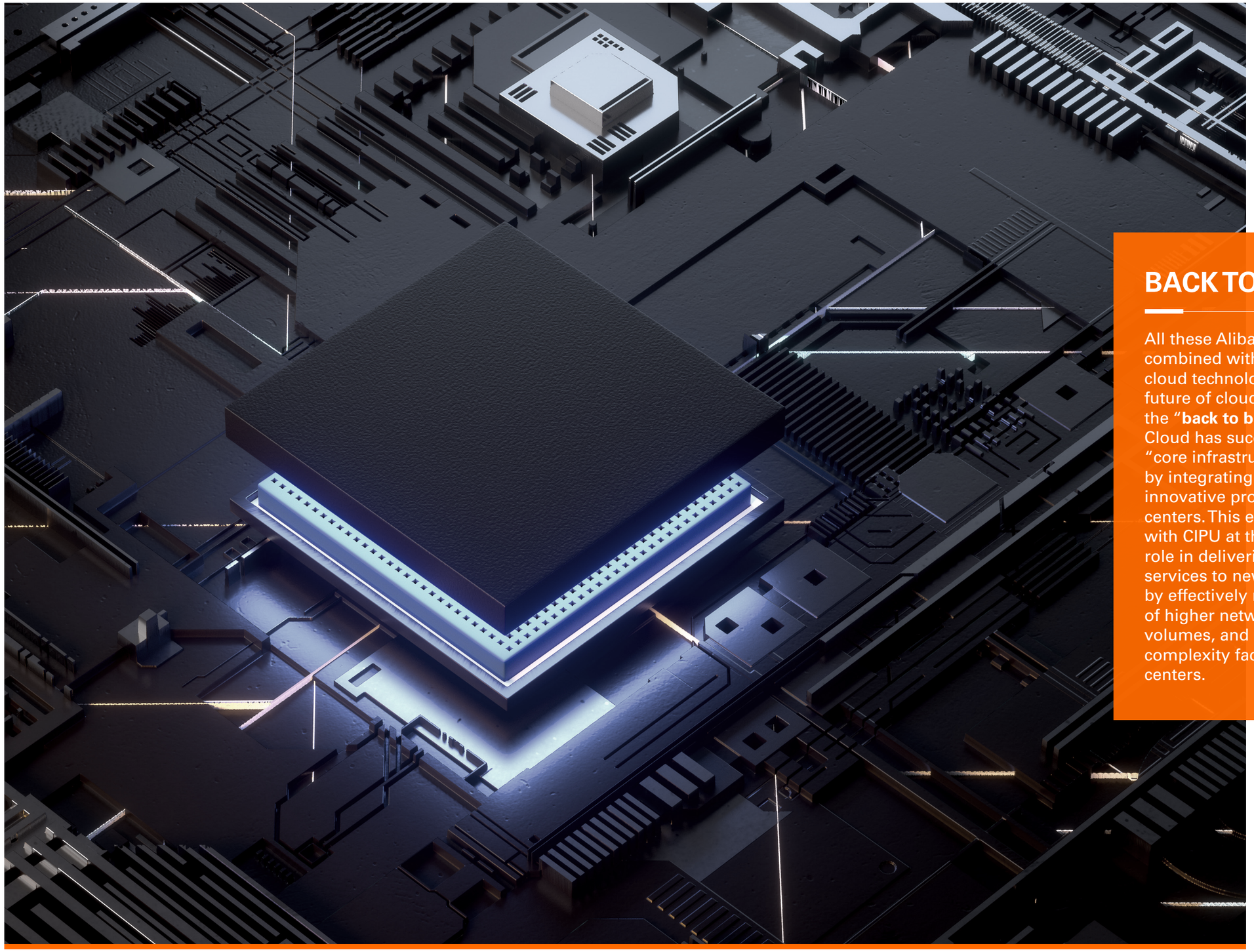


Computing Resource Acceleration

- Computing server plug and become ECS instances
- Hardware based virtualization
- Security hardening and isolation mechanisms

Network Resource Acceleration

- Virtualized network devices
- Hardware acceleration for traffic forwarding
- Hardware based encryption and decryption
- Inclusive elastic RDMA (eRDMA) on the cloud



BACK TO BASIC

All these Alibaba Cloud platforms combined with CIPU are best-in-class cloud technologies that can shape the future of cloud computing. As part of the "**back to basic**" approach, Alibaba Cloud has successfully refreshed these "core infrastructure" components by integrating them with CIPU, an innovative processor for cloud data centers. This entire technology stack with CIPU at the core plays an essential role in delivering enhanced cloud services to new-age digital enterprises by effectively resolving the challenges of higher network latency, high data volumes, and increased system complexity faced by modern cloud data centers.

02

TECH FOR INNOVATION: THE ALIBABA CLOUD ADVANTAGE

By **Selina Yuan**,
Alibaba Cloud Intelligence
International President

We live in an ever-evolving world where staggering technological advancements have created massive opportunities for organizations to achieve digital transformation, gain a competitive advantage, and accelerate business growth. Global organizations striving to maximize value, modernize business models, and improve customer experience, are now focusing on leveraging the transformative power of technology innovation to deal with new emerging challenges and make a global impact.

Organizations across industries univocally acknowledge that cloud computing is the way forward. The latest forecast from Gartner predicts **end-user spending on public cloud services to reach around $600 billion in 2023**. Global industry leaders see cloud as the core foundation for digitalization and a key enabler for empowering growth. Alibaba Cloud, a pioneer in cloud computing, is constantly innovating at the forefront of business intelligence that promotes resilience and sustainability, helping businesses navigate the new digital era. We continue to develop at scale to design a new way ahead in the digital age and drive cloud-native transformation. We continue our efforts to be at the helm of several technological breakthroughs, from delivering unique solutions to help businesses become future-ready to technologically supporting global sports events, such as the Tokyo Summer Olympics and Beijing Winter Olympics.

BUILD ON THE TRANSFORMATIVE POWER OF A NEW-GEN CLOUD INFRASTRUCTURE

With a large-scale global cloud computing infrastructure covering **28 service regions and 85 availability zones across 200+ countries and regions**, we are committed to power innovation to create value and ensure a brighter future.

We help businesses navigate the new digital era with our redefined cloud computing architecture, focusing on the core technology, a move that takes us back to basic. We bring advanced cloud computing architecture with the **Cloud Infrastructure Processing Unit (CIPU)** to meet the increased computational needs on the cloud and growing demands for lower latency. Coupled with the Apsara Cloud operating system, CIPU is expected to become the core of the next generation of cloud computing infrastructure.

ACCELERATE INNOVATION BEYOND BORDERS

With rising global uncertainty, forward-thinking corporations are rapidly accelerating their digital initiatives and proactively planning to successfully tap into the growing opportunities across regions, specifically in Asia-Pacific. We constantly develop new innovative ideas, strategies, and solutions to help companies navigate Asia's dynamic market landscape by leveraging our global infrastructure, local expertise, and comprehensive ecosystem.

New Data Centers in the APAC Region

Last year, we built a new data center in the Philippines and added a third data center in Indonesia. On May 20, 2022, we unveiled our latest data center in Thailand, **marking our 10th data center in Southeast Asia** to meet local demands. These new data centers will serve as infrastructure backbones to support the region's digital economies and local businesses to accelerate their transformation and innovation journey.

Dedicated Regional Teams for Better Local Support

We believe localization is at the heart of internationalization. At Alibaba Cloud, we ensure our regional teams closely connect with the local population to help us better understand local customers. Our local team members play a crucial role in our localization to better support the digital transformation of local industries.

Enabling Digital Ecosystem and Talent to Boost Innovation

Last year, during the Alibaba Cloud Summit 2021, we launched **Project AsiaForward (PAF),** an ambitious initiative to invest in infrastructure, technological innovation, and talent development in APAC. With initial funding and resources of **USD 1 billion**, the project envisions cultivating a digital ecosystem and talent pool in the region by **empowering 100,000 developers** and **enabling the growth of 100,000 technology startups in Asia-Pacific** over three years.

RIDE THE DIGITAL WAVE WITH INNOVATIVE INDUSTRY-TAILORED SOLUTIONS

Alibaba Cloud also supports companies across all the leading industries leveraging innovative technology to create new possibilities. Our comprehensive industry-tailored solutions powered by intelligence and diverse ecosystem help forward-looking enterprises to innovate at the speed of light and accelerate their digital transformation journey.

Revolutionizing the Retail and e-Commerce Industry

Leveraging the industry know-how and practical experience of the leading global retail business, Alibaba Cloud enables an entire era of new retail underpinned by cutting-edge technology. In particular, **NaRaYa** and **Senheng** greatly benefitted from our retail and e-commerce offering by delivering a seamless shopping experience to their customers.

Transforming the Financial Services Industry

We are committed to enabling financial services companies to address accessibility and financial inclusion issues through digitalization. We support traditional financial services organizations and modern-day fintech startups to make their services more accessible, resilient, and secure. We have provided comprehensive financial-grade cloud services to leading brands, such as **CIBI** and **Yunfeng Financial**, to manage their business at scale while remaining compliant with local regulations.

Innovating the Healthcare Industry

Alibaba Cloud's technology-led innovations have been significant in improving the state of the healthcare sector. For instance, our AI-based solutions for medical imaging help doctors diagnose diseases and tumors while improving detection accuracy. In Indonesia, we have worked with **Bumame Farmasi** and **Dexa Group** to digitalize their services, which allowed them to easily cope with the traffic surge of their ever-growing businesses.

Accelerating Digitalization in the Sports Industry

We are committed to accelerating the digital revolution in the sports industry. Our technology innovations powered by intelligence and cloud technology empower sports enterprises to enhance fan engagement, athlete performance, and event operations and safety.

Digitalizing the Agriculture Industry

Alibaba Cloud is stretching its reach from cities to rural areas with its "digital agricultural solution." AI capabilities and remote sensing technologies can revolutionize the agricultural industry by effectively monitoring crop growth and efficiently allocating resources based on accurate information. We have empowered Malaysia's **RegalTech** and **Tanahmu** to build high-performance and reliable platforms to improve the yield rate and efficiency of their farms.

SEIZE SUSTAINABLE TECHNOLOGICAL INNOVATION FOR A GREENER FUTURE

As the global cloud technology leader, we envision contributing to shaping a sustainable future for all. We have taken a proactive stand to continually fuel innovation with a purpose and encourage businesses to embed sustainability across their business models through our sustainable technological innovation.

AI for Improved Solid Waste Management

AI technologies developed by our experts help businesses extract more energy from waste and reduce labor workload. Around 100 waste-burning facilities across 30 cities in China use these technologies and it has led to **generating an additional 360 million more kilowatt-hours (kWh) of electricity every year**.

Green Data Centers for Decarbonization

Thanks to innovative green technologies such as liquid cooling and wind energy being applied at our hyper-scale data centers to help us **reduce the average PUE to as low as 1.09**.

Intelligent Logistics and Sustainable Transport Management

Our experts have developed a driverless robot called Xiaomanlv to transform the transportation and logistics industry and minimize the environmental price of traffic congestion. Self-driving robots powered by electricity consume less power than traditional vehicles, reducing toxic transport-related carbon emissions.

Modern technologies have the vast potential to enable enterprises to conduct business to benefit all and create a big difference globally. Alibaba Cloud is continually exploring ways to innovate and serve our customers, optimize our existing technologies, and unlock massive value for business and society.

03

CIPU: A KEY ENABLER FOR ALIBABA CLOUD PRODUCT ECOSYSTEM

By **Ruirui Huang**,
Director of Cloud Architecture, Chief Architect of Infrastructure Products, Alibaba Cloud Intelligence

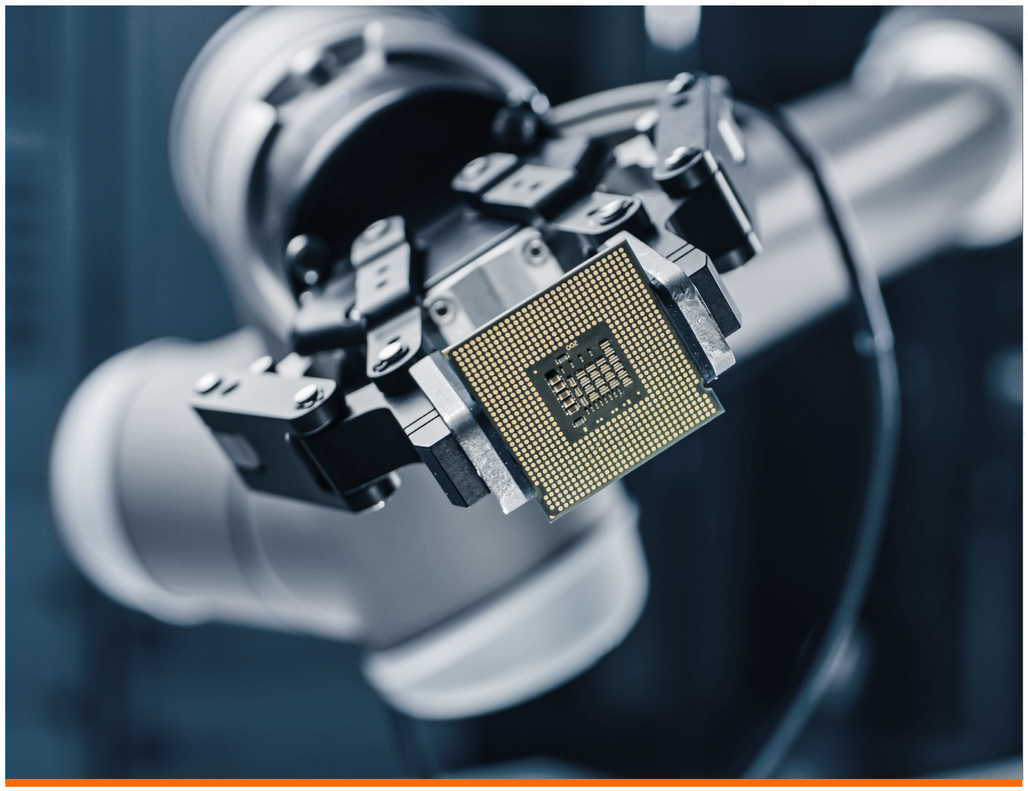
In June this year, Alibaba Cloud unveiled a new cloud infrastructure system designed in-house to power its cloud-native data centers, called Cloud Infrastructure Processing Unit (CIPU). The system is running in a handful of data centers operated by the cloud computing branch of Alibaba Group.

CIPU was developed in response to increased computational needs on the cloud. It helps Alibaba Cloud deliver performance improvements in networking, storage, security and computing power to clients by offloading virtualization functions from servers to dedicated hardware.

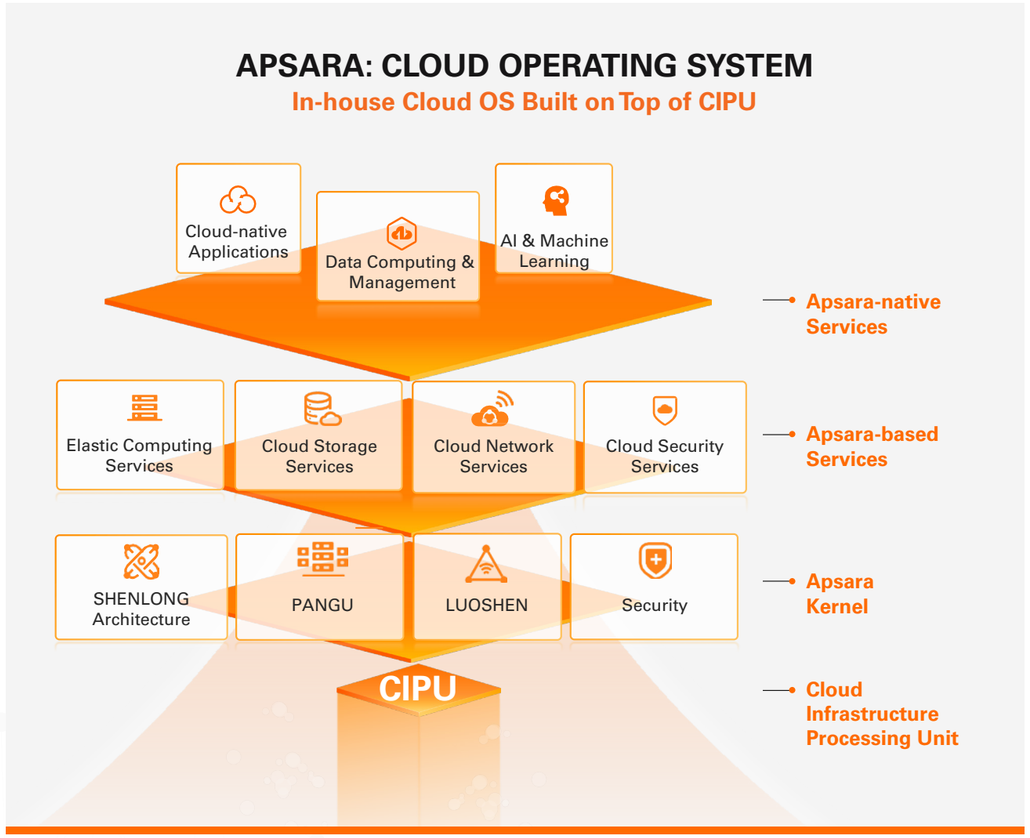
The CIPU comes as a hardware acceleration card that can be attached to servers in data centers. It has enough processing power to offload a wide range of virtualization workloads, such as security, storage, and networking virtualization from the CPU to the hardware.

By taking some computing burdens off of the servers, CIPU can **reduce network latency to as low as five microseconds** and **improve computing power performance in data-intensive AI and big data spark scenarios by 30%.**

Coupled with the Apsara Cloud operating system, the CIPU is expected to become the core of the next generation of cloud computing infrastructure. Let's take a look at how CIPU works along with our core cloud infrastructure.



CIPU: ENABLING ALIBABA CLOUD PRODUCT ECOSYSTEM



CIPU strongly supports the following five cores of Alibaba Cloud products:

LUOSHEN Network

Relying on Alibaba Cloud's robust global infrastructure and CIPU, LUOSHEN delivers reliable and ultra-large-scale network services to customers. As a high-performance cloud network system, LUOSHEN meets the large-scale computing power requirements of big data and AI scenarios.

Apsara Distributed File System

Powered by high-density storage servers, Alibaba Cloud Apsara Distributed File System works together with CIPU and LUOSHEN network to **reduce the storage access latency to 30 microseconds.** This innovative and high-performance storage system enables customers to retrieve data from cloud disks more quickly than their on-premise disks. It supports different types of key storage services from Alibaba Cloud, such as MaxCompute and Block Storage.

SHENLONG Platform

Besides networking and storage services, CIPU supports Alibaba Cloud's innovative computing platform, SHENLONG. Leveraging CIPU's acceleration capabilities, SHENLONG provides highly reliable elastic computing services to customers and supports resource orchestration. CIPU and Apsara Distributed Storage System contribute to the improved speed and performance of SHENLONG computing clusters. Further, CIPU provides security hardening and isolation for the hardware accessing the SHENLONG clusters.

Security Platform

CIPU is an integral part of the cloud-native security platform of Alibaba Cloud. This platform provides end-to-end security capabilities that extend to Alibaba Cloud's integrated software and hardware infrastructure comprising in-house chipsets, systems, and Apsara Distributed Operating System. CIPU also provides encrypted computing capabilities to enhance the overall security of Alibaba Cloud's robust cloud infrastructure.

Apsara Distributed Operating System

This in-house cloud operating system of Alibaba Cloud runs on the CIPU. The computing, storage, and networking resources virtualized using the CIPU are successfully managed and scheduled by the Apsara Distributed Operating System.

04

CREATE NEW POSSIBILITIES WITH ALIBABA CLOUD ENTERPRISE CLOUD SAAS

By **William Xiong**,
Vice President of Alibaba Cloud,
Enterprise Service Cloud

In the past, SaaS offerings and enterprise software were often treated as separate entities – the former mainly consisting of lightweight and relatively simple applications for individuals, while the latter consisting of more complex software offerings ideal for business-specific usage. However, when employed in a combination as a service better known as "enterprise SaaS," it brings together the strength of both offerings. It provides medium and large enterprises with the tools to run businesses efficiently. Enterprise SaaS enables global enterprises to manage mission-critical applications and processes in an on-demand environment, thanks to the flexibility and cost savings offered by the cloud.

Traditional enterprise software solutions often have rigid features and user interfaces designed for a particular business need. Although intuitive at first, these features and interfaces become challenging to use or even obsolete as the needs of a business evolve. Customization can be done but often at a cost, both in terms of time and money. Moreover, operators need to re-familiarize themselves with the updates. On the other hand, SaaS applications can mitigate this issue by offering streamlined interfaces that are easier to use and readily configurable.

Unlike their traditional counterparts, enterprise SaaS offerings are built using modern, more open web-based and cloud-based architecture with cost-efficient tools that enable service providers to rapidly develop and launch new products and services to the market. Unlike traditional enterprise software, they provide on-demand infrastructure and development environments while maximizing productivity and cost savings. What's more, enterprise SaaS applications are not just intuitive; they enable faster upgrades and allow businesses to "mix and match" services from various providers and allow fast time to market. SaaS solutions require minimal training for IT operators and offer easy integration to existing infrastructure.

As part of our strategy to provide enterprises with efficient, inclusive, and value-driven solutions, we have launched a series of enterprise cloud services to help enterprises move to the cloud easily. I'm proud to announce our two recent launches - **Energy Expert** and **Enterprise Mobile Application Studio (EMAS)**. These new enterprise cloud services enable innovation in your business and provide capabilities to help you address the new-age business challenges. While Energy Expert is an AI-powered energy management platform to measure, analyze, optimize energy consumption and manage carbon emissions, EMAS is cloud-native application research and development (R&D) platform that provides enterprises and developers with one-stop application R&D management services. In particular, it provides SuperApp functionality with mini-apps container capabilities, as well as a low-code development framework to accelerate mobile application development and make it much easier.

ENERGY EXPERT

Launched in June 2022, Alibaba Cloud Energy Expert is a sustainability platform that helps customers worldwide measure, analyze, optimize energy consumption, and manage the carbon emissions of their business activities and products. This software-as-a-service offering provides actionable insights and energy-saving recommendations to help businesses accelerate their sustainability journeys. The initiative is well aligned with Alibaba's carbon neutrality pledge to share its energy-efficient technologies with its customers and business partners, uniting them in their collective efforts to reduce carbon footprints. Energy Expert helps customers to automate the carbon accounting and reporting process at a corporate and product level and obtain real-time sustainability impact statistics to make informed decisions. It enables customers to identify the sources of carbon emissions from their daily business activities and the complete life cycle of their products, based on the PAS 2060 and ISO 14064 standards on carbon neutrality. Customers can also quantify their carbon footprint through a prebuilt calculation model leveraging public emission factors datasets and Energy Expert's proprietary datasets. Additionally, it provides visibility into their real-time carbon emission patterns and the progress of their sustainability performance through visualizations on dashboards and online reports.

In addition, the solution provides cutting-edge analytics on energy efficiency and emission forecasts through deep learning-based AI models hosted on Alibaba Cloud. To help customers minimize their overall environmental impact, Energy Expert offers actionable optimization plans with recommendations that balance business growth and ecological impacts. These can include increasing clean energy usage, reducing excessive electricity consumption during peak times, and optimizing the supply chain from sourcing materials to shipping products.

Energy Expert also works with recognized industrial organizations worldwide, such as TÜV Rheinland, to provide authoritative online carbon footprint accounting and certification, helping customers to validate and communicate their energy-saving progress with stakeholders easily.
In China, it was deployed in February this year. So far, it has served over 2,000 companies there, generating **energy savings of over 2 million kilowatt hours per day or reducing 400,000 tons of carbon dioxide emissions since February**.

Energy Expert also played a vital role in the sustainability measures implemented at Alibaba's Xixi headquarters in Hangzhou. With the energy-saving functionalities, such as intelligent control of the air-conditioning system and the installation of roof-top solar energy panels that produce 1.2 million kilowatt hours of electricity per year - the solution has helped to optimize energy efficiency at the headquarters, **reducing power consumption by 30% during non-peak times and saving 17% of air-conditioning energy use in the summer**.

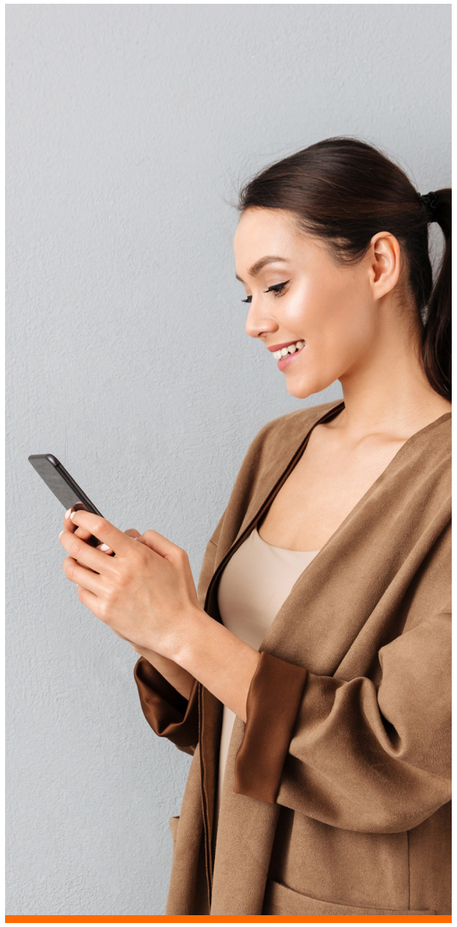
ENTERPRISE MOBILE APPLICATION STUDIO (EMAS)

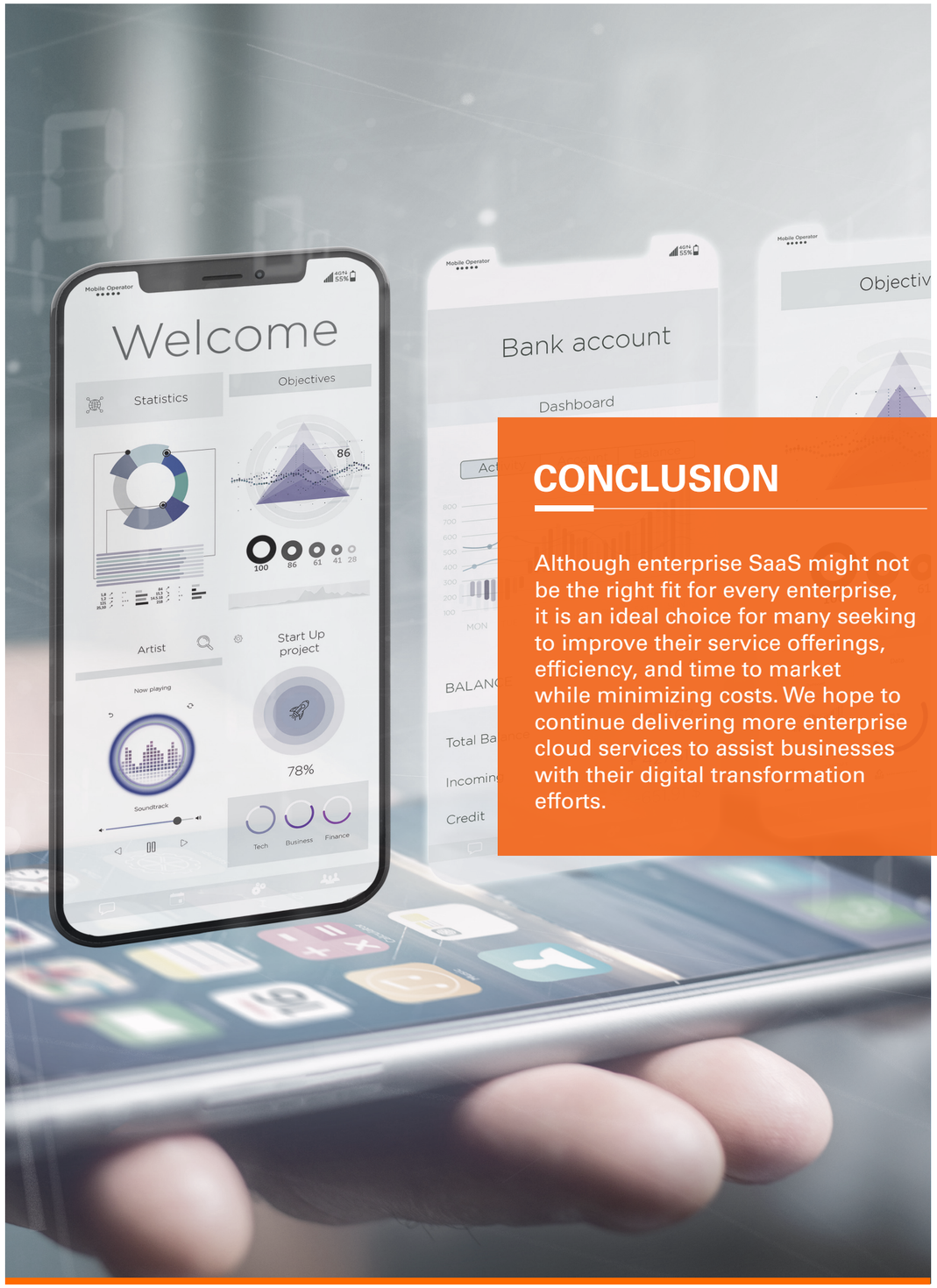
Alibaba Enterprise Mobile Application Studio (EMAS) is a one-stop application research and development (R&D) platform for all-end scenarios, such as mobile, H5, web, and PC applications. Based on Alibaba Cloud's extensive range of cloud-native technologies, EMAS is committed to providing enterprises and developers with end-to-end application R&D capabilities, including development, testing, operations, and maintenance.

EMAS improves the efficiency of businesses by providing complete DevOps support capability and a mature infrastructure to accelerate the pace of business delivery. It lowers the cost for enterprises, thanks to the cost advantage brought by the scale of cloud computing, and it reduces the burden on IT and labor resources comprehensively. EMAS also improves quality control. The closed-loop system of testing + monitoring + logging + hot repair helps users discover, locate and solve problems and comprehensively guarantee the quality of application delivery. Based on Alibaba's business best practices, EMAS also provides expandability through elastic scaling and stable support for peak traffic.

EMAS provides enterprises with innovative R&D tools, such as SuperApp Development with Mini-apps Capabilities, Low-code Development Framework, Mobile DevOps, Mobile Testing, Mobile Hotfix, Crash Analysis, Performance Analysis, and Remote Log. These services help enterprises realize the delivery process, automation, and digitization of the entire R&D process. They also allow application developers to quickly locate and troubleshoot performance issues, improving user experience and customer retention.

Alibaba Cloud's innovative EMAS offering is ideal for real-time resolution of online issues. Bugs and vulnerabilities impacting online apps require prompt resolution, but these processes often require intensive effort and affect service availability. Using the Mobile Hotfix service, patches can be released anytime to solve real-time online problems without impacting the user experience. In addition, EMAS can also intelligently test applications, simulate human operations, support performance baseline comparisons, and quickly detect crashes, unresponsiveness, and other issues, providing businesses with complete automation capabilities. EMAS's Crash Analysis feature can also help classify and analyze common app crash problems on Android and iOS platforms to help companies quickly find and locate online crash problems.





CONCLUSION

Although enterprise SaaS might not be the right fit for every enterprise, it is an ideal choice for many seeking to improve their service offerings, efficiency, and time to market while minimizing costs. We hope to continue delivering more enterprise cloud services to assist businesses with their digital transformation efforts.

05

ENHANCING CUSTOMER VALUE THROUGH CLOUD-NATIVE DATA AND AI TECHNOLOGIES

By **Jans Wei**,
Chief Data Intelligence Architect,
Alibaba Cloud Intelligence

THE IMPORTANCE OF DATA AND AI TECHNOLOGIES

In today's business world, data is the new gold. Practically every transaction generates data, and enterprises are beginning to put in the effort to manage their data more efficiently. To achieve this goal, many key technologies have been developed, including relational databases, data warehouses, data lakes, and data lakehouses. However, these technologies would not be complete without the help of high-performance data analysis and AI technologies to fully extract the value of business data. Businesses in some key sectors like gaming, e-commerce, digital media, autonomous driving, finance, retail, and manufacturing have significantly benefitted from this combination. This includes using high-performance AI technology in real-time and in-depth to forecast demand, optimize inventories, plan resources, optimize user growth, manage risks in real-time, and realize autonomous driving.

THE VALUE OF CLOUD-NATIVE TO BUSINESSES

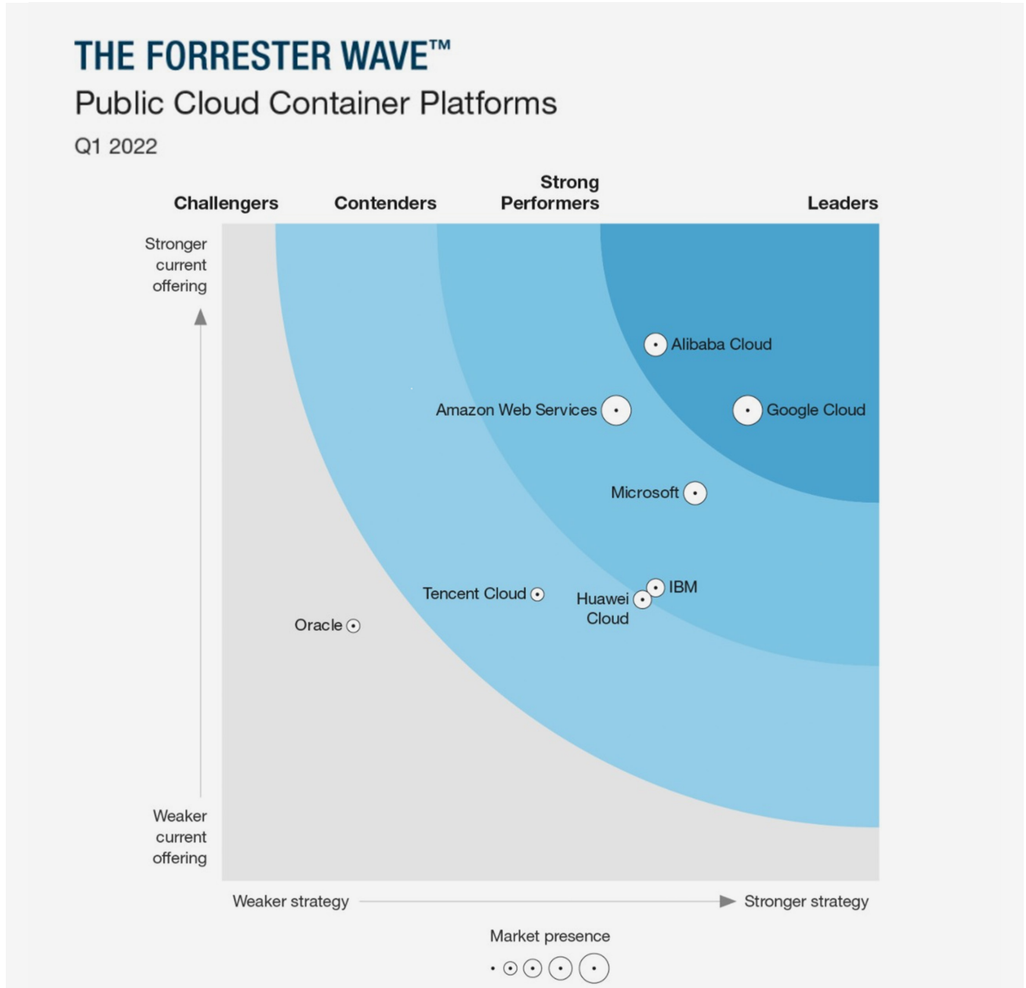
With the constant evolution of business models, coupled with the rise of new digital technologies, businesses are faced with the exponential growth of data and require an intelligent, cost-effective way to manage data. These factors, and the rising cost of proprietary data and AI technologies, have made it inevitable for businesses to migrate to the cloud to take full advantage of the benefits brought by cloud-native technology. From a technical perspective, cloud-native architecture is a collection of architecture principles and design logic based on cloud technologies. In the cloud-native architecture, features that are not business-centric are abstracted from the code base of applications. In particular, this applies to non-business features frequently associated with a cloud-based deployment, such as elasticity, resilience, security, observability, and canary release. This makes cloud-native applications highly automated and more lightweight and agile.

Due to the cost advantages, flexibility, elasticity, and business-friendliness of cloud-native technology, many enterprises have started to adopt it in their architecture for business.

ALIBABA CLOUD'S LEADING CLOUD-NATIVE DATA AND AI TECHNOLOGIES

According to the technology trend of cloud computing, Alibaba Cloud has dedicated investment in the deep research and development of cloud-native data and AI technologies and products too. We have become a global leader and advocate of cloud-native data and AI technologies, owing to our extensive experience supporting Alibaba Group and delivering solutions to enterprise customers worldwide.

Among the eight most significant players evaluated, Alibaba was named a "Leader" in The Forrester Wave™ Public Cloud Container Platforms Q1 2022.

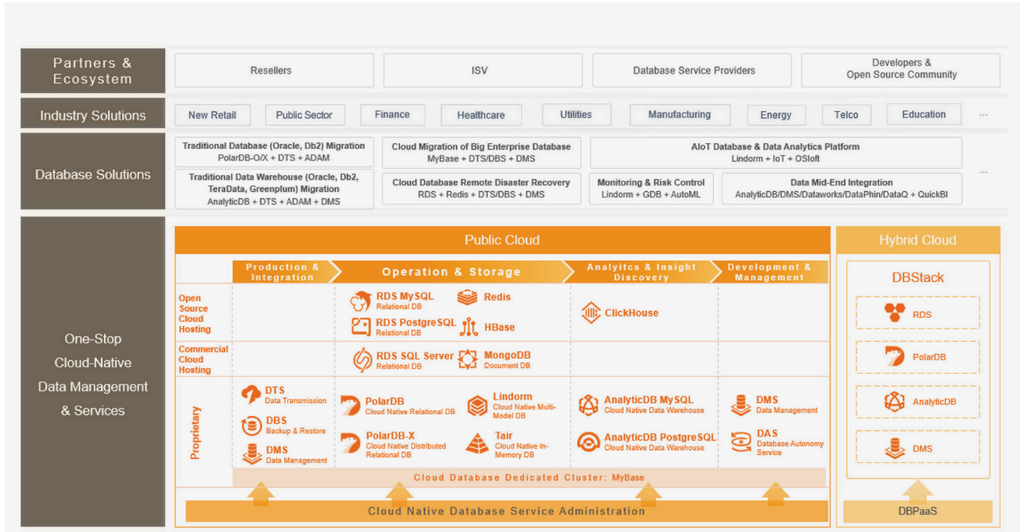


Source: Forrester Research, Inc.

Alibaba Cloud is also one of the first cloud computing companies to propose and implement cloud-native technology. Through continuous research and development, we have formed a comprehensive offering of data and AI products, including databases, big data analytics, and AI products.

Cloud-native Databases

Our comprehensive database offering flexibility supports public and hybrid cloud deployment, which includes database products for product integration, operation and storage, analysis and insight, and database development and management.

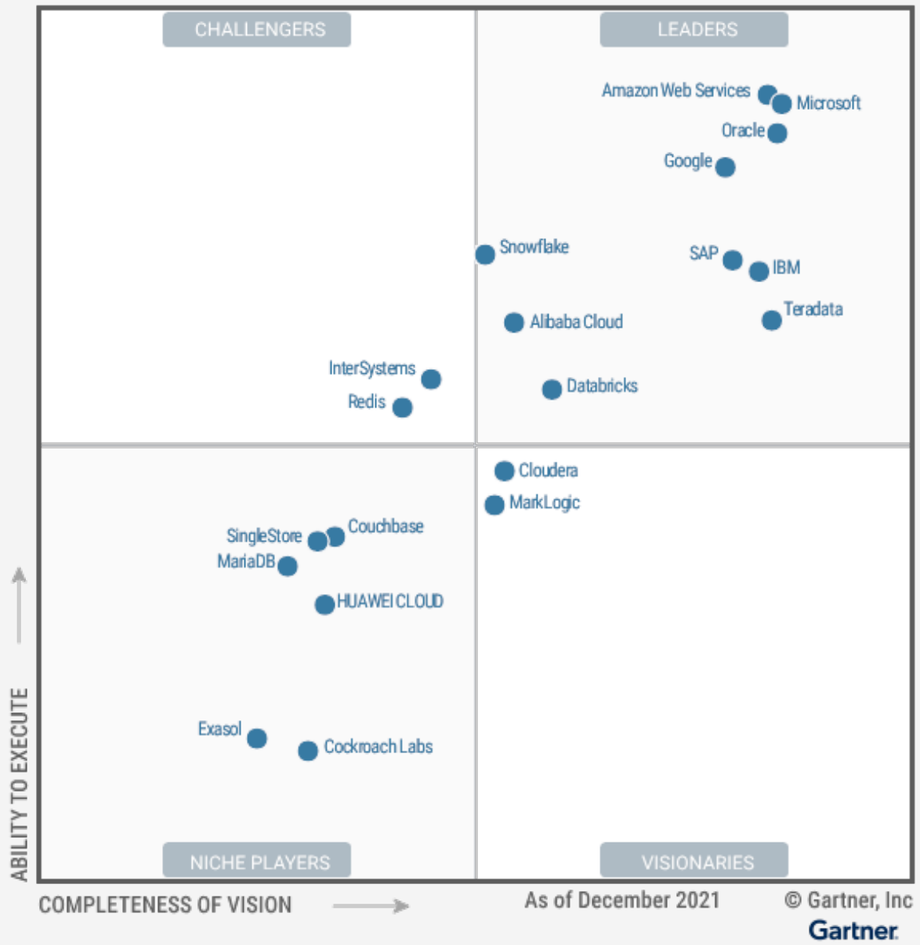


In the field of OLTP, Alibaba Cloud provides the cloud-hosted ApsaraDB RDS database service, the cloud-native relational database PolarDB and the distributed version, PolarDB-X. PolarDB has been used in various scenarios, including supporting Alibaba Group's core systems and powering the annual 11.11 Global Shopping Festival. Furthermore, PolarDB-X can support tens of millions of concurrent scales and hundreds of petabytes of mass storage.

In the field of OLAP, Alibaba Cloud has developed AnalyticDB, a new generation of the cloud-native data warehouse. It has excelled in TPC-DS/TPC-H benchmark tests and supports PB-level ultra-large-scale data.

In the NoSQL field, Alibaba Cloud has launched Lindorm, a cloud-native multi-modal database, and Tair, a cloud-native in-memory database. In addition, we have also built an enterprise-level database ecosystem tool product system, as well as a cloud-native intelligent database management and control platform.

2021 Gartner® Magic Quadrant™ for Cloud Database Management Systems



Cloud Data Warehouse: AnalyticDB

THE FORRESTER WAVE™

Cloud Data Warehouse Q1 2021

Strong Performers

TPC® 1st

TPC-DS benchmark

64% faster than the 2nd

3/4 cost of the 2nd

TPC-H benchmark

3X faster than the 2nd

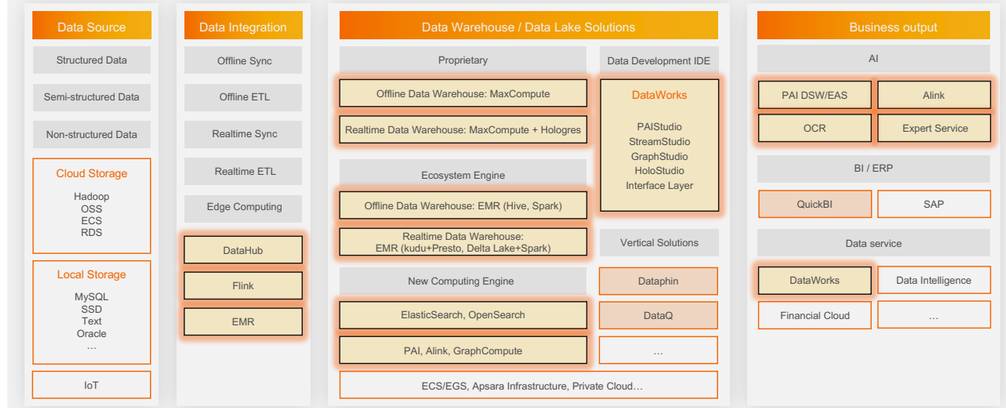
Source: Forrester 2021, TPC 2020

In 2021, Alibaba Cloud was named a Leader in Gartner® 's 2021 Magic Quadrant for Cloud Database Management Systems (DBMS).

In 2021, AnalyticDB was named a Strong Performer in the Forrester Wave™ Cloud Data Warehouse report.

Big Data Analytics

Alibaba Cloud offers simple, easy-to-use, fully managed cloud-native big data analytic services to enable business agility. Our big data + AI integrated platform integrates machine learning capabilities and provides a complete set of cloud-native technology architecture and product systems for the business. Our proprietary big data cloud-native integrated data warehouse supports offline real-time integration, batch-stream integration, lake-warehouse integration, and full-link data governance, ideal for big data analysis in industries such as the Internet, retail, finance, automobile, smart city, and manufacturing.



In March 2021, Alibaba Cloud MaxCompute and DataWorks entered the Forrester Wave™ 2021 Q1 Cloud Data Warehouse Strong Performers Quadrant. Our data warehouse product, MaxCompute, provides real-time and exabyte-level capabilities to support the needs of any data warehouse.

Source: The Forrester Wave™ : Cloud Data Warehouse, Q1 2021

At the same time, Alibaba Cloud, represented by the "Ververica platform," was named a Strong Performer in The Forrester Wave™ : Streaming Analytics, Q2 2021.

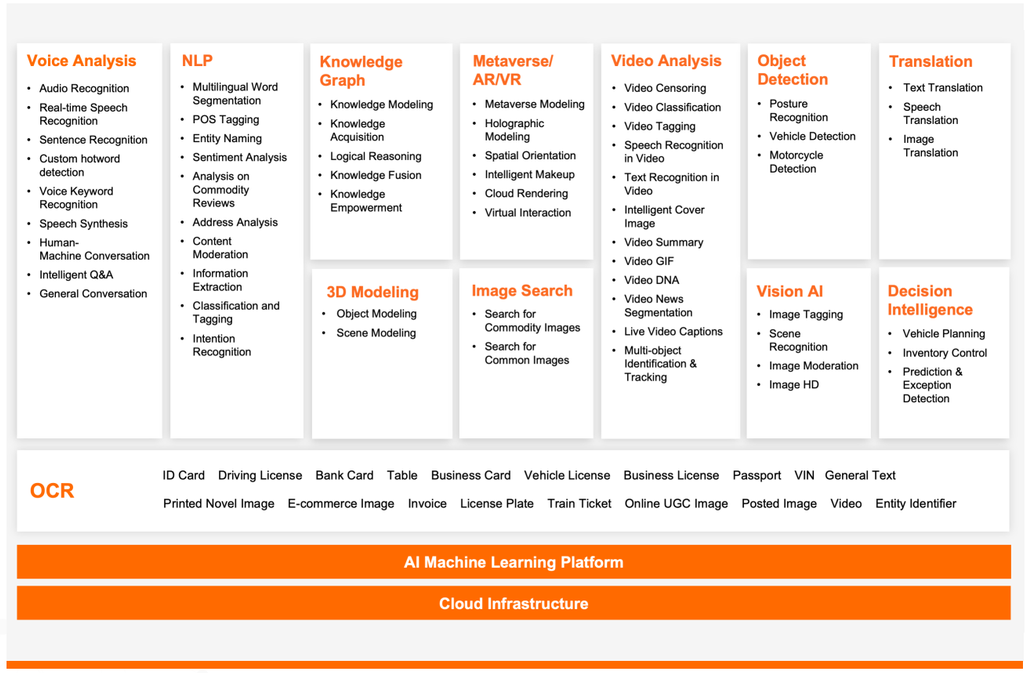
Source: The Forrester Wave™ : Streaming Analytics, Q2 2021

Alibaba Cloud is committed to building a comprehensive cloud-native database product ecosystem, allowing customers to solve the most complex challenges in data management, storage analysis, and computing. We have always adhered to the needs of our customers. We are investing more to develop more products that can solve our customers' pain points, delivering more value to their businesses.

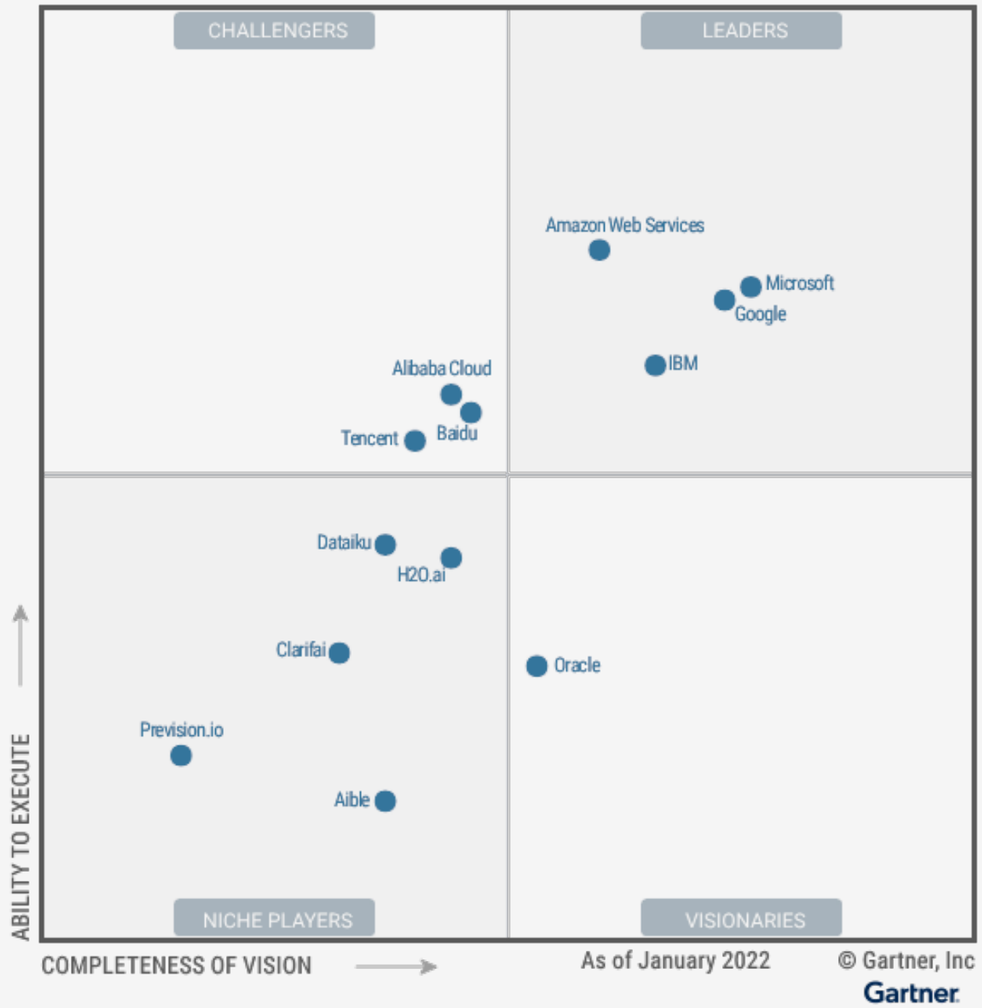
In August 2021, Alibaba Cloud MaxCompute participated in the TPCx-BigBench 100TB and 30TB testing and certification. MaxCompute successfully improved its performance on the 100TB scale to 46,059.27QPM, a significant increase of 73% compared to 26,501.53QPM in 2020.

Artificial Intelligence

Our advanced AI capabilities are a testament to the world-class research and development (R&D) efforts delivered by Alibaba DAMO Research Institute. After years of R&D, we have developed a complete cloud-native AI product technology system, providing rich scenario-based AI SaaS capabilities, including Vision AI, Speech recognition, OCR, NLP, Decision Intelligence, AR/VR/ Metaverse, and Machine Translation. These capabilities can flexibly adapt to the needs of enterprise users and are ideal for intelligent transportation, smart city, advanced medical care, manufacturing quality inspection, O2O retail, and other industries.



In May 2022, Alibaba was named a Challenger in the 2022 Gartner® Magic Quadrant™ for Cloud AI Developer Services.



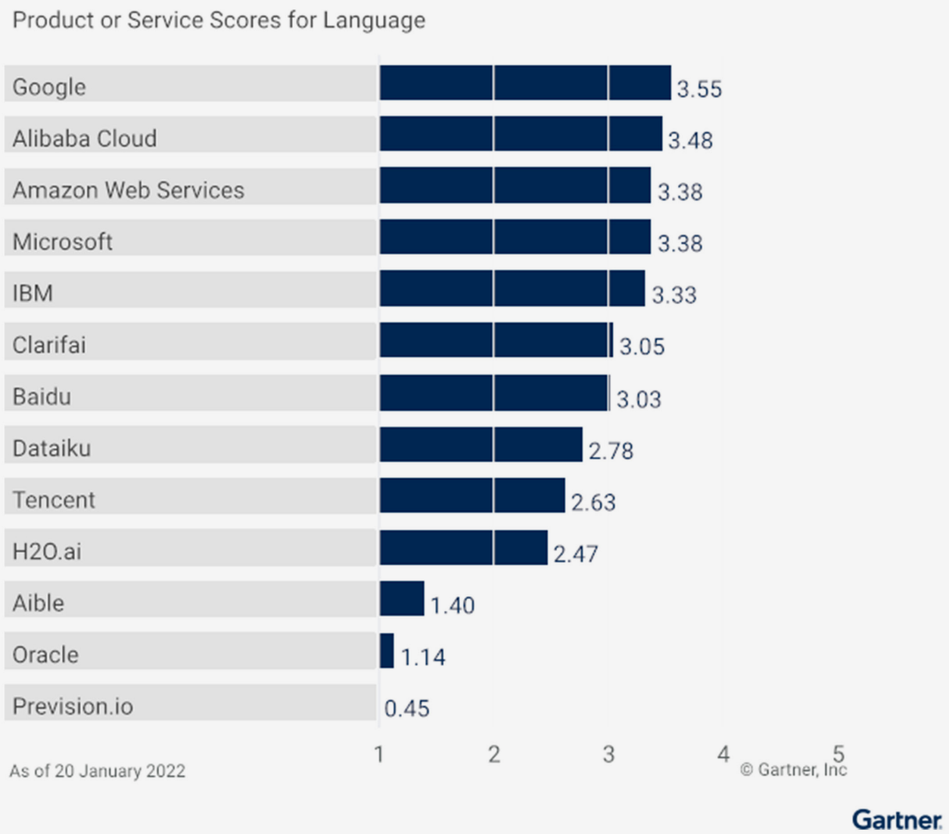
Source: Magic Quadrant for Cloud AI Developer Services, By Van Baker, Svetlana Sicular, Erick Brethenoux, Arun Batchu, Mike Fang, 23 May 2022

At the same time, in May 2022, Alibaba was ranked second for Language Use Case in the Gartner® Critical Capabilities for Cloud AI Developer Services. The scoring items include Speech to text, Language Understanding/Processing, Natural Language Generation/TTS, Translation, Sentiment Analysis, and Text Analytics.

Analysis

Critical Capabilities Use-Case Graphics

Vendor's Product Scores for Language Use Case



Source: Gartner (May 2022)

ACHIEVING BUSINESS INNOVATION WITH DATA AND AI TECHNOLOGIES

Innovation has become the lifeline for enterprises to remain competitive with increasingly fierce market competition. Leveraging data and AI capabilities is the key to business innovation in the digital age. Compared with traditional technologies, cloud-native data and AI technologies bring many advantages, such as cost efficiency and high flexibility, and the barrier to adopting these technologies is much lower. Enterprises only need to pay attention to the core business itself, improving business logic and customer value, rather than focusing on computing technology itself. The AI Model Development, Model Training, Model Service, AI Platform, and Data Platform sections will no longer be a concern for businesses.

Through advanced data and AI capabilities, logistics companies achieve the efficiency improvement of their supply chains, including developing accurate sales forecasts, strengthening data analytics capabilities, enhancing the management of commodity SKUs, and verifying forecast results. Data-driven processing, modeling, and forecasting are greatly simplified from the customer side as they were provided by cloud-native AI.

In addition, industry enterprises across gaming, e-commerce, and digital media are empowered to provide unique personalized services and enhance user experience.

These features are crucial for enterprises in these sectors to remain competitive and are developed based on real-time data analysis and AI technology.

Moreover, customer service and risk control are critical in the financial sector, where services often involve millions of users. Our cloud-native data and AI technologies have helped financial enterprises realize business value efficiently and at low costs.

Many industry enterprises have achieved business success through cloud-native data and AI technologies, including CapitaLand and El Corte Inglés.

“Digitalization has been revolutionizing our society and industries in a profound manner and CapitaLand has proactively embraced changes since 2015. We have continued to deliver success, however, to sustain our competitiveness and leadership in the digital age, we must adapt and keep pace with the times. Our future is filled with challenges. We will continue to develop as an innovative manager for real estate assets and funds. In the new stage and new landscape of China’s economic and social development, we aim to iterate ourselves with the help of digitalization, capitalize on our distinctiveness and strengths, seek new opportunities, drive business profitability, and create more value for our stakeholders.”

Lucas Loh,

President of CapitaLand and CEO of CapitaLand China

“By leveraging Alibaba Cloud’s expertise in retail digitalization as well as by harnessing Alibaba Cloud’s robust solutions, El Corte Ingles provides a more personalized customer experience and pushes for innovation.”

El Corte Inglés, Spain

06

OPENTREK: DRIVING THE INTELLIGENT TRANSFORMATION OF INDUSTRIES

By **Zhenyu Zeng**,
Vice President of Alibaba Cloud

As a leading cloud computing and artificial intelligence technology company, Alibaba Cloud has been keeping pace with the times, exploring and developing innovative technologies. In June this year, Alibaba Cloud launched a new industrial intelligence platform called OpenTrek, which can provide services to customers from all walks of life. This platform adds yet another vital piece of the puzzle to the digital transformation of enterprises.

As you may have guessed, the name "OpenTrek" is partially inspired by the show "Star Trek". "Open" refers to the platform's openness and the goal to serve customers and partners with an open mind. "Trek," on the other hand, refers to the long journey required to help us grow and achieve our ambitions in the industrial intelligence field, from the past to the future. Based on our core strategy of "putting partners first," OpenTrek will encapsulate the industrial intelligence capabilities that Alibaba Cloud has accumulated over the years from practical experience across various industries. By opening up to partners, the public sector, and enterprises, customers can fully explore the value of information and promote digital transformation efficiently and effectively.

THE BIRTH OF OPENTREK

We noticed a small but important detail. The technical solutions used in different sectors, such as transportation and the public sector, all involve similar elements. For example, the 'digital twin' technology was first applied to the transportation field, recreating environments through an array of cameras and radars. However, digital twin technology is now widely used in areas completely different from transportation, such as restoring the production line's complete working status through sensor data. This made us realize that in seemingly unrelated scenarios, many underlying technologies are common across different industries. Therefore, based on the original industry solutions, Alibaba Cloud abstracted the core technical capabilities and packaged them into a new platform named OpenTrek to improve the intelligence capabilities of all industries.

The overall architecture of OpenTrek is not complicated; the bottom layer consists of five core technologies of industrial intelligence: digital twin, simulation deduction, knowledge engineering, decision optimization, and collaborative computing. These are the five typical industry computing scenarios accumulated and summarized by Alibaba Cloud based on industry solutions. Further up the architecture are the data platforms, intelligent engines, APIs, and industry solutions.

GOING THE EXTRA MILE WITH OPENTREK

In the beginning, Alibaba Cloud envisioned the "simplest" model for OpenTrek – providing a proprietary cloud platform, coupled with MaxCompute big data computing services, PolarDB database, and other PaaS products, and handing over all the business of specific scenarios to partners, which is the so-called "customer-led" model. However, we soon discovered that this model was not the best approach to the problem. The traditional cloud migration process is only an upgrade from on-premises or Oracle-based architecture to a cloud-based architecture, which is an efficiency improvement but not a business innovation. On the other hand, customers are only concerned with the end result of "digital transformation" rather than the underlying technical architecture. Taking the transportation sector as an example again, customers require a business system to manage traffic flow and solve the problem of managing large-scale cities without congestion and providing higher traffic efficiency.

Alibaba Cloud has had a similar experience. In the process of delivering a public project, Alibaba Cloud only provided the underlying Apsara cloud platform initially and handed over specific intelligent applications to local software developers. This solution sounded logical initially, but we found it difficult, if not impossible, to implement it effectively. In the end, the developers could only respond passively and reactively to customer needs, whether the requirements are reasonable or not, severely affecting their experience.

Therefore, to overcome these challenges, Alibaba Cloud had to explore a new path by delivering a comprehensive cloud platform, which is now known as the OpenTrek industry intelligence engine. Based on OpenTrek, partners can easily leverage Alibaba Cloud's intelligence and automation capabilities through APIs to create more diverse solutions. Unlike other upper-layer business systems that simply use the cloud platform with Spark or Flink to migrate business applications to the cloud, Alibaba Cloud OpenTrek is an automated, algorithm-driven intelligent engine that can adapt to even unprecedented scenarios in the most innovative industries.

FINDING THE PERFECT BALANCE

Through OpenTrek, we struck a perfect balance of managing the platform's technicalities while providing our partners the required flexibility to adapt to different industry scenarios. We believe this is the crucial factor differentiating Alibaba Cloud from other cloud vendors for industry digitization in the past few years.

The road to achieving industry intelligence is easier said than done, and it is a long journey. Only by going deep into the industry and clearly defining the most challenging problems in various scenarios can we find the best solutions by means of technology.

07

ALIBABA CLOUD: MADE BY THE INDUSTRY, FOR THE INDUSTRY

By **Raymond Xiao**,
Head of International Industry Solutions and Architecture, Alibaba Cloud Intelligence

Alibaba Cloud is the technology and intelligence backbone of Alibaba Group, responsible for the commercialization of all the digital capabilities of Alibaba Group's businesses. As an organization, Alibaba Cloud focuses on providing solutions and digital-native capabilities to our customers that can solve real-world business problems. Years of industry knowledge and best practices from Alibaba Group and its ecosystem partners in retail, logistics, finance, media, and entertainment industries have allowed us to present scenario-oriented industry solutions that can be applied to our customers' daily scenarios in an efficient, scalable, and secure way.

THE FINANCIAL SERVICES INDUSTRY

Financial services is a sector where the trend is moving towards smarter, more personalized, frictionless customer service. Financial organizations are now facing challenges on multiple digital fronts, but Alibaba Cloud has observed the changes impacting the financial sector and has proposed the 5C framework.

Addressing Financial Challenges – The 5C Framework

Alibaba Cloud aims to introduce a digital-native solution landscape for financial services from years of best practices in FinTech from Alibaba and its ecosystem with Ant Group. In the 5C framework, each 'C' represents a capability target for a specific user group in the financial institution:

1. **Digital Onboarding and Servicing** includes software as a service (SaaS) and hybrid cloud deployments to support mobile and digital operations, remote video-based service platforms, and an electronic know-your-customer (eKYC) solution, which helps financial institutions verify users online anytime and anywhere.

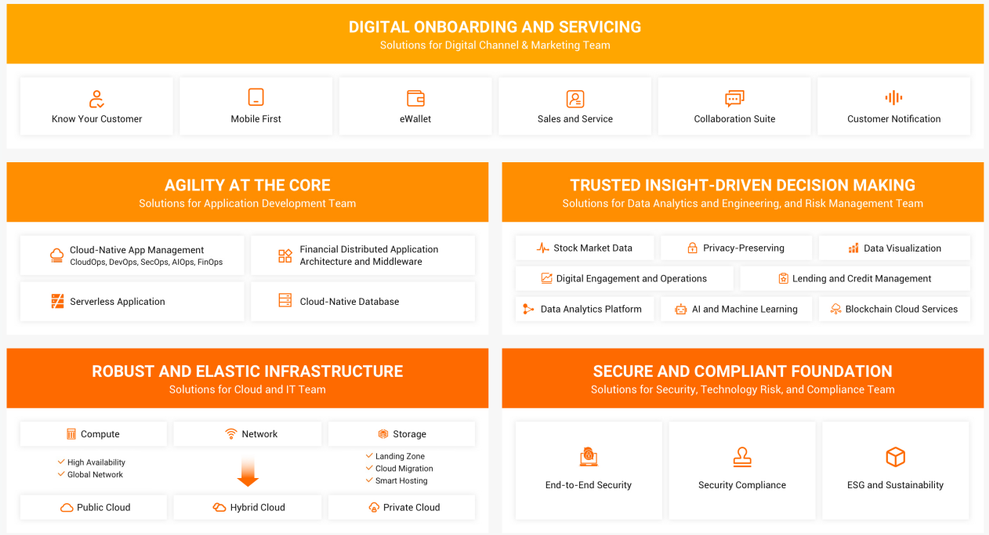
2. **Agility at the Core** offers a wide range of financial-grade cloud tools and middleware to develop cloud-native applications and allows FSIs to manage workloads through Alibaba Cloud platforms, their own premises, or via multi-cloud environments. One of the highlights is the full stack of cloud-native application management, including CloudOps, DevOps/DevSecOps, FinOps, AIOps, and SecOps.

3. **Trusted Insight-Driven Decision-Making** harnesses AI, machine learning, and big data to help FSIs analyze and visualize omnichannel digital engagement intelligence, aiding the discovery of customer insights and enabling intelligence-driven decision-making.

4. **Robust and Elastic Infrastructure** is a cloud infrastructure that is scalable and available worldwide, enabling customers to manage the expansion of their digital capabilities.

5. **Secure and Compliant Foundation** offers end-to-end security on the cloud, including network (Anti-DDoS), applications (Web Application Firewall), data (encryption, recovery, integrity, and management), platforms (e.g., physical security, virtualization, and hardware), and more. Alibaba also adheres to global, regional, and industry-specific compliance that can address some of the latest regulatory challenges.

ALIBABA CLOUD FINANCIAL SOLUTION ROADMAP



THE RETAIL INDUSTRY

The retail sector is undergoing a digital transformation that puts the customer at the forefront of change. Physical and online shopping will become more integrated, and the shopping experience on both sides will become more personalized and customer-centric. However, the sector also faces several challenges. There is increasing pressure to raise revenue and reduce the complexity of organizational structures when finding and retaining employees is becoming more complex. Additional challenges include finding new customers, offering them products they want to buy, and managing inventory.

Based on Alibaba's years of experience and best practices in retail, Alibaba Cloud has come up with a systematic approach and respective solutions to address retailers' challenges in the following 6 areas:

1. **Customer Attraction**: Attract potential customers from the public domain with online/offline advertising across various channels.

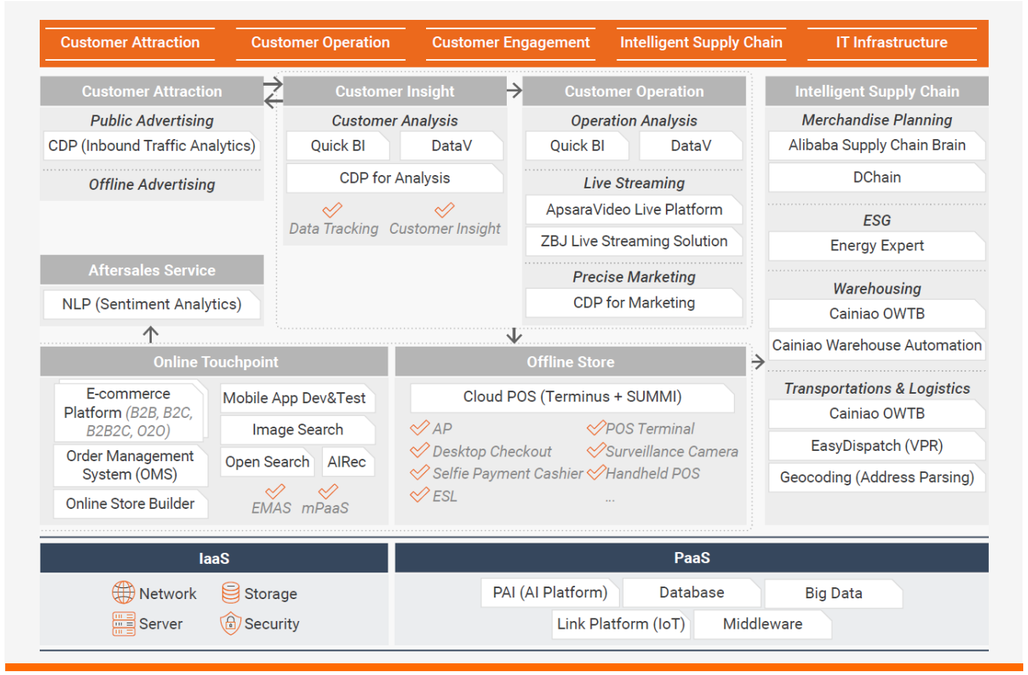
2. **Customer Insight:** Understand the attracted customers and then maintain the relationship by retaining them inside your private pool.

3. **Customer Operation:** Cultivate your members and launch marketing activities to influence their consumption behavior.

4. **Online Touchpoint and Offline Store:** Redirect customers to your online/offline stores to continue with physical consumption.

5. **Intelligent Supply Chain:** Collaborate your supply chain activities during order fulfillment.

6. **Aftersales Service:** Offer round-the-clock aftersales services and listen/speak to your customers proactively.

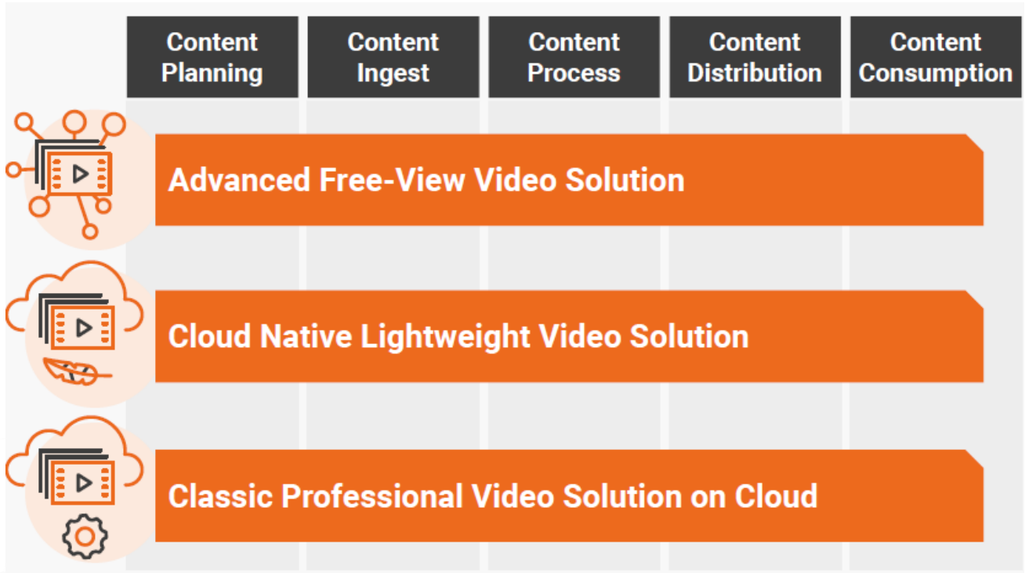


THE GAMING INDUSTRY

The gaming industry is growing in popularity every day. Console and mobile gaming allow friends and strangers alike to play cooperatively or battle for top scores. The end result is hours of fun for everyone, but there are also challenges. Ever-growing open-world games incur high development costs. Connecting with other players city to city, or country to country requires a reliable connection and latency. Players' accounts, in-game upgrades, and data over require security precautions. Alibaba Cloud empowers gaming businesses with cost-effective game development, secure and accelerated global access, high-performance and resilient infrastructure, and personalized gaming services.

THE MEDIA INDUSTRY

Today's viewers expect the best. Advancements in technology, especially 5G network connectivity, have led to a digital media revolution. The rise of streaming services has caused companies to flood the market with new content regularly, and viewers want instant access to that content anywhere and anytime. From video solutions and media storage to video processing and content distribution, Alibaba Cloud's media solutions provide everything a media company needs to deliver a streamlined user experience throughout the lifecycle of content.



08

ALIBABA CLOUD CIPU: A CATALYST FOR ACCELERATED DATA CENTER EFFICIENCY AND PERFORMANCE

By **Yang Hang**,
Senior Technical Expert at Alibaba Cloud

The history of computing goes back over 100 years. Since the early 19th century, the computing landscape globally has witnessed a rapid evolution and modernization. The emergence of disruptive technologies such as cloud, AI, ML, etc., and the growing computing demand for agility and efficiency across enterprises paved the way for a more robust and elastic cloud computing environment. Undoubtedly, cloud computing has emerged as the core driver of rapid digital transformation for organizations across varied industries.

However, specific challenges related to cloud data center infrastructure prevent companies from utilizing their full potential. The CPU-intensive architecture that enables unified resource scheduling and orchestration finds it challenging to meet the demands of continuously growing data and facilitate a low latency network for advanced data processing. Managing large-scale infrastructure and systems is another challenge.

These constraints necessitate a demand to transform the infrastructure of cloud data centers. To move beyond these roadblocks, Alibaba Cloud has unveiled a new architecture – **CIPU (Cloud Infrastructure Processing Unit)**, that puts the **Apsara Distributed operating system** and **new hardware** at its core. This article will take a detailed look at CIPU, its technical characteristics, architecture, and prospects.

UNIQUE CHARACTERISTICS OF CIPU

CIPU is a specialized service processor that brings computing, storage, and network infrastructure onto the cloud. This new architecture enables cloud operators to move to a virtualized storage, computing, and network architecture while accelerating hardware performance. Once compute, storage, and network resources are connected to a CIPU, they can be scheduled via a cloud platform to provide users with clusters that have hardware accelerated high elastic cloud computing power.

The architecture of a CIPU has the following characteristics:

1. I/O Hardware Device Virtualization

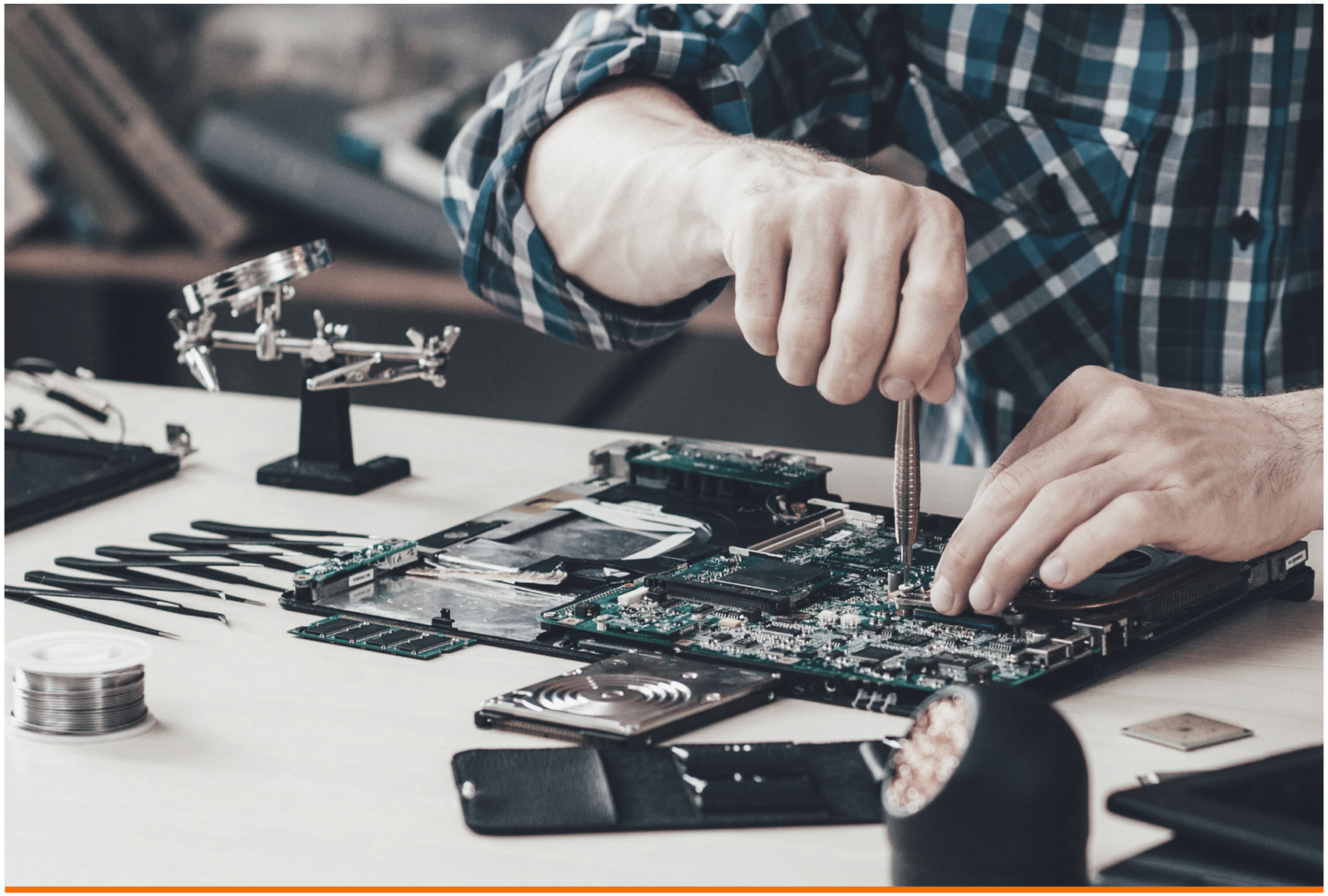
The VT-d technology is integral in helping achieve high-performance I/O hardware virtualization across industry-standard I/O device models such as virtio-net, virtio-blk, and NVMe. Optimizing PCIe protocol layers is also essential to ensure the high performance of I/O devices. The key to I/O hardware device virtualization is reducing the traffic of PCIe transaction layer packets (TLPs) and the number of interrupted guest OSs while balancing the latency requirement. Further, hardware resources in a queue should be pooled flexibly and new I/O services should be fully programmable or flexibly configurable.

2. Hardware Acceleration for VPC Overlay Networks

There's a strong need for hardware-based forwarding acceleration, given the wide gap between network bandwidth and CPU processing capability and the challenge of optimizing DPDK-based pure software network forwarding performance. Hardware-based forwarding acceleration can be implemented using:

- Forwarding technologies based on configurable ASIC such as Intel FXP that has the highest performance per watt (PPA) and lowest forwarding latency.
- Many-core NPU-based technologies provide forwarding flexibility to a degree, but performance per watt and forwarding latency is not comparable to forwarding technologies based on configurable ASIC.
- Forwarding technologies achieved through FPGA reconfigurability logic have a significant advantage in time to market but face challenges in forwarding 400 Gbit/s and 800 Gbit/s services.

All in all, there is a tradeoff at the technical implementation level. But, CIPU enables more in-depth vertical customization based on the vendor's forwarding services, thus obtaining more extreme PPA and minimizing forwarding latency.



3. Hardware Acceleration for EBS Distributed Storage Access

Decoupled storage and compute are necessary to achieve nine 9's data durability for public cloud storage and high elasticity across compute and storage. Alibaba Cloud Elastic Block Storage (EBS) must be able to access distributed storage clusters with high performance and low latency. The storage protocol between the computing initiator and the distributed storage target is highly vertically optimized and customized by cloud vendors. This forms the foundation of how CIPU realizes hardware acceleration for EBS distributed storage access.

4. Hardware Acceleration for Local Storage Virtualization

Although local storage does not have nine 9's data durability and reliability like EBS, it has low cost, high performance, and low latency advantages. After local disks are virtualized, the bandwidth, IOPS, and latency are not attenuated, and at the same time, the capabilities of one-to-many virtualization, QoS isolation, and O&M are obtained. This forms the core competitiveness of hardware acceleration for local storage virtualization.

5. Elastic RDMA

Remote direct memory access (RDMA) networks are critical for HPC, AI, big data, databases, storage, and other data-centric services. Implementing inclusive RDMA capabilities on the public cloud is a crucial capability of CIPU. Elastic RDMA in the cloud must be free from dependence on PFC and lossless networks. At the implementation level of elastic RDMA, the first step is to overcome the low-latency hardware forwarding of VPC. Without the PFC and lossless networks, the deep vertical customization and optimization of transmission protocols and congestion control algorithms become inevitable options for CIPU.

6. Secure Hardware Acceleration

Security is a must for public cloud intrinsic requirements. To continuously improve the competitiveness of cloud services, cloud vendors must embed robust security measures across hardware and perform full encryption of VPC east-west traffic and EBS and local-disk virtualization data. It's also vital to develop hardware-based enclave technologies.

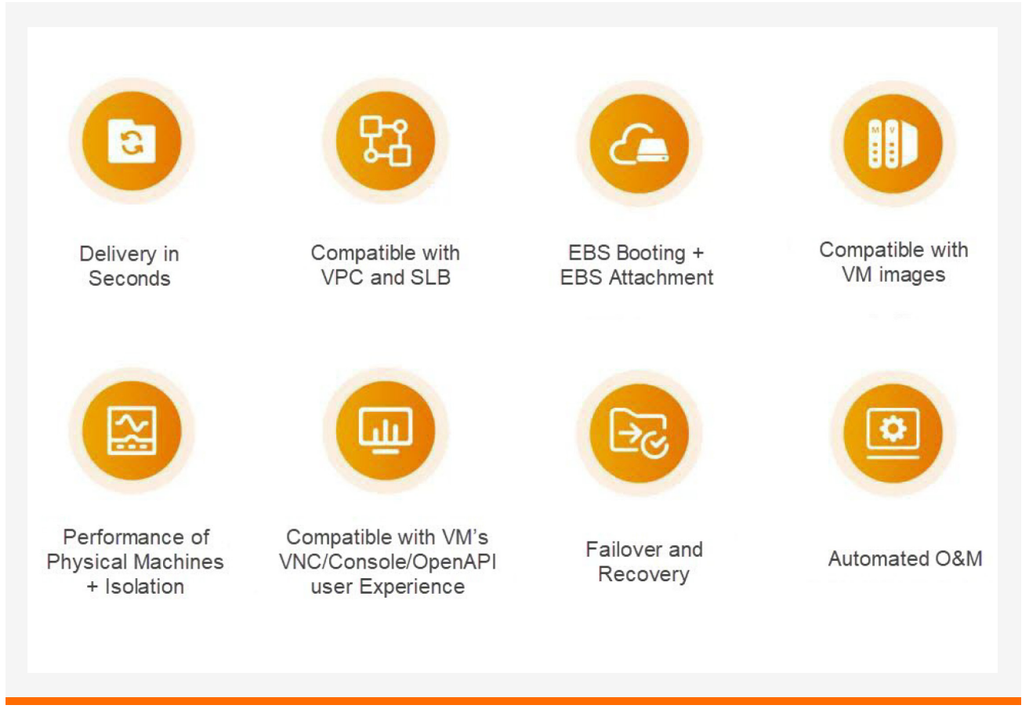
7. Cloud O&M Support

The core of cloud computing is servitization, which allows users to use IT resources without O&M. At the center of the IaaS elastic computing O&M lies the capability to upgrade all components and migrate virtual machines without any potential loss. Therefore, a large amount of software and hardware co-design work between CIPU and the cloud platform is involved.

8. Elastic Bare Metal Support

The following figure shows the eight key business characteristics that must be implemented by elastic bare metal services at the definition level.

For high elastic efficiency, cloud computing requires scheduling compute resources such as elastic bare metal instances, virtual machines, and secure containers in a single resource pool.



9. CIPU Pooling Capability

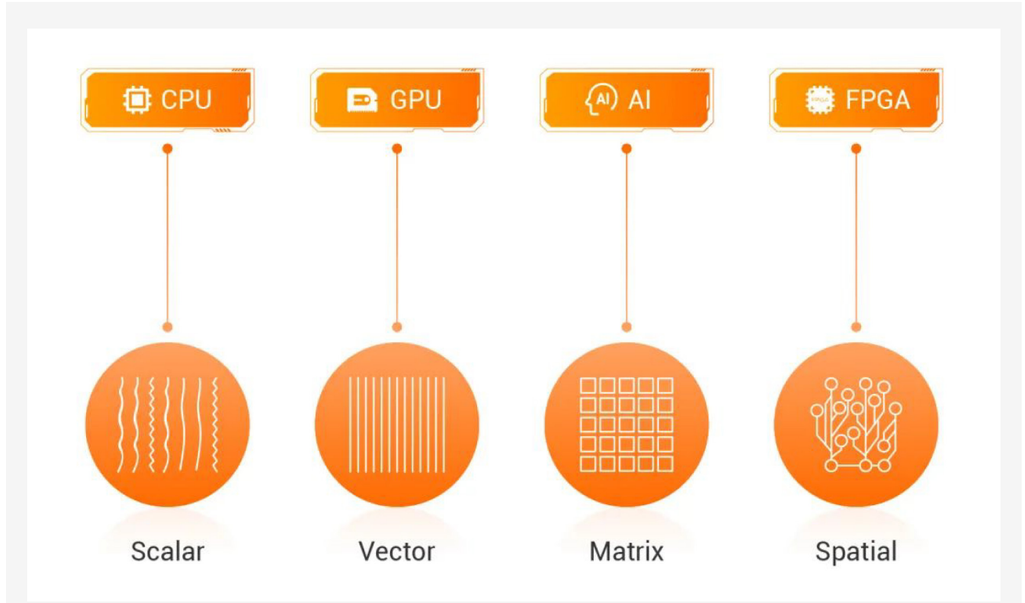
The CIPU pooling technology helps general computing significantly improve the utilization of CIPU resources, improving the core competitiveness in terms of cost. A set of CIPU technology architecture systems can also quickly meet the needs of high-bandwidth services such as AI.

10. Computing Virtualization Support

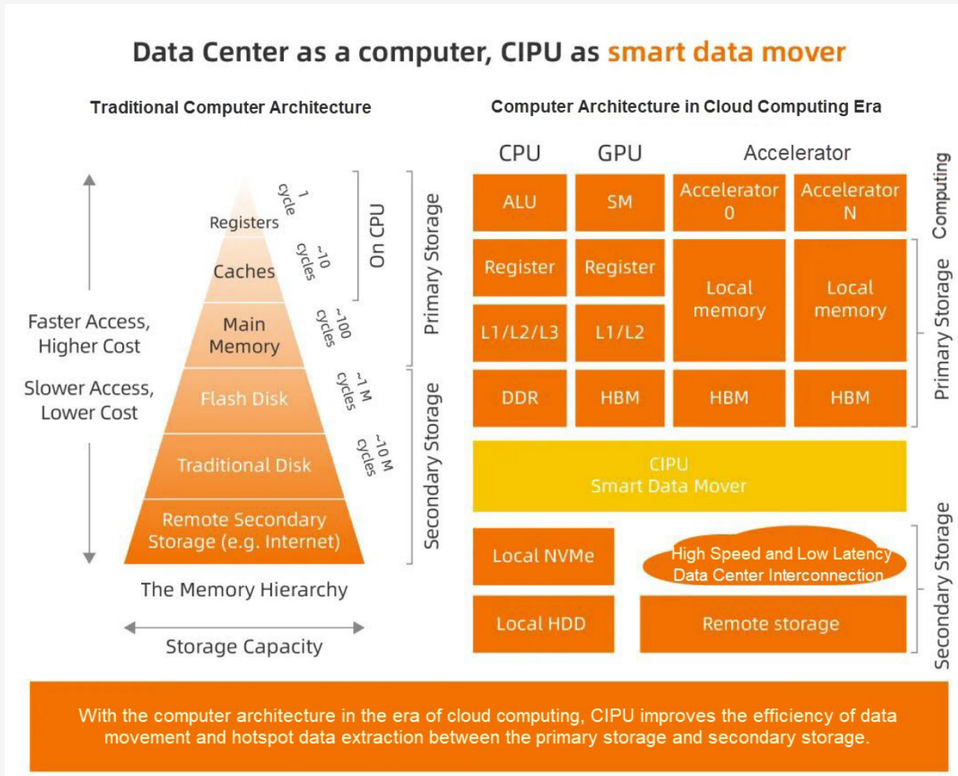
The features of computing virtualization and memory virtualization are enhanced. Cloud vendors have defined many core requirements for CIPU.

CPU ARCHITECTURE ANALYSIS

On the surface, it may appear that CIPU hardware acceleration is achieved by offloading the computing power. But it's not that simple. XEON computing power refers to processing capabilities such as arithmetic logic units (ALUs), hierarchical data cache, and memory access capabilities. For general or scalar computing, XEON's superscalar computing capability is sufficient. It is unrealistic for CIPU to complete the offloading of XEON ALU computing power and GPU stream processors in general scalar computing and AI vector computing. For instance, Intel defines the computing power feature of different workloads and the best-fitting chip.



The next question that comes to mind is about the common characteristics of the most suitable service workloads for the socket of CIPU. The in-depth analysis of the ten attributes of CIPU reveals its common features. Workloads need deep vertical software and hardware co-design during the data movement process to **reduce data movement** and **improve computing efficiency**.



Hence, from the perspective of computer architecture, CIPU's core role is to optimize the access efficiency of data hierarchical cache, memory, and storage between and within cloud computing servers. In a nutshell, CIPU is inline heterogeneous computing which is further equivalent to In-network computing and In-storage computing.

CONCLUSION

Summing it up, CIPU serves as one of the core pillars of the cloud computing strategy at Alibaba Cloud. It builds upon our existing groundbreaking innovations in the cloud and efficiently addresses the challenges and complexities across networking, data processing, and storage in the modern cloud data center. CIPU helps speed up infrastructure functions, optimize overall performance, and reshape the future of cloud data centers.

Alibaba Cloud

www.alibabacloud.com