Alibaba Cloud

Alibaba Cloud

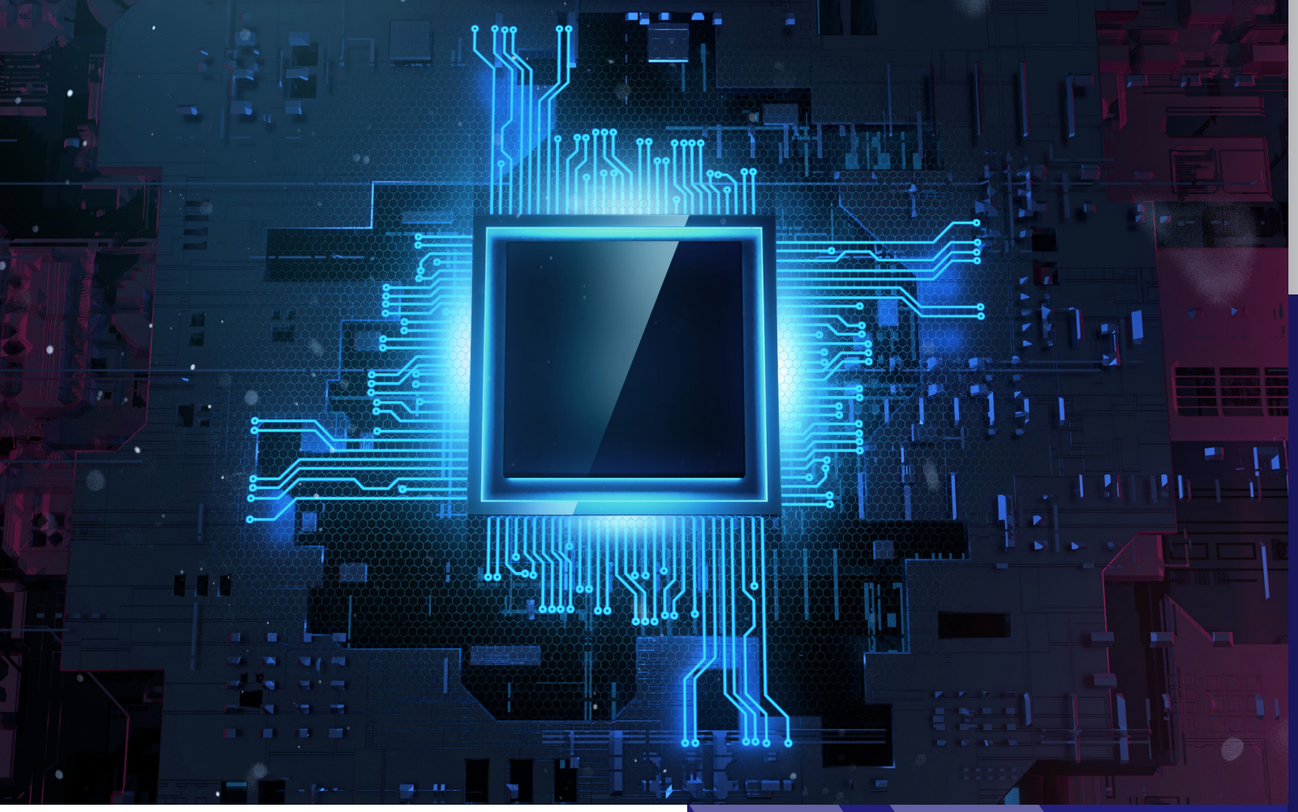
www.alibabacloud.com

2020 Edition 4

Cloud for Business Continuity

Alibaba Cloud

The Fourth Issue



ABOUT US

Editor in Chief / Selina Yuan

Senior Review Editor / John Jiang

Editor / Stephanie Gao

Review Editor / Yan Yang, Raymond Huang, Ning Meng, Cong Guo, Olivia Kang, Qijing Liu, Yanjun Wang, Heyu Guan, Xiaoming Lu, Quan Cheng, Xue Bai, Lan Lai, Lina Zhu

Website Planner / Sandy Zhang, Sue Zhou

Legal Advisor / Ava Zhao

PR Advisor / Crystal Liu

Proofreading Editor / Avex Li

Art Director / Diandian Wang

Designer / Longze Ma

CONTENT

Enable the Powerful Computing Ability —— 3

Cloud for Business Continuity —— 5

12 Years Technical Journey with Alibaba, Alibaba Cloud Infrastructure Leader John Jiang Talks about the Future of Cloud Computing —— 7

40-Year Evolution of Virtualization: from VMware to Alibaba Cloud X-Dragon —— 13

Alibaba Cloud Core Technologies Behind Four World Champions and Inference Performance Five Times Faster Than Its Nearest Rival —— 17

Innovators partner to create better animation production experience —— 23

Alibaba Cloud Releases PrivateLink to Help Enterprises Build Private Network Services —— 27

Alibaba Cloud Releases ALB to Accelerate the Delivery of Enterprise Applications —— 31

Run Kubernetes on Alibaba Cloud, Whose Container Technology Ranks No.1 in Gartner's Public Cloud Container Services Competitive Landscape —— 35

Alibaba Cloud Security Center Named in Gartner Market Guide for CWPP —— 41

Alibaba Cloud Released Industry's First Trusted and Virtualized Instance with Support for SGX 2.0 and TPM —— 43

From "Roughcast House" to "Fine-Decoration House" – Enterprise IT Governance Solutions for On-Cloud Management and Governance —— 45

Management Automation – Enterprises' Inevitable Approach to Cloud Migration —— 51

Enable the Powerful Computing Ability

We have just finished Double 11 shopping festival in 2020 when Alibaba has made a new miracle, with the number of orders per second peaking at 583,000 in only 26 seconds on November 11. 80% of Alibaba's core systems are now running on Alibaba Cloud's container service and ACK, which can scale a workload out to more than 1,000,000 containers in less than an hour. Serverless technology was applied on a big scale for the first time, with new improvements to serverless platform's elastic scaling capabilities, boosting performance by more than 10 times.

When the business is in full bloom, leveraging technologies to turn business ideas into reality is fundamental. Alibaba Cloud is the technical backbone for the whole Alibaba Group, in particular the powerful computing capability.

Alibaba Cloud provides a variety of infrastructure services including "bare metal" servers with no virtualization layer, Kubernetes container services, and a wide array of storage services. Alibaba Cloud also offers network and global acceleration services that allow for building a worldwide private network in minutes.

In AI, Machine Learning and Data Analytics, Alibaba Cloud provides a unique array of both open source and proprietary tools for batch and stream data processing, as well as tools designed to help the demander construct his own data warehouse or data lake with minimum effort.

Since the launch of its ECS (Elastic Computing Service) in 2010, Alibaba Cloud's own computing service has gone through several generations. In 2015, our self-developed public cloud platform Apsara, supported extremely demanding applications like China's 12306 railway ticketing platform which has to deal with huge pressure of hundreds of millions of travelers simultaneously booking train tickets online during China's Spring Festival.

How to improve computing capabilities of traditional server architecture? Alibaba Cloud's self-developed X-Dragon architecture pioneered hybrid software and hardware integrated virtualization technology, which solved many problems associated with resource contention and performance losses due to virtualization.

In Gartner's latest cloud vendor evaluation report, Alibaba Cloud ranks first in the world with a high score of 92.3% in the "computing" category, the best performance ever among all cloud vendors so far. In addition, Alibaba Cloud also ranks second in the world in terms of storage and IaaS capabilities. In May 2020, Gartner released the Global Cloud Security Report, Alibaba Cloud's overall security capabilities ranked second in the world, and 11 security capabilities achieved full scores in Gartner's assessment.

As container-based technology is in higher demand amongst cloud users, containers expect to become a basic part of any cloud deployment. Containers, microservices, serverless technology, and service grids are already in active use in Alibaba and have survived the strong demands and pressure placed on them during Double 11 this year.

Alibaba Cloud has extensive experience and best practices that we are keen to share, in hopes of benefiting more enterprises in their business growth with our powerful computing ability.

Selina Yuan

President of International Business
Alibaba Cloud Intelligence

Cloud for Business Continuity

In today's digital age, IT Services are the lifeblood of any organization, helping businesses create, manage, optimize and access their information and business processes. For savvy organizations, cloud-based solutions provide this competitive edge.

Alibaba Cloud is recognized by Gartner as one of the world's top three cloud computing companies. In this edition, we would like to guide you to explore Alibaba Cloud's expanding range of high-performance cloud products including large-scale computing, storage resources, network solutions, container service for Kubernetes, as well as enterprise IT and security capabilities for users around the world.

High Performance Computing and AI are crucial technologies in a digital world today. They are at the core of major advances and innovation in a wide range of industries. In this edition, we will introduce you to the evolution of virtualization technology in the past 40 years, explaining how the Vmware develops gradually into ECS Bare Metal Instance, a next-generation virtualization technology independently developed by Alibaba Cloud.

ECS Bare Metal Instance features both the elasticity of a virtual server and the high-performance and comprehensive features of a physical server. Compared with its predecessor, the next-generation virtualization technology of these instances excel in supporting standard Elastic Compute Service (ECS) and nested virtualization technology. This enables you to retain the elasticity capability of common ECS while delivering the same user experience as physical servers.

Based on ECS Bare Metal Instance's strong computing capabilities, you will understand better the secret technologies that helped Alibaba Cloud to break the records set by Google and other enterprises, taking first place in four categories in DAWNBench, one of the most authoritative image recognition competitions organized by Stanford University.

With cloud-native databases, storage is never a problem. With unlimited amounts available, you always have enough but you never have too much. Alibaba Cloud is the only Asia Pacific brand listed in Gartner's 2019 Magic Quadrant for Operational Database Management Systems. It is ranked first among database management system providers in Asia Pacific and ranked third in the world in terms of market share. In this edition, you will learn about how our high-performance cloud storage solutions help one of the world's most recognized digital production studios - Animal Logic, who requires a high-performance cloud storage platform that can quickly scale up for backup requirements, especially during peak production periods.

We will also guide give you a glimpse into Alibaba Cloud Container service for Kubernetes (ACK), a fully-managed service compatible with Kubernetes to help users focus on their applications rather than managing container infrastructure. ACK is integrated with services such as virtualization, storage, network and security, providing user a high performance and scalable Kubernetes environments for containerized applications. Alibaba Cloud is a Kubernetes Certified Service Provider KCSP and ACK is certified by Certified Kubernetes Conformance Program which ensures consistent experience of Kubernetes and workload portability.

For multinationals and organizations who have cross-border operations, we will provide with you our network solutions to help connect your business globally with our stable network anytime anywhere. You will get updated of our newly launched product Privatelink and ALB. With private network connections, users of Alibaba Cloud can access services provided by other Virtual Private Clouds (VPCs) through private networks, without additional Internet egress services. This ensures higher security and better network quality by preventing interactive data from going through the Internet. With the launch of ALB, Alibaba Cloud will focus more on facilitating application delivery and ensuring high elasticity, security, reliability, and cost-effectiveness of applications.

Alibaba Cloud is also committed to safeguarding the cloud security for every business. The security services on Alibaba Cloud reduce the heavy lifting required to tackle key enterprise security challenges in the cloud, and each of these challenges can be solved with the use of one or multiple Alibaba Cloud security services and products. You can quickly establish robust, end-to-end protection to address application security, data security, and platform security for new or migrated applications alike, and easily audit and govern the ongoing security posture, all by leveraging a comprehensive suite of enterprise security services and products on the platform.

In this edition, you will learn the story of how Alibaba Cloud Security Center, Selected as Gartner CWPP Global Market Guide, provides a unified security management system giving round the clock protection, and how it supports the Intel SGX technology allowing customers to have another option to run their most sensitive applications while keeping the applications and data protected.

Finally, we would like to introduce you to our new enterprise IT management platform – which serves as a model for enterprise users to construct a complex cross-account enterprise IT governance system on Alibaba Cloud.

Cloud-based solutions are bringing substantial benefits of cost savings, scalability and agility to businesses. Responding to global challenges with technology, that's how we hold on certainties in an uncertain world today.

12 Years Technical Journey with Alibaba, Alibaba Cloud Infrastructure Leader John Jiang Talks about the Future of Cloud Computing

Overview: Alibaba Cloud Magazine had an interview with John Jiang, Alibaba Partner, Senior Fellow of Cloud Infrastructure of Alibaba Cloud Intelligence, on his view of the future of cloud computing infrastructure. The story is an excerpt from the interview.

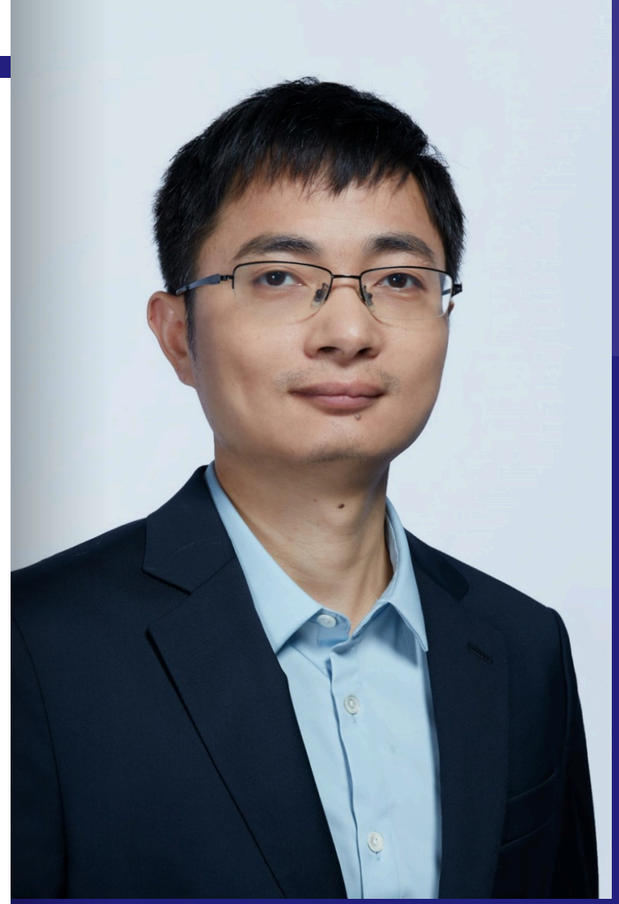
Mr. John Jiang joined Taobao in 2008 and led Taobao Mall's engineering team to design and implement its high availability architectural system. In 2012, he became the general manager of the Alibaba middleware product line and the technical leader of the high availability architecture team. He later joined Alibaba Cloud in December 2017, and now leads the cloud infrastructure product team on research and development of the Apsara cloud operating system. He is also a partner of Alibaba Group.

As a Senior Fellow of Cloud Infrastructure, you lead your team in developing cloud computing infrastructure products and services. What is your view on the development trend of cloud computing infrastructure technology in the next few years?

John Jiang: Cloud computing has gone through several stages. **In the first stage, cloud computing provides scalable IT resources.** This is the reason why enterprises choose cloud computing. Cloud computing help enterprises deal with fluctuations in business demand. In this stage, there is no big difference between the computing performance of cloud resources and that of traditional IT architecture.

In the second stage, Internet applications empower enterprises in digital transformation. Alibaba has accumulated rich technical expertise and industry know-how in various fields, such as new retail, new finance, new manufacturing, logistics, digital entertainment, and navigation. We have extensive experience in ensuring data consistency because the trading system of our e-commerce platform demands high data consistency with zero-tolerance to inventory and transaction amount inaccuracy. Global shopping festivals with massive amounts of transactions are even more complex. I believe that many enterprises are faced with similar complex business scenarios. Alibaba Cloud is one of the cloud service providers that can deal with data consistency, complex business scenarios and ultra-large traffic at the same time.

In the third stage, customers can build applications natively on the cloud, and thus evolving from the Internet technology to the next generation. Alibaba Cloud's technical capability are the combination of Internet-based IT technologies and cloud-based technologies. The former is the virtualization of computing, network, and storage resources. The latter refers to hardware, virtualization and application being synergized as the broader sense of cloud-native technology. A typical case in point is Alibaba Cloud's sixth-generation enhanced Elastic Compute Service (ECS) instances, which are built on top of the X-dragon architecture. The enhanced ECS instances can deliver a much better performance and improved cost performance ratio for running I/O-intensive businesses. Here comes why.



John Jiang
Alibaba Partner
Senior Fellow of Cloud Infrastructure,
Alibaba Cloud Intelligence

Firstly, the I/O capability will play a critical role in the performance of future Internet-based systems as the difference in chip's computing capability will diminish in time. Plus, most Internet-based systems are built on multi-node clusters and remote access, which depends on I/O capability.

Secondly, CPUs can be more customizable and can access memory resources in a more efficient manner to meet specific requirements. For example, Alibaba Cloud AEP instances are specially designed for large memory scenario.

Thirdly, memory will be managed as a resource pool. This is also what Alibaba Cloud is doing.

Last but not least, a unified technical system will provide a better user experience. In the 5G era, edge computing will develop rapidly. Alibaba Cloud will provide a unified technical system to ensure consistent user experience on both edge and cloud. In the future, artificial intelligence (AI), big data and high-speed network connection solution will also be deployed on edge. This unified solution will be highly applicable to high traffic scenarios such as live streaming, intelligent manufacturing etc.

Therefore, application systems running on Alibaba Cloud are two to three times higher in performance than traditional IT architecture with the same server configuration. Cloud-native architecture improves the performance of applications and reduces the hardware costs of computing power to support specific task requirement. For example, Alibaba Cloud storage systems can support 1 million input/output operations per second.

The performance gap will be further deepened in the next two to three years. This only opens up the new cloud-native era. It will go beyond the cloud-native architecture to the entire system.

How does the outbreak of COVID-19 affect the evolution?

John Jiang: Enterprises that provide digital services but have not adopted cloud computing yet, deployed their business only on on-prem IT systems. When the pandemic began, enterprises have to deal with a surge in demand for computing resources. One example is the enterprises engaged in live streaming and online education. Demand for online learning services soared during the pandemic. However, these enterprises had difficulty in expanding their business because they either could not find adequate computing resources or the servers cannot be shipped in due to disrupted logistic systems. If these enterprises used cloud computing technologies, they could easily access computing resources and scale business as needed.

Another group of enterprises are those who have not digitalized their traditional businesses. Compared with the first group, this type of enterprise might take a harder blow. For instance, if retailers or restaurants in retail or catering industry did not offer services or products on a mobile app, their services might have been interrupted because most people were requested to stay at home during the pandemic. Brick-and-mortar stores and supermarkets that sell fresh food cannot continue the business without online ordering and digitally-enabled delivery system.

The examples above are only a reflection of the fact that the digital transformation worldwide is accelerated by the pandemic. The pandemic also pushes the evolution of cloud computing infrastructure. Cloud service providers must speed up the R&D of cloud services to meet the ever-changing requirements of customers.

You announced some novel cloud-native products at the Apsara Conference held in September. How do you understand cloud-native technology?

John Jiang: I believe that cloud-native technologies are consistent with the cloud computing infrastructure trend I mentioned earlier. Cloud-native means our customers develop their IT technology directly on cloud, where our full range of technologies and services are fully accessible to them.

High scalability on cloud. Enterprise in their infantry stage with limited budget can use a small amount of cloud resources while they can flexibly adjust the amount of required resources based on business requirements when in full swing.

More stable technology on cloud. Alibaba Cloud provides computing capabilities that are as stable as Unix server and guarantees world-class SLA of computing and storage for customers.

Professional security management on cloud. Cloud has zero-tolerance to security issues which will be addressed by one-stop systemic security solutions. The security solutions crystallized from our own comprehensive business applications keep customers from any concern for security risks.

Comprehensive products on cloud. Customers can always find mature technologies that meet their business requirements on Alibaba Cloud. Both proprietary services and open source are available such as EMR, cloud-native remote procedure call (RPC), open-source Dubbo service. Such mature technologies help enterprises build better Internet-based systems.

Lastly and also importantly, enterprise IT governance capability on cloud. Alibaba Cloud enables customers to operate and manage their assets on cloud more effectively and flexibly with strong governance functionalities including accounts, permissions, cost, compliance managements and so on.

Each year, Double 11 is a major test of the supporting technologies. Last year, core systems of Double 11 were entirely running on the cloud. What are the new technologies that Alibaba Cloud provides to support Double 11 2020?

John Jiang: Double 11 going on cloud has gone through four phases. **In the first phase, Double 11 boosts the development of cloud computing.** The traffic during normal operations is only one-thirtieth of the peak time on November 11. Therefore, a large number of computing resources become idle after Double 11, which was then used for cloud computing.

In the second phase, there are a full range of technologies available on cloud. The Internet-based technologies that Alibaba Cloud uses now are the best practices validated by the business of Alibaba Group in the fields such as e-commerce and finance. For example, Enterprise Distributed Application Service (EDAS) and PolarDB are widely used in the technical system that supports the businesses of Alibaba Group.

In the third phase, the core systems adopt ready-to-use cloud products. In 2019 the core systems were deployed on the cloud which means the shopping festival not only adopts cloud technology but also uses our own cloud products directly.

In the fourth phase, which is this year's double 11, the system is going cloud-native via a large deployment of cloud-native technologies. Compared with previous years, Alibaba Cloud has made major breakthroughs in both products and technologies. Cloud-native products were fully utilized in supporting Double 11 2020. Big data technologies applied in search and advertising all used cloud products. We also optimized storage to deliver high I/O. In addition, the X-Dragon architecture and enhanced sixth-generation instance families are heavily used to provide high performance and reliable computing capability. ECS Bare Metal Instance is a computing platform designed for cloud-native scenarios. It supports 100 Gbps network bandwidth, 24 million packets per second (PPS) for forwarding, and 1 million input/output operations per second (IOPS) of disks. It can also add 500,000 vCPU cores within 3 minutes for scale-out. As proved by real business of Alibaba Group, an ECS bare metal instance improve the number of QPS by 30%, lower the latency by 96.3%, and significantly improve resource utilization compared with a physical machine. Besides, the hybrid deployment technology of Alibaba Cloud provides a unified resource pool for flexible deployment across services and the unified scheduling capability for large-scale load shifting.

Our four key Cloud-Native technologies supported breakthroughs in both scale and innovation:

1. Ultra Large Container Cluster and Mesh Cluster

X-Dragon and Container Service for Kubernetes (ACK) combined can scale up to one million containers in 1 hour during transaction peak time. In terms of container resources allocation and management, online and offline hybrid deployment utilization rate is 50%. Computing resources cost supporting every 10,000 transactions during peak time is reduced by 80% compared to 4 years ago.

2. The Largest Real-Time Computing Platform in China

Big data platform handles data volume up to 1.7EB per day, which amounts to7 billion people worldwide processing 230 high-resolution pictures per person. Flink consumes stream data up to 4 billion requests per second during peak time. Cloud-native database PolarDB's read/write performance is 50% higher than last year. PolarDB TPS reaches 140 million, 60% higher than last year.

3. Cloud-Native Middleware

Alibaba Cloud's middleware service framework deployed over 10 billion QPS

4. Large Scale Deployment of Serverless in Core Business

Serverless technology significantly enhances the efficiency and stability, increasing scalability by 10 times.

Double 11 this year with such a large scale of cloud-native technologies and products deployment, is a new milestone achieved. Double 11 is the largest trial ground for Alibaba Cloud products, technologies and services. For enterprise customers, **the 12-year practices of supporting Double 11 shopping festival is the best testimony to Alibaba Cloud product capabilities. Standing the test of time and history events, Alibaba Cloud is confident about supporting rapid business growth and innovation for our customers.**

In the past, the IT resources invested to meet our own business demand during peak time boosted the development of cloud computing. At present, with its continuous development, cloud computing serving as the technical backbone is cementing fast growth of Alibaba Group across all of its businesses. I also believe it is the path that all Internet companies in the world will go down.

40-Year Evolution of Virtualization: from VMware to Alibaba Cloud X-Dragon

Background Information on X-Dragon

In 2018, Alibaba migrated its entire business to cloud by deploying all the core trading systems on Alibaba Cloud. This migration represented a significant milestone.

This migration wouldn't be possible without X-Dragon. Before the migration, Alibaba was using a vast number of physical machines. It was a great challenge for Alibaba Cloud to support such a business volume, including e-commerce, finance, and logistics, especially the massive transactions during Double 11 shopping festival. But we made it. And the hero behind the scene is X-Dragon.

History of Virtualization

For many years virtualization has been the focus of research institutions and large IT companies.

In the history of virtualization, 1974 is a critical year in the formation of the earliest virtualization theory. In 1974, the thesis Formal Requirements for Virtualizable Third Generation Architectures was published. This thesis defined virtualization and described the requirements for implementing virtualization. This thesis laid the theoretical foundation for the rapid evolution of virtualization in the next 40 years.

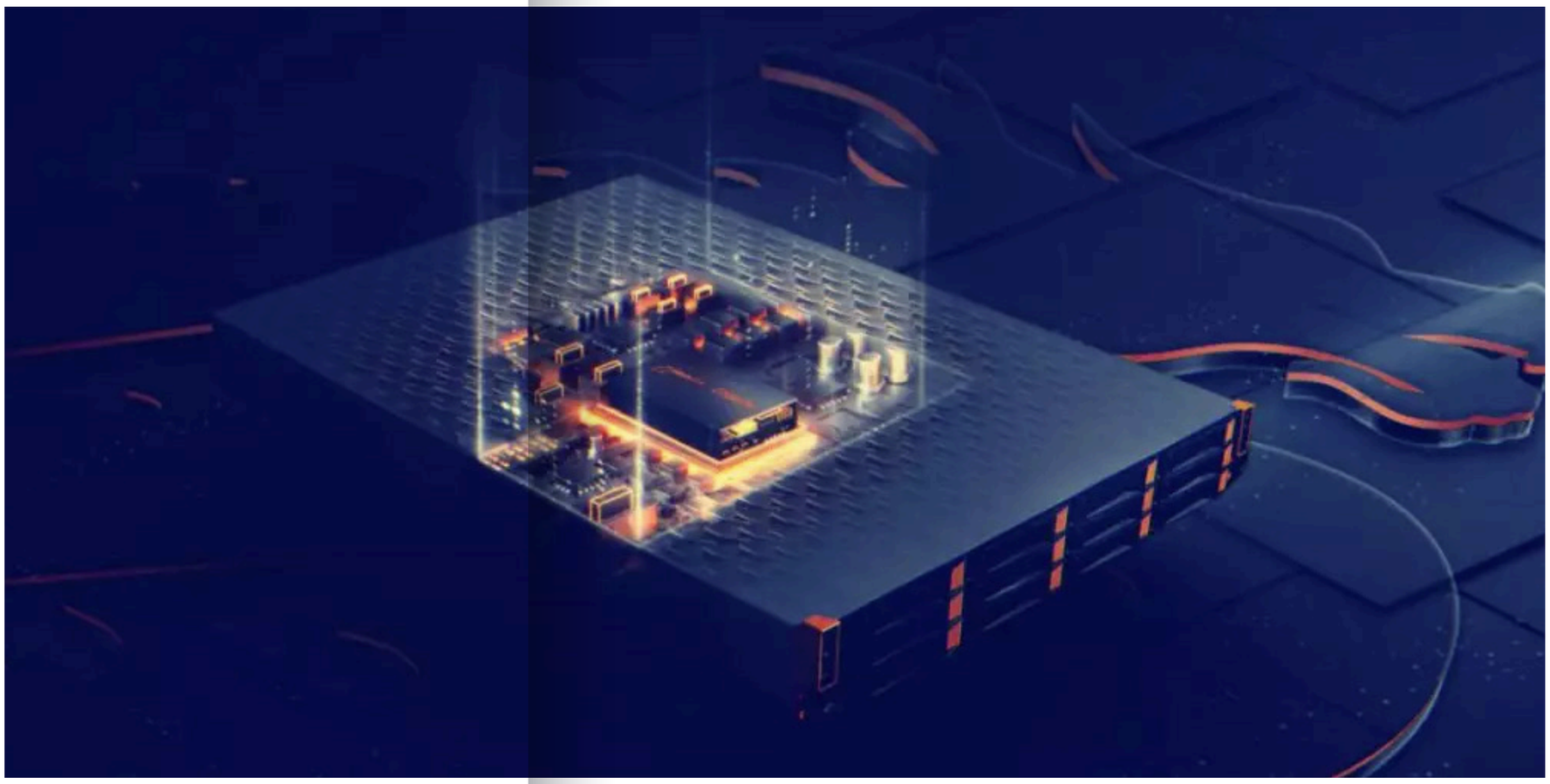
Another important year is 1997, when a professor from Stanford University founded VMware. The establishment of VMware put the theoretical research of virtualization into practice. VMWare developed binary translation techniques to implement full virtualization, but it was mainly for PC. In 2005, the virtualization of Internet data centers (IDCs) for cloud computing started to accelerate. That year, the chip giant, Intel, released VT-x, followed by AMD with its AMD-V released in 2006. These technologies enabled x86 CPUs to better support virtualization based on the expanded instruction set and CPU design. In 2009, Alibaba Cloud was founded. It was impractical to use a commercial software such as VMware to implement cloud computing. At first, Alibaba Cloud chose popular open source virtualization software Xen.

In 2014, Alibaba Cloud switched to Kernel-based Virtual Machine (KVM). Alibaba Cloud performed in-depth customization on both Xen and KVM based on their business requirements.

In 2014 and 2015, when Alibaba Cloud began to serve other large enterprises, it faced the needs to reduce costs and improve service capabilities.

During this time, the virtualization technology of Alibaba Cloud cannot keep up with the pace of its cloud computing business. In addition, Alibaba planned to migrate its entire business to cloud at that time, which posed a greater challenge to Alibaba Cloud. The virtualization technology must be innovated.

For this purpose, Alibaba Cloud started the exploration of its next generation of virtualization technology in 2015, set up the X-Dragon project in 2016, and launched the X-Dragon architecture in 2017. Traditional virtualization technologies tried to adapt software to the given servers and computing architecture. On the contrary, X-Dragon defines a new computing architecture for more convenient virtualization implementation. It employs an innovative design that integrates hardware and software. In the X-Dragon architecture, all performance-critical parts are implemented by hardware and other non-performance-critical parts such as the control plane are implemented by software. This ensures both flexibility and performance. Compared to traditional virtualization technologies, X-Dragon was purpose-built for the cloud. It brings better isolation of resource, lower virtualization overhead, and much higher performance. Last but not least, it supports a new type of cloud computing instance type called Elastic Compute Service Bare-Metal. This is a key feature that made it easier for Alibaba to migrate its business to the cloud.



Evolution of X-Dragon

First-Generation X-Dragon: Virtualization of Bare Metal Servers

The first-generation X-Dragon provides a solution to support virtualization of Elastic Compute Service (ECS) bare metal instances that are similar to physical servers. Yet different from traditional physical servers, ECS bare metal instances are integrated with the cloud computing infrastructure. For example, ECS bare metal instances can make use of pooled cloud storage resources, network resources, and databases. In short, ECS bare metal instances enjoy similar "cloud" experience to virtual machine instances while retaining the high performance of physical servers.

Second-Generation X-Dragon: Support for Virtual Machines

The second-generation X-Dragon supports both bare metal and virtual machine instances. It was adopted on massive scale to implement the 6th ECS of Alibaba Cloud.

The second-generation X-Dragon provides Dragonfly, which is a lightweight hypervisor and consumes very few resources. As a result, all computing resources on the physical machine can be allocated to virtual machines. Virtual machines are isolated from each other by using hardware queues. This way, virtual machines that reside in the same IDC do not affect each other.

Third-Generation X-Dragon: Extreme Performance

The third-generation X-Dragon was launched at Alibaba Apsara Conference 2019, providing the highest performance in the industry. The key performance factors of the third-generation X-Dragon, such as the storage and network performance, are almost five times higher than other similar computing architectures.

The third-generation X-Dragon has the following new features:

1. Builds the entire data plane into the chip, including storage and networking. This greatly improves performance.

2. Enables telecommunication-grade quality of service (QoS) management. For example, the third-generation X-Dragon precisely controls the number of packages and the size of bandwidth per second based on the requirements of storage and network. In the past, this feature was available only for telecommunication devices.

3. Implements an enhanced fusion network. The third-generation X-Dragon ensures low network latency that is close to the network latency between bare metal servers.

4. Provides enhanced hardware queues. The third-generation X-Dragon supports 1,024 storage queues and 1024 network queues, and further enhances the isolation among queues.

The third generation X-Dragon is now adopted in the Enhanced 6th generation ECS and High Frequency 7th Generation ECS. It provides 25 million PPS and 1 million IOPS, giving a huge performance boost of the above new ECS instances.

In addition, X-Dragon supports migration from VMware-based private cloud to Alibaba Cloud. Many data centers still use VMware-based private cloud. If customers want to migrate business from VMware-based private cloud to Alibaba Cloud, it is possible to with X-Dragon because traditional cloud computing servers do not support VMware. If a customer uses OpenStack based on KVM, the customer can also migrate its entire business in the OpenStack-based private cloud to ECS bare metal instances of Alibaba Cloud, thanks to the help of X-Dragon.

Today, X-Dragon is fully adopted for all of Alibaba's businesses and all the Public Cloud computing services of Alibaba Cloud. All the new servers added are powered by X-Dragon.

During the outbreak of COVID-19, Alibaba Cloud has provided a dozen of public research institutions with high-performance computing power, which is also supported by ECS bare metal instances based on X-Dragon.

Alibaba Cloud Core Technologies Behind Four World Champions and Inference Performance Five Times Faster Than Its Nearest Rival

Overview: How did Alibaba Cloud take the top place in four rankings in the image recognition field? The Alibaba Cloud heterogeneous computing team shares the technical secrets that allowed Alibaba Cloud to place first in the competitions.

Recently, the latest results of the DAWNBench ImageNet competition held by Stanford University showed that Alibaba Cloud surpassed Google and Facebook to take the top place in four rankings.

It took Alibaba Cloud only 158 seconds to train ResNet-50 on 128 V100 GPUs and reach top 5 93% accuracy. Alibaba Cloud reached top 5 accuracy of no less than 93% when classifying the 10,000 images in a validation set, with an inference performance more than five times faster than its closest competitor.

In the field of heterogeneous computing, Alibaba Cloud provides world-class capabilities in integrating Artificial Intelligence (AI) software and hardware to maximize the performance and minimize the costs of training and inference.

How did Alibaba Cloud achieve this?

What Does Such Achievement Mean

DAWNBench, designed by Stanford University, is a benchmark suite and competition for end-to-end deep learning training and inference performance. DAWNBench was announced at the 2017 Neural Information Processing Systems (NIPS) Conference and has gained wide support from the industry. The competition results have become one of the most influential and authoritative rankings in the field of AI.

Performance and cost are the two most important metrics in AI computing. The latest results from DAWNBench demonstrate the world-class performance optimization capabilities of Alibaba Cloud in training and inference based on the integration of software and hardware.

According to the heterogeneous computing AI acceleration team of Alibaba Cloud, such superior performance comes from the proprietary Apsara AI Acceleration (AIACC) engine, proprietary AI chip Hanguang 800 (also known as AliNPU), and heterogeneous computing cloud services of Alibaba Cloud.

AIACC, an AI acceleration engine independently developed by Alibaba Cloud, is the first in the industry to uniformly accelerate mainstream AI computing frameworks such as TensorFlow, PyTorch, MXNet, Caffe, and Kaldi. AIACC includes the training acceleration engine AIACC-Training and inference acceleration engine AIACC-Inference.

As the first AI chip independently developed by Alibaba, Hanguang 800 is the most powerful AI inference chip in the world and is primarily used in cloud vision processing. The chip surpasses all other existing AI chips in terms of performance and energy efficiency. In the industry-standard ResNet-50 test, Hanguang 800 reached an inference performance of 78,563 images per second (IPS), which is four times better than the leading AI chip in the industry. Hanguang 800 also achieved an energy efficiency ratio of 500 IPS/W, 3.3 times better than the chip in second place. AIACC-Inference fully leverages the ultra-high computing capability of Hanguang 800, making the chip a model of how Alibaba Cloud can achieve unprecedented performance optimization through the integration of hardware and software.

Alibaba Cloud provides heterogeneous computing cloud services that integrate heterogeneous computing devices such as GPUs, field programmable gate arrays (FPGAs), and NPUs. This gives users access to heterogeneous computing services in the form of cloud computing services.

The rise of AI has popularized heterogeneous computing as a means to accelerate the performance of AI computing. Alibaba Cloud provides heterogeneous computing services based on a wide array of cloud-based acceleration instances to accelerate AI computing in an inclusive, elastic, and simple manner.

A New ResNet-50 Training Record in ImageNet

The most important competition in the image recognition field is ResNet-50 training on ImageNet dataset.

In the latest rankings, AIACC-Training beat other competitors in ResNet-50 training, demonstrating the highest performance and cost efficiency. This proved the superiority of AIACC in distributed training capabilities compared to other training acceleration engines. This also proved the ability of AIACC to help users improve training performance while reducing computing costs.

Stanford DAWN

Training Time — All Submissions

Objective: Time taken to train an image classification model to a top-5 validation accuracy of 93% or greater on ImageNet.

Rank	Time to 93% Accuracy	Model	Hardware	Framework
1 Mar 2020	0:02:38	ResNet50-v1.5 *Apsara AI Acceleration(AIACC) team in Alibaba Cloud* source	16 ecs.gn6e-c12g1.24xlarge (AlibabaCloud)	AIACC-Training 1.3 + Tensorflow 2.1
2 May 2019	0:02:43	ResNet-50 *ModelArts Service of Huawei Cloud* source	16 nodes with InfiniBand (8*V100 with NVLink for each node)	Moxing v1.13.0 + TensorFlow v1.13.1
3 Dec 2018	0:09:22	ResNet-50 *ModelArts Service of Huawei Cloud* source	16 * 8 * Tesla-V100(ModelArts Service)	Huawei Optimized MXNet

The new world record in training performance is 2 minutes and 38 seconds, which was the time required to reach a top 5 accuracy of 93% in ResNet-50 training. The training was performed in a cluster that included 128 V100 GPUs, provided by 16 cloud service instances designed for heterogeneous computing: ecs.gn6e-c12g1.24xlarge. In addition, a 32 Gbit/s virtual private cloud (VPC) was used as the communication network.

When the previous world record was set, the cluster for training included 128 V100 GPUs and used 100 Gbit/s InfiniBand as the communication network, which provided a bandwidth that is three times larger than the 32 Gbit/s VPC used to set the new world record. Heterogeneous computing cloud services typically use a 32 Gbit/s VPC as their network configuration. Alibaba Cloud also chose a 32 Gbit/s VPC to better close to the actual scenarios of end-users.

The heterogeneous computing team of Alibaba Cloud faced a major challenge due to the huge difference in physical network bandwidth between the 32 Gbit/s VPC and 100 Gbit/s InfiniBand. To overcome this, the team made in-depth optimizations in the following two areas:

First, the team optimized the model. The team adjusted hyperparameters and improved optimizers to reduce the iterations required to reach 93% accuracy. The team also tried to improve the performance of individual instance as much as possible.

Second, the team optimized distributed performance. The team used AIACC-Training, as a distributed communication library to fully exploit all the potential of the 32 Gbit/s VPC.

Stanford DAWN

Training Cost — All Submissions

Objective: Total cost of public cloud instances to train an image classification model to a top-5 validation accuracy of 93% or greater on ImageNet.

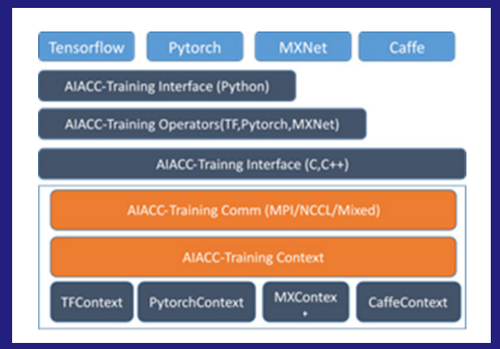
Rank	Cost (USD)	Model	Hardware	Framework
1 Mar 2020	$7.43	ResNet50-v1.5 *Apsara AI Acceleration(AIACC) team in Alibaba Cloud* source	1 ecs.gn6e-c12g1.24xlarge (AlibabaCloud)	AIACC-Training 1.3 + Tensorflow 2.1
2 Sep 2018	$12.60	ResNet50 *Google Cloud TPU* source	GCP n1-standard-2, Cloud TPU	TensorFlow v1.11.0
3 Mar 2020	$14.42	ResNet50-v1.5 *Apsara AI Acceleration(AIACC) team in Alibaba Cloud* source	16 ecs.gn6e-c12g1.24xlarge (AlibabaCloud)	AIACC-Training 1.3 + Tensorflow 2.1

Together, these two optimizations allowed the team to overcome a seemingly insurmountable performance barrier and helped Alibaba Cloud set a new world record with relatively low network bandwidth.

AIACC-Training

AIACC-Training is a proprietary communication engine developed by Alibaba Cloud for distributed deep learning training. It provides support for TensorFlow, PyTorch, MXNet, and Caffe. At the infrastructure as a service (IaaS) layer, AIACC-Training provides acceleration libraries that can be integrated and are compatible with open source libraries.

AIACC-Training has been extensively deployed in the production environments of multiple AI and Internet companies to significantly improve the cost performance of heterogeneous computing services. AIACC-Training provides differentiated computing services at the software layer. The following figure shows the architecture of AIACC-Training.



AIACC-Training played a critical role as the distributed backend during DAWNBench deep learning training.

A New Record on Inference Performance: Five Times Faster Than the Nearest Competitor

In the inference contest, DAWNBench requires participants to classify the 10,000 images in an ImageNet validation set. The classification model must reach a top 5 accuracy of at least 93%.

The average time and cost for inference per image are calculated, with the batch size set to 1. For the previous performance record, the average inference time was less than 1 ms, which far exceeded the speed of human visual responses.

In the latest rankings, Alibaba Cloud used AliNPU cloud service instances ecs.ebman1.26xlarge designed for heterogeneous computing to reach the highest inference performance, which was more than five times faster than that of the nearest competitor.

Alibaba Cloud also used GPU cloud service instances ecs.gn6i-c8g1.2xlarge designed for heterogeneous computing to set the record for the lowest inference cost, which has still not been surpassed. As a result, Alibaba earned first place in both the performance and cost rankings.

Stanford DAWN

Inference Latency — All Submissions

Objective: Latency required to classify one ImageNet image using a model with a top-5 validation accuracy of 93% or greater.

Rank	1-example Latency (milliseconds)	Model	Hardware	Framework
1 Mar 2020	0.0739	ResNet26d *Apsara AI Acceleration(AIACC) team in Alibaba Cloud && Alibaba T-Head* source	Alibaba Cloud [ecs.ebman1.26xlarge]	Pytorch+AIACC-Inference+HGAI
2 Feb 2020	0.3880	ResNet101 *AI Cognitive Computing team in Alipay Group* source	Alibaba Cloud Npu	tensorflow+NpuInference
3 Mar 2020	0.3926	MIVT-NET-v2 *Machine Intelligence in Alibaba Cloud* source	Alibaba Cloud [ecs.gn6i-c8g1.2xlarge]	HIE
4 Feb 2020	0.4662	ResNet26 *PAI: Platform of A.I. in Alibaba Cloud* source	Alibaba Cloud [ecs.gn6i-c8g1.2xlarge]	PAI-Blade + TensorRT

Stanford DAWN

Inference Cost — All Submissions

Objective: Average cost on public cloud instances to classify 10,000 validation images from ImageNet using of an image classification model with a top-5 validation accuracy of 93% or greater.

Rank	Cost (USD)	Model	Framework	Hardware
1 Oct 2019	$0.00	ResNet26d *Apsara AI Acceleration(AIACC) team in Alibaba Cloud* source	Pytorch+AIACC-Inference	Alibaba Cloud [ecs.gn6i-c8g1.2xlarge]
2 Jun 2019	$0.00	ResNet50 *InferenceX Team of Didi Cloud* source	ifx	Didi Cloud [1 P4 / 16 GB / 8 vCPU]
3 May 2018	$0.01	ResNet50 *Perseus AI Cloud Acceleration team in Alibaba Cloud* source	TensorFlow 1.12.2	Alibaba Cloud [ecs.gn5i-c8g1.2xlarge]
4 Dec 2018	$0.02	ResNet50 *Perseus AI Cloud Acceleration team in Alibaba Cloud* source	TensorFlow 1.10.0	Alibaba Cloud [ecs.gn5i-c8g1.2xlarge]

AIACC-Inference

As Alibaba Cloud serves customers and strives to win every DAWNBench contest, Alibaba Cloud constantly refines the inference optimization technologies for heterogeneous computing. Alibaba Cloud developed the model acceleration engine AIACC-Inference based on customer requirements to help customers optimize models under mainstream AI frameworks, such as TensorFlow, PyTorch, MXNet, and Kaldi.

To optimize a model, AIACC-Inference analyzes the computation graph of the model and fuses the compute nodes in it. This reduces the number of compute nodes and improves the efficiency of computation graph execution.

AIACC-Inference also supplies the options of FP32, FP16, and Int8 accuracy models. Currently, AIACC-Inference supports common image classification and target detection models, as well as Natural Language Processing (NLP) models and Generative Adversarial Networks (GANs), such as Bert and StyleGAN, etc.

Model and Framework Optimization

In the last version that Alibaba Cloud submitted, the base model was replaced with the simpler ResNet26d model, which attracted a great deal of attention in the industry.

To further improve model accuracy and simplify the model, Alibaba Cloud adjusted the hyperparameters and introduced more data enhancement methods. By combining AugMix and JSD loss with RandAugment, Alibaba Cloud improved the accuracy of the ResNet26d model to 93.3%, an increase of more than 0.13%.

In addition, AIACC-Inference had further optimized 1x1, 3X3 and 7x7 convolution kernels, and added some new OP fusion mechanism in AIACC-Inference, which can achieve a performance speedup of 1.5-2.5 times compared with TensorRT, formerly the fastest inference library in the industry.

Optimization Based on AliNPU

Alibaba Cloud optimized the inference engine based on the architectural features of AliNPU.

AliNPU uploads and downloads data stored in the uint8 data type. This requires Alibaba Cloud to insert the Quant and Dequant operations to restore data before and after entry to the engine. However, these operations cannot be accelerated by AliNPU on CPUs and take up a large amount of the inference time. When these operations are performed during preprocessing and postprocessing, the inference latency can be reduced to 0.117 ms.

Based on the small size of the inference model used and the 4 Gbit/s empirical GPU bandwidth, it takes 0.03 ms to upload 147 KB data to AliNPU when an image is imported. Therefore, Alibaba Cloud has introduced the preload mechanism to the framework to pre-fetch the data to AliNPU, which further reduces the average inference latency to 0.0739 ms.

Summary

Based on the hard work of the heterogeneous computing AI acceleration team, Alibaba Cloud tops four world rankings of the DAWNBench ImageNet competition held by Stanford University. By using the AIACC engine developed by Alibaba Cloud, the team dramatically optimized the performance of heterogeneous computing GPU instances, which topped the rankings in training performance, training cost, and inference cost. The team also optimized the performance of heterogeneous computing AliNPU instances which set the record of the highest inference performance. These achievements demonstrate Alibaba Cloud's leading capabilities in software and hardware integration and performance optimization.

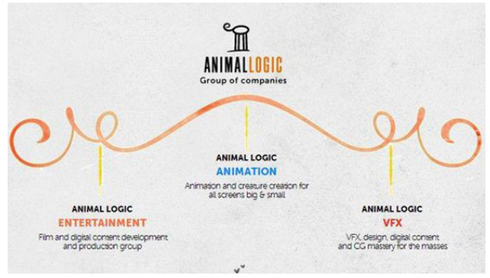
The AIACC engine has been deployed by many Alibaba Cloud key customers of heterogeneous computing on the public cloud, improving their application performance by 2 to 10 times.

In the future, Alibaba Cloud will persistently fully leverage leading capabilities of software and hardware integration and performance optimization. The solution winning the DAWNBench ImageNet competition will be open to customers, fully benefiting them with optimized high performance at a minimal cost. Alibaba Cloud will continue to strengthen capabilities of software and hardware integration and performance optimization, delivering higher performance and more cost-efficient heterogeneous computing AI acceleration services. These solutions from Alibaba Cloud can help advance the AI industry to a higher level.

Animal Logic Partners with Alibaba Cloud, Leads the Innovation of the Visual Effects Industry

Alibaba Cloud and Animal Logic, one of the world's leading independent creative digital studio, announced a partnership to meet the growing and demanding requirements of media production. Animal Logic has started backing-up its on-premise production content onto Alibaba Cloud's world-class computing platform. Through this partnership, the parties hope to drive efficiencies of media production by utilizing more cloud computing technologies.

As one of the world's most recognized digital production studios which has created animation and visual effects for award winning films including Peter Rabbit, The LEGO Movie 2, Captain Marvel and Happy Feet, Animal Logic requires a high-performance cloud storage platform that can quickly scale up for backup requirements, especially during peak production periods, in which 150TB of data can be generated in a 24-hour period. Animal Logic required a partner that could not only provide them with secure data protection capabilities, but also secure backup solutions for the increasingly large amount of data generated within a short period of time.



"Our partnership with Alibaba Cloud will provide us with the best technology to secure our content more efficiently," says Darin Grant, Animal Logic's Chief Technology Officer. "With Alibaba Cloud, we have the capability to backup large amounts of data and in turn, operate seamlessly even during our busiest times."

Alibaba Cloud, recognized by Gartner as one of the world's top three cloud computing companies, can meet the demanding requirements set by Animal Logic by providing its first-class storage solutions. To protect storage safety, and enhance the approach to Disaster Recovery, Alibaba Cloud designed a comprehensive storage gateway for Animal Logic and created private networks to minimize disturbance during data transfer. The data during transfer will also be encrypted to enhance the level of security.

Alibaba Cloud and Animal Logic will continue working on future cross-regional storage and backup projects, large scale rendering farm requirements and migration of legacy applications to the cloud. These solutions will support Animal

Logic's multi-cloud, whilst retiring a number of critical legacy applications, and ageing on-premise hardware.

"Alibaba Cloud is thrilled to announce our partnership with Animal Logic, in which we will bring increased efficiency and safety to the artists at Animal Logic and in turn, push the boundaries of animation and VFX to a new level," said Raymond Ma, General Manager for Alibaba Cloud in Australia and New Zealand.

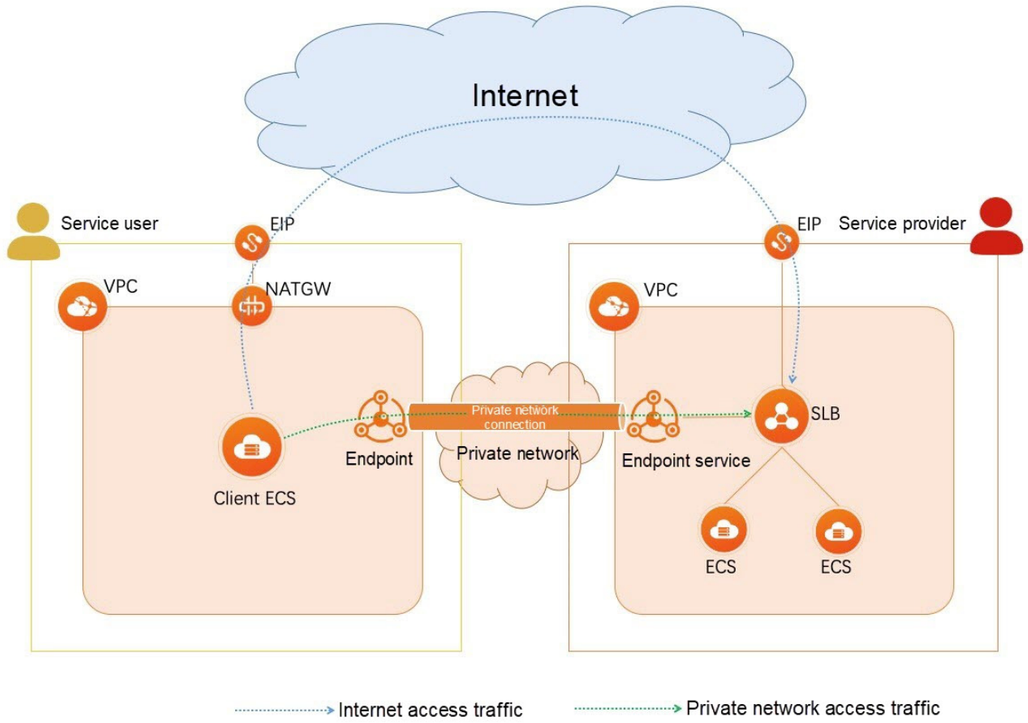
Alibaba Cloud Releases PrivateLink to Help Enterprises Build Private Network Services

Overview: At the Apsara Conference 2020, Zhu Shunmin, Researcher of Alibaba Cloud's intelligent network products, introduced the PrivateLink, a product for private network connection. PrivateLink uses Alibaba Cloud's private network for business interaction. With private network connections, users of Alibaba Cloud can access services provided by other Virtual Private Clouds (VPCs) through private networks, without additional Internet egress services. This ensures higher security and better network quality by preventing interactive data from going through the Internet.

What is PrivateLink?

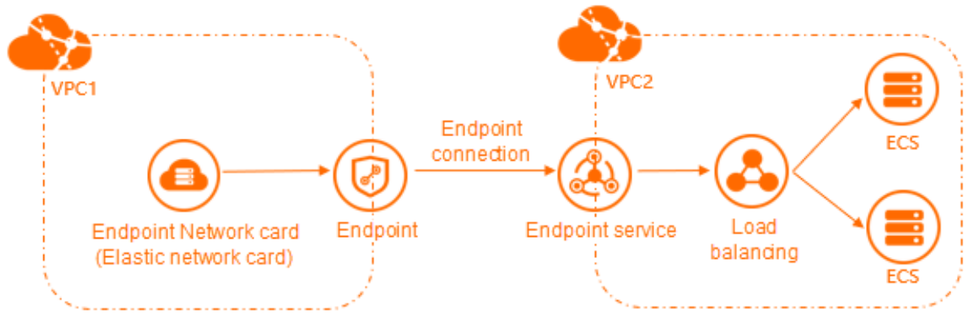
In the past, enterprises needed to create Internet egresses to provide on-cloud services or access resources of other business networks. Enterprises used products, such as Enterprise Information Portal (EIP) based on elastic public networks, Server Load Balancing (SLB) for public networks, and gateways for Network Address Translation (NAT), to create connections and provide on-cloud services.

However, as the number of enterprises on the cloud gradually increases, enterprises also gradually want to provide services on the cloud network. They hope to provide services and achieve mutual access through the internal network of Alibaba Cloud. By doing so, they can solve problems, such as relatively low security and high network latency. Fortunately, PrivateLink can provide private network connections within the cloud.



What Scenarios May Require PrivateLink?

There are a large number of business scenarios on Alibaba Cloud, such as enterprise internal services, inter-enterprise on-cloud services, and on-cloud enterprise ecosystems. PrivateLink can be applied to establish secure and stable private connections between VPC and Alibaba Cloud's services. This provides a flexible configuration to meet the needs of different scenarios.



Scenario 1: Sharing Cloud Services Across VPC Networks

Through PrivateLink, the SLB service of one VPC can be shared with other VPC, achieving cross-VPC private access of the SLB service.

As shown above, to achieve private access to the SLB service in VPC2, the SLB service needs to be added into the endpoint service as a service resource first. Then, endpoints for accessing the SLB service need to be created in VPC1. Thus, VPC1 can access the SLB service in VPC2 through endpoints.

Practice 1: An enterprise-level SaaS cloud service allows services to be provided on Alibaba Cloud's intranet. Enterprises or individuals can access service resources across regions and share the low-latency, high-availability, and high-security network of Alibaba Cloud.

Practice 2: Large and medium-sized enterprises or multinational companies set up the service releasing layer at the enterprise level. Each subsidiary and overseas office can access service resources through PrivateLink, achieving multi-account, multi-VPC fast interconnection, and security. PrivateLink can help these enterprises and companies migrate all their business to the cloud, making business usage and interconnection more convenient and reliable.

Scenario 2: Sharing On-Cloud Services in One VPC With a Local Data Center

Through PrivateLink, the SLB service in one VPC can be shared with a local data center, achieving on-cloud access to the SLB service in off-cloud private networks.

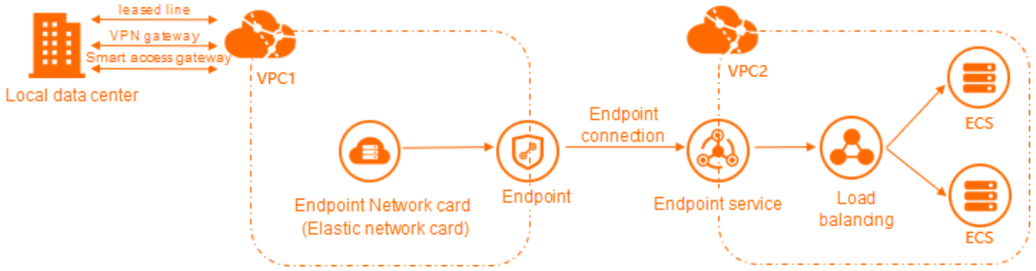
As shown above, to achieve private access in a local data center with the SLB service in VPC2, the SLB service needs to be shared with VPC1 first. Then, VPC1 will be connected with a local data center through a leased line, VPN gateway, or Smart Access Gateway (SAG.) In this way, private access from a local data center with on-cloud SLB services can be achieved.

Scenario Practice: When Independent Software Vendors (ISV) and System Integrators (SI) construct cloud ecosystems with enterprises, more of their offline services are migrated to the cloud. They build their own services on Alibaba Cloud or connect their local Internet Data Center (IDC) services to Alibaba Cloud. By doing so, they can help their long-term enterprise users to achieve more efficient and high-quality cloud migration.

What Are the Benefits of PrivateLink?

Communication in Private Networks

Alibaba Cloud network services provide stable, secure, reliable, low latency, and high-quality network communication. Alibaba Cloud network has more than 21 regional data centers, 63 availability zones, and over 120 Point of Presence (PoP) nodes globally. Through PrivateLink, access traffic is forwarded within the Alibaba Cloud intranet, which avoids potential risks caused by public network access.



Security and Reliability

When accessing on-cloud services through PrivateLink, users can add security group rules to Elastic Network Interfaces (ENIs) that are used to access services in a VPC. This provides enhanced security protection and control measures, so traffic stays within the Alibaba Cloud intranet. Therefore, the possibility of data leakage can be greatly reduced, and network security issues, such as attacks, can also be avoided.

Ultra-Low Latency

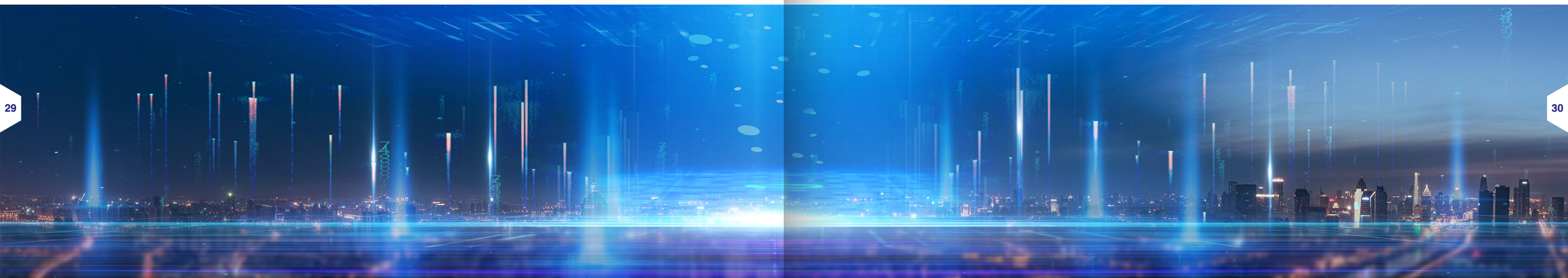
When accessing on-cloud services through PrivateLink, access requests are forwarded in the same availability zone with lower latency and jitter. At the same time, the underlying layer of the Alibaba Cloud network has high availability and reliability.

Simple Management

When accessing on-cloud services through PrivateLink, networks of service providers, and service users can be planned separately. There is no need to worry about address collision. By separately planning networks, the routing configuration can be simplified. Cross-account service access is also supported, which simplifies account management and security.

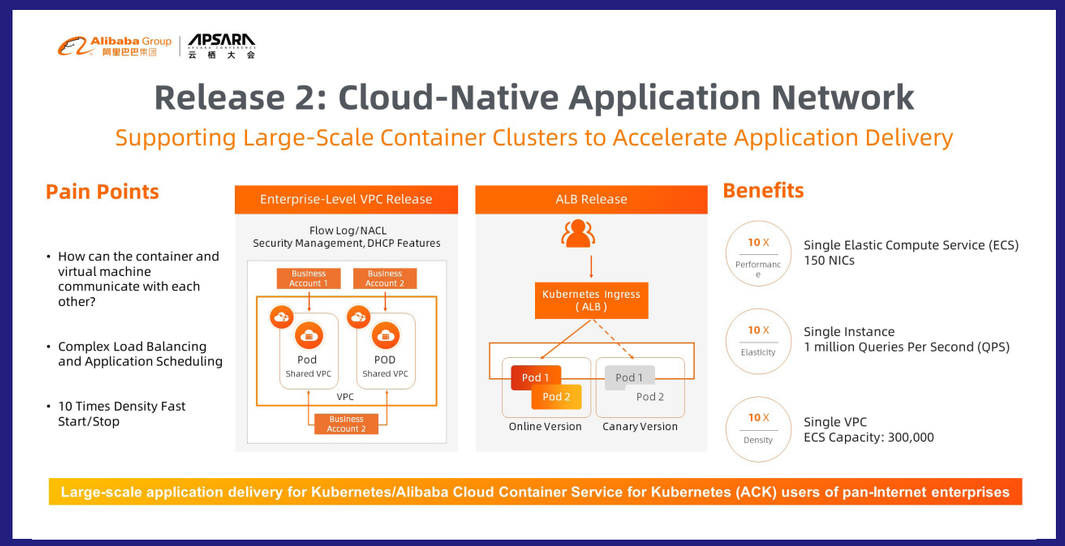
Summary

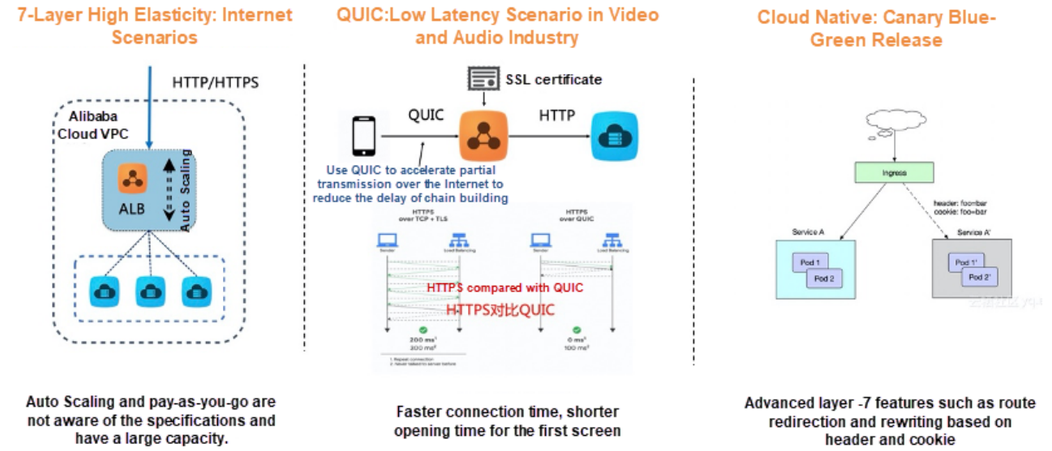
Network development is the most important thing for enterprises when migrating to the cloud. The Alibaba Cloud network provides enterprises with various cloud network services. Users can select connection services in public or private networks based on business characteristics. With these services, users can optimize business modes and access quality. They can also comprehensively allocate usage costs and simplify O&M management. PrivateLink provides more stable and secure network services and brings about new opportunities in terms of business modes as well. For more information, please check the websites below:



Alibaba Cloud Releases ALB to Accelerate the Delivery of Enterprise Applications

Recently, Zhu Shunmin, a researcher with Alibaba Cloud Network Services, released a variety of new network products at the Apsara Conference 2020. One of the products is the Application Load Balancer (ALB.) Positioned at the application layer, ALB provides superior performance. It is secure, reliable, cloud-native, and out-of-the-box. It supports auto scaling, the Quick UDP Internet Connection (QUIC) protocol, content-based advanced routing, built-in Distributed Denial of Service (DDoS) protection, cloud-native applications, flexible billing, and other product capabilities. It meets many diversified application-layer load requirements.





The Business Scenarios of ALB

When seeing the letters ALB, many people will think of the classic SLB service. Alibaba Cloud Server Load Balancer (SLB) released nearly ten years ago, providing users in various industries with powerful and stable load balancing capabilities. It distributes large amounts of concurrent traffic, prevents Single Point of Failures (SPOFs), and improves service availability. However, as enterprises and Internet businesses develop rapidly, business forms and demands are constantly changing. As a result, traditional load balancing cannot meet the requirements of many business scenarios. There is an urgent need for new product design to meet the requirements for high performance, elasticity, multi-protocol layer-7 forwarding, security, and cloud-native, such as Internet service, big e-commerce promotions, audio and video service, mobile Internet applications, gaming, financial services, and cloud-native applications.

Scenario 1: Big E-Commerce Promotions

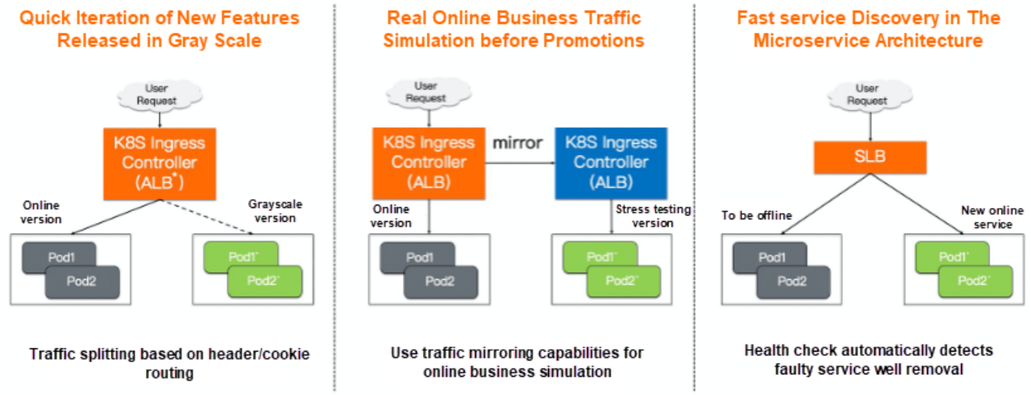
E-commerce companies need to carry out big promotions during festivals and major events. Live broadcasting is essential, but the peak traffic volume cannot be estimated before the event. In addition, load sharing must be performed based on the region, time period, and payment process. Although traditional load balancing features unified and flexible scheduling capabilities, it still lacks elasticity, scalability, and high performance. In addition, it cannot achieve real-time elasticity, high concurrency, and large capacity. It cannot complete multi-protocol (HTTP/HTTPS/HOST/URL/Cookie/Http Method) layer-7 forwarding. A new application load balancing (ALB) product is required to meet business needs.

Scenario 2: Live Streaming Service

For various video services, such as long videos, short videos, live videos, and online education, the demand for back-end resources increases. Millions of new connection requests, automatic elastic scaling, and traffic forwarding mechanisms based on user profiles make the current layer-4 SLB unable to fully meet the business development needs. By leveraging the transmission and forwarding capabilities of QUIC, ALB can quickly establish connections, shorten latency, distribute request traffic, and ensure high-traffic services.

Scenario 3: Cloud-Native Applications

In recent years, many businesses have been migrated to the cloud. Cloud vendors provide unified Infrastructure as a Service (IaaS) capabilities and cloud services, which greatly improves resource reuse at the IaaS layer. Users also want to unify systems at higher layers of IaaS. In this way, resources and products can be continuously reused to further reduce the operating costs of enterprises. This is where cloud-native architecture focuses. Based on the cloud-native architecture, ALB meets user requirements in scenarios, such as phased release, traffic simulation, and microservices.



The Benefits of Alibaba Cloud ALB

- **Ultra-High Performance:** Based on "Cloud Network Management 2.0" developed by Alibaba Cloud, ALB provides flexible, elastic, and ultra-high-performance instance types. It provisions super layer-7 processing capability so that a single instance can process as high as 1 million Queries Per Second (QPS.)

- **Secure and Reliable:** It provides four-level, high-availability, and disaster recovery and is integrated with Anti-DDoS Protection and Web Application Firewall (WAF) to ensure business security.

- **Low Latency:** In addition to existing protocols, such as HTTP, HTTPS, and Windows SharePoint Services (WSS), ALB supports QUIC, providing an extremely low-latency experience for video and live streaming users.

- **Cloud-Native:** ALB is deeply integrated with Alibaba Cloud Container Service for Kubernetes (ACK), Serverless App Engine (SAE), and Kubernetes (K8S), and uses the official cloud-native Ingress gateway.

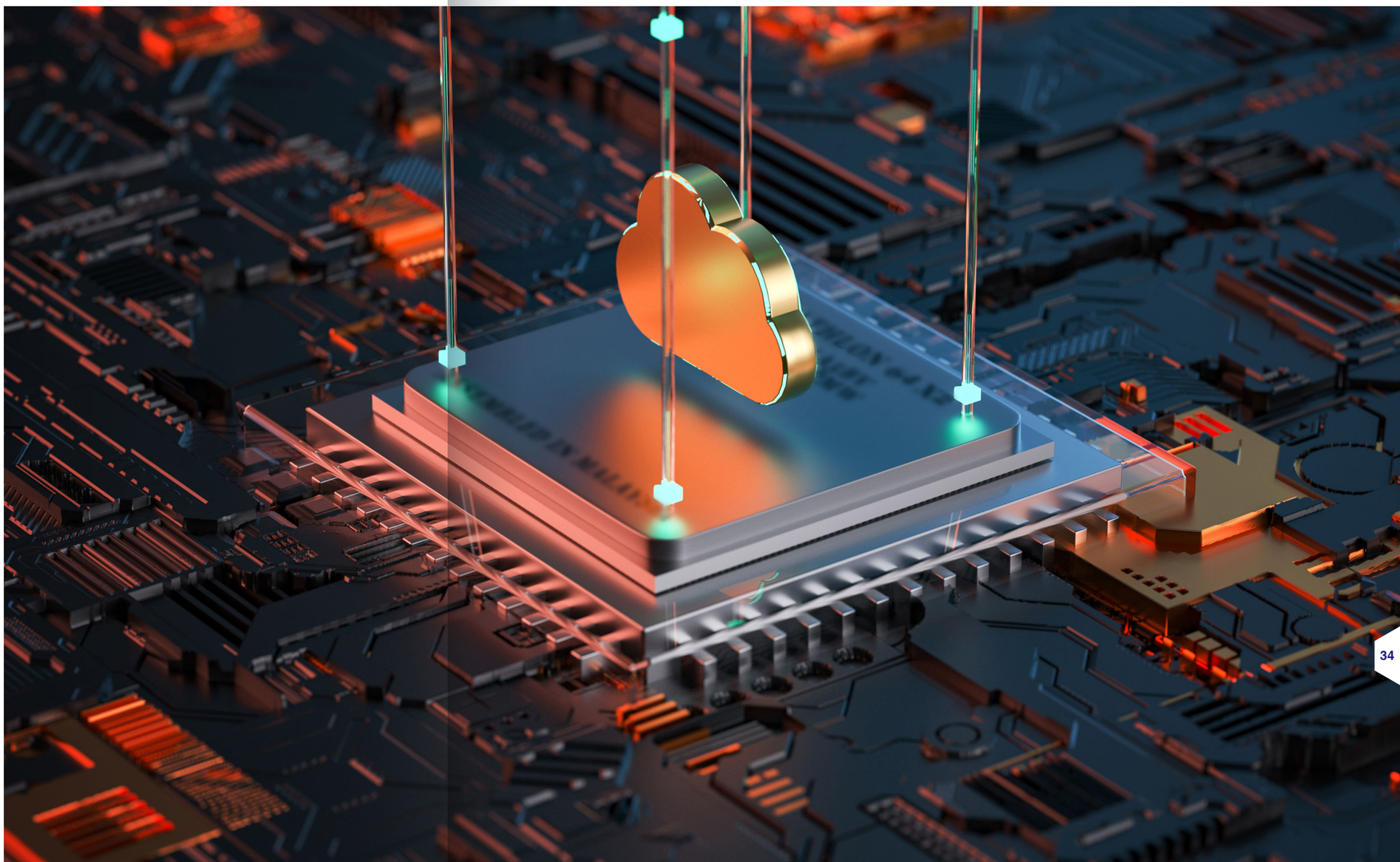
- **Personalized Route Forwarding:** It provides various advanced forwarding, such as Header, Cookie, and Method, and can better meet the personalized routing forwarding requirements of users.

- **Out-of-the-Box:** You can create instances in seconds. ALB delivers an out-of-the-box experience and a complete monitoring and logging service to mine access log data with one click.

- **Flexible Billing:** ALB uses the pay-as-you-go billing method and a dual-active disaster recovery mechanism to provide a more stable service experience and higher elasticity.

Summary

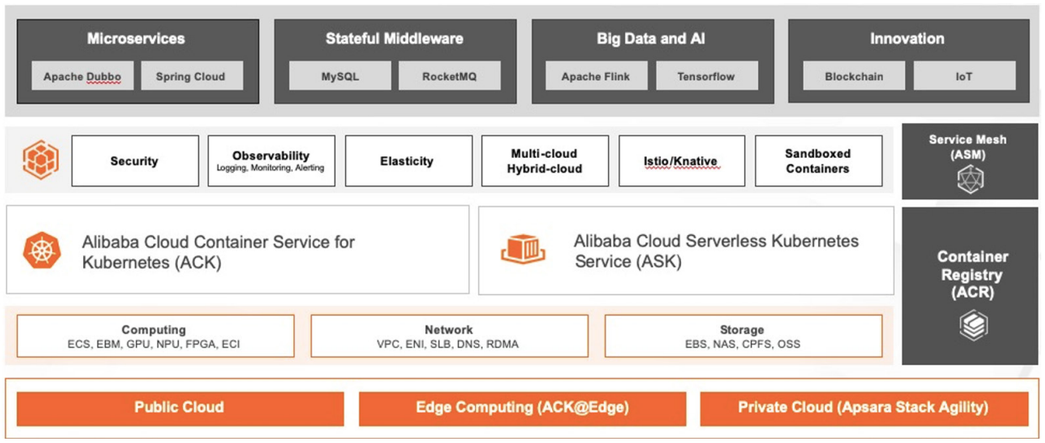
Network is critical for service migration to the cloud. As more complex enterprise applications are migrated to the cloud, the requirements for cloud networks are becoming higher. The top priority is concurrent high-traffic scheduling and application load balancing capabilities. With the launch of ALB, Alibaba Cloud will focus more on facilitating application delivery and ensuring high elasticity, security, reliability, and cost-effectiveness of applications.

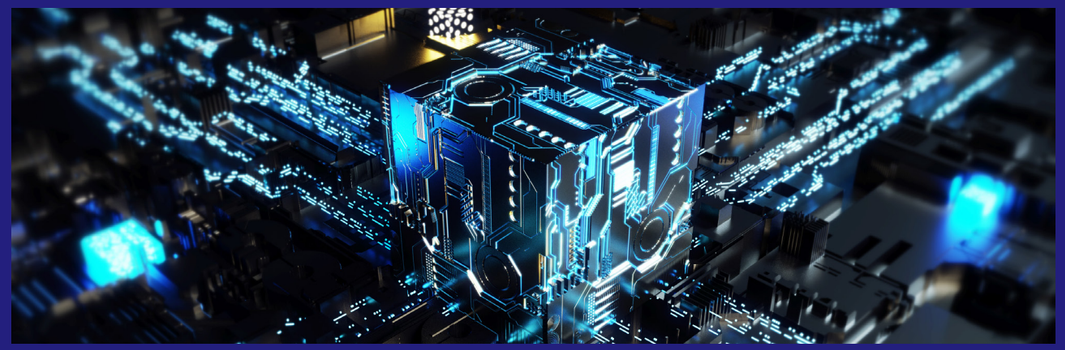


Run Kubernetes on Alibaba Cloud, Whose Container Technology Ranks No.1 in Gartner's Public Cloud Container Services Competitive Landscape

Introduction to ACK

Alibaba Cloud ACK is short for Container Service for Kubernetes. It provides Managed Kubernetes cluster, Dedicated Kubernetes cluster, and Serverless Kubernetes Cluster. ACK is deeply integrated with the high-performance compute, network, storage, and security services of Alibaba Cloud. It is globally available in 21 regions. It provides the best optimized containerized runtime on Cloud, makes it easy to run containerized applications and manage containerized applications life-cycle On Alibaba Cloud. ACK is fully tested through the massive traffic demands of Alibaba Group's large e-commerce platform Double 11, with the application auto-scaling in just a few seconds. ACK is an enterprise-level application running platform and one of the best Kubernetes services in the industry.





Global Recognition

Recently, Gartner released the "2020 Competitive Landscape: Public Cloud Container Services". According to the report, Alibaba Cloud and AWS have the richest product layout, covering 9 product capabilities, and tied for the first place.

According to Gartner analyst comments, Alibaba Cloud has rich and comprehensive container product portfolio and has a strong performance in the global market. It has good technology development strategies in 9 product areas, including serverless containers, service mesh, secure sandbox containers, hybrid cloud and the edge.

Now Alibaba Cloud Container is available in 21 regions around the world, and its service scale has grown by more than 400% in several years, supporting dozens of thousands of clusters with millions of containers.

Besides, Alibaba Cloud Container also takes the lead in China. At the end of September, Alibaba Cloud became the first cloud service provider to pass the container scale performance test of CAICT, and obtained the highest level of certification—"excellent" level, leading the development of China's container technology.

Application Life-cycle Management

Taking a central role, ACK has a full and diverse portfolio to easily manage the applications life-cycle.

1. Once an application is developed, deployment is the first step in Application life-cycle management. Whether in development environment, stage environment or production environment, ACK + ACR (Container Registry) automatically deploy applications in the Kubernetes cluster. ACR automatically synchronize docker image cross different regions, such as between Shanghai region and Singapore region, which solve problems that some images cannot be pulled in China. It also provides a feature to accelerate image building overseas. That enables faster application deployment.

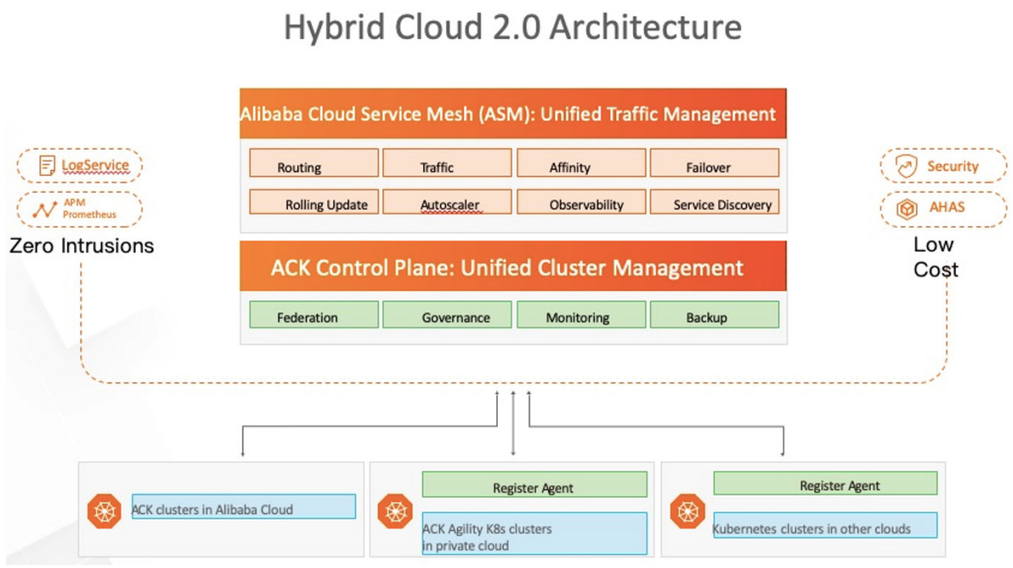
2. Monitoring, logging, tracing applications , reporting and analyzing how the application is doing is important when running applications. ACK is deeply integrated with LogService, ARMS(Application Real-Time Monitoring), Prometheus Monitoring, CloudMonitor, etc to make it easy to observe your applications.

3. Autoscale is one of the benefits of cloud, ACK makes it easy to scale down or scale up pods or nodes of applications based on CPU and Memory, as well as other metrics such as ingress latency, ingress QPS, SLB active connection, etc.. It supports ECS, EBS(ECS Bare Metal), GPU Instance, Preemptible Instance and ECI(Elastic Container Instance).

4. Sometimes your applications are running on serverless architecture, ACK provides a serverless cluster, which scale automatically and you do not need to manage worker nodes or master nodes. It also supports ECS, GPU Instance, and Preemptible Instance. It is paid by on-demands. It can scale 1000 pods in one minute, and scale to zero pod when there is no traffic.

Hybrid/Multi-Cloud Cluster Management

Hybrid deployment has become a common choice for enterprises to migrate their workloads to the cloud. However, the adoption of hybrid cloud brings a new challenge: There is huge difference in terms of capabilities and security requirements between on-premises and cloud-based infrastructures. And so we arrive at the question: how can you manage both of them effectively at the same time? To address this issue, ACK has provided the application-centric hybrid cloud 2.0 architecture. You can use the "Register Cluster" feature for unified traffic and cluster management of on-premise and other cloud Kubernetes clusters. This is such a great feature because you can easily manage your clusters from different cloud vendors, different runtime environments with unified governance, observability, scheduling, and deployment.



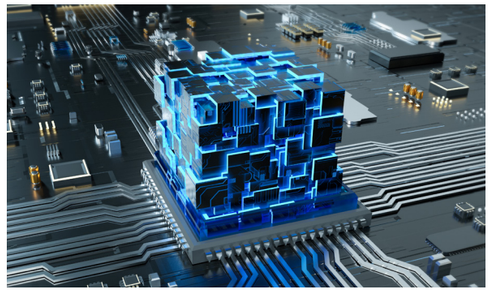
Managed Service Mesh

Over the past few years, Service Mesh has risen and become a standard component of the cloud-native stack. Istio is one of the popular solutions. ASM(Alibaba Cloud Service Mesh) is a managed Service Mesh platform. It is compatible with the open-source Istio service mesh of the Istio community. With ASM, you can manage services in a simplified manner. For example, you can use ASM to route and split inter-service traffic, secure inter-service communication with authentication, and observe the behavior of services in meshes.

GPU Sharing to Increase GPU Utilization, Saving GPU Cost

Nowadays, there are more and more AI jobs running on Kubernetes Cluster, but GPU is expensive, in most time GPU utilization is very low. By default Kubernetes infrastructure enforces exclusive GPU usage, preventing sharing GPUs across pods. However if you want to use the sharing capabilities of NVIDIA GPUs to increase GPU utilization in a cluster, ACK can provide GPU sharing solution to solve the problem. For example, you can run multiple inference tasks on the same GPU at the same time. Meanwhile, ACK ensures it is fully isolated, which means the GPU usage of each application is not affected by the other.

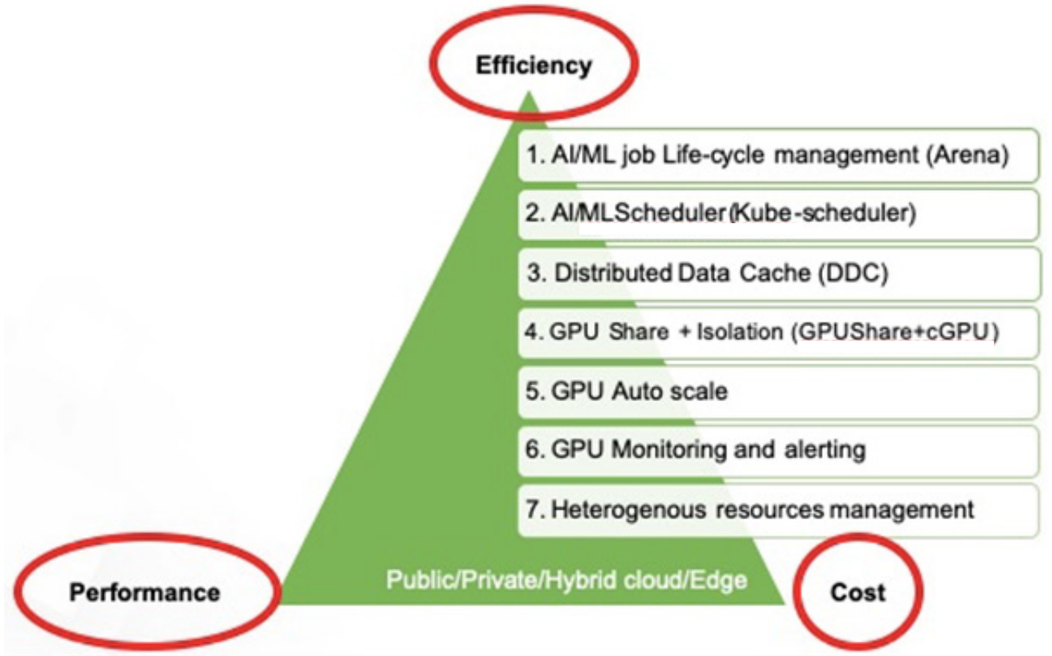
Additionally, ACK has a series of features to enable you to get efficiency, performance, cost at the same time. Arena is a command-line interface to run and monitor the machine learning training, inference jobs, and easily check their results also GPU utilization in real-time. GPU scheduler which uses the Kubernetes scheduler extender mechanism is responsible for determining whether a single GPU device on the node can provide enough GPU Memory when the global scheduler Filter and Bind. Distributed Data Cache (DDC) can accelerate remote data reading for AI training jobs by getting rid of GPU/CPU waiting time. It supports horizontal scaling and different storage backends including OSS/HDFS/NAS. Moreover, ACK DDC supports multiple cache layers including RAM/SSD/HDD and supports preload once, read by many jobs.

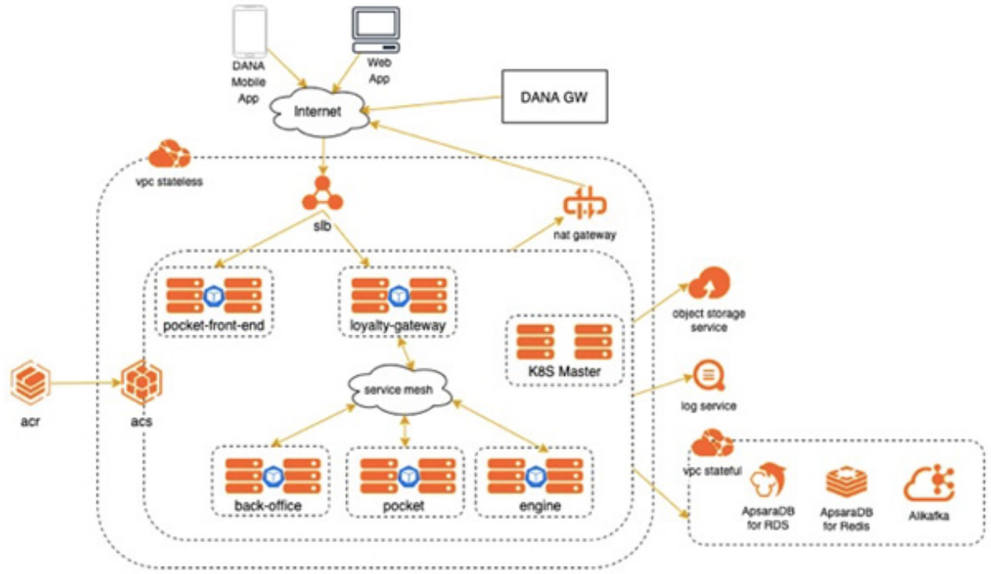


Edge Computing

As emerging technology like 5G and the Internet of Things(IoT) rise, some computing ability transfer to edge, such as collecting IoT data to do some analysis in edge, or optimizing video or image in the edge. So Edge side will meet operation requirements, such as deployment application in edge computing, edge node management. ACK provides the edge computing management from cloud to edge side. Therefore it is easy to deploy applications to edge computing, group edge nodes to define tag or tolerance and manage edge node.

ACK is fully embracing opensource, it contributes lots of components to Kubernetes and CNCF community, such as Autoscaler, GPU Sharing, and scheduling, Arena, OpenYurt.





Case study

Alibaba Cloud container services are used by more than 10 thousand customers around the globe, from a variety of different industries.

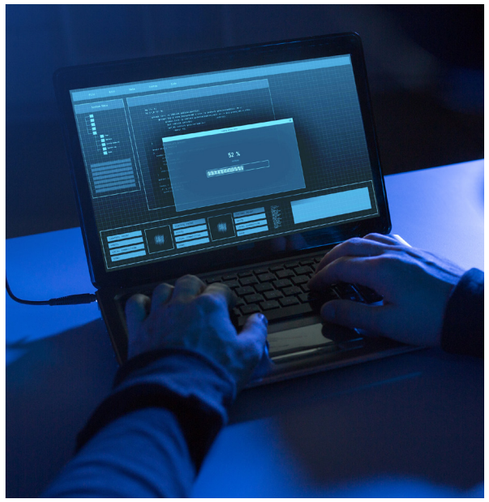
DANA is one of the top five digital wallet service providers in Indonesia, they provide mobile payment gateway for both online and offline transactions.

Its Merchant Portal application is running on Alibaba Cloud. They implement CI/CD Platform, AutoScaling with HPA, Logging, and Monitoring applications to simplify the operation of IT infrastructure, increase scalability to make the application robust.

Alibaba Cloud Security Center Named in Gartner Market Guide for CWPP

Overview: Security Center, developed by Alibaba Cloud, is a solution that integrates server security, container security, cloud security posture management (CSPM), and a security O&M center. Centralized security management based on a single platform enables basic security protection against viruses, ransomware, and hacking at the server level. More importantly, centralized security management closes the loop of security O&M through overall automation from security enhancement and threat detection to investigation response and active defense. Offering comprehensive security protection by default, this solution frees the security O&M staff of a company from an unmanageable number of alerts and automatically fixes security issues, which is especially beneficial to customers who are short of O&M specialists.

Recently, the leading international research and advisory firm Gartner released the Market Guide for Cloud Workload Protection Platforms (CWPP), a guide that recommends cloud service providers to enterprise users. Different from other vendors whose security services are selected in a specific field, Alibaba Cloud is named as a global provider for "broad, multi-OS capabilities" due to the diversified and comprehensive features of the Security Center and its ability to support multiple cloud platforms.



Enterprises Need Integrated Cloud-Native Security Solutions

As digital transformation deepens, enterprises must manage increasingly diversified cloud assets. As the complexity of security threats increases, enterprises must invest great efforts in alert analysis, threat detection, and virus detection and removal. The traditional approach of building a security system by stacking up discrete security services cannot provide the agility needed by cloud-based business and cannot cope with the complexity of such business. Enterprises planning for the future need integrated solutions. Traditional security services designed to resolve individual issues are evolving into all-in-one solutions to streamline security O&M for enterprises.

Security Center is exactly the solution that integrates server security, container security, and a security O&M center. **Centralized security management based on a single platform** enables basic security protection against viruses, ransomware, and hacking at the server level. More importantly, centralized security management closes the loop of security O&M through overall automation from security enhancement and threat detection to investigation response and active defense, freeing the security O&M staff from a large number of alerts and automatically fixes security issues, which is especially beneficial to customers who are short of professional security O&M specialists.

Integrate Cloud Native Capabilities to Effectively Address Security Pain Points

Unlike other server security services, Security Center is a service built on cloud-native capabilities. Its vulnerability and baseline repair feature integrates the snapshot capability of cloud computing. Snapshots can be automatically created when fixing vulnerabilities, to avoid the irreversible impact on business that is caused by repair failure.

To resolve a single security issue, traditional security services usually require operations that involve multiple services, which is a tedious and unsustainable process. Based on cloud-native capabilities, Server Guard elegantly resolves the issue of service fragmentation. Enterprises on the cloud can use the network security modules of Alibaba Cloud that are integrated into Security Center to block cyberattacks such as vulnerability exploitation and brute-force attacks with least operations. Enterprises can also configure custom policies for rapid responses through automated security orchestration.

Use Adaptive Risk Assessment Technology to Tackle the Challenge of Wasting Resources on Massive Low-risk Vulnerabilities

Security Center provides Carta-Inspired Vulnerability Management based on the continuous adaptive risk and trust assessment technology (CARTA), one of the top 10 Gartner security technology projects in 2019. The greatest challenge in vulnerability repair is not the identification of weaknesses, but the manpower needed for vulnerability repair.

Based on the thorough analysis and evaluation of vulnerabilities and risks of cloud-based assets, Security Center time points within the lifecycle of a vulnerability when the vulnerability is most likely to be exploited by hackers and evaluates the possibility of the vulnerability being exploited on servers. This allows users to judge the necessity of fixing vulnerabilities, avoid low-risk vulnerability fixes that consume much time and effort, and focus on the vulnerabilities and risks that truly pose security threats.

Perform Built-in Cloud Security Posture Management to Eliminate Potential Security Risks

The highly automated and self-service features of Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) on a public cloud make proper cloud configurations and compliance increasingly important. An improper configuration may expose thousands of systems or large amounts of sensitive data. The cloud security posture management (CSPM) of Security Center is a risk assessment capability designed for cloud platforms. Unlike CWPP that focuses on server security, CSPM helps users identify configuration risks on a cloud platform, analyze and manage infrastructure security configurations.

Security Center supports real-time risk assessment of the following five types of cloud platform configurations: identity authentication, network access, data security, log auditing, and monitoring and alerting. Based on the best practices of cloud security configuration from the Alibaba Cloud security team, users can detect and fix improper security configurations by using the capabilities of Server Guard without training.

Alibaba Cloud Released Industry's First Trusted and Virtualized Instance with Support for SGX 2.0 and TPM

Recently, Alibaba Cloud announced its support for SGX 2.0 and released a virtualized ECS instance based on SGX 2.0 and TPM.

In 2015, Alibaba Cloud launched the Data Protection Proposal, making it one of the first cloud service providers to do so. In this proposal, Alibaba Cloud stated that it would never make use of user data without approval. Alibaba Cloud also proposed that the platform had the responsibility and obligation to help its customers ensure the privacy, integrity, and availability of user data. Over the past five years, Alibaba Cloud has held fast to its proposal and released various data security products and services, such as transparent logging, sensitive data protection, and key management. In addition, Alibaba Cloud is also the first enterprise in the Asia-Pacific region to deploy cryptographic computing, exploring chip-level protection capabilities of data security.

Virtualized ECS Instances based on SGX 2.0 and TPM

Recently, Alibaba Cloud announced its support for Software Guard Extensions (SGX) 2.0, and released the industry's first virtualized ECS instance based on SGX 2.0 and Trusted Platform Module (TPM).

The virtualized ECS instance released this time has two value-added features:

- **Larger EPC memory:** Compared with the EPC's memory limitation of 256MB for the first generation of SGX services, the EPC memory based on SGX 2.0 can reach up to 1TB. Larger EPC memory can remove the memory restriction that hinders the development of big data related applications.
- **Alibaba Cloud's DCAP-based remote attestation service:** Users can directly use the remote attestation service provided by Alibaba Cloud. Moreover, the service can be customized according to users' needs, helping users achieve better performance and gain better experience.

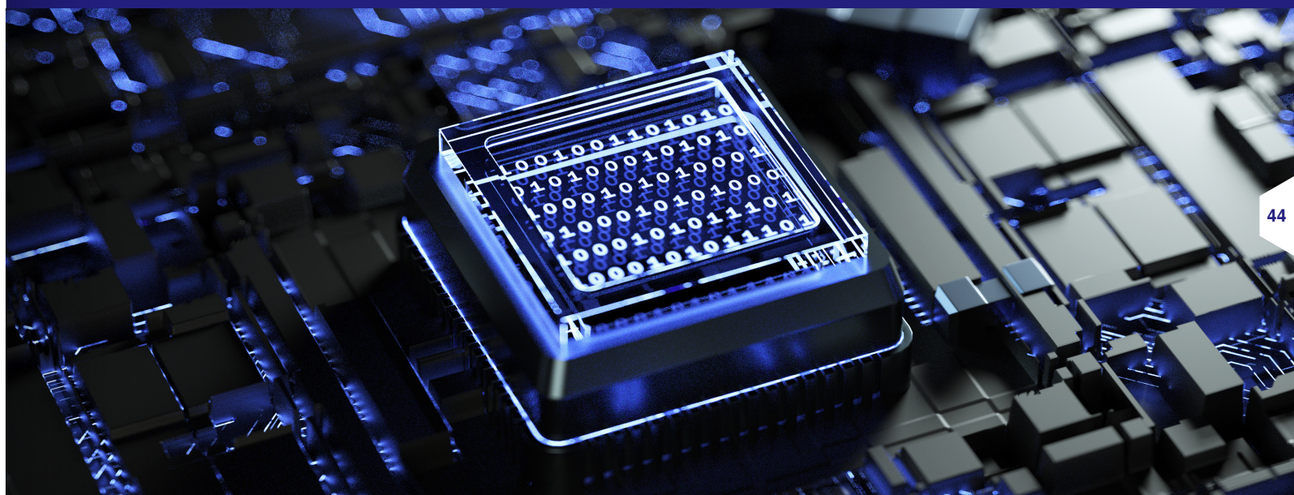
This instance fundamentally meets enterprises' needs of efficient computing with gigabyte of data, such as machine learning and artificial intelligence. The instance also provides higher-level data protection in new financial and large-scale internet usage scenarios. In addition, the instance also provides efficient and stable remote attestation service based on native advantages of Alibaba Cloud as a cloud service provider.

Cultivating the Growth of SGX Security Technology

In 2017, Alibaba Cloud was the first to launch chip-level SGX-based cryptographic computing solution, and it was also the first cloud service provider to commercialize the SGX technology. On November 2019, Alibaba Cloud jointly held the industry's first Application Contest Based on Chip-level Encryption with Zhejiang University. Through this contest, Alibaba Cloud strives to seek for and cultivate more SGX application developers in Chinese universities and enterprises, and to explore new business scenarios.

In addition, Alibaba Cloud also hopes to jointly build a new ecosystem and a new force in the security technology field of SGX, through the combination of industry, university and research. In the same year, as the only cloud service provider in Asia-Pacific region, Alibaba Cloud was listed as a typical vendor in Gartner's Report on Maturity Curve of Cloud Security Technology. Alibaba Cloud gained this title for its several practices in cryptographic computing. In Gartner's Global Security Capability Assessment Report, Alibaba Cloud has reached High level in the assessment of trusted execution environment for cryptographic computing.

Alibaba Cloud's accumulation and exploration of SGX 2.0 encryption technology will further improve protection capabilities of chip-level data security of the cloud infrastructure. This will help cloud developers and users build a more reliable execution environment with higher data protection capabilities.



From "Roughcast House" to "Fine-Decoration House" – Enterprise IT Governance Solutions for On-Cloud Management and Governance

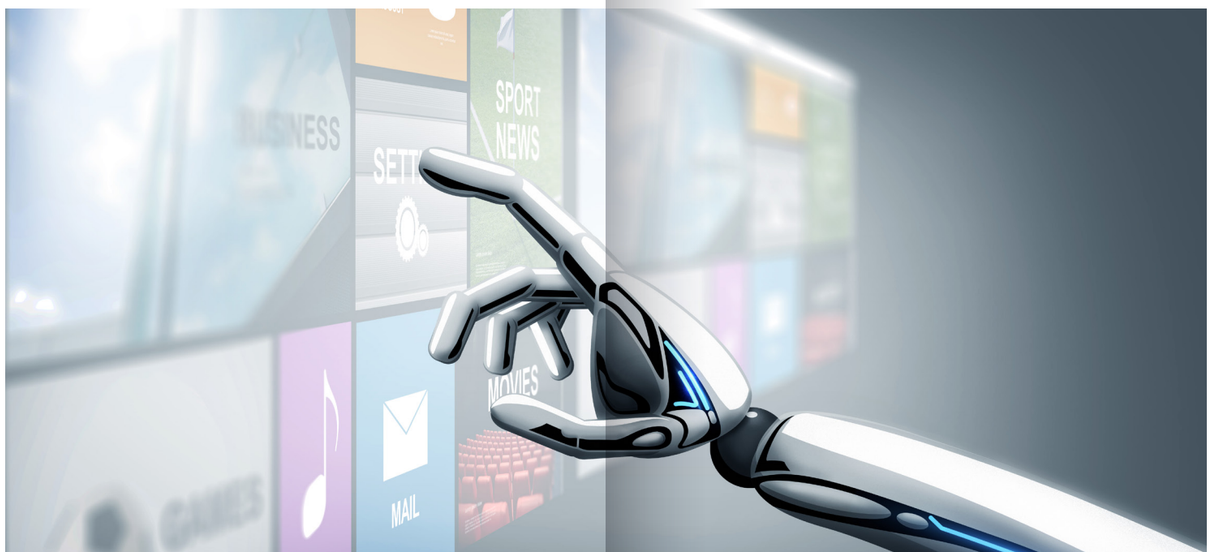
Overview: This article discusses how the Alibaba Cloud Open Platform Team can help enterprises innovate, manage, and make good use of the cloud using Cloud-Native capabilities.

The Challenges of Enterprises' Migration to the Cloud

With the rapid development of cloud technology in recent years, the concept of cloud-native is generally understood and accepted. More enterprises are choosing to migrate to the cloud to implement digital transformation. From moving traditional applications to the cloud or developing new products and businesses based on cloud-native technology, enterprises hope to utilize cloud technology for flexible innovation of their business at a low cost and to maximize the value of cloud migration.

However, with the increasing adoption of cloud technology, business and resource types and scales are increasing. Enterprises are also encountering new problems:

- How can we ensure identity security on the cloud?
- How can we isolate resources and permissions for multiple projects?
- How can we ensure network isolation and security among different business?
- How can we allocate spending on the cloud to different business teams?



These problems can affect the stability and development speed of business, cause security risks, and endanger the foundation of enterprises' survival. Therefore, before migrating to the cloud, enterprises need to plan and create a secure, controllable, and compliant "Landing Zone" for each business to be migrated to the cloud, except for adapting business applications to the cloud environment. By doing so, business developers are allowed to focus on their own business for quick iteration and innovation of the business based on cloud capabilities in the Landing Zone. Developers can take efficiency and controllability into account to achieve the maximum value of cloud migration.

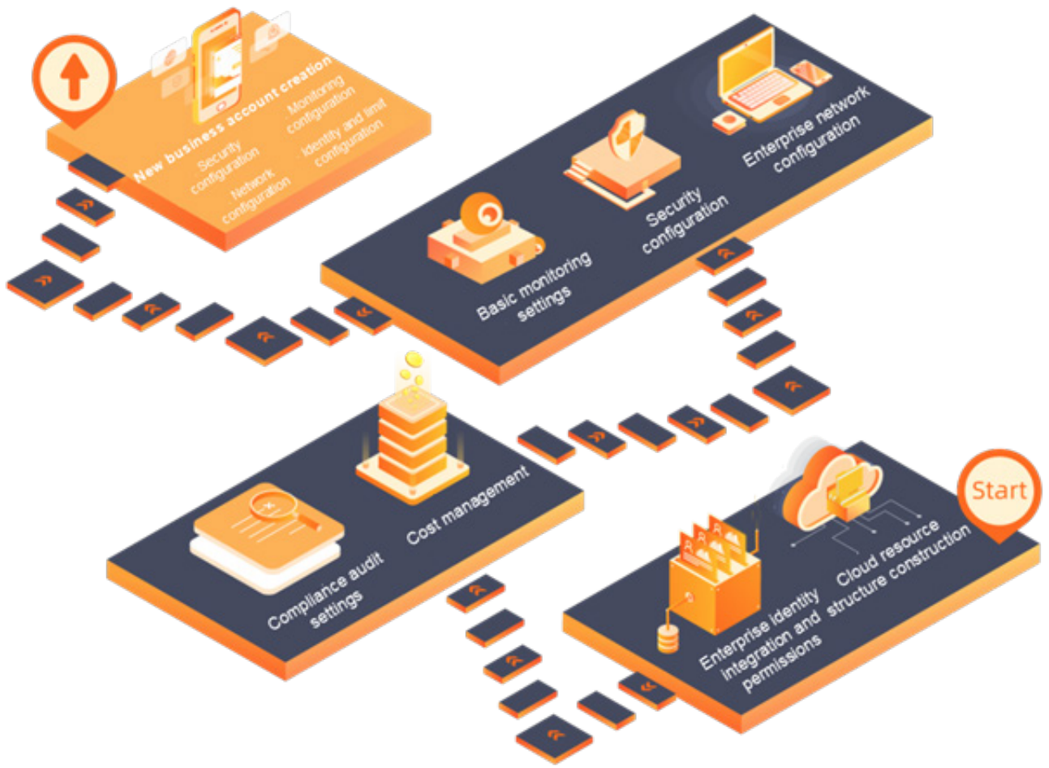
The key procedure of this part of the work lies in the improvement of **enterprises' IT governance** infrastructure.

Overview of Enterprise IT Governance

Enterprise IT Governance is a series of strategies, principles, and implementation processes that guide enterprise IT planning and operation, which allows IT personnel to control business risks at the IT level. In addition, Enterprise IT Governance can also ensure efficient and stable operation of enterprise business. A complete set of on-cloud Enterprise IT Governance infrastructure includes the following features:

- **Unified Framework:** Enterprises need to plan a unified IT governance architecture and apply relevant standards to specific business for the management and governance of each business.
- **Up-to-Date Compliance:** In the early stage of cloud migration, enterprise IT governance requirements should be met. In later stages, up-to-date compliance should also be provided automatically to ensure continuous business iteration and the rapid growth of new businesses.
- **Separate Management:** When the IT governance architecture is established, business teams can conduct O&M by themselves to reduce the pressure on the IT O&M team, except for maintaining the IT infrastructure.

To maximize the value of cloud migration, enterprises don't need to spend a lot of effort into learning on-cloud capabilities. More importantly, they need to conduct unified planning and implementation in the early stage. Instead of creating a poor "roughcast house," this way, a secure and controllable Landing Zone can be created for business on the cloud. In recent years, many enterprise customers of Alibaba Cloud have also been troubled by these problems and they turned to Alibaba Cloud for the best practices. For helping these enterprises quickly access Alibaba Cloud, the Alibaba Cloud Open Platform Team summarized the best practices based on several enterprises IT governance capabilities and pain points in enterprises' cloud migration. The team released the Enterprise IT Governance solution and three sets of specific implementation plans for enterprises in different sizes, as well as automated tools for quick implementation. Now, let's take medium- and large-sized enterprises and multinationals as examples to learn the design concept of the Enterprise IT Governance solution.



The Design Concept of the Enterprise IT Governance Solution

This solution serves as a model for enterprise users to construct a complex cross-account enterprise IT governance system on Alibaba Cloud. The framework includes the following aspects:

- **Enterprises' On-Cloud Resource Structure:** The first step for enterprises' cloud migration is to construct the infrastructure of on-cloud resources through multiple accounts. Based on the infrastructure, enterprises can carry out effective permission control, compliance audits, network planning, and financial trusteeship. By using various methods provided by Alibaba Cloud to organize resources, enterprises can easily and effectively build on-cloud resource architecture and copy it for organizing and dividing various business lines. By doing this, resources can form a clear "tree" and enterprises can lay the foundation for subsequent governance of other aspects.

- **Identity Integration:** Enterprises usually have their own identity management system and it is essential for enterprises to log on to Alibaba Cloud through their own management system. The Role-Based Single Sign On (SSO) of Alibaba Cloud allows enterprises to map employee identities or user groups to Alibaba Cloud roles with specific permissions to facilitate organizational management. Except for identity management, enterprises also need to assign different permission policies to different roles to minimize permissions. This solution provides a series of best practices for preset roles and permission policies as well as SSO automated tools to help enterprises quickly complete SSO configuration.

- **IT Compliance and Audit:** IT compliance and audit is the key to achieve "efficiency" and "controllability" in the enterprise IT governance process. Besides, it has also become one of the core requirements of enterprise IT governance, particularly after classified protection compliance became a mandatory requirement for enterprises' cloud migration.

Compliance and audit can be implemented in three ways:

- **Preventive Management:**It refers to forbidding non-compliant operations, such as changing basic configurations of the solution, connecting to public networks, and creating unencrypted disks, thus complying with the corporate compliance principles.

- **Detective Management:** For some suggested compliance principles, enterprises can set detective rules instead of preventive management and continuously monitor resources. When non-compliant resources are discovered, the solution can send an alert, and these resources can be recorded and fixed automatically.

- **Long-Term Storage of Audit Log:** Logs of on-cloud operations, resource changes, and network traffic can be stored for a long time in case of auditing.

- **Fees and Costs:** Cost Analysis is the basic demand for enterprises' cloud migration. It is a prerequisite for enterprises to be assured if they can calculate spending and make the costs more predictable. The larger the size of an enterprise is, the more attention needed to be paid to the budget and spending of each business and department. There are two cost allocation modes, namely Showback and Chargeback, according to the type of enterprise. Besides, there are several common methods, such as account-based cost allocation and tag-based cost allocation, according to the planning of enterprise on-cloud resource structure.

- **Network Planning, Security Protection, and Monitoring:** Network architecture is crucial for an enterprise, which is related to business operation, application calls, business expansion, and information security. This part mainly includes enterprise IP address planning, network connection, and access control. The focus is to plan which security domains of the enterprise network can be interconnected, which service

can access or be accessed by the Internet, and how to control horizontal and vertical traffic for ensuring information security. Furthermore, enterprises need to set unified monitoring and alerting rules for relevant network resources and business resources to detect and resolve business problems in advance.

- **New Account Baseline:** When an enterprise carries out new business through a new account, it is also needed to meet the requirements of enterprise IT governance principles. To do so, enterprises need to implement the design principles mentioned above when using a new account, such as identity integration, initializing network architecture, configuring security protection, and conducting monitoring and warning. At the same time, enterprises should hold the account compliance baseline in combination with preventive management to avoid misoperations that may result in non-compliance and risks to enterprises.

Solution Implementation

With the design concept of the solution, the next step is how to construct and implement the infrastructure according to the characteristics and development stages of enterprises, assisting enterprises to quickly turn the "roughcast house" into a "fine-decoration house". It is impossible for an implementation solution to perfectly match the demands of every enterprise in real practices. Enterprises must customize and combine different solutions based on their own demands and design principles. These three representative solutions mentioned above are the best solutions proposed by Alibaba Cloud for start-ups, medium- and large-sized enterprises, and multinationals. For more information, you can visit the Alibaba Cloud Open Platform website. For start-ups, operation steps and codes that are automatically generated can be obtained on the official website to implement such a solution. As for other enterprises, please contact your Alibaba Cloud sales representative or service manager.

During the implementation process, the ideal state is full automation. Based on the concept of Infrastructure as Code (IaC) and several tools, including Terraform, the Alibaba Cloud Open Platform provides automated deployment scripts and codes and makes them available open on the Aliyun Landing Zone Github to help you quickly deploy a solution or integrate it into the internal automation process system.

Summary

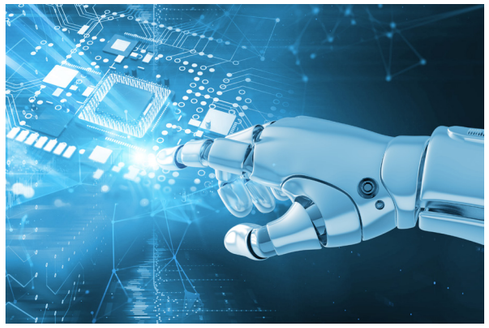
With the arrival of the Cloud-Native era, enterprises will face more new challenges on the cloud. The Alibaba Cloud Open Platform Team will continue to optimize products and solutions, accumulate additional best practices, and help enterprises manage and make good use of the cloud, allowing enterprises to innovate more quickly based on Cloud-Native capabilities.

Management Automation – Enterprises' Inevitable Approach to Cloud Migration

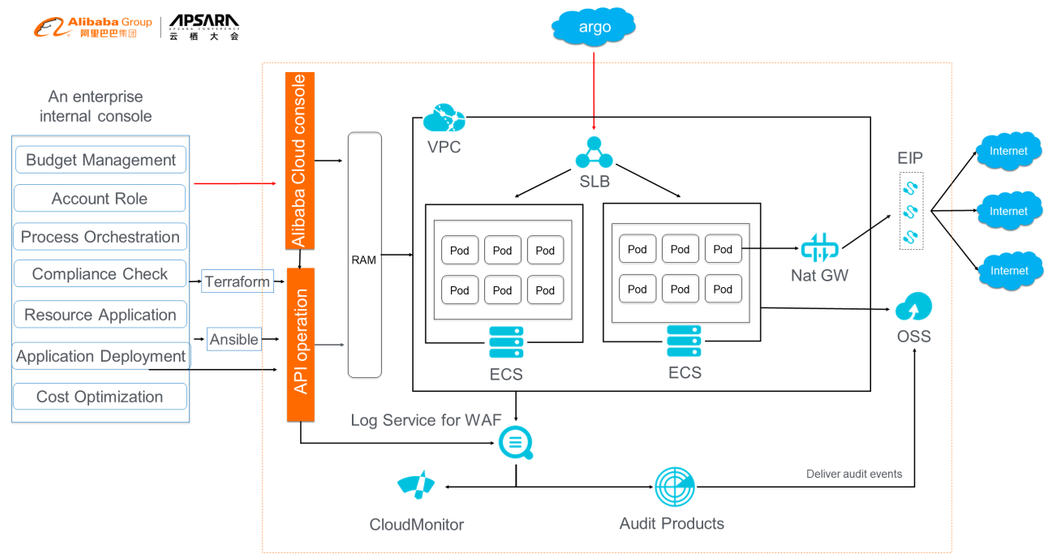
> **Overview: The management automation on cloud resources can reduce financial costs and increase enterprises' efficiency and competitiveness by lowering technical thresholds.**

Why Do We Need Automation on the Cloud?

While serving customers, we found that foreign customers are more dependent on automation tools than domestic customers. It is widely acknowledged that the technology orientation, high labor costs, and high compliance requirements in management boost the demand of foreign companies for the automation of IT systems. For business-oriented domestic companies with relatively sufficient employees that are at another development stage compared to foreign companies, they tend to employ more inexpensive employees to do the work that should be done by the IT system.



However, with the constant maturity of cloud computing, it is an inevitable trend for enterprises to migrate their business to the cloud. Under such circumstances, if domestic enterprises keep their old-fashioned ideas, their business operation will be negatively affected. The management automation on cloud resources can reduce financial costs and increase enterprises' efficiency and competitiveness by lowering technical thresholds.



Automation Needed by Enterprise Customers

Which dimensions of customers' management automation on the cloud do we need to focus on? From a customer case, let's learn the requirements of an enterprise's cloud migration:

In the picture above, the customer wanted more than just the development of programming automation in the O&M field. The first thing the customer considered was how to manage budgets and staff. After communicating with the customer, we made a list of main requirements for cloud migration:

1. Organization Management

Many enterprises have their own account and permission systems, which need to be interconnected with on-cloud systems. On Alibaba Cloud, enterprises can use Resource Access Management (RAM) (including identity management, permission management, and other components), resource management (including resource directories, resource groups, resource sharing, tags, and other components), and other products under the enterprise IT governance product line to interconnect those systems.

2. Orchestration Automation of Infrastructure

Alibaba Cloud has already provided more than 200 cloud services and more than 10,000 OpenAPIs. Resource orchestration tools, such as Terraform and Resource Orchestration Service (ROS), can help customers efficiently manage resources on the cloud and reduce the complexity of management with the concept of IaC.

3. Orchestration Automation of Application Programs

Open-source O&M tools, such as Ansible, Puppet, and Chef can be used for application deployment. Currently, Alibaba Cloud primarily supports Ansible and provides Operation Orchestration Service (OOS). The Open Application Model (OAM) specification was recently released as well, which further simplifies the application deployment process.

4. Security Requirements

Without automation, it is often too late to fix security loopholes manually. Powered by RAM and other security products, Alibaba Cloud's OpenAPI system provides a high-level of security to prevent various security issues.

5. Compliance Requirements

Compliance, on the one hand, requires external compliance, such as compliance of audit data and financial data. On the other hand, it requires compliance of internal data. Alibaba Cloud provides customers with ActionTrial and Config, as well as the compliance capabilities of industries cloud. This topic will be described subsequently.

6. Monitoring Requirements

When monitoring the resources on the cloud, customers need to connect the monitoring system with operations of enterprises, including data integration and data visualization. Cloud Monitor is a useful tool for automatic monitoring on Alibaba Cloud. In addition to its visual interface, Cloud Monitor can connect to systems of customers through OpenAPI.

7. Cost Requirements

When monitoring the resources on the cloud, customers need to connect the monitoring system with operations of enterprises, including data integration and data visualization. Cloud Monitor is a useful tool for automatic monitoring on Alibaba Cloud. In addition to its visual interface, Cloud Monitor can connect to systems of customers through OpenAPI.

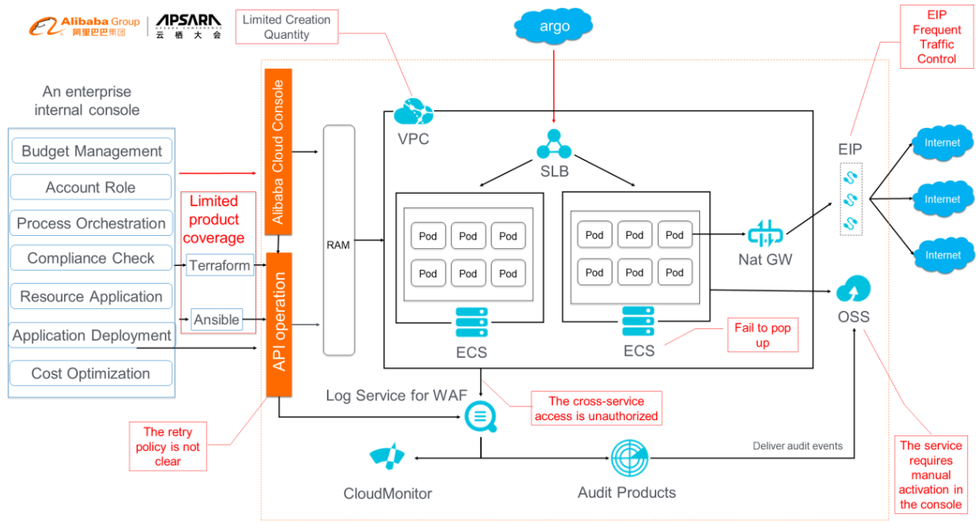
8. Situation Awareness

Customers can reserve resources in advance and quickly allocate resources based on the current resource usage and historical records, or according to prior planning. This requires cloud computing products to be capable of rapid scaling as well as perceiving resource usage and planning.

Aiming at the enterprise scenarios mentioned above, I would like to introduce the sample solution launched by the Alibaba Cloud Open Platform team, which is integrated with preceding capabilities. The solution not only defines best practices for migrating enterprises' IT to the cloud, but it also provides the automation codes for Terraform. You can download the latest codes from Github. Please visit this website and share your opinions with us.

Upgrade the OpenAPI Automation Capabilities

What technical problems with automation besides functions did customers encounter in the past? Again, let's take a customer case as an example:



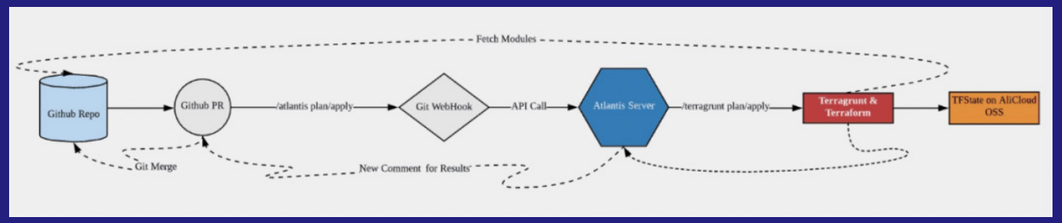
As shown in the picture above, Alibaba Cloud had several long-standing defects in terms of basic automation capabilities:

- Insufficient coverage of orchestration products, such as Terraform, makes rapid orchestration impossible for some products.
- Many ambiguous calling strategies at the OpenAPI level affect the efficiency optimization of the client. For example, the throttling threshold is not transparent, and the caller has some problems with unknown causes.
- For important resources, it is difficult for customers to know the quota limit of resources. Therefore, customers can only raise the demand through tickets with limited response speed.
- Due to some historical reasons, many Alibaba Cloud products need to be manually activated, which becomes a stumbling block on the automation process.
- Customers must manually authorize the inter-access among Alibaba Cloud products in the console. This stops the progress of automatic connection.

To solve these issues, Alibaba Cloud has made efforts to eliminate barriers that affect user experience and made some achievements.

Supporting Terraform for Products

WeWork is a company that focuses on the joint office community. It has chosen Alibaba Cloud as its partner and has carried out in-depth cooperation with Alibaba Cloud in basic resources, global network, security, IoT, big data, and other aspects. According to Yu Liang, Director of O&M, the infrastructure team of WeWork built a manageable self-service portal based on Terraform with less than two people in a few months. This portal can be fully deployed automatically within seconds. It can also support the infrastructure O&M of over 40 business systems with a three-person team, ensuring WeWork's security and compliance.

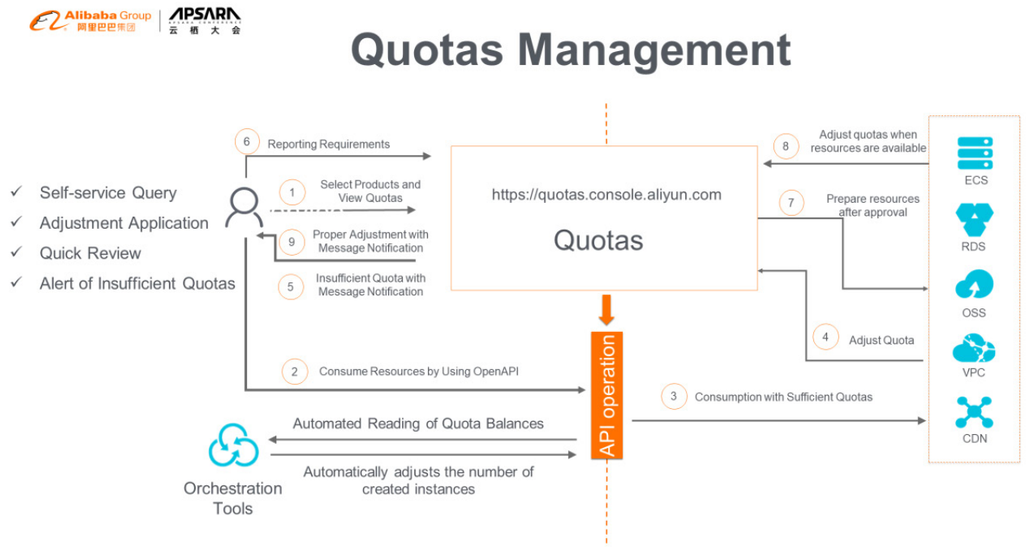


WeWork manages Terraform based on Github and Atlantis

Currently, the number of products supported by Alibaba Cloud's Terraform has increased from 40 to 53, and the number of resources has grown to 249. It can meet the needs of most scenarios. Alibaba Cloud will launch some tools in the second half of this year, such as cloud-based Terraform workflows and the ability to visually write Terraform templates. The former can reduce the extra burden of customers in building and managing their own Terraform workflows, and the latter can improve the user experience while lowering usage costs.

Quota Management

Quota management is another major problem in the process of automation. Users often want to know how many quotas they have, how many quotas they have used, how to increase quotas, and how to manage quotas in a more refined manner. To resolve the issue that users cannot quickly obtain and adjust quotas, Alibaba Cloud provides a quota center at this address. The following picture shows the main workflow of the quota center:



The Quota Center Mainly Solves Three Problems:

- **Product Quotas Request:** After logging on to the corresponding page, users can check the quota settings and use quotas of 15 cloud products.
- **Self-Service Application for Adjusting Quotas:** Users can directly submit an application for adjusting quotas to the corresponding cloud product administrator at the quota center. The administrator will make a quick decision on whether the application should be approved or not according to the real situation of customers.
- **Providing OpenAPI and Alerts for Getting Quotas:** The application on the client-side may need to acquire quota information in real-time to determine the next operation process. When the quota is insufficient, the application sends an alert to users to adjust the operation strategies in time.

Hundreds of enterprise customers have applied for quota increases through the quota center since its launch. In the future, more cloud products will be able to solve quota issues in the quota center.

Automation Activation for Cloud Products

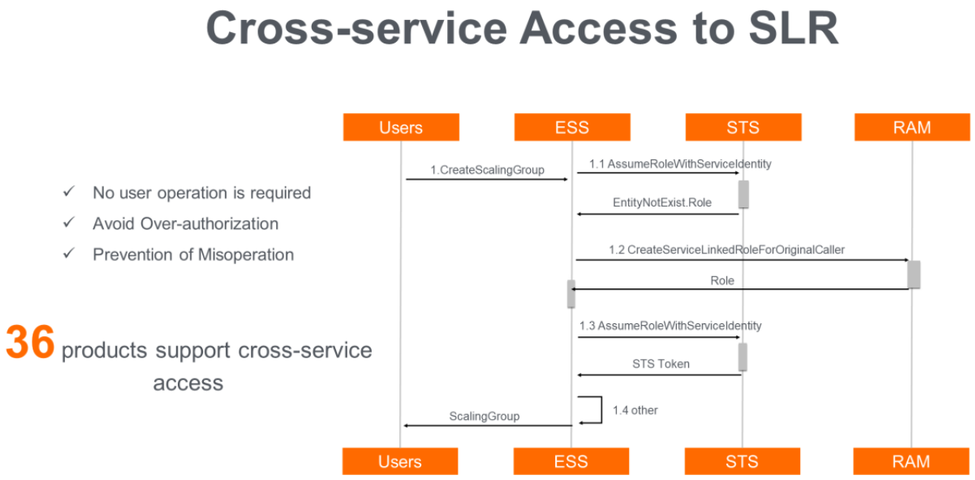
Many cloud products must be activated manually in the Alibaba Cloud console, which restricts the customers' automation process in some cases. For this difficulty in the automation process, Alibaba Cloud has upgraded some related products. Among the products that need to be manually activated in the past, 13 of them have been completely exempted from activation, and 9 of them have been provided with OpenAPI automatic activation. In addition, we will continue to upgrade products that need to be manually activated in the second half of this year to achieve 100% automation in the activation process.

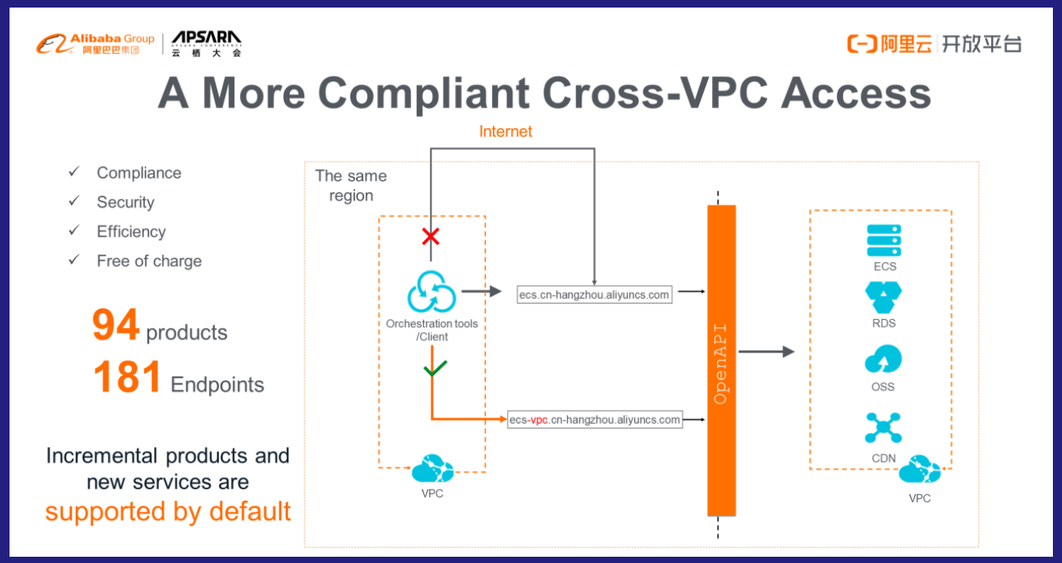
Alibaba Cloud's Terraform Provider has supported the automatic activation of these products. Users only need to add a DataSource corresponding to the cloud product activation in the template. Then, users need to set **enable = "On"** to run the terraform apply to enable automatic activation. For example, codes for activating log service Terraform automatically are listed below:


```
1. Data  "alicloud_log_service" "open" {
2. Enable  = "On"
3. }
```

Cross-Service Access to SLR

In real business scenarios, users may encounter a situation where they need to access the resources of cloud service B to use with cloud service A. For example, when you export images from ECS to OSS, you need to call the OSS upload interface of the customer directly from the backend of ECS. These resources belong to the customer, but they are not managed by the same cloud service. Essentially, this process requires obtaining user identities and permissions. In the past, to perform this operation, you had to create a service role and get permission granted through RAM on the quick authorization page (console.) This process cannot be operated automatically.





To meet such needs, Alibaba Cloud has upgraded its OpenAPI access compliance capability, as shown in the following picture:

In the past, customers would go through the public network when accessing OpenAPI, as shown in the picture. However, if customers need to access Alibaba Cloud OpenAPI in a VPC network, they can now change the target endpoint to xxx-vpc.[RegionId].aliyuncs.com, when calling OpenAPI in a public cloud environment. Thus, all traffic destined for this target domain name is forwarded to the internal network of Alibaba Cloud instead of a public network. This enhances the security of specific industries.

Summary

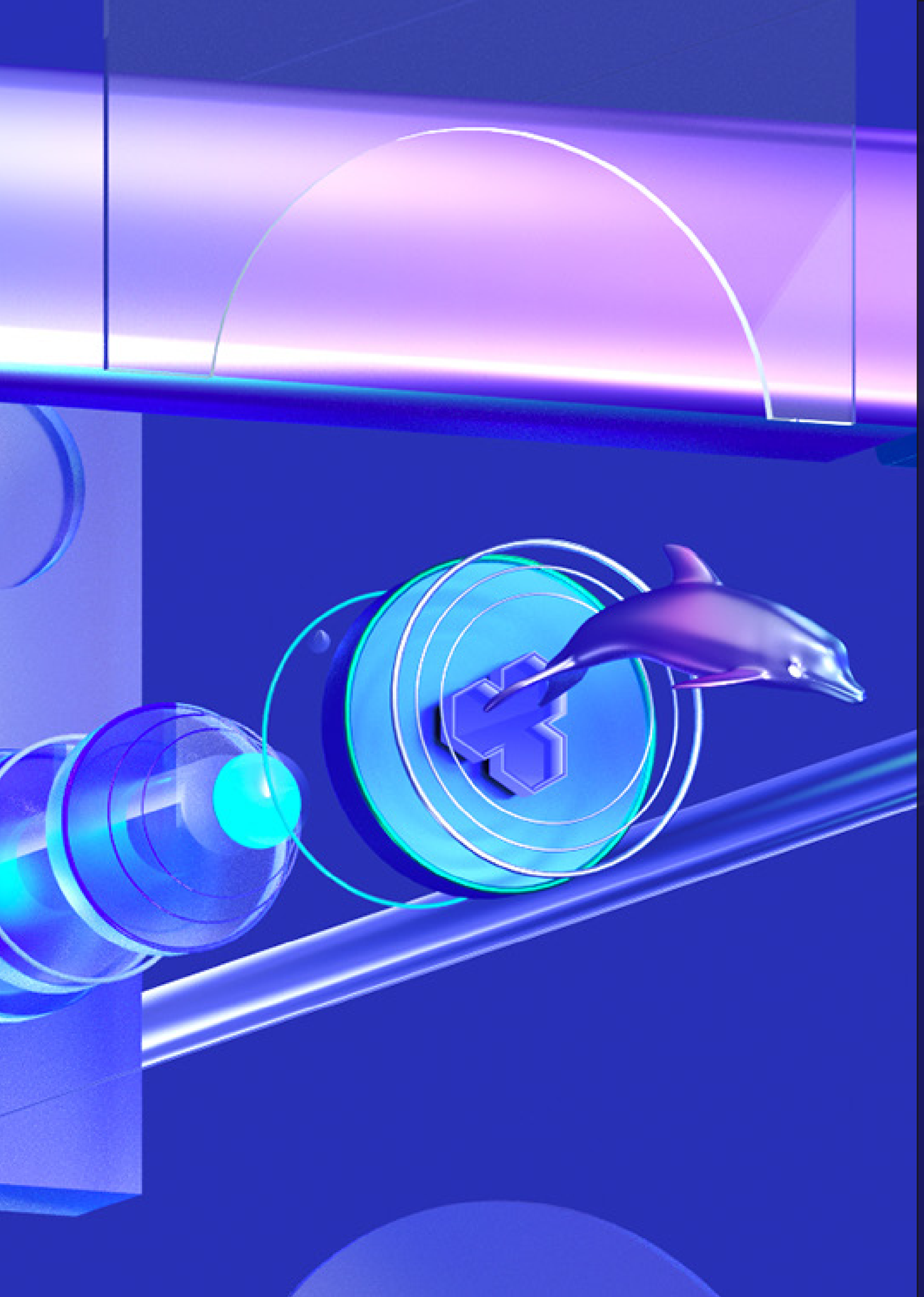
The automation capability is an important topic for enterprises' large-scale migration to the cloud. Even small and medium-sized enterprises can benefit from this capability. On the one hand, enterprises need to choose proper integration tools based on their real situations. On the other hand, they need to make plans and designs related to financial and property laws before cloud migration. Alibaba Cloud will keep improving on-cloud enterprise automation capabilities and help customers achieve business success.

The flowchart above shows that the Service Linked Role (SLR) mechanism does not require user intervention. A sub-user with product management permission can trigger the SLR creation of the related product. At the same time, the modification and deletion are strictly controlled to avoid misoperations.

Currently, up to 36 products support SLR and more products will be supported in the second half of this year. At that time, automatic cross-service access will no longer be a problem on Alibaba Cloud.

OpenAPI Access Compliance

In the compliance field, operation audit and resource audit are generally performed in common scenarios. However, the industry supervision principle is also an important reference factor. For example, in the finance cloud industry, cross-network callings must be made under controllable and secure conditions. This requires that cloud-based network callings must comply with supervision requirements.



Alibaba 11.11 Global Shopping Festival: Behind the Scenes Week

Join our one-week digital series, that covers exciting announcements and behind the scenes technical deep-dive sessions, on how Alibaba Cloud and modern cloud native technologies, have empowered the 2020 Alibaba 11.11 global shopping festival, the world largest online shopping festival. Learn from product experts and gain actionable takeaways to help supercharge your digital journey.

Agenda

Dec 1 - Double 11 Insights

Dec 2 - Cloud Native - Building Modern Applications

Dec 3 - Managing Infrastructure and Media

Dec 4 - Database Management

Date

Dec. 1 - Dec. 4 (10AM - 1PM SG/HK | 9AM - 12PM Jakarta)

Register now for free

alibabacloud.com/campaign/singles-day-double-11-2020