CORE CONCEPTS AND BEST PRACTICES

Alibaba Cloud

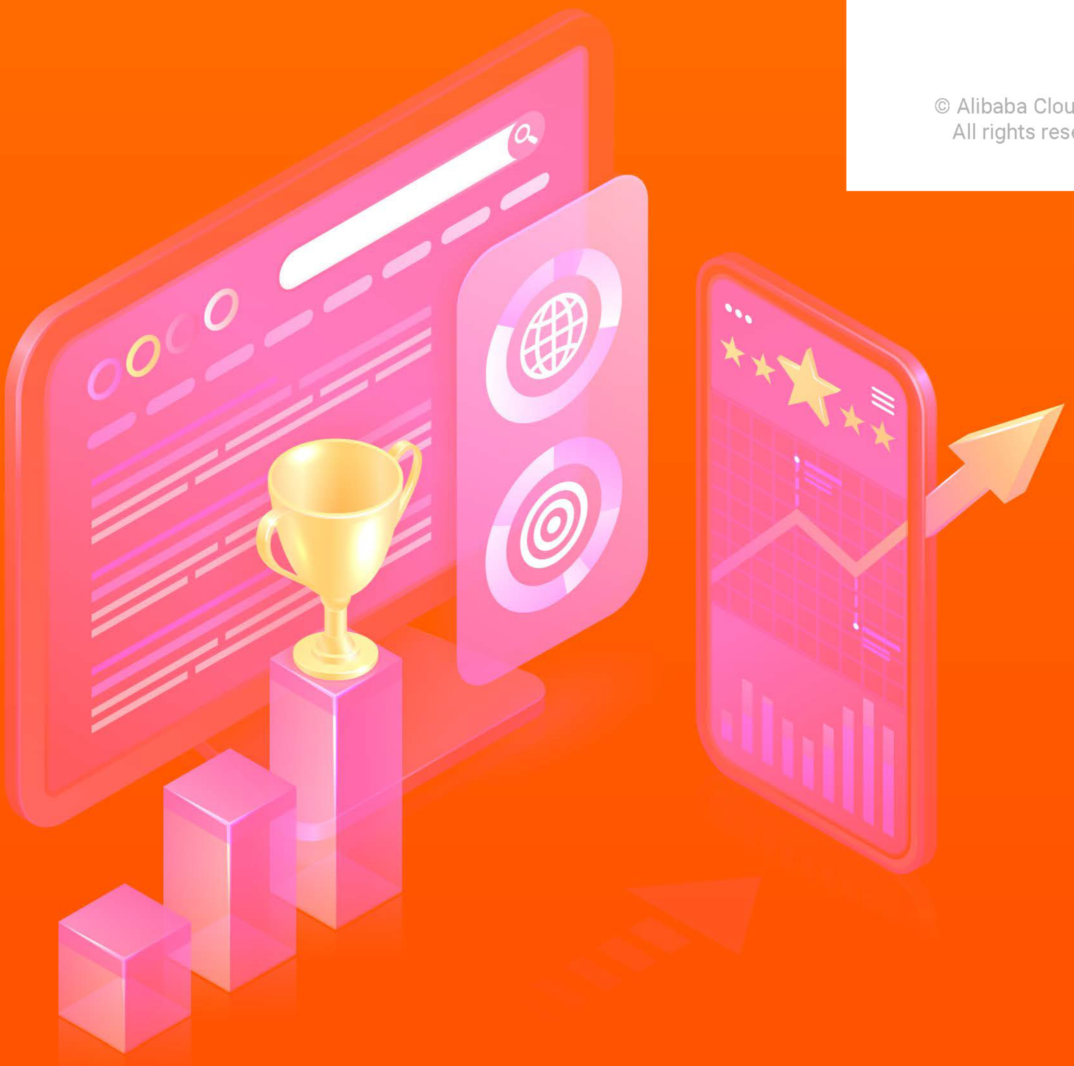
AUTHORS

Oliver ARAFAT
Dr. Ye HUANG
Wei TONG

SPONSORS

XinJia GE
Lei MA
QingSong JIANG
Hai HE
YingQiao SONG
Qian YUAN

alibabacloud.com



LEGAL NOTICES

Alibaba Cloud reminds you to carefully read through and completely understand all content in this section before you read or use this document. If you read or use this document, it is considered that you have identified and accepted all contents declared in this section.

1. You shall download this document from official website of Alibaba Cloud or other channels authorized by Alibaba Cloud. This document is only intended for legal and compliant business activities. The contents in this document are confidential, so you shall have the liability of confidentiality. You shall not use or disclose all or part of contents in this document to any third party without written permission from Alibaba Cloud.
2. Any sector, company, or individual shall not extract, translate, reproduce, spread, or publicize, in any method or any channel, all or part of the contents in this document without the written permission from Alibaba Cloud.
3. This document may be subject to change without notice due to product upgrades, adjustment, and other reasons. Alibaba Cloud reserves the right to modify the contents in this document without notice and to publish the document in an authorized channel from time to time. You shall focus on the version changes of this document, download and get the updated version from channels authorized by Alibaba Cloud.

4. This document is only intended for product and service reference. Alibaba Cloud provides this document for current products and services with current functions, which may be subject to change. Alibaba Cloud makes the best efforts to provide an appropriate introduction and operation guide on the basis of current technology, but Alibaba Cloud does not explicitly or implicitly guarantee the accuracy, completeness, suitability, and reliability of this document. Alibaba Cloud does not take any legal liability for any error or loss caused by downloading, using, or putting trust in this document by any sectors, company, or individuals. In any case, Alibaba Cloud does not take any legal liability for any indirect, consequential, punitive, occasional, incidental, or penalized damage, including profit loss due to use of or trust to this document (even if Alibaba Cloud has notified you it is possible to cause this kind of damage).
5. All content, including, but not limited to images, architecture design, page layout, description text, and its intellectual property (including but not limited to trademarks, patents, copyrights, and business secrets), in this document are owned by Alibaba Cloud and/or its affiliates. You shall not use, modify, copy, publicize, change, spread, release, or publish the content from official website, products, or programs of Alibaba Cloud without the written permission from Alibaba Cloud and/or its affiliates. Nobody shall use, publicize, or reproduce the name of Alibaba Cloud for any marketing, advertisement, promotion, or other purpose (including, but not limited to, a separate or combined form to use the name, brand, logo, pattern, title, product or service name, domain name, illustrated label, symbol, sign, or similar description that may mislead readers and let them identify it comes from Alibaba Cloud and/or its affiliates, of or from Alibaba Cloud, Aliyun, Wanwang, and/or its affiliates) without the written permission from Alibaba Cloud.
6. If you find any error or mistake in this document, directly contact Alibaba Cloud.

CONTENTS

Foreword 1

Introduction 3
- Why Alibaba Cloud 4
- Who Should Read This Book 7
- How To Read This Book 8

Governance 9
- Account Management 9
- Notification Management 15
- User and Permission Management 16
- Links 24
- Billing Management 24

Interacting with Alibaba Cloud 27
- API Model 27
- Using the Command-Line Interface 29
- Alicloud-Vault 30
- Using the OSSUtil CLI 31
- Infrastructure as Code 32
- Debugging with API-Explorer 35

Infrastructure Essentials 37
- Elastic Compute Service (ECS) 37
- Storage 40
- Network Performance 44

Securing your system 47
- Security Center 47
- Account Security 49
- Account Auditing 50
- Network Security 51

Architecting for High Availability and Fault-Tolerance on Alibaba Cloud 55
- Regions and Availability Zones 56
- Deployment Sets 56
- Built-in Redundancy of Alibaba Cloud Services 57
- Multi-Site High Availability 62
- Disaster Recovery 64

Global Cross-Border Integration 66
- VPC-Peering with Cloud Enterprise Network 66
- Hybrid Networks 68
- Global Accelerator 69
- Solution Case Study: Private Global Internet Accelerator 71

Conclusion 78

FOREWORD

In recent years, more and more enterprises have started their digitization journey in order to enhance business efficiency and ensure continuous success. Cloud technology has been proven in practice to be pivotal for many global well-known enterprises and organizations, serving both as the technical foundation as well as the innovative backbone, to enable new business models for businesses of all sizes and industries. I trust one of the most important elements of the “digital era” is cloud computing and its relevant technologies, which is undoubtedly providing tremendous opportunity for sustainable growth. However, there exists a high demand and barrier for cloud vendors to remain competent in the highly competitive global market.

Alibaba Cloud, born from Alibaba Group, has developed its own sophisticated technology stack and a wide scope of product portfolio, including but not limited to computing, storage, security, Big Data, artificial intelligence, and hybrid cloud services. Through its considerable product capabilities, Alibaba Cloud has supported the business development and operation of many global enterprises from various geographic regions worldwide.

It is also noteworthy that most of Alibaba Cloud's offering has been battle-tested by Alibaba Group's own business scenarios, including some of the most challenging cases, such as supporting Alibaba's annual global shopping festival – Double 11. Alibaba Cloud's cutting-edge offering is not only complete feature-wise, but also robust and reliable in extreme use cases, capable of serving hundreds of millions of customers during peak periods while ensuring optimal customer experience. The success of Alibaba Group's usage of Alibaba Cloud is a testament to the maturity of the cloud portfolio and the performance and feature breakthroughs of the products.

This book aims to share a glimpse into Alibaba Cloud's core offering including our computing and storage services, as well as topics enterprise architects often care about and are interested in, such as IT governance and management, building high-availability and fault-tolerance systems, global deployment of applications, and seamless integration of services. By reading this book, we sincerely hope you get to know Alibaba Cloud better, and we hope you will continue learning about more sophisticated topics like Big Data, Artificial Intelligence, Media Services, and Internet of Things, from our website at www.alibabacloud.com.

Welcome to learn more about Alibaba Cloud!

Selina Yuan
President of Alibaba Cloud Intelligence International Business

INTRODUCTION

Technology and in particular today's Public Cloud platforms are evolving fast at an incredible pace while books are rather static with "release cycles" measured in years. Many technology books are already outdated they day they are published. This begs the question, why would anyone would want to read - let alone write - a book about one of the fastest moving targets in the information technology industry: Alibaba Cloud?

Having spent many years working with different cloud platforms it has become clear that the underlying principles and concepts of just about any cloud platform do not change half as fast as new services and features are being added about almost every single day. These core concepts serve as the foundation for just about anything that's being built upon and offered by the platform. A thorough understanding and mastery of these concepts is crucial for successfully building and operating anything between a simple web application and the next big thing with millions of users spread around the globe. Many of those concepts are valid for years and are only very carefully touched upon by the product groups that are building the cloud platform.

This book will dive deep into the core concepts of Alibaba Cloud, and best practices that have stood the test of time and have been successfully applied at many customer projects. You will learn how and why things are working the way they are, and how and why you should do things in a certain way when using the Alibaba Cloud platform. Once you got the hang of the underlying principles, and the "Ali-way" of doing things, performing tasks such as building, testing, monitoring, and operating will become much more enjoyable. It will also help you reason about the behavior and limitations of our big portfolio of managed services in the area of Big Data, IoT, Analytics, Compute, Storage, and Network.

Over the past few years we had the joy of working with many different kinds of customers in different verticals, ranging from small startups, to big enterprise players. As broad and diverse as this range is, the questions and uncertainties raised by each customer when starting to build and operate real-world workloads on Alibaba Cloud remain the same. This book is the result of our accumulated experience we have gained during these years. We hope that you will enjoy reading it and that it will help you in reaching your full potential with cloud-based solutions.

WHY ALIBABA CLOUD

There are multiple strategic reasons to include Alibaba Cloud into your IT strategy, three of which we will detail in the subsequent sections.

Fast-Paced Innovator that Supports World's Most Demanding E-Commerce Applications

Alibaba Cloud supports one of the highest demanding online events on earth reliably each year: Double 11, the world's biggest shopping festival with a gross merchandized volume of more than 74.1 billion USD in 11 days in 2020. To give you a better idea on the scale of this event let's drop some numbers on the resources and workloads that are exclusive to this event (2020):

- 800+ million participating customers
- 2.32 billion orders and deliveries orchestrated, coordinated, and delivered.
- 583k orders per second during peak times which the highest traffic peak ever witnessed in the world
- 250k+ participating brands
- 1.7 billion network attacks that were successfully mitigated

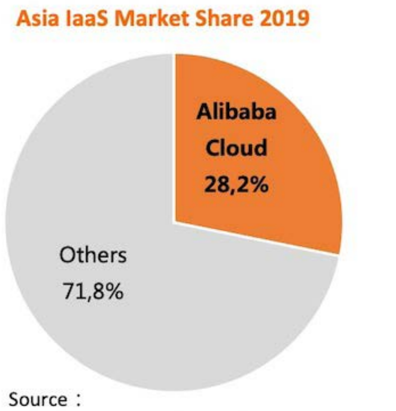
Alibaba Cloud is the technology backbone of Alibaba Group and has been constantly pushing the limits of today's technologies to support our own core business. This fast-paced innovation and battle-proven technology is provided to our millions of customers, which are often large enterprises by themselves with millions of end-customers. The constant stream of innovation is productized constantly and released frequently as new services and features to build upon and to reliably support business-critical application infrastructure by the millions of our customers world-wide.

Second to None Business and Technology Partner for China

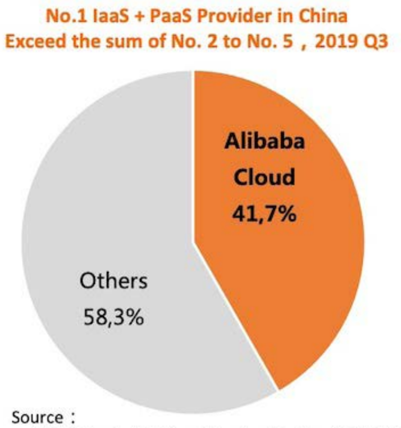
Alibaba Group and its cloud division is an international company with Chinese roots. As such we are second to none as a technology and business partner for your IT and innovation projects in South

East Asia and China in particular. Alibaba Cloud is the clear market leader with 28.2% market-share in Asia Pacific, well ahead any of its cloud competitors such as AWS, Microsoft Azure, or Google Cloud as depicted in below graphic, which is the Market Share for IT Services report by Gartner from 2019.

For Mainland China the market-share proportions are even more drastic. Alibaba Cloud took over 41.7% in Q3 of 2019, which is larger than the next five cloud competitors combined.

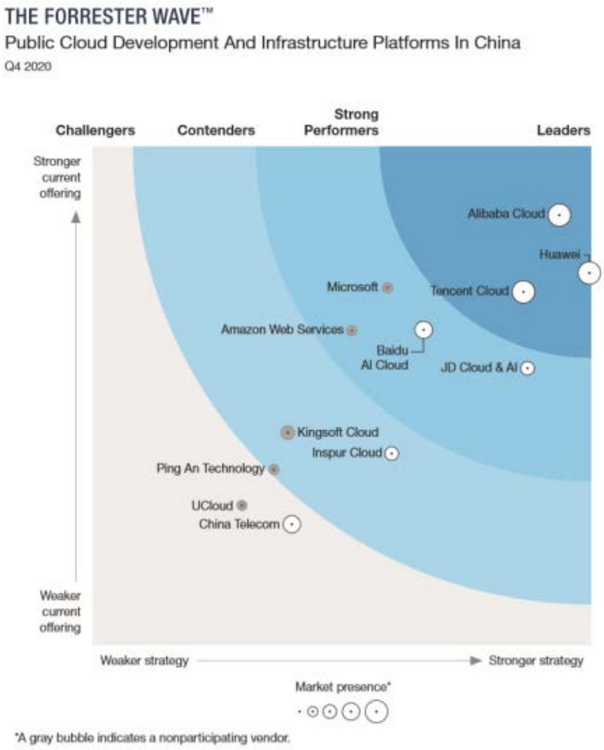

Asia IaaS Market Share 2019

Source：
- Gartner Market Share: IT Services，2019


No.1 IaaS + PaaS Provider in China
Exceed the sum of No. 2 to No. 5，2019 Q3

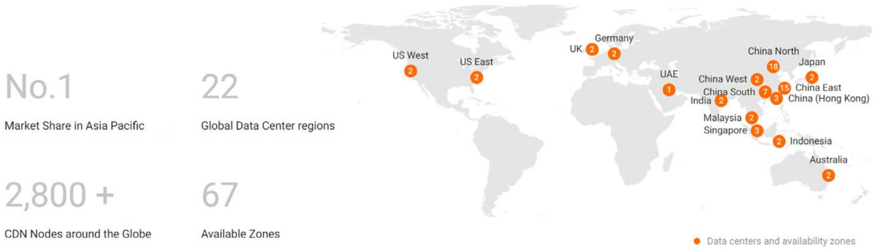
Source：
- IDC China Public Cloud Service Tracker, 2019 Q3

Forrester Wave also recognizes Alibaba Cloud as the leader in product offerings and market performance in its *Full-stack public cloud developments platforms in China* report, which is depicted below.


THE FORRESTER WAVE™
Public Cloud Development And Infrastructure Platforms In China
Q4 2020

*A gray bubble indicates a nonparticipating vendor.

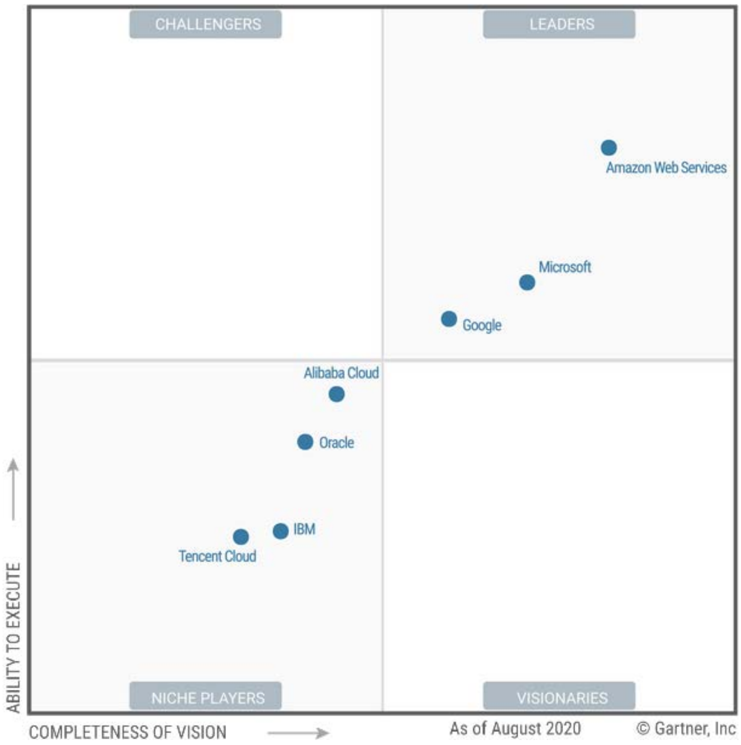
With its strong presence in the Asia Pacific region with (as of this writing) 16 regions, 10 of which are in Mainland China, and one additional region in Hong Kong, Alibaba Cloud is the number one choice for global, scalable, and secure application platforms that empower your digital innovation strategy in the APAC regions. In total, Alibaba Cloud has 22 regions world-wide which also includes North-America (2) and Europe (2) which lets you also benefit from the innovation capabilities of Alibaba Cloud in your local regions while fully complying to the local law and regulation such as GDPR. Below figure gives you an overview about our global footprint of 22 region and 67 availability zones as of 2020.



Note that all of these 22 regions can be managed from one single account. Many services also provide cross-region integration capabilities that let you easily replicate and copy data to and from any of our regions. We will look in depth at the different cross-regional services and features in chapter Global Cross-Border Integration.

Geo-Politically Balanced Multi-Cloud Strategy

Below graphic shows the Gartner Magic Quadrant for IaaS Worldwide from 2020. If you compare it to the quadrants of previous years you will notice that both the number and geographic heterogeneity has been drastically reduced. In total, there are only seven players left, five of which are headquartered in the USA. This leaves Alibaba Cloud and Tencent Cloud as the only hyperscalers that are not under American jurisdiction. A multi-cloud strategy that is aimed at minimizing regulatory and legal risks associated with a multi-regional deployment can be crucial for businesses. Alibaba Cloud is a reliable and trusted business and technology partner for implementing such a strategy and can help you mitigate these risks.



Magic Quadrant for Cloud Infrastructure as a Service, Worldwide

WHO SHOULD READ THIS BOOK

This book is intended for anyone who wants to get a thorough understanding about the core concepts of Alibaba Cloud along with best-practices and migration strategies. If you already have basic knowledge in cloud computing, it will be helpful, but it is not necessary. We will start with a thorough discussion on the principles of Account Management, Authentication and Authorization, Billing Management, and our geography-based data center architecture. After that we will look at the different concepts and approaches for securing, auditing, and monitoring your accounts and workloads, as well as modern approaches on automating different aspects of your system ranging from service provisioning and configuration to build- and deployment-pipelines. We will then continue this book with a deep discussion on our best-practices to design for high-available and fault-tolerant systems on Alibaba Cloud. Finally, we will devote the last chapter to our recommended approaches and services at how to reliably connect and integrate geographically disperse applications on a global (and hybrid) cross-border network.

If you are an architect, a consultant, an administrator, or a technical decision maker who just wants a better knowledge of the Alibaba Cloud core concepts, unique selling points, and best-practices, this book is for you. Many scenarios are also supported by code examples. We do not make any assumptions regarding the reader's level of knowledge.

HOW TO READ THIS BOOK

Here is a glance at what's in each chapter:

- **Chapter 1: Introduction** is where you are right now.
- **Chapter 2: Governance** focuses on the fundamental aspects of Account Management, User and Permission Management, and Billing Management.
- **Chapter 3: Interacting with Alibaba Cloud** dives into the various means on how to programmatically manage and debug our service portfolio. We will discuss the general API model of Alibaba Cloud and tools that support you in your daily work.
- **Chapter 4: Infrastructure Essentials** gives a focused rundown on the very essentials on the Alibaba Cloud Infrastructure services concepts such as Compute, Network, and Storage.
- **Chapter 5: Securing Your System** discusses proven best-practices and methodologies to secure your account, and networking environment. Additionally, we will also look at how to audit your system to quickly analyze who has done what at which point in time.
- **Chapter 6: Architecting for High-Availability and Fault-Tolerance on Alibaba Cloud** looks in depth at proven practices and recommendations to make your system and application architecture robust and resilient against sudden unpredictable traffic spikes and service interruptions. It will also look at our geography-based data center architecture, and our Service Level Agreements (SLAs).
- **Chapter 7: Global Cross-Border Integration** explores the various networking services Alibaba Cloud offers to implement and manage global and hybrid cross-border network integration projects, and also looks briefly at the practical implications of the Cyber Security Law and ICP licensing.

GOVERNANCE

Cloud governance is the development and implementation of controls to manage access, budget, and policies for security, compliance, and even high-availability and resiliency across your workloads in the cloud. In this chapter, we will focus on the core concepts and according best-practices to implement according controls and policies for most of the before mentioned aspects.

ACCOUNT MANAGEMENT

There are currently two different types (or *memberships*) of Alibaba Cloud accounts: *Individual Account* and *Enterprise Account*. There is no difference in terms of functionality. They differ, however, in terms of real-name verification requirements, free tier offering, discount eligibility, and whether they can be added to a Resource Directory account. Let's quickly go through each of them.

- **Real-Name Verification:** Each account needs to be verified with a real identity if you want to create cloud resources in Mainland China. This is a strict requirement. Resources include OSS buckets and ECS instances for example. Basically, all resources that either expose a public IP address or a public domain name. The verification differs from what kind of identity proof needs to be submitted for each account type. For *Individual Account* you need to submit a proof of your personal identity which could be a copy of your passport. For *Enterprise Account* you need to submit an excerpt from the commercial register. This process is completely supported by the Web Portal and usually takes 3-4 business days to complete.
- **Free-Tier Offering:** Based on the account type the free trial offering differs in terms of how many free credits you get and what kind of products you are eligible for. Please see https://www.alibabacloud.com/campaign/free-trial for details. The *Enterprise Account* has a much bigger free product range and also more free credits.

- **Discount Eligibility:** Throughout the year Alibaba Cloud often has special promotion campaigns which include discounts on certain products. These discounts often differ depending on the cloud account type. Usually, *Enterprise Accounts* are eligible for higher discounts.
- **Resource Management Integration** For Multi-Account management, Alibaba Cloud is providing a service called Resource Management. It allows to manage multiple accounts under one master account and to aggregate metrics, logs, RAM users, and billing data for example. Please refer to next section to get an overview about Alibaba Cloud Resource Management Service.

If possible, we strongly recommend to choose an *Enterprise Account* since it usually provides higher discounts during promotion campaigns, and also has a stronger free-trial offering. As mentioned before, there is no difference in functionality otherwise other than Resource Management support.

Multi-Account and Resource Management

Managing resources in large-scale settings is very complex and needs to take into account the following questions:

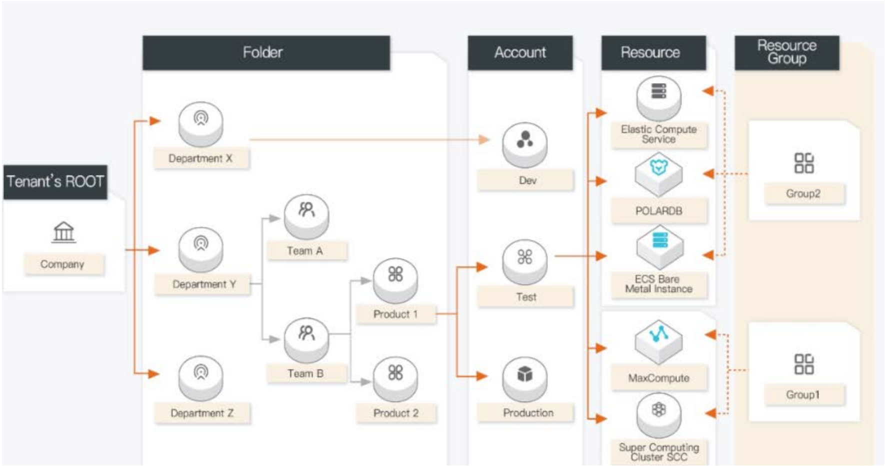
- How can enterprise decision makers have a top-down overview of the use of all resources?
- How can enterprises align the resource structure with the business management structure to match different management strategies?
- How can enterprises enable their different branches or departments to implement different procurement, usage, and regulatory requirements in the cloud?
- How can enterprises organize scattered cloud accounts according to the business structure for effective management?

Alibaba Cloud Resource Management Service enables you to build an organizational structure for resources based on your business needs. It is comprised of different services that let's you hierarchically organize and manage accounts, billing and settlements, users and permissions, and resources and thus provides solutions to before mentioned questions. In particular, it consists of:

- **Resource Directory** Enables you to quickly build a business structure according to the needs of the enterprise and to carry out overall management of resources on it.

- **Resource Group** Helps you to group related services across different regions and assign dedicated permissions. Also allows you to view billing statements by resource group.
- **Resource Sharing** Allows you to share the resources under your account with other accounts.

Below picture puts all of these three services in perspective and context:



Resource Management

In principal there are two types of cloud accounts in such a settings: The Master or Root account, and so-called Member accounts.

The *Master* or *Root Account* is a regular Enterprise Account and should be solely used to initially activate and configure Resource Management. It should not be used for any LoB or project workloads. It is highly advised to use string passwords and MFA to secure this account and highly limit administrator access. It is the root node in your Resource Directory structure.

Member accounts exist in two flavors: *Cloud Accounts* and *Resource Accounts*. While *Cloud Accounts* have all the characteristics of a regular Alibaba Cloud account such as a dedicated RAM-based user management and a root user, *Resource Accounts* can be thought of as container-like entities for managing cloud services without dedicated RAM-users and root account. Thus they are safer and more convenient but they fully depend on Resource Directory. They can be programmatically created via the API `CreateResourceAccount` (https://www.alibabacloud.com/help/doc-detail/159392.htm). They can be anytime "promoted" to a regular cloud account which is not dependent on Resource Directory. We recommend to work with *Resource Accounts* if possible.

It should be noted that you can also invite existing Alibaba Cloud accounts. As of this writing, the invited accounts need to have the same legal business entity as the master account. Support for inviting accounts with different legal entities is coming soon. The master of account has full control over the invited accounts.

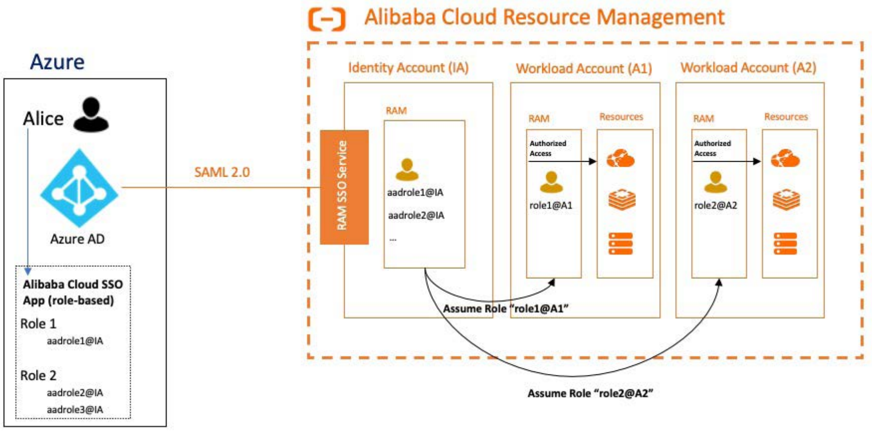
Single-Sign On

Alibaba Cloud supports SAML 2.0-based single sign-on (SSO), which is also known as identity federation. You can implement SSO between Alibaba Cloud and your Identity Provider (IdP), such as Microsoft Azure Active Directory, based on SAML 2.0. Alibaba Cloud provides the following two SAML 2.0-based SSO methods:

- **User-based SSO** The RAM user identity that you can use to log on to the Alibaba Cloud Management Console is determined based on an SAML assertion. After you log on to the Alibaba Cloud Management Console, you can access Alibaba Cloud resources as a RAM user. For more information, see Overview of user-based SSO.
- **Role-based SSO** The RAM role that you can use to log on to the Alibaba Cloud Management Console is determined based on an SAML assertion. After you log on to the Alibaba Cloud Management Console, you can use the RAM role that is specified in the SAML assertion to access Alibaba Cloud resources. For more information, see Overview of role-based SSO.

For integrating with Azure AD hosted in Mainland China please be aware that of this writing 3rd-party Enterprise application feature is not available yet. This means that you cannot use the Azure AD Enterprise application for directly configuring Alibaba Cloud role-based SSO in Azure AD.

A common setup for managing a multi-account setup with SSO we recommend the following approach as depicted in below figure:



Role-based SSO Setup with Resource Directory

The root account can serve as the identity account but they can also be separated if needed. This account serves as the landing page for each identity. From there a subsequent role can be assumed that grants according rights in a member account.

Billing Settlement is quite flexible. For any member account you can choose between three options:

1. Master Account: Consolidate bill to the master account
2. Other Account: Consolidate bill to another account within the resource Directory
3. Self-Pay: Account pays by itself with the payment method configured in this account

This enables you to configure any settlement structure that can be expressed in a hierarchical tree with any node working as a billing account.

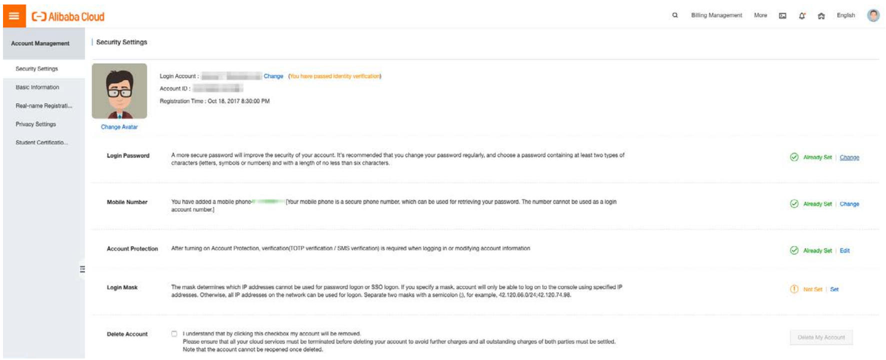
Moreover, for any multi-account management approach it is vital to have an aggregated view and consolidated management access to auditing, configuration, and security. Alibaba Cloud Resource Management integrated with other cloud services to account for that. In particular it integrates with

- ActionTrail to give you an aggregated view on all audit logs and events across all member accounts
- Cloud Config to enable you to track and record configuration changes for resources
- Security Center to help you protecting all your member accounts and workloads from a unified security solution

Security Settings

Now that we have discussed the two principal account types of Alibaba Cloud and discussed multi-account management, let's look at the core concepts of managing a cloud account. Each cloud account has exactly one and only one root user. You are specifying the login name and password during initial cloud account creation. Each root user also has an associated mobile phone number which can be re-used across different root users and cloud account respectively at most six times.

In section **User and Permission Management** we will look in detail at how to create additional users and define according permissions. Below screenshot shows the options of the *Security Settings* administration page which is only available for the root user and accessible directly via https://account-intl.console.aliyun.com.



Security Settings

Login Account : [redacted] Change (You have passed identity verification)
Account ID : [redacted]
Registration Time : Oct 18, 2017 8:30:00 PM

Login Password	A more secure password will improve the security of your account. It's recommended that you change your password regularly, and choose a password containing at least two types of characters (letters, symbols or numbers) and with a length of no less than six characters.	Already Set | Change
Mobile Number	You have added a mobile phone·[redacted] [Your mobile phone is a secure phone number, which can be used for retrieving your password. The number cannot be used as a login account number.]	Already Set | Change
Account Protection	After turning on Account Protection, verification(TOTP verification / SMS verification) is required when logging in or modifying account information	Already Set | Edit
Login Mask	The mask determines which IP addresses cannot be used for password logon or SSO logon. If you specify a mask, account will only be able to log on to the console using specified IP addresses. Otherwise, all IP addresses on the network can be used for logon. Separate two masks with a semicolon (;), for example, 42.120.66.0/24;42.120.74.98.	Not Set | Set
Delete Account	□ I understand that by clicking this checkbox my account will be removed. Please ensure that all your cloud services must be terminated before deleting your account to avoid further charges and all outstanding charges of both parties must be settled. Note that the account cannot be reopened once deleted.	Delete My Account

Account Management - Security Settings

On this page you can easily change the login password and your mobile phone number. The mobile number is important as it is used as a second authentication factor for changing the password, and also for setting up MFA for your root user. Last but not least you can also define a login mask that lets you explicitly whitelist IP ranges from which (SSO) login is allowed. Per default, all IP ranges are allowed.

BEST PRACTICES

1. Enable *Account Protection*, that is MFA, which currently supports Time-based One-time Password (TOTP) and SMS verification
2. Define a *Login Mask* that only allows login from a specific IP range, e.g. the outbound IP range of your corporate network, thus blocking illegal login attempts from unknown IP ranges.
3. Choose a password that is sufficient in both complexity and length, make sure to activate password rotation, and restrict session duration. Consult your security advisor on your company's password and security guidelines.
4. Never activate *Access Key ID* and *Access Key Secret* for your root user. You can check at https://usercenter.console.aliyun.com whether according keys have been defined. Deactivate them immediately since the root account is not recommended for any programmatic use. Think of it as the root user on Linux systems which has universal access rights to each and everything and whose rights cannot be restricted. For the day to day work the root account should never be used at all!
5. Activate ActionTrail to fully audit your account. Please see section **Account Auditing** of chapter **Securing your System** for details.

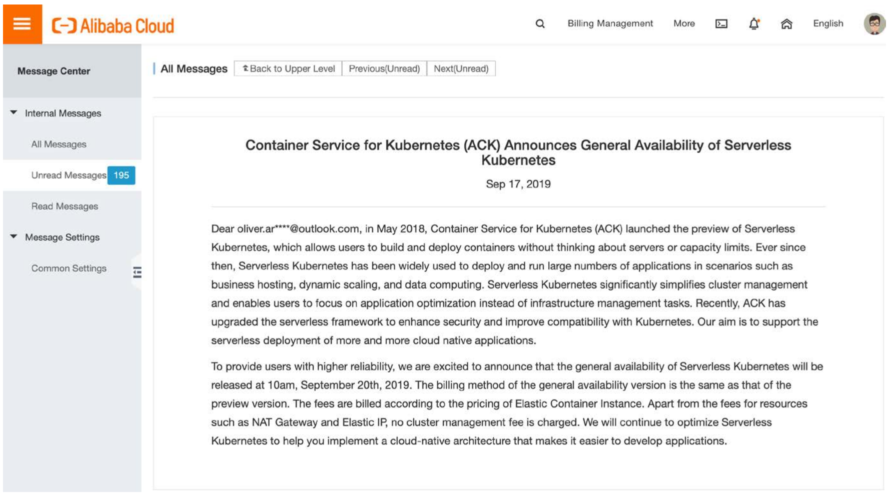
In the **best-practices section of User and Permission Management** we will look at more recommendations to keep your cloud resources secure.

NOTIFICATION MANAGEMENT

The so-called *Message Center* which is available at https://notifications-intl.console.aliyun.com provides means to get proactively notified about important incidents and updates on and about the Alibaba Cloud platform. These notification messages are divided into five distinct groups:

- **Account Message:** This is about notifications about account expenses. For example, you can subscribe to notifications about overdue bills.
- **Product Message:** This includes updates on product upgrades, configuration changes, price changes, new product launches, but also notifications when services are about to expire and need to be renewed.
- **Fault Message:** Notifications in this group mostly deal with service interruptions, malfunctioning, and system maintenance windows. Really useful to get proactively notified about such kind of unplanned events.
- **Activity Message:** This is about marketing campaigns and promotions, offline events, and new product and regions launches.
- **Service Message:** This includes notifications on recommended easy-to-use tools, tutorials and lessons-learned to help you use our cloud products even more efficiently.

Each notification can be configured to be delivered either via email or as internal message. While the email receivers can be freely configured, internal messages work just like an email inbox in your Alibaba Cloud web portal which is only accessible by the root user. Below screenshot shows an example of the internal message UI with a notification from the *Product Message* group which announces the general availability of our Serverless Kubernetes offering.



Container Service for Kubernetes (ACK) Announces General Availability of Serverless Kubernetes

Sep 17, 2019

Dear oliver.ar****@outlook.com, in May 2018, Container Service for Kubernetes (ACK) launched the preview of Serverless Kubernetes, which allows users to build and deploy containers without thinking about servers or capacity limits. Ever since then, Serverless Kubernetes has been widely used to deploy and run large numbers of applications in scenarios such as business hosting, dynamic scaling, and data computing. Serverless Kubernetes significantly simplifies cluster management and enables users to focus on application optimization instead of infrastructure management tasks. Recently, ACK has upgraded the serverless framework to enhance security and improve compatibility with Kubernetes. Our aim is to support the serverless deployment of more and more cloud native applications.

To provide users with higher reliability, we are excited to announce that the general availability of Serverless Kubernetes will be released at 10am, September 20th, 2019. The billing method of the general availability version is the same as that of the preview version. The fees are billed according to the pricing of Elastic Container Instance. Apart from the fees for resources such as NAT Gateway and Elastic IP, no cluster management fee is charged. We will continue to optimize Serverless Kubernetes to help you implement a cloud-native architecture that makes it easier to develop applications.

Message Center Notification

BEST PRACTICES

While we believe that each notification is valuable for our customers we recommend to activate at least the following ones:

- Account Message - *Notifications of Account Expenses*
- *Product Message - Alibaba Cloud DNS High Risk Notification*
- *Product Message - Notifications of Product Expiration*
- *Product Message - Product Overdue Payment, Suspension, and Imminent Release Notifications*
- *Product Message - Notifications of Product Release*
- *Product Message - Notifications about product upgrades, configuration changes, and price changes*
- *Product Message - New Product Function Launch and Function Removal Notifications*
- *Product Message - Security Notice*
- *Fault Message* - **Activate all notifications**

USER AND PERMISSION MANAGEMENT

Resource Access Management (RAM) is the cloud service which provide means to create additional users (so-called RAM users), and roles with according policies (sometimes also referred to as permissions) that define the access rights on Alibaba Cloud services and specific resources. The interface (i.e. the set of APIs) which is used to manage your cloud resources is usually referred to as OpenAPI which you can interactively explore with the OpenAPI Explorer at https://api.aliyun.com/.

Let's break down the different terms we just mentioned and explain what they exactly mean.

Root User

The initial single sign-in identity that has complete access to all Alibaba Cloud services and resources in the account. This identity is called the Alibaba cloud account root user and is accessed by signing in with the email address and password that you used to create the account. Follow the best-practice and use it only to create your first RAM user. Never use it for day to day tasks, and never use it to access your Alibaba Cloud resources.

There are, however, some tasks only the root user can do:

- Modify root user details on the *Security Settings* administration page at https://account-intl.console.aliyun.com
- Delete the Alibaba Cloud account which is accessible via *Security Settings* administration page

- Initial activation of cloud services: most services are deactivated by default and need to be explicitly activated by the root user. This action is irreversible, meaning an activated service can never be deactivated again.
- Real-name verification of your account which is needed for creating any kind of resources in Mainland China regions which have a public IP and/or a public domain name.

Most other things (such as whitelisting port 25 for an ECS instance, Reverse DNS entry for ECS) are usually requested by opening an according support ticket which is not restricted to the root user, though.

RAM Users and Policies

A RAM user is sometimes also referred to as *Sub-Account*. It is a user account that is used for web-based login to the Alibaba Cloud portal and/or programmatic access to the OpenAPI. They can't access anything in your account until you give them permission. All permissions need to be explicitly granted. You give permissions to a user by creating an identity-based policy, which is a policy that is attached to the user or a group to which the user belongs. The following example shows a JSON policy that allows the user to perform all TableStore actions (ots:*) on the Books table in the 123456789012 account within the eu-central-1 region.

```
{
  "Version": "1",
  "Statement": {
    "Effect": "Allow",
    "Action": "ots:*",
    "Resource": "acs:ots:eu-central-1:123456789012:table/Books"
  }
}
```

After you attach this policy to your RAM user, the user only has those TableStore permissions. Most users have multiple policies that together represent the permissions for that user. The evaluation policy is as follows:

- By default, all requests are implicitly denied (except for requests by the root user)
- An explicit *Allow* overrides this default
- Any explicit *Deny* overrides any allows.

There are two access modes you can define: Console Password Logon and Programmatic Access The first one is used for web-based login where each actions are being done from the Alibaba Cloud portal. In terms of account protection it follows the same recommended guidelines regarding password security and rotation, and MFA. The latter one is meant for being used in combination with our command line interface (CLI) tools such as `aliyun`^aliyun or `ossutil`^ossutil or with our various SDKs^sdk. As such it does not provide MFA but relies on long-term credentials (Access Key ID and Access Key Secret) to programmatically sign requests to the CLI tools or the OpenAPI.

Roles

A RAM role is an RAM identity that you can create in your account that has specific permissions. An RAM role is similar to an RAM user, in that it is an Alibaba cloud identity with permission policies that determine what the identity can and cannot do in the cloud account. However, instead of being uniquely associated with one person, a role is intended to be assumable by anyone who needs it and is defined as an authorized principal to assume it. Also, a role does not have standard long-term credentials such as a password or access keys associated with it. Instead, when you assume a role, it provides you with temporary security credentials for your role session which consist of an *AccessKeySecret*, and *AccessKeyId*, and a *SecurityToken*. In this case, the *AccessKeyId* is always prefixed with `STS.`.

Below is an example role definition that allows every RAM user (yes, this is a little bit counter-intuitive since we specify `root` as the principal) to assume it who is allowed to invoke `sts:AssumeRole` and whose request originates from within a certain IP range.

```
{
    "Statement": [
        {
            "Action": "sts:AssumeRole",
            "Effect": "Allow",
            "Principal": {
                "RAM": [
                    "acs:ram::<UID>:root"
                ]
            }
            "Condition": {
                "StringEquals": {
                    "acs:SourceIp": [
                        "47.234.42.0/24"
                    ]
                }
            }
        }
    ],
    "Version": "1"
}
```

For example the user can manually invoke `sts:AssumeRole` with the CLI just like that:

```
aliyun sts AssumeRole --RoleArn acs:ram::<UID>:role/myrole
--RoleSessionName service
```

As you can see you can also specify a role session name which is used as the caller id which is logged in ActionTrail for example. Usually, you should restrict access to a specific group of users. An individual user can be specified as `acs:ram::<UID>:user/myuser`.

There are three principal types we currently support:

- RAM user from a trusted Alibaba Cloud account. This is what we have just discussed. Note that you can also specify RAM users from other accounts by simply changing the account id accordingly. This enables you to implement cross-account access permissions.
- Cloud services such as ECS, RDS, Function Compute, etc. This enables these services to automatically assume roles and thereby getting according permission when they are configured to access other cloud resources.
- Identities from other Identity Providers such as Active Directory that are being used in Single Sign-On scenarios.

Resources

We have already briefly touched upon what a cloud resource is in the previous sections but let's recap: a cloud resource is any instance of a particular cloud service. For example, an ECS instance is a resource of the ECS service. Likewise, an OSS bucket and an OSS object is a resource of the OSS service. Each resource on Alibaba Cloud has a unique identity that you can use to define very fine granular permissions. Instead of granting full access to each and every ECS instance in your account you can only grant access to a particular ECS instance. This is where the `Resource`field of a RAM policy comes into play as already shown section **RAM Users and Policies**. The identifier of an Alibaba cloud resource is always structured like this:

```
acs:<service-name>:<region-id>:<account-id>:<resource-name>
```

You can also specify a wildcard (*) expression for the individual parts of a resource. For example, if you like to specify all ECS instances in all regions you would write

```
acs:ecs:*:<account-id>:*
```

If you like to specify a specific object such as myfile.dat in a specific bucket mybucket in eu-central-1, you would write:

```
acs:oss:eu-central-1:<account-id>:mybucket/myfile.dat
```

Since a bucket is always bound to a particular region you can also omit the region like so

```
acs:oss::<account-id>:mybucket/myfile.dat
```

Best Practices

To help secure your Alibaba Cloud account follow these recommendations for Alibaba Cloud Resource Access Management:

- Lock Away Your Root Credentials: You use an access key (an access key ID and secret access key) to make programmatic requests to Alibaba Cloud. However, do not use your root user access key. The access key for your root user gives full access to all your resources for all Alibaba Cloud services, including your billing information. You cannot reduce the permissions associated with your Alibaba Cloud account root user access key.

Therefore, protect your root user access key like you would your credit card numbers or any other sensitive secret. Unless you do not absolutely need a root access keys never create them in the first place. If you do have one, delete them. You can check at https://usercenter.console.aliyun.com whether according keys have been defined.

- **Create Individual RAM Users:** Don't use your Alibaba Cloud account root user credentials to access any cloud service or resource, and don't give your credentials to anyone else. Instead, create individual users for anyone who needs access to your cloud account. Create a RAM user for yourself as well, give that user administrative permissions, and use that RAM user for all your work. By creating individual RAM users for people accessing your account, you can give each RAM user a unique set of security credentials. You can also grant different permissions to each RAM user. If necessary, you can change or revoke a RAM user's permissions anytime. If you give out your root user credentials, it can be difficult to revoke them, and it is impossible to restrict their permissions.
- **Use Groups to Assign Permissions to RAM Users:** Instead of defining permissions for individual RAM users, it's usually more convenient to create groups that relate to job functions (administrators, developers, accounting, etc.). Next, define the relevant permissions for each group. Finally, assign RAM users to those groups. All the users in an RAM group inherit the permissions assigned to the group. That way, you can make

changes for everyone in a group in just one place. As people move around in your company, you can simply change what RAM group their RAM user belongs to.

- **Grant Least Privilege:** When you create RAM policies, follow the standard security advice of granting least privilege, or granting only the permissions required to perform a task. Determine what users (and roles) need to do and then craft policies that allow them to perform only those tasks. Start with a minimum set of permissions and grant additional permissions as necessary. Doing so is more secure than starting with permissions that are too lenient and then trying to tighten them later. Alibaba Cloud RAM service comes with a set of pre-defined policies which are called *System Policies*. You can find them in your cloud account portal here: https://ram.console.aliyun.com/policies Usually, for each service there are two policies defined: One that gives full access, and one that only gives read access to any resource in any region. While this may work for some scenarios you might need more fine-granular access policies that are specific to certain resource, to a certain region, or to some specific actions. For these scenarios you can define a so-called *Custom Policy*. Some built-in system policies you might want to consider (in potentially modified versions) for your account governance are:

  - `AliyunSupportFullAccess` which grants access to Support Center via Management Console which includes filing support tickets.
  - `AliyunBSSFullAccess` which grants full access to Billing System via Management Console.
  - `AliyunBSSReadOnlyAccess` read-only access to Billing System via Management Console.
  - `AliyunBSSOrderAccess` which grants permission to view, pay, and cancel orders on Billing System.
  - `AliyunSTSAssumeRoleAccess` which grants access to the API AssumeRole of Security Token Service(STS).
  - `AliyunRAMFullAccess` which grants full access to the RAM service which includes rights to create, modify and delete users, and specify account policies such as password policies.
  - `AliyunNotificationsFullAccess` which grants full access to Alibaba Message Center via Management Console.
  - `AliyunMarketplaceFullAccess` which grants full access to Alibaba Cloud Marketplace via Management Console.
  - `AliyunBeianFullAccess` which grants full access to the Alibaba Cloud ICP Filing system at https://beian.aliyun.com

- **Configure a Strong Password Policy for Your RAM-Users:** If you allow users to change their own passwords, require that they create strong passwords and that they rotate their passwords periodically. On RAM Settings page of the Alibaba Cloud portal at https://ram.console.aliyun.com/settings, you can create a password policy for your account. You can use the password policy to define password requirements, such as minimum length, whether it requires non-alphabetic characters, how frequently it must be rotated, password history checks, and so on.
- **Enable MFA** For extra security, we recommend that you require multi-factor authentication (MFA) for all users in your account. With MFA, users have a device that generates a response to an authentication challenge. Both the user's credentials and the device-generated response are required to complete the sign-in process. If a user's password or access keys are compromised, your account resources are still secure because of the additional authentication requirement. Please check https://www.alibabacloud.com/help/doc-detail/119555.htm for details.
- **Use Service-Roles for Applications That Run on Alibaba Cloud Services:** Applications that run on Alibaba Cloud services such as ECS instances or Function Compute need credentials in order to access other Alibaba Cloud services. To provide credentials to the application in a secure way, use RAM service-roles. A service-role is an entity that has its own set of permissions, but that isn't a user or group. Roles also don't have their own permanent set of credentials the way RAM users do. In the case of Function Compute for example, RAM dynamically provides temporary credentials to the Function Compute instance, and these credentials are automatically rotated for you. When you launch an ECS instance, you can specify a role for the instance as a launch parameter. Applications that run on the ECS instance can use the role's credentials when they access other cloud resources through Secure Token Service (https://www.alibabacloud.com/help/doc-detail/28756.htm). The role's permissions determine what the application is allowed to do.

- **Use Roles to Delegate Permissions:** Don't share security credentials between accounts to allow users from another Alibaba Cloud account to access resources in your cloud account. Instead, use RAM roles. You can define a role that specifies what permissions the RAM users in the other account are allowed. You can also designate which Alibaba Cloud accounts have the RAM users that are allowed to assume the role. See https://www.alibabacloud.com/help/doc-detail/93745.htm for details on how to configure accordingly.
- **Do Not Share Access Keys:** Access keys provide programmatic access to Alibaba Cloud. Do not embed access keys within unencrypted code or share these security credentials between users in your Alibaba Cloud account. For applications that need access to Alibaba Cloud, configure the program to retrieve temporary security credentials using a RAM role. To allow your users individual programmatic access, create a RAM user with personal access keys.
- **Rotate Credentials Regularly:** Change your own passwords and access keys regularly, and make sure that all RAM users in your account do as well. That way, if a password or access key is compromised without your knowledge, you limit how long the credentials can be used to access your resources. You can apply a password policy to your account to require all your RAM users to rotate their passwords. You can also choose how often they must do so.
- **Use Policy Conditions for Extra Security:** To the extent that it's practical, define the conditions under which your RAM policies allow access to a resource. For example, you can write conditions to specify a range of allowable IP addresses that a request must come from. You can also specify that a request is allowed only within a specified date range or time range. You can also set conditions that require the use of MFA (multi-factor authentication). For example, you can require that a user has authenticated with an MFA device in order to be allowed to terminate an ECS instance. Please see https://www.alibabacloud.com/help/doc-detail/100680.htm#h2-url-8 for details on the syntax of the `Condition` field of an according RAM policy.
- **Enable ActionTrail:** You can use the Alibaba Cloud service *ActionTrail* to determine the actions users have taken in your account and the resources that were used. The log files show the time and date of actions, the source IP for an action, which actions failed due to inadequate permissions, and more. Keep in mind that these logs are based exclusively on OpenAPI calls, i.e., the management APIs of Alibaba Cloud. Application specific logs can be recorded and analyzed with *Log Service*, for example.

LINKS

- The Official RAM documentation at https://www.alibabacloud.com/help/product/28625.htm which discusses many aspects of this chapter in more detail and also comes with a Tutorial section that focuses on many common scenarios by giving concrete examples.
- The official ActionTrail documentation at https://www.alibabacloud.com/help/product/28802.htm which explores various scenarios such as security analysis, resource change tracking, and compliance audit.
- The official Log Service documentation at https://www.alibabacloud.com/help/product/28958.htm which discusses how to collect, store and analyze logs from your applications.

BILLING MANAGEMENT

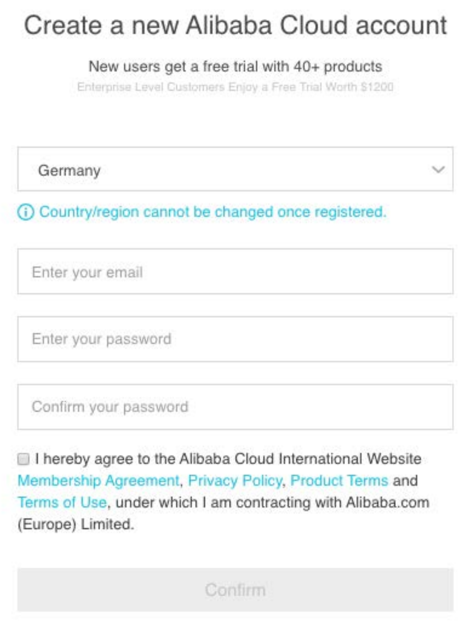
Alibaba Cloud Billing Management is the service that you use to pay your Alibaba Cloud bill, monitor your usage, and budget your costs.

Alibaba Cloud automatically charges the credit card you provided when you signed up for a new account with Alibaba Cloud. Charges appear on your credit card bill monthly. You can view or update credit card information, and designate a different credit card for Alibaba Cloud to charge, on the Payment Methods page in the Billing Management console. Alibaba Cloud also supports the payment ia credit lines which is a common option for enterprise customers.

Depending on the region you choose during initial account creation, you are contracting with a different legal identity of Alibaba Cloud. Outside of Mainland China, we provide the following legal entities for contracting:

- Alibaba.com (Europe) Limited
- Alibaba Cloud (Singapore) Private Limited
- Alibaba Cloud US LLC
- Alibaba Cloud (India) LLP
- Alibaba Cloud (Malaysia) Sdn. Bhd

Be careful in choosing the region. It cannot be changed afterwards and your contracting legal entity depends on it.



Create a new Alibaba Cloud account

New users get a free trial with 40+ products

Enterprise Level Customers Enjoy a Free Trial Worth $1200

Germany

ⓘ Country/region cannot be changed once registered.

Enter your email

Enter your password

Confirm your password

□ I hereby agree to the Alibaba Cloud International Website Membership Agreement, Privacy Policy, Product Terms and Terms of Use, under which I am contracting with Alibaba.com (Europe) Limited.

Billing Management - Billing Address

Payment Methods

If you choose U.S. Dollars as payment currency, we support payment by credit card, debit card, or PayPal. In addition, we also support payment by invoice which needs to be explicitly activated by your responsible Alibaba Cloud business manager. If you choose Indian Rupee as payment currency, we support payment by Paytm Wallet. If you choose Malaysian Ringgit as payment currency, we support payment by credit card or debit card. Each account may have multiple payment methods registered at the same time, but there can only be one default payment method to be used for all your payments. For example, if you have used a credit card to pay for a prepaid service, this credit card will be your default payment method and you can only use it to make other purchases. You can only use another registered payment method to make payments if you do not have any other existing products or services currently being billed to another payment method in effect.

Payment Failure

If, for reasons attributed to you or your registered payment method, we cannot bill you or otherwise process your payment, we will notify you by email to your registered email address and request you to resolve the problem. In this case, you will be able to continue to use your products for another 15 days. After this 15-day period, if the issue has not been resolved and payment has not been made, Alibaba Cloud shall have the right, to suspend your service until payment has been processed or to terminate your services without any liability to you. Prior to any service suspension or termination, an email will be sent to your registered email address.

After the termination of your services, Alibaba Cloud shall have the right, to release all your instances without any liability to you 15 days after the date of termination of your services.

INTERACTING WITH ALIBABA CLOUD

Alibaba Cloud provides different means to interact programmatically and to debug and troubleshoot your system. Before we look at more detail at the various options let's first discuss the general API model of Alibaba Cloud since this is the very fundament on which every supporting technology is based on.

API MODEL

Alibaba Cloud provides a web-based management API to manage the entire lifecycle and configuration of the various cloud resources available on our platform such as ECS and VPC. These management APIs are referred to as **OpenAPI**. They are only routable with public internet access. In addition there is also the service specific endpoint which let you use the actual service's functionality such as executing a SQL query or writing data to an OSS bucket. This endpoint usually differs from the OpenAPI endpoint and is provided as an public endpoint routable from the public internet, and an internal endpoint which is only routable from a VPC. Please refer to the documentation of that particular service or look at the web-console of a provisioned service to get the endpoint schema.

Alibaba Cloud performs authentication on each access request. Therefore, each request, whether being sent by HTTP or HTTPS, must contain signature information. Every service performs symmetric encryption using the `Access Key ID` and `Access Key Secret` to authenticate the client's request. Both keys are issued by the service *Resource Access Management Service (RAM)*. The `Access Key ID` indicates the identity of the client. The `Access Key Secret` is the secret key used to both encrypt and verify the signature string on the client side and on the server, respectively. A detailed example based on RDS can be found here: https://www.alibabacloud.com/help/doc-detail/26225.htm

At the time of this writing, Alibaba Cloud APIs are either exposed as RPC-style endpoints or REST-based endpoints. Which style is used depends on the service and can be looked up at the API documentation of the particular service. ECS for example uses RPC-style, while Container Registry uses REST-based style.

- RPC-based APIs have the following format: https://Endpoint/?Action=xx&Parameters So if you need to query one or more VPCs, the action is `DescribeVpcs`
- REST-based APIs follow the (surprise!) REST principles, e.g. GET https://Endpoint/resource/

The endpoint is usually structured like this: `{service abbreviation}.{regionId}.aliyuncs.com` but there are some exceptions to that. For instance the Object Storage Service is structured like this `mybucket.oss-{regionId}.aliyuncs.com`

For each service there is always a default endpoint `{service abbreviation}.aliyuncs.com`, and a region specific endpoint as described above. From a functional perspective it does not matter which one you are going to call, but network-latency wise it does. So make sure to always call the endpoint in the region which is closest to your system, and avoid calling the standard endpoint since this is usually located in Singapore region or in some rare cases in Chinese regions which come with high latency and an unreliable network connection.

The endpoints we just discussed are exposed to the public internet. In many scenarios, however, communication happens from inside a Virtual Private Network (VPC) with no outbound (inbound) internet access for security and isolation reasons. Calling these public endpoints in such environments with no outbound internet access is thus not possible.

Both HTTP and HTTPS are supported for all endpoints. We recommend to send requests over HTTPS for a higher level of security.

Now that we have taken a look at how Alibaba Cloud Management APIs work let's discuss important technologies and methodologies that built upon that.

USING THE COMMAND-LINE INTERFACE

The Alibaba Cloud CLI is a tool to manage and use Alibaba Cloud resources through a command line interface. It is written in Go and built on the top of Alibaba Cloud OpenAPI. It is developed and hosted on Github at https://github.com/aliyun/aliyun-cli. CLI access the Alibaba Cloud services through OpenAPI (see previous section). Before using Alibaba Cloud CLI, make sure that you have activated the service (see chapter 2) to use and known how to use OpenAPI. Also make sure that your machine has public internet access (e.g. through a NAT-Gateway, Elastic IP/Instance-bound IP), otherwise OpenAPI cannot be called.

Alibaba Cloud CLI provides an interactive configuration experience by running

```
$ aliyun configure
Configuring profile 'default' ...
Aliyun Access Key ID [None]: <Your AccessKey ID>
Aliyun Access Key Secret [None]: <Your AccessKey Secret>
Default Region Id [None]: eu-central-1
Default output format [json]: json
Default Language [en]: en
```

This command will create the file `$HOME/.aliyun/config.json` which you can also manually create, of course. It is structured like this:

```
{
   "current": "default",
   "profiles": [
          {
                  "name": "default",
                  "mode": "AK",
                  "access_key_id": "LTAI4FxdfaoqTCJWKi******",
                  "access_key_secret":
"Hv1PvFJHiYQKbES6wB8jrFIW******",
                  "sts_token": "",
                  "ram_role_name": "",
                  "ram_role_arn": "",
                  "ram_session_name": "",
                  "private_key": "",
                  "key_pair_name": "",
                  "expired_seconds": 0,
                  "verified": "",
                  "region_id": "eu-central-1",
                  "output_format": "json",
                  "language": "en",
                  "site": "",
                  "retry_timeout": 0,
                  "retry_count": 0
          }],
   "meta_path": ""
}
```

You can add multiple profiles with different configurations. Switching between profiles is as easy as

```
aliyun configure set --profile <profile-name>
```

Environment variables always take precedence over the configuration file. The following variables are considered:

```
ALIBABACLOUD_REGION_ID > ALICLOUD_REGION_ID > REGION
ALIBABACLOUD_ACCESS_KEY_ID > ALICLOUD_ACCESS_KEY_ID > ACCESS_
KEY_ID
ALIBABACLOUD_ACCESS_KEY_SECRET > ALICLOUD_ACCESS_KEY_SECRET >
ACCESS_KEY_SECRET
```

Note that your credentials are stored in plain-text so make sure that access to this directory is properly secured and use it in combination with the open-source tool Alicloud-Vault. So especially when running on ECS instances we do not recommend to use the Access Key mode (AK) of Alibaba Cloud CLI. Instead, we recommend to use the mode `EcsRamRole` which instructs the Alibaba Cloud CLI to get an STS-token from the metadata service and thus assume the role that was assigned to that particular ECS instance. Assuming that the role named *ecs_role* is to be assumed you would configure the CLI interactively as follows:

```
$ aliyun configure --mode EcsRamRole --profile myprofile
```

See https://github.com/aliyun/aliyun-cli#configure-authentication-methods for details.

Last but not least the Alibaba Cloud CLI command and usage structure is as follows:

```
$ aliyun <product> <api> [--parameter1 value1 --parameter2
value2 ...]
```

Putting `help` after either the *product* or *api* gives you detailed information about the available actions and required and optional parameters. A simple

```
$ aliyun help
```

will print out all available services.

Alibaba Cloud take the following precedence.

ALICLOUD-VAULT

Alicloud-Vault is another handy tool for many development scenarios. It is open-source and developed at https://github.com/alibabacloud-de/alicloud-vault. It is a vault for securely storing and accessing Alibaba Cloud credentials in development environments and takes many inspirations from https://github.com/99designs/aws-vault.

Alicloud-Vault stores RAM credentials in your operating system's secure keystore and then generates temporary credentials from those to expose to your shell and applications. It's designed to be complementary to the Alibaba Cloud CLI tools, and is aware of your config file and profiles in `~/.aliyun/config`.

It uses Alibaba Cloud's STS service to generate temporary credentials via the AssumeRole API call. These expire in a short period of time, so the risk of leaking credentials is reduced. Note that not all services support STS. You can find the currently supported services here: https://www.alibabacloud.com/help/doc-detail/135527.htm

Alicloud-Vault then exposes the temporary credentials to the sub-process through environment variables in the following way

```
$ alicloud-vault exec jonsmith -- env | grep ALICLOUD
ALICLOUD_VAULT=jonsmith
ALICLOUD_REGION_ID=us-east-1
ALICLOUD_ACCESS_KEY_ID=%%%
ALICLOUD_ACCESS_KEY_SECRET=%%%
ALICLOUD_STS_TOKEN=%%%
ALICLOUD_SESSION_EXPIRATION=2020-03-06T10:02:33Z
```

Please refer to the official site at https://github.com/alibabacloud-de/alicloud-vault for detailed information on how to use it and configure it properly.

USING THE OSSUTIL CLI

While `Alibaba Cloud CLI` gives you great control over a wide array of services for OSS Alibaba Cloud is providing a dedicated command line utility called `ossutil` which is hosted on Github at https://github.com/aliyun/ossutil. It provides *both* convenient access to the *OpenAPI* and fine-grained control to the *Service API* of OSS which allows you to generate signed URLs, disabling CRC64 during data transmission, etc.

The configuration file is stored in `$HOME/.ossutilconfig`. If you need to store in a different place you can so by using the `-c` option with which you can specify a custom configuration file path. A configuration file has the following structure:

```
[Credentials]
        language = EN
        endpoint = oss.aliyuncs.com
        accessKeyID = your_key_id
        accessKeySecret = your_key_secret
        stsToken = your_sts_token
[Bucket-Endpoint]
        bucket1 = endpoint1
        bucket2 = endpoint2
        ...
[Bucket-Cname]
        bucket1 = cname1
        bucket2 = cname2
        ...
[AkService]
        ecsAk=http://100.100.100.200/latest/meta-data/Ram/security-
credentials/<your RAM role name>
```

Let's break down this structure:

- The **Credentials** section define the default settings that should be used for every call unless otherwise specified (see below sections). The *endpoint* option overrides the default OSS endpoint. By default, the Internet address `oss.aliyuncs.com` directs to the Internet endpoint of China East 1 (Hangzhou), and the intranet address `oss-internal.aliyuncs.com` directs to the intranet endpoint of China East 1 (Hangzhou).
- The **Bucket-Endpoint** section lets you define endpoints for specific buckets. This is needed if the location of your bucket differs from your default endpoint you have specified in the Credentials section (or from the default endpoint if you haven't specified an endpoint at all).
- The **Bucket-Cname** section lets you define CNAMEs for your bucket endpoints. This is especially usefull and needed if you are exposing your buckets through a Content Delivery Network (CDN) service which you like to use in combination with `ossutil`, for example.
- The **AkService** section is required if you need to use a RAM role bound to an ECS instance to perform operations on OSS. When you configure this option, you only need to set *ecsAK* to the RAM role bound to the ECS instance. After configuring the AkService option, you do not need to configure the accessKeyID, accessKeySecret, and stsToken options. If these options are configured, the configurations of these options instead of the AkService option take effect.

The priority of endpoint selection is defined as follows: `--endpoint > Bucket-Cname > Bucket-Endpoint > endpoint > default OSS endpoint`` That is, CLI option takes precedence over Bucket-Cname takes precedence over Bucket-Endpoint, ...,

INFRASTRUCTURE AS CODE

Wikipedia defines Infrastructure as Code (IaC) as follows:

```
Infrastructure as code is the process of managing and
provisioning computer data centers through machine-readable
definition files, rather than physical hardware configuration or
interactive configuration tools.
```

Or to put it in simpler terms:

```
Infrastructure as Code (IaC) means to manage the entire
lifecycle and configuration of your IT infrastructure using
(declarative) configuration files.
```

With IaC, your infrastructure's configuration takes the form of a code file which uses a declarative description. *Declarative* means you are solely defining the desired state, not how to get to this state. This is usually taken care of by the IoC-Technology such as Terraform and thus much less error-prone and readable and understandable. Since it's just text, it's easy for you to edit, copy, and distribute it. You can *and should* put it under source control, like any other source code file.

Benefits

Let us now dive into some of the benefits your organization will reap by adopting an IaC solution in combination with cloud-based technologies.

- **Speed** The first significant benefit IaC provides is speed. Infrastructure as code enables you to quickly set up your complete infrastructure by running a script. You can do that for every environment, from development to production, passing through staging, QA, and more. IaC can make the entire software development lifecycle more efficient.
- **Consistency** Manual processes result in mistakes, period. Humans are fallible. Our memories fault us. Communication is hard, and we are in general pretty bad at it. As you've read, manual infrastructure management will result in discrepancies, no matter how hard you try. IaC solves that problem by having the config files themselves be the single source of truth. That way, you guarantee the same configurations will be deployed over and over, without discrepancies.
- **Accountability** This one is quick and easy. Since you can version IaC configuration files like any source code file, you have full traceability of the changes each configuration suffered. No more guessing games about who did what and when.
- **Higher Efficiency** By employing infrastructure as code, you can deploy your infrastructure architectures in many stages. That makes the whole software development life cycle more efficient, raising the team's productivity to new levels. You could have programmers using IaC to create and launch sandbox environments, allowing them to develop in isolation safely. The same would be true for QA professionals, who can have perfect copies of the production environments in which to run their tests. Finally, when it's deployment time, you can push both infrastructure and code to production in one step.
- **Lower Cost** One of the main benefits of IaC is, without a doubt, lowering the costs of infrastructure management. By employing cloud computing along with IaC, you dramatically reduce your costs. That's because you won't have to spend money on

hardware and build or rent physical space to store it. But IaC also lowers your costs in another, subtler way, and that is what we call "opportunity cost."

Every time you have smart, high-paid professionals performing tasks that you could automate, you're wasting money. All of their focus should be on tasks that bring more value to the organization. And that's where automation strategies—infrastructure as code among them—come in handy. By employing them, you free engineers from performing manual, slow, error-prone tasks so they can focus on what matters the most.

Terraform

Terraform is a tool for building, changing, and versioning infrastructure safely and efficiently. Terraform can manage existing and popular service providers as well as custom in-house solutions. It is hosted at https://www.terraform.io/

Configuration files describe to Terraform the components needed to run a single application or your entire datacenter. Terraform generates an execution plan describing what it will do to reach the desired state, and then executes it to build the described infrastructure. As the configuration changes, Terraform is able to determine what changed and create incremental execution plans which can be applied.

The infrastructure Terraform can manage includes low-level components such as compute instances, storage, and networking, as well as high-level components such as DNS entries, SaaS features, etc.

Terraform is extensible and thus supports a wide range of different cloud vendors. The official Alibaba Cloud Terraform code-repository is hosted at https://github.com/terraform-providers/terraform-provider-alicloud. The complete documentation can be found at https://www.terraform.io/docs/providers/alicloud/. So while you cannot re-use the same template code across different cloud providers you can re-use your general Terraform knowledge.

We highly recommend to look at the official examples provided at https://github.com/terraform-providers/terraform-provider-alicloud/tree/master/examples to get started easily. For example, the following script defines the desired state of a VPC with a certain network range you would like to create:

```
provider "alicloud" {
}

resource "alicloud_vpc" "main" {
  name       = "my-vpc"
  cidr_block = "192.168.0.0/16"
}
```

The Terraform CLI then provides a huge set of options to manage the entire state and lifecycle of your resources and configurations which are described in your template code.

Resource Orchestration Service

Resource Orchestration Service (ROS) is Alibaba Cloud's native IoC offering. It provides support for the most important services and features of Alibaba Cloud and will extend support even further in the near future. Just like with Terraform you build a template and use it to create all of the necessary resources, collectively known as a ROS stack. This model removes opportunities for manual error, increases efficiency, and ensures consistent configurations over time. You can either store check out your ROS Templates locally to your machine or upload them to Alibaba Cloud Object Storage Service (OSS) and then use ROS via the browser console, *Alibaba Cloud CLI*, or APIs to create a stack based on your template code. ROS aims to provide the best possible support for Alibaba Cloud services and features. Its application is, however, solely bound to Alibaba Cloud. Knowledge cannot be re-used across different cloud providers, albeit that it shares many concepts and its syntax with AWS Cloud Formation.

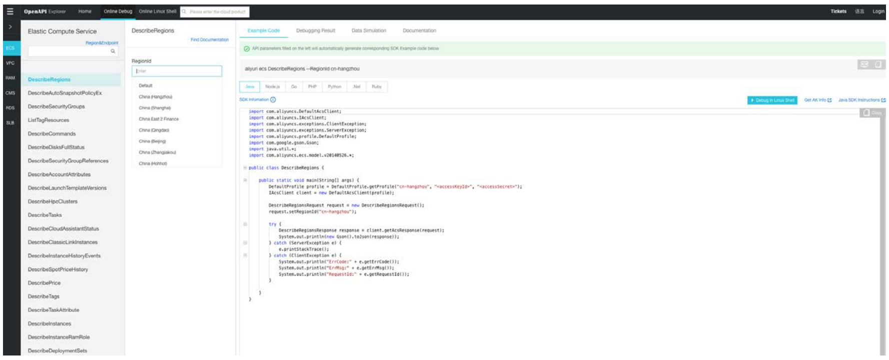
The following ROS script defines the desired state of a VPC with a certain network range:

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "EcsVpc": {
      "Type": "ALIYUN::ECS::VPC",
      "Properties": {
        "VpcName": "my-vpc",
        "CidrBlock": "192.168.0.0/16"
      }
    }
  },
}
```

The Alibaba Cloud CLI lets you then manage the entire state and lifecycle of your resources and configurations which are described in your template code. You will find detailed information on how to use the *Alibaba Cloud CLI* with ROS here: https://www.alibabacloud.com/help/doc-detail/137399.htm

DEBUGGING WITH API-EXPLORER

Sometimes you want to be able to call and explore the OpenAPI easily using a convenient graphical interface. This is where OpenAPI Explorere comes into play which you can find at https://api.aliyun.com. Make sure to have an active login session running in your browser since the OpenAPI Explorer will use the Access Key of the current user to temporarily access and operate on the user's resources. This means that permission-wise you are of course restricted to your assigned permissions. Below figure gives you an idea on how OpenAPI Explorer looks like.



API parameters filled on the left will automatically generate corresponding SDK Example code below

```
aliyun ecs DescribeRegions --RegionId cn-hangzhou
```

```
import com.aliyuncs.DefaultAcsClient;
import com.aliyuncs.IAcsClient;
import com.aliyuncs.exceptions.ClientException;
import com.aliyuncs.exceptions.ServerException;
import com.aliyuncs.profile.DefaultProfile;
import com.google.gson.Gson;
import java.util.*;
import com.aliyuncs.ecs.model.v20140526.*;

public class DescribeRegions {

    public static void main(String[] args) {
        DefaultProfile profile = DefaultProfile.getProfile("cn-hangzhou", "<accessKeyId>", "<accessSecret>");
        IAcsClient client = new DefaultAcsClient(profile);

        DescribeRegionsRequest request = new DescribeRegionsRequest();
        request.setRegionId("cn-hangzhou");

        try {
            DescribeRegionsResponse response = client.getAcsResponse(request);
            System.out.println(new Gson().toJson(response));
        } catch (ServerException e) {
            e.printStackTrace();
        } catch (ClientException e) {
            System.out.println("ErrCode:" + e.getErrCode());
            System.out.println("ErrMsg:" + e.getErrMsg());
            System.out.println("RequestId:" + e.getRequestId());
        }

    }
}
```

Interacting - OpenAPI Explorer

As you can see you explore and search for specific actions and immediately get a graphical interface which lets you conveniently specify all parameters. It also gives you a nice graphical representation of the resulting request and response values, which of course is great for debugging and troubleshooting. What is also very practical is the *Example Code* section for different programming languages such as Java, NodeJS, Go, etc. that instantly generates the code that will create and submit the according action you have selected.

One of the highlights is the Data Simulation feature. It lets you automatically create a Mock endpoint with mocked return values. As such it simulates the results of a real OpenAPI call, which you can use to replace the real OpenAPI request address in a development environment and lets you simulate different data scenarios of OpenAPI.

The mock endpoints take the following form:

```
https://api.aliyun.com/mock/<service/<action>
```

So for mocking the `DescribeTasks` of ECS service you can simply call

```
https://api.aliyun.com/mock/Ecs/DescribeTasks
```

which returns according mock data. No authentication and signing of the request is needed for calling the mock API.

INFRASTRUCTURE ESSENTIALS

This chapter gives a focused and condensed rundown on the very essentials of the Alibaba Cloud Infrastructure service concepts such as Compute, Network, and Storage, almost like a cheat-sheet.

ELASTIC COMPUTE SERVICE (ECS)

ECS is the infrastructure service by Alibaba Cloud that provides customers compute power as virtual machines. The security and compliance is a shared responsibility between Alibaba Cloud and the customers. Alibaba Cloud is responsible for "Security of the Cloud". That is, it is responsible for protecting the infrastructure that runs all of the services offered in the Alibaba Cloud. This infrastructure is composed of the hardware, software, networking, and facilities that run Alibaba Cloud services. For ECS, Alibaba Cloud's responsibility includes everything up to the hypervisor of the host machines that powers the ECS service. Everything above is the customer's responsibility which includes the guest operating system and all security configuration tasks such as configuration of Alibaba Cloud provided firewall (called a security group) on each network interface of an ECS instance. We recommend to consult our Security Whitepaper for a detailed discussion on this topic. It is freely available at https://www.alibabacloud.com/help/doc-detail/42435.htm

Availability Service Level Agreement (SLA)

ECS comes with a Monthly Single-Instance Availability SLA of 99.975%, and provides a Monthly Multi-Zone Availability SLA of 99.995% if your application is deployed on at least 2 ECS instances spread across two different Availability Zones.

An ECS instance is considered *Unavailable* if the disconnection between an ECS instance configured with access permitted rules and any IP address over TCP or UDP in the inbound and outbound directions lasts for more than one minute.

Instance Families and Instance Types

The ECS service is organized by so-called *instance families* which in turn consist of different *instance types*. An instance family describes the fundamental characteristics and use-cases for instances types of this family. Some are optimized for network-intense applications, others are designed for memory or compute-intense workloads. As such they usually differ in terms of Core-to-RAM ratio, maximum persistent disk IOPS, network performance (as both in bandwidth and PPS), and the type of disks they support. Please consult the official documentation at https://www.alibabacloud.com/help/doc-detail/25378.htm for detailed numbers.

Note that ECS configurations and types can be updated, however, not arbitrarily. For instance family and type changes only certain types are supported (see below documentation link). As a rule of thumb, you can always change instance families between g6, c6, and r6.

As of this writing there exist 13 different instance families on Alibaba Cloud (please consult documentation at https://www.alibabacloud.com/help/doc-detail/108490.htm for details). The character in brackets denotes the abbreviation of the instance family.

- General Purpose (g)
- Compute Optimized (c)
- Memory Optimized (r)
- Big Data (d)
- Local SSDs (i)
- High-Clock Speed (hfc)
- Compute Optimized with GPU (gn, vgn)
- Visualization Compute with GPU (ga)
- Compute Optimized with FPGA (f)
- Bare Metal (ebmg)
- Super Computing Cluster (scc)
- Burstable (t)
- Dedicated Hosts (ddh)

The new generations of Alibaba Cloud x86-based ECS instances are equipped 2.5 GHz Intel ® Xeon ® Platinum 8269CY (Cascade Lake) processors with Turbo Boost up to 3.2 GHz. The newest generation of the General Purpose G family now also support burstable network bandwidth. Also note that the 6th generations of g, c, r families are now running on Alibaba Cloud X-Dragon hypervisor architecture which provides predictable and consistent high performance and reduces virtualization overheads.

Deployment Sets (https://www.alibabacloud.com/help/doc-detail/91258.htm) gives you control on the distribution strategy. You can use a deployment set to distribute your ECS instances to different physical servers to guarantee high availability and set up underlying disaster discovery. When you create ECS instances in a deployment set, Alibaba Cloud will start the instances on different physical servers within the specified region based on your configured deployment strategy. Right now, Alibaba Cloud only provides "High Availability" strategy. As an effect all the ECS instances within your deployment set are strictly distributed across different physical servers within the specified region. The high availability strategy applies to application architectures where several ECS instances need to be isolated from each other. The strategy significantly reduces the chances of services becoming unavailable.

Each and every ECS instance has one operating system disk and can have up to 16 data disks each 32TB in size at max.

Dedicated Hosts (DDH)

ECS instances as discussed in the previous section share the underlying physical servers with different tenants. For scenarios that require strict isolation of the underlying resources DDH is a specialized solution for enterprise customers. DDH provides a dedicated hosting environment for a single tenant based on the virtualization technology of Alibaba Cloud. It offers flexible and scalable services that enable you to enjoy exclusive use of all resources provided by a physical server.

A very useful feature for cost-efficiency is the ability to over-provision a DDH. This is only possible with certain DDH instance types such as the `ddh.c6s`, `ddh.g6s`, or `ddh.r6s`. The memory-optimized type ddh.r6s allows you to provision 416 vCPUS on a 52 core machine for example. This way you can fully utilize your compute capacity in case your workloads have a lot of idle time. See https://www.alibabacloud.com/help/doc-detail/68564.htm for a complete list of all DDH instance types and their respective specifications.

In case the physical machine is malfunctioning failover to a healthy instance is managed automatically by Alibaba Cloud. A new healthy DDH will be assigned automatically from the shared pool to the your account, and after the migration is done, the crashed DDH will be removed from your account. The DDH ID will remain the same after the migration, the machine ID will be different, though. Note that automatic failover is not supported for DDHs with local storage (i.e. `ddh.i2`).

System Events and Live Migration

In previous sections we already talked about the *Shared Responsibility Model* of Alibaba Cloud. Sometimes, maintenance activities on our services such as ECS executed by Alibaba Cloud also affect your system. Thus, in order to react to and handle such kinds of events gracefully you need a way to get notified and possibly define appropriate actions. Please welcome, ECS System Events.

A system event is a scheduled and recorded maintenance event of ECS service. System events occur when updates, invalid operations, unexpected system failures, or unexpected hardware or software failures are detected on your ECS instance. Moreover, you will receive notification about the details of the event in the console when it occurs, including the event response plan and event cycle.

When an ECS user receives a notification from Alibaba Cloud, he or she can acknowledge the planned underlying maintenance for ECS instance by system event. The user can then choose the appropriate time window to execute the system event as well as operation activities according to individual business needs. By providing users this flexibility, users can reduce the impact on system reliability and business continuity.

Normally, when there is maintenance activity planned on the physical server, the ECS instance will be live migrated to another server to maintain the health of ECS instance with minimal performance impact. Note, that data stored on local SSDs (also sometimes referred to as ephemeral disks) is no migrated so make sure that your application accounts for that.

The official documentation at https://www.alibabacloud.com/help/doc-detail/66574.htm gives a deep-dive on the various system event types and event statuses, and also explains how to modify scheduled restart times of ECS instances.

For example, to view the all *unsettled events* (i.e. scheduled events that have not yet been executed) you can easily grab them via the CLI as follows:

```
aliyun ecs DescribeInstancesFullStatus --RegionId
<TheRegionId> --InstanceId.1 <YourInstanceId> --output
cols=EventId,EventTypeName
```

STORAGE

This section will give a short overview about the fundamental aspects of storage on Alibaba Cloud. In particular, we will focus on storage services that you can use in combination with ECS, and Object Storage.

ECS Mountable Storage

Generally speaking, there exist two kinds of storage options you can use in combination with ECS instances:

1. Persistent Storage
2. Ephemeral Storage

Persistent Storage keeps your data stored independently from the lifecycle of the ECS instance whereas Ephemeral Storage is closely bound to the lifecycle of your ECS instance. That is, data stored on persistent storage won't get lost in cases of ECS restarts or hot migration, or when you delete an ECS instance. It exists independently and can be also mounted easily between different ECS instances. Ephemeral Storage on the other hand will get lost in case of ECS reboots, hot migration (to be more precise it is not guaranteed to be kept) and ECS deletion. Don't use it for long-term storage. Usually, it is a lot faster since this kind of storage is directly attached to the physical host system whereas persistent storage lives in a different dedicated cluster. So every access needs to cross the network.

For *Persistent Storage* Alibaba Cloud offers the following options: **Cloud Disk** Cloud Disk is a high-performance, low latency block storage service for Alibaba Cloud ECS. It supports random or sequential read and write operations. Block Storage is similar to a physical disk. You can format a Block Storage device and create a file system on it to meet the data storage needs of your business. One Cloud Disk can be mounted to *at most* one ECS instance. The ECS instance and the Cloud Disk need to be in the same availability zone. Cross-AZ communication is not supported.

The disks come in three tiers: *Enhanced SSD*, *Standard SSD*, and *Ultra Disk*. Basic Disks are no longer for purchase but only exist for backwards compatibility reasons. Each of these tiers share the same maximum capacity of 32.768 GB and data reliability of 99.9999999%. They differ, however, in maximum IOPS, maximum throughput, and single-channel random write access latency. The maximum IOPS of *Enhanced SSD* is 1 million and the maximum throughput 4000 MB/s with an access latency of around 0.2 ms making it one of the most performant persistent disk options out of every cloud provider. Please consult https://www.alibabacloud.com/help/doc-detail/25382.htm for details.

Each piece of data is also replicated three times across the block storage cluster in the same availability zone to ensure a data reliability of 99.9999999% during read and write operations. Thus, any extra redundancy mechanism such as RAID 1 is not recommended.

Cloud Disk also supports encryption based on the industry standard AES-256 and directly integrates with Alibaba Cloud Key Management Service (KMS). Both system and data disks are supported. The key management infrastructure of Alibaba Cloud conforms to the recommendations in (NIST) 800-57 and uses cryptographic algorithms that comply with the (FIPS) 140-2 standard. When encryption is enabled every snapshot you take is also encrypted. Please note that currently, key rotation is not natively supported.

Shared Cloud Disk Shared Block Storage is a block-level data storage service that features high concurrency, high performance, and high reliability. It supports concurrent reads and writes on multiple ECS instances, and provides data reliability of up to 99.9999999%. In a traditional cluster architecture, multiple computing nodes access the same copy of data to provide services. To prevent service disruptions due to single point of failures, you can use Shared Block Storage to ensure access to the data, achieving high availability. We recommend that you store business data in Shared Block Storage devices and use a cluster file system such as General Parallel File System (GPFS) to manage these devices. Data consistency can be guaranteed between multiple front-end computing nodes during concurrent read/write operations.

Note that a single Shared Block Storage device can be attached to a maximum of eight ECS instances in the same zone and region at the same time. Cross-AZ Shared Cloud Disks are not supported.

Network Attached Storage (NAS) NAS is a distributed file system that features shared access, scalability, high availability, and high performance. Based on POSIX file APIs, NAS is compatible with native operating systems. This ensures data consistency and exclusive locks during shared access. It provides data reliability of 99.999999999% (eleven nines).

NAS supports standard protocols, such as NFS V3.0 and NFS v4.0. and SMB 2.1 and later versions, with corresponding support for Windows 7, Windows Server 2008 R2 and all later versions of Windows, but does not support Windows Vista, Windows Server 2008 and earlier versions. NAS provides data consistency and file locking based on POSIX file APIs.

NAS comes in three different service-tiers: *NAS Capacity*, *NAS Performance*, and *NAS Extreme*. They are suited for different workloads and differ in terms of IOPS, latency, and throughput. While *NAS Capacity* supports volume sizes as big as 10PB with a relatively high latency of 10ms, *NAS Performance* optimizes for throughput and latency but has a reduced volume size of 1 PB at max. See https://www.alibabacloud.com/help/doc-detail/61136.htm for details.

For *Ephemeral Storage* Alibaba Cloud offers **Local Disks**.

Local disks are disks that are attached to the same physical machine that hosts their ECS instance. Local disks provide local storage and access for ECS instances. Local disks are cost-effective and provide high random IOPS, high throughput, and low latency. As discussed previously it is not suited for long-term storage since in case of reboots or hardware failures data may get lost. Data redundancy must be implemented at the application layer by yourself, as well as any encryption.

Note that local disks are only available with certain instance families. These include `i2, i2g, i2ne, i2gne, gn5`, and `ga1`. Please consult https://www.alibabacloud.com/help/doc-detail/63138.htm for a detailed overview of the performance characteristics.

OSS

The storage options in the previous section where all members of the so-called block-level storage which supports random read/write patterns making it suitable for any kind of computing where you need fast and efficient access. Alibaba Cloud Object Storage Service (OSS) in contrast is a so-called Object Storage. Files are not split into evenly sized blocks of data but organized in a flat object hierarchy that is composed of the content, meta data, and a globally unique identifier. Random read / write patterns are not supported. It scales very well, though, is much cheaper than block-level storage and particularly suited to store large amounts (PB-level) amounts of data for batch analyses in the area of AI and Big Data.

Data is organized in so called buckets which is a FQDN and globally unique. A bucket has unlimited capacity. A single object can be at most 48,8 TB in size. OSS provides strong consistency per default. That is after you have created or updated an object every read operation will *always* get the most recent version.

It also supports asynchronous cross-region replication. It uses our own internal network. So data is not replicated across the public internet. There is no dedicated bandwidth, however. If you need dedicated bandwidth and predictable data copy duration we recommend to use Cloud Enterprise Network (https://www.alibabacloud.com/help/product/59006.htm) which lets you configure a dedicated bandwidth between 2 Mbits and 10Gbits.

For migration scenarios where you need to copy over data from other Object Storage vendors such as AWS S3, or Azure Blob Storage, Alibaba Cloud offers the fully managed service *Data Online Migration* (https://www.alibabacloud.com/help/product/94157.htm). It supports a wide array of different third-party object storage services including self-hosted solutions that offer a public HTTPS endpoint.

There are two performance limits which need to be taken into consideration when designing your system based on OSS. The capacities described below are reserved at account level, meaning they are shared between your individual buckets.

- **Bandwidth:** Per default, 10 Gbit/s are reserved for both inbound and outbound in Mainland China regions, whereas 5 Gibt/s are reserved outside of Mainland China regions. These limits are soft-limits. That is they can be increased by opening an according ticket. For the public endpoint the bandwidth is mainly limited by the local bandwidth of the client and the quality of the network provided by operators. So if possible it is recommended to use the internal endpoint if possible.
- **Operations:** OSS can sustain 2000 operations per second per partition (downloading, uploading, deleting, copying, and obtaining metadata are each counted as one operation, while deleting or enumerating more than one files in batch is considered as multiple operations). OSS automatically and constantly partitions your data into up to 65,536 partitions based on the prefix of your filename. So make sure to follow according best-practices outlined at https://www.alibabacloud.com/help/doc-detail/64945.htm when defining a naming schema for your object names.

So if you are experiencing performance bottlenecks make sure to look at both aspects when troubleshooting your system.

NETWORK PERFORMANCE

Let's quickly define *outbound* and *inbound* traffic:

- Inbound refers to network traffic that is sent from the public internet to any Alibaba Cloud service (i.e. traffic flows *into* the cloud)
- Outbound refers to network traffic that is sent from any Alibaba Cloud service to the public internet (i.e. traffic *leaving* the cloud)

ECS - External Performance

Inbound traffic is at *minimum* 100MBits. It will be at *most* as high as EIP Bandwidth.

Outbound traffic is capped by the EIP bandwidth. Bandwidths greater or equal 1Gbits can only be saturated by multiple threads.

Note that the maximum default EIP bandwidth is 200 Mbits, the maximum instance-bound public IP bandwidth is 100 Mbits.

In order to increase that you have to add your EIPs (no instance-bound public IPs are supported) to a shared internet bandwidth package which can be as high as 1Gbits (see https://www.alibabacloud.com/help/doc-detail/55784.htm for details). There are no additional costs for a shared bandwidth internet package. This way you can increase you outbound bandwidth to up to 1 Gbits. This bandwidth can only be saturated by multiple threads, though. You can also create multiple bandwidth packages of course.

To further increase the external network performance you can also use multiple ENIs (Elastic Network Interfaces) and bind up to 10 EIPs to up to 10 private ip addresses of a single ENI in NAT mode. By assigning multiple bandwidth packages to these EIPs you can further increase the network throughput of a single instance. Please check https://www.alibabacloud.com/help/doc-detail/88991.htm for further details on ENI and the different supported modes such as *Cut-Through mode* and *Multi-EIP to ENI mode*.

ECS - Internal Performance

Both inbound and inbound bandwidth limits of an ECS instance are usually the same and do not differ. They depend on the instance family and type.

For regular instance types they vary between 0.5 Gbits and 25 Gbits.

The new generation of G6 instance types also provide burstable bandwidth capacity which allows the bandwidth to go up as three times as high as the base bandwidth for a short amount of time. See https://www.alibabacloud.com/help/doc-detail/108490.htm for details on the exact numbers.

Super Compute Cluster (SCC) types instances can achieve 2x25 Gbits via RDMA over Converged Ethernet (RoCE).

Service Load Balancer (SLB) Network Performance

The outbound (SLB to Internet) network performance depends on the region the SLB is created. It ranges from 1Gbit/s up to 5 Gbit/s. You can get a detailed breakdown by region at https://www.alibabacloud.com/help/doc-detail/85966.htm

For intranet network performance the limit is always 5 Gbit/s, independently from the region the SLB is running in.

SECURING YOUR SYSTEM

System and application security is a very important and complex topic that spans many different aspects and layers of cloud-based applications. From account security to proper rights management with according auditing, to networking security of the entire application and platform services being used, to securing the software that runs on the cloud. A comprehensive discussion would be a book on its own. This is why we will focus on the most important aspects and leave the reader with further reading suggestions to dive deeper into the topics of interest.

SECURITY CENTER

Security Center is a unified security management system that dynamically identifies and analyzes security threats, and generates alerts when threats are detected. It provides ransomware protection, anti-virus protection, web tamper protection, and compliance assessments to ensure the security of cloud resources and local servers. This allows you to automate security operations, responses, and threat tracing, and meet regulatory compliance requirements.

Security Center meets the standards of ISO 9001, ISO 20000, ISO 22301, ISO 27001, ISO 27017, ISO 27018, ISO 29151, ISO 27701, BS1 0012, CSA STAR, and PCI DSS.

Security Center can be used in many ways that help you protecting your cloud environment and your ECS instances in particular. For one, it allows you to whitelist and blacklist certain IP-ranges which you consider eligible to connect to your publicly available ECS instances. In case there are connection attempts (e.g. via SSH or RDP) from IP-ranges you have not explicitly whitelisted you get automatic alarm notifications (e.g. by email) that list in detail from which IP and region at what time a connection request was established. All of these events are stored persistently and may also be forwarded to SLS for easy analysis.

Another useful feature is that it automatically assesses what kind of software packages and versions are installed on your ECS machines and whether they are currently affected by any publicly known exploits and bugs. It comes with a one-click feature that lets you patch vulnerabilities instantly and in a coordinated way across fleets of ECS instances.

It also enables you to do automated configuration assessments of your services in use: from your general account protection, to your networking environment, up to the configuration of individual database services. In case risks such as unprotected, public endpoints are discovered it will automatically raise an alarm with suggestions helping you to mitigate the risk appropriately and in a timely manner before attackers can exploit it. Misconfigured databases and applications servers are one of the most often found penetration risks for any IT system.

An additional feature which is also well integrated into Alibaba Cloud Container Registry service is the Image Security Scanner.

This service can identify more than 120,000 historical vulnerabilities and detect the latest vulnerabilities. It also provides vulnerability repair solutions to implement one-stop vulnerability management. makes vulnerability fixes easier. It is a One-stop Vulnerability Management covering the entire lifecycle of vulnerability fixes, making vulnerability management easier. It can identify more than 120,000 historical vulnerabilities and also detect the latest vulnerabilities along with suggestions on how to fix them.

Last but not least, Security Center also comes with an anti-virus protection feature which is based on the automated analysis of a massive numbers of virus samples, virus persistence, and attack methods.

Note that only a subset of these features is available in the Basic edition which is free for all Alibaba Cloud assets. Security Center comes in three tiers: Basic, Advanced, and the Enterprise Edition. The pricing model works on a monthly subscription basis charged by the number of protected assets. It can also be used with any kind of external workloads given that they have access to the public internet to be able to communicate with the public Security Center endpoints. For both internal and external workloads a small agent called 'AliyunDun` needs to be installed on the machines. For detailed installation instructions please refer to https://www.alibabacloud.com/help/doc-detail/68611.htm

Please note that there is a kernel bug in combination with KVM: https://bugs.launchpad.net/ubuntu/+source/linux/+bug/1858760 which is exploited by `AliyunDun` (and basically any process that uses loops in combination with process suspension by using usleep() for example.)

Older instance types are affected since they are running on KVM. Our most recent instance types do not run on KVM anymore but on our own hypervisor technology. Thus, they are not affected by this bug.

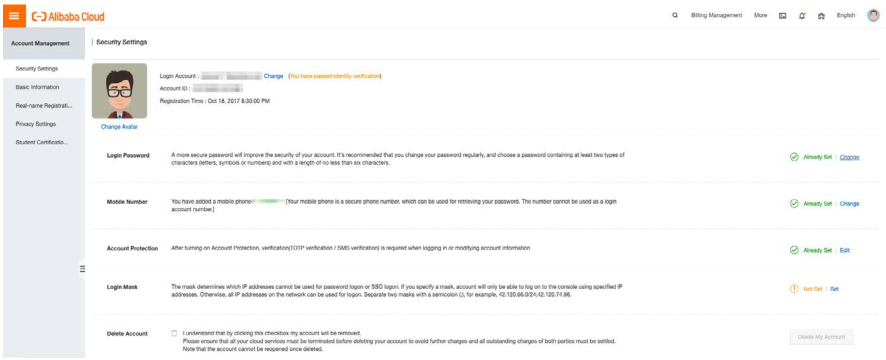
So there are currently two solutions if this bug is affecting your system:

- Either downgrade the Kernel version to <= 4.15.0-70
- Or consistently use the newest generation types of the 6th generation and above (e.g. G6).

ACCOUNT SECURITY

Each cloud account has exactly one and only one root user. You are specifying the login name and password during initial cloud account creation. Each root user also has an associated mobile phone number which can be re-used across different root users and cloud account respectively at most six times.

Below screenshot shows the options of the *Security Settings* administration page which is only available for the root user and accessible directly via https://account-intl.console.aliyun.com.



Security Settings

Login Account : [redacted] Change (You have passed identity verification)

Account ID : [redacted]

Registration Time : Oct 18, 2017 8:30:00 PM

Login Password — A more secure password will improve the security of your account. It's recommended that you change your password regularly, and choose a password containing at least two types of characters (letters, symbols or numbers) and with a length of no less than six characters. — Already Set | Change

Mobile Number — You have added a mobile phone[redacted] [Your mobile phone is a secure phone number, which can be used for retrieving your password. The number cannot be used as a login account number.] — Already Set | Change

Account Protection — After turning on Account Protection, verification(TOTP verification / SMS verification) is required when logging in or modifying account information — Already Set | Edit

Login Mask — The mask determines which IP addresses cannot be used for password logon or SSO logon. If you specify a mask, account will only be able to log on to the console using specified IP addresses. Otherwise, all IP addresses on the network can be used for logon. Separate two masks with a semicolon (;), for example, 42.120.66.0/24;42.120.74.98. — Not Set | Set

Delete Account — □ I understand that by clicking this checkbox my account will be removed.
Please ensure that all your cloud services must be terminated before deleting your account to avoid further charges and all outstanding charges of both parties must be settled.
Note that the account cannot be reopened once deleted. — Delete My Account

Account Management - Security Settings

On this page you can easily change the login password and your mobile phone number. The mobile number is important as it is used as a second authentication factor for changing the password, and also for setting up MFA for your root user. Last but not least you can also define a login mask that lets you explicitly whitelist IP ranges from which (SSO) login is allowed. Per default, all IP ranges are allowed.

Best Practices

1. Enable *Account Protection*, that is MFA, which currently supports Time-based One-time Password (TOTP) and SMS verification
2. Define a *Login Mask* that only allows login from a specific IP range, e.g. the outbound IP range of your corporate network, thus blocking illegal login attempts from unknown IP ranges.
3. Choose a password that is sufficient in both complexity and length, make sure to activate password rotation, and restrict session duration. Consult your security advisor on your company's password and security guidelines.
4. Never activate *Access Key ID* and *Access Key Secret* for your root user. You can check at https://usercenter.console.aliyun.com whether according keys have been defined. Deactivate them immediately since the root account is not recommended for any programmatic use. Think of it as the root user on Linux systems which has universal access rights to each and everything and whose rights cannot be restricted. For the day to day work the root account should never be used at all!
5. Activate ActionTrail to fully audit your account. Please see section Account Auditing of chapter Securing your System for details.

ACCOUNT AUDITING

Account Auditing is the process of recording and reliably storing all events that are based on direct or indirect invocation of the Alibaba Cloud Management APIs. It captures which user executed what API with which payload at a specific point in time. In short, it records and stores *Who* did *What When*.

Alibaba Cloud ActionTrail is a service that monitors and records the actions of your Alibaba Cloud account, including the access to and use of cloud products and services through the Alibaba Cloud console, API operations, and SDKs. ActionTrail records these actions as events. You can download these events from the ActionTrail console or configure ActionTrail to deliver these events to Log Service Logstores or Object Storage Service (OSS) buckets. Then, you can perform behavior analysis, security analysis, resource change tracking, and compliance auditing based on these events.

Per default, ActionTrail stores any such event for up to 90 days. By defining so-called Trails you can automatically store them on LogService and OSS for an extended period of time of many more months or even years. Such trails can be configured with filters based on specific regions and EventTypes (*Write, Read, All*). Multi-

Account Management features of *Resource Directory* are being used, ActionTrail allows to apply such a trail to all member accounts to collect all audit logs at a central place.

NETWORK SECURITY

In this section we will briefly discuss the fundamental serviced and features that allow you to secure your applications network-wise. This includes network isolation, firewall rules, flow logs, but also defense mechanisms against Level-4 and Level-7 attacks such as malformed packet attacks, SYN flooding, DNS request flooding, connection exhaustion attack, but also SQL injections, XSS attacks, etc. by using Alibaba Cloud Anti-DDoS and Web Application Firewall. Let's first look at network isolation and protection with Virtual Private Cloud (VPC) and according best-practices.

Virtual Private Cloud (VPC)

A VPC allows you to define a private logically isolated network on Alibaba Cloud. It defines network range, routing tables, subnets (aka VSwitch in Alibaba Cloud) which are bound to an according availability zone, and also security groups that define inbound and outbound communication rules for your VPC-resources such as ECS instances.

Ideally, you should design your subnets according to your architecture tiers, such as the database tier, the application tier, the business tier, and so on, based on their routing needs, such as public subnets needing a route to the internet gateway. You should also create multiple subnets in as many availability zones as possible to improve your fault-tolerance. Each availability zone should have identically sized subnets, and each of these subnets should use a routing table designed for them depending on their routing need. Distribute your address space evenly across availability zones and keep the reserved space for future expansion.

Always use Elastic IP (EIP) instead of instance-bound IPs for all resources that need to connect to the internet. The EIPs are associated with an Alibaba Cloud account instead of an instance. They can be assigned to an instance in any state, whether the instance is running or whether it is stopped. It persists without an instance so you can have high availability for your application depending on an IP address. The EIP can be reassigned and moved to Elastic Network Interface (ENI) as well. Since these IPs don't change, they can be whitelisted by target resources.

Monitoring is imperative to the security of any network, such as Alibaba Cloud VPC. Enable ActionTrail and VPC Flow Logs to monitor all activities and traffic movement such as allowed and dropped requests. This can be monitored on VPC, VSwitch and network interface level. ActionTrail will record all activities, such as provisioning, configuring, and modifying all VPC components. The VPC flow log will record all the data flowing in and out of the VPC for all the resources in VPC. Additionally, you can set up config rules with the Alibaba Cloud Config service for your VPC for all resources that should not have changes in their configuration.

Be sure to connect these logs and rules with Alibaba Cloud CloudMonitor to notify you of anything that is not expected behavior and control changes within your VPC. Identify irregularities within your network, such as resources receiving unexpected traffic in your VPC, adding instances in the VPC with configuration not approved by your organization, among others.

For every resource you provision or configure in your VPC, follow the least privilege principle. So, if a subnet has resources that do not need to access the internet, it should be a private subnet and should have routing based on this requirement. Similarly, security groups should have rules based on this principle. They should allow access only for traffic required.

In order to keep your VPC and resources in your VPC secure, ensure that most of the resources are inside a private subnet (aka VSwitch) by default. If you have instances that need to communicate with the internet, then you should add a Server Load Balancer (SLB) to a VPC and add all instances behind this SLB in the private subnet. Also, use NAT Gateway to access public networks from your private subnet. Alibaba Cloud NAT gateway is a fully managed, highly available, and redundant component.

Anti-DDoS

Alibaba Cloud provides 4 different offerings:

- **Anti-DDoS Basic**

  If your workloads are running on Alibaba Cloud you'll get a mitigation capacity of 5 Gpbs fo free. Available in any of our 22 cloud regions. Note, that there is no SLA on this mitigation capacity. It is delivered on a best-effort basis. For production workloads we recommend to use any of the below mentioned service tiers.

- **Anti-DDoS Pro**

  This service covers our 10 regions in Mainland China only. It has a maximum mitigation capacity of currently 600Gbps per instance. It supports workloads on and off Alibaba Cloud. The complete mitigation capacity of this offering is around 8 Tbps.

- **Anti-DDoS Premium**

  This service covers the rest of our regions outside of Mainland China. It comes with a mitigation capability of 2 Tbps through anycast technology. It supports workloads on and off Alibaba Cloud.

The latter two work by always routing the traffic to the closest mitigation center and from there to the origin. This may result in slightly higher latencies. If workloads are latency-sensitive we recommend to use Anti-DDoS Origin which is explained further below. The price model for the latter two is subscription-based (1 month minimum), that is pre-payed, and is based on the following values:

- Basic Protection Capacity in Mbps

  This is payed upfront. Available configurations are documented at: https://www.alibabacloud.com/help/doc-detail/67901.htm

- Burstable Protection Capacity

  In case you would like to be able to react to bursts you can also choose to buy additional burstable protection capacity. This is charged on a PAYGO basis if there are actual peaks. If there are no peaks higher than your basic protection no additional charges occur.

- Clean Bandwidth Capacity

  This is the clean bandwidth your Anti-DDoS service needs to be able to handle (excluding attack traffic). If your regular bandwidth demands gets higher packet drops may occur.

The fourth offering is called **Anti-DDoS Origin**. It works differently from Anti-DDoS Pro and Premium in that traffic is always routed directly to the origin. Only in case of attacks is the traffic redirected to our global scrubbing centers. This works by announcing according BGP routes to redirect the traffic. For customers outside of Mainland China this is also supported for origins off Alibaba Cloud. For this scenarios customers need to have a BGP network and AS, however. In the event of an attack traffic is then routed either through a GRE tunnel or a cross connect to the origin.

Web Application Firewall

Alibaba Cloud WAF is a web application firewall that monitors, filters, and blocks HTTP traffic to and from web applications. Based on the big data capacity of Alibaba Cloud Security, Alibaba Cloud WAF helps you to defend against common web attacks such as SQL injections, Cross-site scripting (XSS), web shell, Trojan, and unauthorized access, and to filter out massive HTTP flood requests. It protects your web resources from being exposed and guarantees your website security and availability.

WAF comes into different editions (Pro, Business, Enterprise, Exclusive) which support different scales and features. A detailed description can be found at https://www.alibabacloud.com/help/doc-detail/58487.htm

After a website is connected to WAF, it uses the default protection policies to protect the website against common Web attacks (such as SQL injections and XSS) and HTTP flood attacks. You can enable more WAF features and adjust the protection policies based on your actual business needs. A complete list of the protection policies are listed at https://www.alibabacloud.com/help/doc-detail/96868.htm

If you want to deploy WAF together with CDN and Anti-DDoS Pro or Anti-DDoS Premium, we recommend that you deploy components in the following sequence: client, Anti-DDoS Pro or Anti-DDoS Premium, CDN, WAF, SLB, and origin server.

ARCHITECTING FOR HIGH AVAILABILITY AND FAULT-TOLERANCE ON ALIBABA CLOUD

Highly available systems are reliable in the sense that they continue operating even when critical components fail. They are also resilient, meaning that they are able to simply handle failure without service disruption or data loss, and seamlessly recover from such failure. High availability is commonly measured as a percentage of uptime. The number of "nines" is commonly used to indicate the degree of high availability. For example, "four nines" is indicative of a system that is up 99.99% of the time, meaning it is down for only 52.6 minutes during an entire year.

High-Availability is a qualitative measure which is affected by every component of a system, and by the way how this component is setup and how it is integrated with other components. As such, it is a broad topic that needs to be looked at many different levels of abstractions and also has a great overlap with disaster recovery.

Alibaba Cloud provides different services for each level of abstraction: from workloads distribution on physical server level (deployment sets) to data centers and availability zones, to the built-in redundancy of basic services such as Service Load Balancers and VPN-Gateways, to sophisticated DNS-based load-balancing to account for cross-regional high-availability, and of course managed replication and synchronization services such as Data Transmission Service that allows for disaster recovery strategies for your database workloads.

Let us look into the various services and options in more detail.

REGIONS AND AVAILABILITY ZONES

Availability Zones (AZs) are distinct locations within an Alibaba Cloud Region that are engineered to be isolated from failures in other Availability Zones. They provide cost-free, low-latency network connectivity to other Availability Zones in the same Alibaba Cloud Region. Each region is completely independent. By launching your workloads in different AZs you are able to achieve the greatest possible fault tolerance. More details on Alibaba Cloud's current regions and AZs can be found here: https://www.alibabacloud.com/help/doc-detail/40654.htm

Your workload distribution across availability zones directly influences the Availability Service Level Agreement (SLA). If your workload is only run on one ECS instance for example, your availability SLA will be 99.975% of monthly uptime. If your workload is deployed on at least 2 ECS instances across two or more AZs then your availability SLA will be 99.995%. You'll find detailed information on our services' SLA here: https://www.alibabacloud.com/help/doc-detail/42436.htm

DEPLOYMENT SETS

A deployment set is a policy that controls the distribution of ECS instances on the physical server hosts. As of now, it supports the so-called *High-Availability* policy. If it used, all the ECS instances within your deployment set are strictly distributed across different physical servers within the specified region. The high availability policy applies to application architectures where several ECS instances must be isolated from each other. The policy significantly reduces the chances of service being unavailable. When you create ECS instances in a deployment set, you can create up to seven ECS instances in each zone. This limit varies with your ECS usage. You can use the following formula to calculate the number of ECS instances that can be created in an Alibaba Cloud region: *7 × Number of availability zones*. Deployment sets do not incur any additional costs. For more details please consult: https://www.alibabacloud.com/help/doc-detail/91258.htm

BUILT-IN REDUNDANCY OF ALIBABA CLOUD SERVICES

Many of the services on Alibaba Cloud provide built-in redundancy out of the box to provide a high degree of high-availability and according SLAs. We will quickly discuss the most commonly used ones here:

Service Load Balancer (SLB)

Server Load Balancer (SLB) is a traffic distribution and control service that distributes inbound traffic among several backend servers, namely ECS instances, based on configured forwarding rules and works both on TCP/UDP and HTTP(S) level. SLB expands the serving capacity of applications and enhances their availability.

Deployed in clusters, SLB can synchronize sessions among node servers to protect the SLB system from single points of failure (SPOFs). This improves redundancy and guarantees service stability. Layer-4 SLB uses the open source software Linux Virtual Server (LVS) and Keepalived to achieve load balancing. Layer-7 SLB uses Tengine to achieve load balancing. Tengine, a Web server project based on Nginx, adds advanced features dedicated for high-traffic websites.

Requests from the Internet reach the LVS cluster through Equal-Cost Multi-Path (ECMP) routing. Each LVS in the LVS cluster synchronizes the session to other LVS machines through multicast packets, thereby implementing session synchronization among machines in the LVS cluster. At the same time, the LVS cluster performs health checks on the Tengine cluster and removes abnormal machines to guarantee the availability of layer-7 SLB.

An SLB is always deployed in two AZs. If a primary zone becomes unavailable, SLB rapidly switches to a secondary zone to restore its service capabilities within 30 seconds. When the primary zone becomes available, SLB automatically switches back to the primary zone. The Availability SLA is a monthly uptime of at least 99.99%.

Please note that the layer 7 SLB is deployed as a cluster of usually 8 Tengine nodes (see https://tengine.taobao.org/ for details on Tengine). This has some very important implications for the maximum number of *Queries per Second (QPS)* a certain SLB instance can sustain. For example, an *slb.s1.small* can sustain 1000 QPS at max. These 1000 QPS refer to the overall cluster limit, not an individual node of the cluster! So in case of persistent HTTP connections (which is the default setting starting with HTTP 1.1) the requests are always sent to the same node. Thus, for a particular client you may need to divide the overall QPS limit by the number of nodes (which is usually 8). So for an *slb.s1.small* this would result in 1000 / 8 = 125 QPS per node.

VPN-Gateway

VPN Gateway is an Internet-based service that securely and reliably connects enterprise data centers, office networks, or Internet-facing terminals to Alibaba Cloud Virtual Private Cloud (VPC) networks through encrypted connections. VPN Gateway supports both IPsec-VPN connection and SSL-VPN connection. By design, an instance of VPN-Gateway is redundantly setup, meaning there is an invisible failover instance in case the primary goes down.

For configuring redundant IPSec connections with both an Active and a Standby tunnel the local gateway needs two public IP addresses. Be sure to configure the healthcheck on the VPN-Gateway as well and then define according weight values. 100 for the active tunnel and 0 for the standby tunnel. When the active tunnel is unavailable all traffic between the on-premises data center and the VPC is then automatically directed to the standby tunnel.

RDS and PolarDB

RDS is short for *Relational Database Service* and it supports different relational database engines such as MySQL, PostgreSQL, SQL Server, MariaDB, PPAS (Oracle/EnterpriseDB). Below is a quick break-down and discussion of the available editions and configurations and how it may impact your availability Recovery Point Objective (RPO).

RDS comes in four different editions:

ENTERPRISE

The Enterprise Edition offers enterprise-level reliability with a Recovery Point Object (RPO) of 0, and supports the database engines MySQL 5.7 and 8.0. It consists of one primary instance and two secondary instances. Your primary and secondary instances can be deployed in three different data centers in the same city to support cross-zone disaster recovery which makes it suitable for finance, securities, and insurance industries that require high data security. It comes with an availability SLA of 99.99% monthly uptime on dedicated instances. If deployed on general purpose instances the monthly availability SLA is 99.95%.

HIGH-AVAILABILITY

Your database system consists of one primary instance and one secondary instance. Data is synchronously replicated from the primary instance to the secondary instance. If the primary instance breaks down unexpectedly, your database system automatically fails over to the secondary instance. Secondary instance cannot be accessed. To scale horizontally for read operations you can add up to 10 read-replicas. It comes with an availability SLA of 99.99% monthly uptime on dedicated instances. If deployed on general purpose instances the monthly availability SLA is 99.95%.

HIGH-PERFORMANCE (AKA POLARDB)

This edition is also known as PolarDB. A cloud-native database service which separates the compute and storage layer. It supports high scalability, large auto-incrementing storage space, low primary/ secondary latency, and fault recovery within several seconds. It allows you to expand the storage to up to 100 TB and scale out an individual cluster to up to 16 nodes. You can create a snapshot on a database of 2 TB in size within 60 seconds. The monthly availability SLA is 99.99%. It also comes with Global Database Replication feature that allows you to replicate data to read-only nodes in other regions including Mainland China over Alibaba Cloud's private backbone network.

BASIC

The edition only provides a single master node and is not designed for high-availability. Its use-case is mainly for personal learning, small-sized websites, and development and test environments for small- and medium-sized enterprises.

RDS supports two different instance types:

GENERAL PURPOSE

A general-purpose instance exclusively occupies the memory resources allocated to it, but shares CPU and storage resources with the other general-purpose instances that are deployed on the same server. CPU resources are moderately reused among general-purpose instances that are deployed on the same server to increase CPU cost-effectiveness. The same configuration might lead to higher compute and storage performance compared to a dedicated instance. It is not, however, consistent over a longer period of time.

DEDICATED

A dedicated instance exclusively occupies the CPU and memory resources allocated to it. Its performance remains stable for a long term and is not affected by the other instances that are deployed on the same server.

RDS supports two different deployment methods:

SINGLE-ZONE DEPLOYMENT

Indicates that the primary and secondary instances are located in the same zone. Replication may be faster, but zone faults will have a serious impact on the entire database setup.

MULTI-ZONE DEPLYOMENT

Indicates that the primary and secondary instances are located in different zones for cross-zone disaster recovery. Replication is slower, but the database setup is resilient against individual zone faults. Note that this mode is currently not supported by all regions.

RDS supports different storage types:

LOCAL SSD (AVAILABLE FOR ENTERPRISE EDITION AND HIGH AVAILABILITY ONLY)

This is the recommended storage type. A local SSD resides on the same server as the database engine and therefore reduces I/O latency. The maximum possible size is directly related to the RDS instance being used. Upgrades of instance types may take very long time, however, since data might be needed to get copied over to a new physical database server in case the original one does not have enough capacity. Both logical and physical backups are supported.

ENHANCED SSD (AVAILABLE FOR HIGH-AVAILABILITY)

It is also a recommended storage type. This new SSD product is designed by Alibaba Cloud based on next-generation distributed block storage architecture. It integrates 25 Gigabit Ethernet and remote direct memory access (RDMA) technologies to provide super high performance at low latency. An enhanced SSD can process up to 1 million random read/write requests per second. Depending on the DB engine and instance type is supports up to 32 TB storage capacity.

STANDARD SSD (AVAILABLE FOR HIGH-AVAILABILITY AND BASIC)

A standard SSD is an elastic block storage device that is designed based on a distributed storage architecture. You can store data on a standard SSD to separate computing from storage. It is cheaper than ESSD but does not provide the high performance characteristics.

HIGH-PERFORMANCE DISTRIBUTED STORAGE (AVAILABLE FOR HIGH-PERFORMANCE)

Sharing the same group of data copies among multiple DB servers, rather than storing a separate copy of data for each DB server, significantly reduces your storage cost. The distributed storage and file system allows automatically scaling up database storage capacity, regardless of the storage capacity of each single database server. This enables your database to handle up to 100 TB of data at max. Storage capacity is bound by an instance-specific soft-limit which can be increased via ticket, however. See https://www.alibabacloud.com/help/doc-detail/68498.htm for details.

For every storage option except *Local SSD*, backups are done as snapshots including the binary log to guarantee consistency.

Object Storage Service (OSS)

OSS uses the data redundancy mechanism that is based on erasure coding and multiple replicas to store copies of each object in multiple devices across different facilities within the same region. This way, data durability and availability are ensured in case of hardware failures. Object operations in OSS are highly consistent. For example, when a user receives an upload or a copy success response, the uploaded object can be read immediately, and the copies of the object have been written to multiple devices for redundancy. To ensure complete data transmission, OSS checks for errors when packets are transmitted between the client and the server by calculating the checksum of the network traffic packets. The data redundancy mechanism of OSS can prevent data loss when two storage devices are damaged at the same time. After data is stored in OSS, OSS regularly checks whether copies of the data are lost and recovers the lost copies to ensure the durability and availability of data. OSS periodically verifies the integrity of data to detect data corruption caused by errors such as hardware failures. If data is partially corrupted or lost, OSS reconstructs and repairs the corrupted data by using the other copies.

In order to account for disaster recovery scenarios OSS also provides Cross-Region Replication (CRR). It allows you to asynchronously replicate each object written to a bucket in region A to another bucket in region B. Replication traffic is being transmitted over Alibaba Cloud's private backbone network. There are no bandwidth guarantees, though.

OSS is designed for an availability of at least 99.995% and a durability of 99.9999999999% (12 nines). The SLAs depend on the storage tier (*Standard, Infrequent Access, Archive* in combination with *Locally redundant (LRS)* or *Zone Redundant (ZRS)*) and are documented on our SLA page at: https://www.alibabacloud.com/help/doc-detail/42438.htm

Cloud Enterprise Network (CEN)

Alibaba Cloud provides a high-performance and low-latency private network through the Cloud Enterprise Network (CEN) service. This private network provides a secure cloud computing environment to meet your network needs. The loss of network packets during the network transmission may be caused by many factors, including the network stream collision, underlying network (Layer-2) errors, and other network malfunctions. The Alibaba Cloud transmission network is optimized and maintained to ensure that data can be transmitted across regions with a 99th percentile (P99) of per-hour packet loss lower than 0.0001%.

In addition, it has a minimum of four sets of independent redundant links between two network instances to ensure uninterrupted service should some links be disconnected. This is backed by a monthly availability SLA of 99.95%.

Global Accelerator

Global Accelerator (GA) provides access points worldwide. It is designed to accelerate transmission of network traffic. The GA service ensures high-quality Border Gateway Protocol (BGP) bandwidth and high service reliability. This allows businesses to accelerate global connections to Internet-facing services. Backed by the reliable and congestion-free global network of Alibaba Cloud, GA provides a high-speed network experience and ultra-low transmission latency for users across different regions. It is redundantly setup across at least two availability zones of a service and accelerated region. It is also backed by a monthly availability SLA of 99.95%.

MULTI-SITE HIGH AVAILABILITY

Global Traffic Manager

Global Traffic Manager (GTM) allows to route user access traffic of an application service to different IP addresses. GTM supports access addresses of both Alibaba Cloud products and non-Alibaba Cloud products, and helps you build a hybrid cloud application quickly and conveniently.

GTM uses the DNS smart resolution and the application service's running status health check to direct the user access request to the most appropriate IP address. GTM provides smart resolution based on the network area and health check based on ping, TCP, or HTTP(S). It can be used to build same-city multi-active and remote disaster recovery services flexibly and quickly.

GTM can be used for architecting high-available and resilient globally distributed applications related to the following scenarios:

- IP disaster recovery using a primary-standby architecture
- Multiple active IP addresses of an application service
- Cross-region load balancing
- Geo-regional IP Resolution

Multiple IP addresses of an enterprise application service may be distributed in data centers of different carriers or manufacturers in different regions. In this case, a single IP address cannot bear all users' access requests and it is very hard to build a load balancing architecture for the application service. There is no other service than DNS that can simply and efficiently organize IP addresses of multiple data centers to provide service for customers. However, DNS itself cannot sense the availability of IP addresses. Thus, it cannot route the access traffic of a user to the available IP address of the application service quickly and efficiently in case of faults or disasters. GTM allows you to define health checks from different monitoring nodes of Alibaba Cloud, and different balancing policies such as Weighted Round Robin to load balance your application on a network layer 3 level. In addition, it also allows for routing traffic to different regions based on the client's ISP in use.

Please check the following links for a detailed discussion on the necessary configurations and step by step instructions: https://www.alibabacloud.com/help/doc-detail/87478.htm and https://www.alibabacloud.com/help/doc-detail/87477.htm

Data Transmission Service

In cases where you also need to synchronize, replicate or change track data between different databases and stores in potentially different regions (both on and off cloud) we strongly suggest to use Data Transmission Service (DTS). In particular, it suited for the following scenarios:

DATABASE MIGRATION WITH MINIMIZED DOWNTIME

DTS can help you migrate data with minimized downtime. Your applications remain operational during migration. The only downtime is when you switch your application to the new target database.

GEO-REDUNDANT READ REPLICAS

In this case, you can build a secondary deployment in a different region to increase the availability of your application. DTS continuously replicates changes between these two geo-redundant deployments and keep the regional replicas in sync. If a failure occurs in the primary region, you can switch user requests to the secondary region.

ACTIVE GEO-REDUNDANCY

DTS performs two-way real-time data replication between business units to keep the regional replicas in sync. Simple cache updating with the change tracking mode of DTS, you can implement a simple cache updating mechanism by subscribing to the changes committed to the primary database and updating the cache in near real time.

For a more detailed discussion on the architectural setup of DTS please refer to https://www.alibabacloud.com/help/doc-detail/26598.htm

For MySQL migrations scenarios make sure that the source database has the following parameters set as follows:

```
log_bin = ON
binlog_format = ROW
```

DISASTER RECOVERY

A short overview over the available services and 3rd party support for implementing disaster recovery strategies and requirements.

Hybrid Backup Recovery (HBR)

Elastic Compute Service (ECS) Disaster Recovery is a scheme provided by Alibaba Cloud Hybrid Backup Recovery (HBR) to serve the needs of key enterprise applications and guarantee business continuity. It features disaster recovery with a second-level or minute-level recovery point objective (RPO) and recovery time objective (RTO). It can be used to reliably backup different storage services including OSS, NAS, CSG, and also ECS.

For ECS, HBR also provides ECS Disaster Recovery which lets you continuously backup and replicate application data of ECS instances into another region. In case the region goes offline or you want to be able to quickly start your application in another region additionally you can failover and launch the new environment respectively. More details can be found here: https://www.alibabacloud.com/help/doc-detail/62362.htm

3rd party Integration

Alibaba Cloud is also well integrated into 3rd-party offerings such as Hystax which allows for Disaster Recovery, Cloud Backup, and Continuous Data Protection from multiple sources including Bare metal, KVM, VMWare, and other cloud providers. More information on this offering can be found on the company's website: https://hystax.com/disaster-recovery-to-alibaba/

GLOBAL CROSS-BORDER INTEGRATION

In this chapter we will explore the various networking services Alibaba Cloud offers to implement and manage global and hybrid cross-border network integration projects, and also look briefly at the practical implications of the Cyber Security Law and ICP licensing. When we say *global cross-border networks*, we are referring to a network that spans at least two international geographies, countries, or cloud regions, including Mainland China. Such a network can either be entirely cloud-based, or it can also encompass external networks from third-party vendors or on-premises networks. In the latter case we call it a hybrid (cross-border) network. Let's look at each of these network setups in more detail and conclude this section with an overview and use-case for Alibaba Cloud Global Accelerator, which complements the network service portfolio by enabling accelerated public endpoints world-wide via Alibaba Cloud's own private backbone network.

VPC-PEERING WITH CLOUD ENTERPRISE NETWORK

Cloud Enterprise Network (CEN) allows you to peer up to 20 arbitrary Virtual Private Cloud (VPC) networks per region on Alibaba Cloud with each other. So theoretically you can peer up to *20 x number of cloud regions* with one another per CEN instance. Routing rules can be defined on a fine-granular basis with so-called routing maps. Note that 20 VPC per regions is considered a soft-limit which can be increased by contacting our support.

The pricing model works on a subscription basis where you buy one or multiple bandwidth packages with a pre-defined full-duplex (send and receive at full bandwidth) bandwidth anywhere between 2Mbps and 10Gbps. Depending on the geographies you would like to connect you need different bandwidth packages. For example, in order to peer a VPC in our Frankfurt region with a VPC in our Shanghai region, you need to buy a bandwidth package *Mainland China - Europe*. Any combination (except *Australia - Australia* since it only has one cloud region) of the following geographies are available and only differ in the price per Mbit.

- Mainland China
- Europe
- North America
- Asia Pacific
- Australia

CEN-peered VPCs benefit from a reliable, redundant, low latency and almost zero packet loss connection via Alibaba Cloud's backbone network. This is also true for connections to and from Mainland China. For example, we typically see 150ms network latency between our Frankfurt region and our Beijing region.

CEN can also be used for intra-region (i.e. VPCs are all in the same cloud region) peering where no bandwidth package is needed and hence is for free as traffic is not charged separately.

An interesting feature of CEN bandwidth packages is that they can be scaled up and down dynamically at any time to account for changing bandwidth requirements. This can also be automated based on time or different network metrics such as the average usage of bandwidth over a certain timespan. Please refer to the open-source CEN-Scaler at https://github.com/arafato/CEN-Scaler which enables to automatically scale Cloud Enterprise Network (CEN) bandwidth packages based on different metrics and timing events. It ships together with Alibaba Cloud Function Compute code and Terraform templates that set up all necessary configurations and services to get you going fast.

Keep in mind, however, that at the time of this writing downscaling a bandwidth packet during a subscription period is only supported in INTL Portal. It is not supported in Domestic Portal (only upscale).

Quick Facts Bandwidth Scaling

The change of the bandwidth is effective immediately. Meaning, the updated bandwidth can be used immediately by your applications, and it takes immediate effect on your bill.

The billing granularity is in seconds. The length of the billing period depends on the subscription type.

For monthly subscription it is always exactly 30 days independent from the calendarian length of a particular month.

For a yearly subscription it is always 12 months each 30 days

The effective billing in both up- and downscale scenarios is always based on the remaining time of the billing period, and will result in a dedicated item on the bill. So, in case of a monthly subscription this is the time in seconds until the end of the month, in case of a yearly subscription this is the time in seconds until the end of the year.

Be aware that an upscale might result in very large billing items since the price is calculated upfront until end of the month (monthly subscription) or even until end of year (yearly subscription). Thus, make sure that the customer has a large enough credit limit on his Alibaba Cloud account, even though he might never actually spend it.

Every up- and downscale action results in an additional item on the customer's bill. In case of an upscale it is a billing item, in case of a downscale it is a refund. In below picture you can see an according excerpt from a real billing where the customer scaled up the bandwidth for roughly 2 days and then downscaled the bandwidth again.

See https://www.alibabacloud.com/help/doc-detail/130927.htm for details and please consult the CEN-Pricing Document that provides in-depth examples of CEN bandwidth pricing calculations.

HYBRID NETWORKS

Alibaba Cloud provides and supports multiple ways to connect your on-premises or any other cloud-vendor network to Alibaba Cloud.

Express Connect

Alibaba Cloud Express Connect helps you build internal network communication channels that feature enhanced cross-network communication speed, quality, and security. Express Connect also helps you mitigate network instability and data breaches. It allows you to connect a leased line to any of the Alibaba Cloud access points. We recommend to contact any of our official network service providers NSP who will help you to establish one or more physical connections and connect your on-premises data center to an Alibaba Cloud VPC. A full list of NSPs access points can be found at https://www.alibabacloud.com/help/doc-detail/96019.htm

VPN Gateway

Alibaba Cloud VPN Gateway is an Internet-based service that securely and reliably connects enterprise data centers, office networks, or Internet-facing terminals to Alibaba Cloud Virtual Private Cloud (VPC) networks through encrypted connections. VPN Gateway supports both IPsec-VPN connection and SSL-VPN connection. You must configure the Maximum Transmission Unit (MTU) limit of the local VPN Gateway (that is your premises) to not more than 1,400 bytes. We recommend that you set the MTU to 1,400 bytes.

Smart Access Gateway

Smart Access Gateway (SAG) is a software-defined wide area network (SD-WAN) solution developed by Alibaba Cloud based on cloud-native technologies. SAG provides a more intelligent, reliable, and secure approach for enterprises to migrate their workloads to Alibaba Cloud. It comes with different available configurations of hardware devices and allows for zero touch provisioning and native integration into Alibaba Cloud.

Equinix Platform and Equinix ECX Fabric

Equinix Platform and Equinix ECX Fabric enables an easy and private connection to Alibaba Cloud and supports among other locations Frankfurt, Dubai, Hongkong, Jakarta, London, Singapore, Sydney, Tokyo, Chicago, Dallas and Denver. More information can be found here: https://www.equinix.com/platform-equinix/

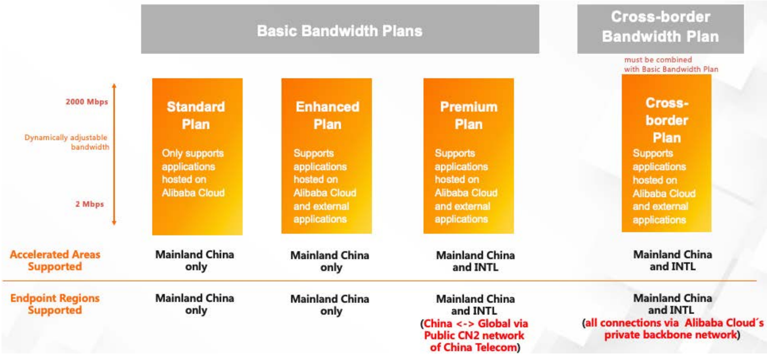
All of these solutions for last-mile connectivity can be used in combination with CEN which allows you to connect on-premises networks with each other that may be in Mainland China and overseas regions in a reliable and performant way. While the last mile is for example IPSec, the rest is being transmitted over Alibaba Cloud's private backbone network. This allows for fully automatable and quick cross-border network integrations world-wide.

GLOBAL ACCELERATOR

Global Accelerator (GA) can provide access points worldwide and accelerate transmission of public network traffic. The GA service guarantees high-quality Border Gateway Protocol (BGP) bandwidth and high service reliability and helps businesses accelerate global connections to Internet-facing services. Backed by the reliable and congestion-free global network of Alibaba Cloud, GA provides high-speed networking experience and ultra-low transmission latency for users across regions. GA assigns an accelerated IP address to each acceleration region in an acceleration area. Clients from an acceleration region can connect to the access point nearest to the clients through the accelerated IP address. The access point receives client requests and forwards the requests over the Alibaba Cloud global network. GA then automatically selects routes and distributes client requests to the optimal endpoints to avoid network congestion and reduce network latency. Endpoints can be IP addresses or domain names of origin servers.

Let's quickly discuss the bandwidth package options in more detail. Below figure gives an overview about the current options. For accelerating areas and/or endpoints to and from Mainland China, Alibaba Cloud provides two options:

1. *Cross-Border Plan + Standard or Enhanced Plan.* By choosing this option traffic is routed over Alibaba Cloud's private backbone network once the request arrives at the nearest access point. Access points for accelerated regions are located in Mainland China. Thus companies using this configuration need a valid ICP filing or license.
2. *Premium Plan.* With this configuration the access points are located in Hong Kong which does not require an ICP filing or license. Traffic to and from Mainland China is routed via the public CN2 network operated by China Telecom with only minimal additional latency while still offering exceptional and reliable network connection quality. This is especially useful in scenarios where companies need more time to get an ICP license but already want to deliver their digital services or websites to Mainland China via an optimized network connection. Still, we recommend this as a temporary solution. For serious business intentions we recommend to acquire an ICP license.



	Basic Bandwidth Plans			Cross-border Bandwidth Plan must be combined with Basic Bandwidth Plan
	Standard Plan Only supports applications hosted on Alibaba Cloud	Enhanced Plan Supports applications hosted on Alibaba Cloud and external applications	Premium Plan Supports applications hosted on Alibaba Cloud and external applications	Cross-border Plan Supports applications hosted on Alibaba Cloud and external applications
Accelerated Areas Supported	Mainland China only	Mainland China only	Mainland China and INTL	Mainland China and INTL
Endpoint Regions Supported	Mainland China only	Mainland China only	Mainland China and INTL (China <-> Global via Public CN2 network of China Telecom)	Mainland China and INTL (all connections via Alibaba Cloud´s private backbone network)

Global Accelerator - Bandwidth Packages

SOLUTION CASE STUDY: PRIVATE GLOBAL INTERNET ACCELERATOR

This chapter describes how to accelerate http(s)-requests to any SaaS application and public service from workloads running on Alibaba Cloud by using Cloud Enterprise Network and a redundant setup of a nginx-based reverse proxy that will also do DNS resolution locally to mitigate any DNS spoofing attacks.

Introduction

From local startups to Small-Medium Enterprises (SMEs) to Multi-National Companies (MNCs), running and operating globally distributed IT-landscapes, -processes and -workflows for the most business-critical applications has become the new normal in our globalized economy. With employees and customers distributed all over the world many have realized the need to provide global communication and service platforms that are not operated in local silos but rather in a global and coherent way that maximizes synergies and cost-efficiency. The bandwidth of scenarios is very broad and diverse. It starts with pretty basic things such as easy-to-use file shares between employees located in Europe and China, smooth online video calls and meetings between multiple countries such as Munich, Shanghai, and Sidney, and finally ends at very complex scenarios such as globally interconnected SAP workloads and globally distributed IoT-Platforms that manage and analyse millions of devices in different countries and legislations in real-time.

Challenges

Maintaining a reliable Quality of Service (QoS) for globally distributed workloads that need to interact with each other to support business-critical workflows and processes is challenging, however, due to the following reasons:

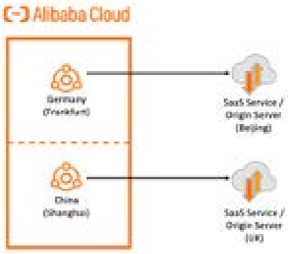
- Reliability: Public internet bandwidth of many regions and countries such as Mainland China is limited and thus not reliable resulting in high packet loss and high roundtrip latency.
- Flexibility: While multi-national telco providers provide private, reliable and high-bandwidth offers (e.g. leased lines) they are usually not very flexible. They require commercial long-terms commitments of multiple months or even years, are hard to change configuration-wise, do not support automation and modern DevOps practices very well, and also do not scale easily with changing bandwidth requirements.

- Security: To decrease the attack surface the entire networking traffic should not be routed through the public internet until it reaches its local internet breakout.
- Compliance: Global workloads need to adhere to the local law and regulations of the respective countries they operate in. In multi-national setups it is crucial to account for the individual compliance laws and regulations such as the GDPR in Europe and the Cyber Security Law in Mainland China. In particular, it is important that any of such regulations are also met by the network connectivity providers.

Every point can be easily a blocker or at least a reason to reduce speed for any project that includes multi-national interconnected IT-workloads.

Example Scenario

In the remainder of this article we would like to focus on a running example that we often find with many of our customers. As depicted in below figure, workloads deployed on Alibaba Cloud need to communicate with external public services that are located in a different geographic region of the world. In our example, an application in the German region of Alibaba Cloud needs to communicate with an external public server (or service) which is located in Beijing. The other way round is of course also possible which is depicted at the lower side of this picture where an application deployed in Shanghai region needs to communicate with a public service in the United Kingdom (UK).


Alibaba Cloud
Germany (Frankfurt)
SaaS Service / Origin Server (Beijing)
China (Shanghai)
SaaS Service / Origin Server (UK)

Example Scenario - Cross-Border Access to External Public Services

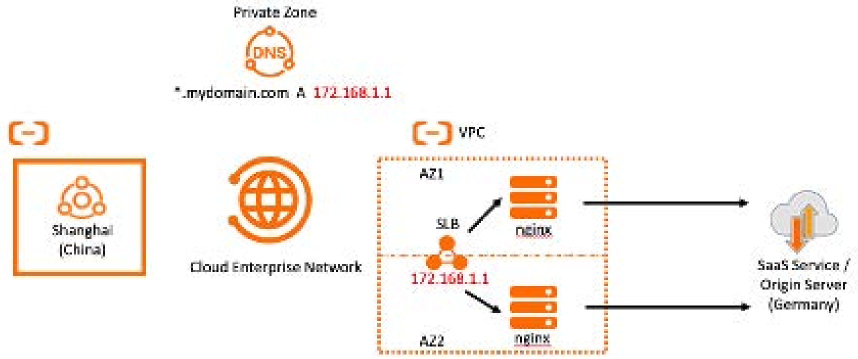
While this seems not utterly complex and easy to implement at first glance some important considerations have to be taken into account:

- Public internet bandwidth to Mainland China is limited and not reliable often resulting in packet loss and high latency
- Dedicated private lines and bandwidth are expensive and are usually hard to easily scale up and down with your demands
- Networking operators, services and data design need to adhere the respective local legislations and laws such as GDRP, Cyber Security Law (CSL) and MPLS 2.0

Solution Design

Alibaba Cloud makes it simple to address these challenges and to quickly setup an environment where our applications and workloads can reliably communicate with these public services by building on top of Alibaba Cloud's world-class and battle-proven networking services. At the same time Alibaba Cloud ensures that its cloud platform meets the MLPS 2.0 baseline and GDPR.

The three major building blocks will be Alibaba Cloud Enterprise Network (CEN), a DNS Private Zone configuration, and a redundant pair of reverse proxies (nginx) which are run on our Elastic Compute Service (ECS) across two availability zones and exposed to the application over an internal load balancer (SLB). Throughout the remainder of this section we will focus on only one networking direction for the sake of simplicity. The other networking direction can be setup analogously.



Solution architecture for a private global internet accelerator

The solution works by defining the domain names you like to have DNS-resolved and forwarded by the proxy in a Private Zone and have them pointed to the proxy or internal SLB IP address respectively. In our example this would be *.mydomain.com (be sure to turn off recursive proxy resolution in PrivateZone if you are using wildcards). Every other domain will be resolved locally and be forwarded. So any http-request against such a domain will be forwarded to the proxy. DNS resolution and subsequent http-forwarding will then be handled by nginx. We will look at the required configuration files for nginx later in this article.

Let's briefly look at Cloud Enterprise Network (CEN). It is a highly-available network built on the high-performance and low-latency global private network provided by Alibaba Cloud. By using CEN, you can establish private network connections between Virtual Private Cloud (VPC) networks in different regions, or between VPC networks and on-premises data centers which is routed over Alibaba Cloud's private backbone network. CEN supports automatic route distribution and learning, which speeds up network convergence, improves the quality and security of cross-network communications, and interconnects all network resources. The Alibaba Cloud transmission network is optimized and maintained to ensure that data can be transmitted across regions with a 99th percentile (P99) of per-hour packet loss lower than 0.0001%. Bandwidth of CEN can be scaled up and down anytime from 2Mbps up to 10Gbps and will be charged on a second-granularity. It can also be automated based on different metrics via scripts. For a detailed discussion and solution on this please refer to https://github.com/arafato/CEN-Scaler Once we attach both VPCs to the CEN instance traffic is automatically routed between these two VPCs. Be sure to not use overlapping IP-ranges, otherwise there might be routing issues and errors.

The second building block will be NGINX (https://www.nginx.com/) which will act as our reverse proxy. It will proxy any request from the client application in Frankfurt region to the destination which also includes DNS requests. We will set up a redundant pair of NGINX servers across two availability zones and distribute traffic between the two using a Service Load Balancer (SLB) in front of them.

Before we look at detail at the configuration let's think about the necessary configuration of the ECS instances in more detail with a specific focus on the Outbound and Inbound internet bandwidth which is important for our scenario. Let's quickly define inbound and outbound traffic:

- Inbound refers to network traffic that is sent from the public internet to any Alibaba Cloud service (i.e. traffic flows into the cloud)
- Outbound refers to network traffic that is sent from any Alibaba Cloud service to the public internet (i.e. traffic leaving the cloud)

Inbound traffic is at minimum 100MBits. It will be at most as high as EIP Bandwidth. Outbound traffic is capped by the EIP bandwidth. Bandwidths greater or equal 1Gbits can only be saturated by multiple threads. Note that the maximum default EIP bandwidth is 200 Mbits, the maximum instance-bound public IP bandwidth is 100 Mbits.

In order to increase that you have to add your EIPs (no instance-bound public IPs are supported) to a shared internet bandwidth package which can be as high a 1Gbits (see https://www.alibabacloud.com/help/doc-detail/55784.htm for details). There are no additional costs for a shared bandwidth internet package. This way you can increase your outbound bandwidth to up to 1 Gbits. This bandwidth can only be saturated by multiple threads, though. You can also create multiple bandwidth packages of course.

To further increase the external network performance you can also use multiple ENIs (Elastic Network Interfaces) and bind up to 10 EIPs to up to 10 private ip addresses of a single ENI in NAT mode. By assigning multiple bandwidth packages to these EIPs you can further increase the network throughput of a single instance. Please check https://www.alibabacloud.com/help/doc-detail/88991.htm for further details on ENI and the different supported modes such as Cut-Through mode and Multi-EIP to ENI mode.

Let's go through step by step through the configuration:

1. Install and configure nginx on both ECS instances Depending on your Linux distribution you are using commands may vary. Please consult your distribution documentation on how to install nginx.

After it has been installed add the following configuration to `/etc/nginx/nginx.conf` Make sure to make a backup of the file prior to the modification.

```
# Configure stream forwarding of https protocol stream {
  map $ssl_preread_server_name $backend_pool {
      # Explicitly configure allowed domain names
      ~.*\.mydomain\.com $ssl_preread_server_name:$server_port;
     default "";

  }

  server {
    listen 443;
    ssl_preread on;
    resolver 8.8.8.8;
    proxy_pass $backend_pool;
  }
}
```

This will allow to resolve and forward the explicitly specified domain names (e.g. *.mydomain.com) to be resolved by any DNS name server you may want to use. In our case, it is the Google DNS name servers which is available under 8.8.8.8, but could be any other DNS name servers which are reachable by the proxy of course. Often you also need more control over the HTTP headers and possibly error codes you would like to return as a result of any http-call to your proxy. In order to do this you can modify the server.location block in your nginx configuration-file as follows:

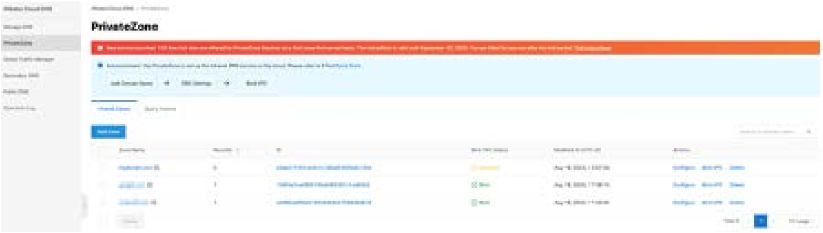
```
server {
…
    location / {
          set $is_allowed 0;
if ($host ~ '.*\.mydomain\.com') {
            set $is_allowed 1;
}
if ($is_allowed = 0) {
            return 404;
}
```

```
        proxy_set_header Host $host;
        proxy_set_header Accept-Encoding "";
        proxy_set_header X-Real-IP $remote_addr;
        proxy_set_header X-Forwarded-For $proxy_add_x_forwarded_
for;
        proxy_set_header Cookie $http_cookie;

        resolver 8.8.8.8;
        proxy_pass http://$host:$server_port$request_uri;
      }
}
```

You can modify this more specific to your requirements, of course. Also do the same configuration on the second ECS instance. After that restart the nginx-server to make the new configuration effective. Depending on your distribution this can be done for example via `$ systemctl restart nginx`

2. Load Distribution with SLB In order to distribute requests equally on both proxy machines you also need to setup an internal SLB in the same region as the proxy machines. It will not be exposed to the public internet. Then add the two ECS instances with the nginx proxy running to the standard server group and create a TCP listener that forwards any request to the proxies on ports 80 and 443. For DNS requests we recommend to use a UDP listener which is listening and forwarding on port 53. The web-based wizard will guide you through the entire process. If not sure which values to choose just go with the default values for now and adapt later for specific requirements.
3. Private Zone Now it is time to configure the Private Zone which is part of the Alibaba Cloud DNS service.



Solution architecture for a private global internet accelerator

Simply add the zones you want to forward, add according entries that point to the internal proxy IP (or in case of a redundant setup to the internal IP of the SLB). Then bind the VPC where the client machines are deployed to this Private Zone. The configuration is then effective immediately.

4. Security Groups Configuration Once this has been done, there is one last thing left to configure. You have to grant the client machines explicitly access to the proxy machines. You do that by defining a so-called security group which is associated to the proxy machines.

Obviously, this needs to be adapted with your specific configuration. But the point is, you need to explicitly grant the client machines (aka "Authorization Object") access to the proxy machines over all required ports which are usually 80, 443 for http(s) and 53 for DNS.

If all of this set up you can check if the setup is working by doing a ping-request against a public endpoint like so

```
$ ping my.endpoint.com
```

Which should resolve into the internal IP-address of the proxy or SLB IP. You can then send requests over the proxy to the according public endpoints of yours.

Conclusion

In this section we looked at how to build a private internet accelerator based on Alibaba Cloud Enterprise Network service and nginx to accelerate DNS and http(s) requests over Alibaba Cloud's private high-performance, low-latency network to external endpoints. We discussed how to setup each one of these components, what to keep in mind when selecting the according ECS instances in terms of networking bandwidth and what kind of challenges can be easily addressed, both from a technical point of view and a compliance point of view.

CONCLUSION

Throughout this book we touched upon a broad variety of topics and services that we consider to be the core concepts of Alibaba Cloud. Most of it serves as a starting point for you to dive deeper into the relevant topics and to explore what Alibaba Cloud has to offer in more detail by yourself. While we believe that the content in this book gives a comprehensive overview about the most important topics for a fast ramp-up, there are many other topics that have not been covered in this book. This includes our platform capabilities in the realm of Big Data, Artificial Intelligence, Media Services, Internet of Things, and also Digital Finance. One example in the Big Data and analytics realm that is worth mentioning is Alibaba Cloud's Realtime Compute for Apache Flink. This product is powered by Ververica and developed by Alibaba Cloud based on Apache Flink, and it is an enterprise-level, high-performance system that is used to process big data in real time. It is officially released by the founding team of Apache Flink and has a globally uniform commercial brand. This system is fully compatible with the APIs of open source Flink and provides a wide range of value-added features for enterprises that are jointly developed by Alibaba Cloud and Ververica. Also, we did not cover cloud-native technologies on our cloud platform. For example, Alibaba Cloud provides a state-of-the-art Kubernetes service, which was ranked by Gartner as the leading provider for bare-metal based Kubernetes clusters. Additionally, it strongly focuses on open-source standards and provides fully managed versions of Prometheus. These features, as well as our Container Registry's support for CNNF's DragonFly (https://d7y.io/) and reliable cross-region replication (including Mainland China), enable our customers to quickly and efficiently build highly-scalable cloud-native applications.

Thus, we would like to conclude this chapter with a list of recommended links to the topics that were not or only partly covered in this book to enable our readers to efficiently ramp-up on these fields via self-study.

In general, we recommend our technical documentation as the best source for learning about our cloud services in detail.

The international documentation is available at https://www.alibabacloud.com/help while the domestic documentation can be found at https://help.aliyun.com. At the time of writing, the domestic documentation is more comprehensive, up to date, and covers the entire service portfolio. It is, however, only available in Chinese. The state of the international documentation is catching up quickly, though.

Another great source for self study is the Alibaba Cloud Academy which can be found here: https://edu.alibabacloud.com/ It provides free e-learning courses and also so-called Clouder courses, which lets you quickly understand an Alibaba Cloud product or solution architecture while earning a digital certification. The free e-learning courses feature a broad selection from infrastructure Big Data and Analytics, and Internet of Things. It also includes an elaborate course that focuses exclusively on cloud-native. This course was jointly developed by CNCF and Alibaba Cloud and is completely free of charge. It is available here: https://edu.alibabacloud.com/certification/university-cloudnative

Last but not least we also recommend our official Alibaba Cloud blog at https://www.alibabacloud.com/blog. It is constantly updated with technical articles that feature great insight into the technology landscape of Alibaba Group in general, and also hands-on articles on common solution architectures and their respective implementation on Alibaba Cloud. Be sure to check this blog out for the latest updates on Alibaba Cloud technology and announcements.

We hope that this book will be valuable to you by helping you to get the most out of Alibaba Cloud and to be successful with your technical endeavours and digital business. For any kind of feedback please do not hesitate to contact us at https://www.alibabacloud.com/about/contact-us

ABOUT

Established in 2009, Alibaba Cloud (alibabacloud.com), the digital technology and intelligence backbone of Alibaba Group, is among the world's top three IaaS providers, according to Gartner. It is also the largest provider of public cloud services in China, according to IDC.

Alibaba Cloud provides a comprehensive suite of cloud computing services to businesses worldwide, including merchants doing business on Alibaba Group marketplaces, start-ups, corporations and public services.

Alibaba Cloud is the official Cloud Services Partner of the International Olympic Committee.

www.alibabacloud.com/contact-sales