ALIBABA CLOUD

Alibaba Cloud

Performance Efficiency Pillar White Paper

A Well-Architected Framework

June 2020

Version: 1.0.0

阿里云

Legal Disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.
2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company, or individual in any form or by any means without the prior written consent of Alibaba Cloud.
3. The content of this document may be changed due to product version upgrades, adjustments, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and the updated versions of this document will be occasionally released through Alibaba Cloud-authorized channels. You shall pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.
4. This document is for reference only as a guide for Alibaba Cloud's products and services. The document is provided as is, regardless of the "current situation", "defectiveness", or "current function" of Alibaba Cloud's products and services. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not bear any liability for any errors or financial losses incurred by any organizations, companies, or individuals arising from their download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, bear responsibility for any indirect, consequential, exemplary, incidental, special, or punitive damages, including lost profits arising from the use or trust in this document, even if Alibaba Cloud has been notified of the possibility of such a loss.
5. By law, all the content in Alibaba Cloud documents, including but not limited to images,

page design, and texts are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of the Alibaba Cloud website, product programs, or content shall be used, modified, reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. Without Alibaba Cloud's prior written consent, no individual or entity can use, publish, or copy Alibaba's names, including but not limited to individual or combined forms, such as "Alibaba Cloud", "Aliyun", "Wanwang", or any other brands used by Alibaba Cloud and/or its affiliates for marketing, advertising, sales promotions, or any other purposes. The following are also protected as they pertain to all brands and subsidiaries: all marks, patterns, or any similar company names, corporate names, trademarks, product or service names, domain names, graphic designs, symbols, logos, or ways in which third-parties can recognize Alibaba Cloud and/or its affiliated companies through specified descriptions.

6. Please contact Alibaba Cloud directly if you discover any errors in this document.

Table of Contents

1. Summary

This white paper focuses on the performance efficiency pillar in the well-architected framework of Alibaba Cloud. The aim of the document is to help you follow best practices in the design, delivery, and maintenance in Alibaba Cloud environments.

2. Performance Efficiency

The performance efficiency pillar focuses on the efficient use of computing resources to meet various requirements and to maintain this efficiency as requirements change and technologies evolve. This white paper provides in-depth best practice guidance for you to design performance-efficient architectures on Alibaba Cloud.

3. Design Principles

In a cloud environment, we recommend that you follow these principles to achieve desired performance efficiency:

- Prioritize advanced services and technologies that are native to the cloud

In the era of data technology (DT), computing needs are diverse, and requirements for technical solutions will also continue to expand and change. Compared with relying on your internal IT team to learn new technologies and create full-stack implementations of them, it is more efficient to use the technical solutions that cloud service providers offer. In this way, you can benefit from the knowledge and technical know-hows that cloud service providers have accumulated through their experience. You can directly use the cloud-native "out-of-the-box" services and professional technologies and managed services that cloud service providers offer. This allows you to focus on developing your business and applications without worrying about how they are supported by underlying technologies or about resource allocation and management.

- Use serverless architectures compliant with cloud-native standards

In a cloud environment, serverless architectures can better host your business and applications, and free you from operating or maintaining any server systems. We recommend that you embrace industry standards related to cloud-native technologies to build services that meet delivery standards and technical architecture standards. By doing so, you can easily host your applications on the service stack of your cloud computing provider. This makes access more efficient and offers a unified architectural abstraction.

- Achieve efficient global deployment

With several clicks in the console offered by your cloud service provider, you can deploy resources and applications in multiple regions around the world. Your customers will have a better experience with less latency by accessing the resources that are nearest to them. Cloud computing also reduces your operational cost.

- Make full use of the elasticity and performance stability of cloud computing resources

Cloud computing providers offer resources that are elastic and stable. We recommend that you make full use of this advantage in your architecture design to build cloud applications capable of auto scaling as well as fast start and stop. This ensures performance efficiency under high resource utilization.

- Improve experiment efficiency

We recommend that you use virtualized and automated ways to configure and deploy resources. This allows you to quickly experiment various combinations of resources and ultimately find the configuration that maximizes performance efficiency.

4. Definitions

Performance efficiency in the cloud environment is divided into four dimensions of best practices:

- Custom Choice
- Measurement
- Monitoring
- Trade-offs

Based on these four dimensions, you can build a high-performance architecture solution based on Alibaba Cloud's products and services. First, the customer should start from their own business scenarios, and select the cloud-based resources and the corresponding architecture construction scheme based on the design principles and best practices in this article. Then, the customer should periodically test and measure the above choices, and use data to ensure that the existing architecture design gives full play to the underlying capabilities and performance advantages provided by the cloud platform. Finally, it is often necessary to achieve higher performance requirements through trade-offs in architecture design, such as application layer caching or asynchronous message queues.

4.1. Custom choice

The best solution for a particular system will vary depending on your workload type, and it is generally necessary to combine multiple solutions. A good architecture system may employ a variety of solutions and implement many functions to improve performance levels.

In Alibaba Cloud, all resources exist in the form of virtualization, and are delivered through different types and configuration schemes. This means that customers can more easily find proper solutions that meet their own needs, and can also establish more applicable options that are often difficult to achieve in an intranet.

4.1.1. Computing

4.1.1.1. Manage instances

Elastic Compute Service (ECS) instance

An ECS instance is a cloud server, that is, a virtual machine sever. ECS instances provide different instance types and meet the most cost-effective requirements in different business scenarios based on different CPU-to-memory ratio or network/storage performance metrics. The latest ECS instances are based on the third-generation X-Dragon architecture developed by Alibaba Cloud. The instances have upgraded their stability and performance in all aspects, fully releasing technical benefits and meeting cloud customers' core demands for reducing costs and increasing efficiency. With the X-Dragon architecture, the instances successfully build plenty of virtualization features into its dedicated hardware, so the virtualization overheads are greatly reduced and a stable, predictable, and high performance is achieved.

ECS instances have the following benefits:

- **Cutting-edge scalability and ready to use all the time**

Alibaba Cloud has the largest IAAS scale in the Asia Pacific, with large resource pools and global availability zones, allowing customers to purchase and use it whenever they want. Proved by Alibaba Cloud's large-scale practices during double 11, powered by the large-scale intelligent scheduling technology, proprietary X-Dragon platform technology, and distributed file system, ECS instances have the honor of fastest startup speed, comparing with global cloud vendors. The instances with 160,000 cores can starts up in five minutes for a customer in a region.

- **Excellent performance and high stability**

Upgraded stability and other performance, CPU core turbo frequency of up to 3.2 GHz, leading the world; Network bandwidth of up to 25 GB, six million PPS, storage bandwidth of 16 GB, supporting up to one million IOPS. On the basis of massive amounts of cloud server fault data we have collected for many years, by using professional algorithms and machine learning model of Alibaba DAMO Academy, the instances can accurately predict 99% of faults. Even if failure prediction fails, ECS instances can implement the fault avoidance. This can effectively avoid your business losses and ensure that you will not experience any downtime.

- **Comprehensive scenario coverage and various specifications**

Alibaba Cloud provides enterprise-level instances with different CPU-to-memory ratios such as 1:1, 1:2, 1:4, 1:8, 1:16, and 1:32 to meet the business needs of different scenarios. Different business applications shall use varied instance types, which can optimize your resource utilization. Additionally, it provides network enhanced specifications for network-intensive application scenarios such as NFV/SD-WAN, telecom forwarding and other applications, with PPS up to 13.5 million. We have introduced storage enhanced instances for I/O-intensive applications, memory enhanced instances for memory application databases, and burst instances for applications in telecom traffic peak and valley. For some applications that are not sensitive to computing power and ask for a higher cost-efficiency ratio, Alibaba Cloud provides shared instances, AMD instances, and AEP instances.

GPU instances

A Graphics Processing Unit (GPU) is added to ECS instances, which is a heterogeneous computing instance. By using the general-purpose CPU computing capability and the intensive computing acceleration capability of the graphics computing unit (GPU), developers can combine computing platforms and software libraries, such as CUDA, OpenGL, DirectX, and FFmpeg, and can fully utilize the high-concurrency computing power provided by GPU instances in their own applications and business systems. In addition, combined with the AIACC acceleration engine provided by Alibaba Cloud, GPU instances can further improve the cluster acceleration efficiency and ease of use when establishing large-scale deep learning training clusters.

ECS bare metal instances

ECS bare metal instance is based on the X-Dragon architecture developed by Alibaba Cloud. By using the independently developed chips, system software, and a redefined server hardware architecture, Alibaba Cloud provides a world-leading innovative computing product that deeply integrates the features of physical servers and virtual machines. It provides features of scalable resources, minute-level delivery, fully automated O&M, and physical machine performance without compromising, lossless in functions and hardware-level Isolation. It is fully compatible with Alibaba Cloud ecosystem products, meeting the cloud migration requirements for key enterprise applications and high-load applications, and achieving cloud migration without any obstruction.

ECS bare metal instances have the following benefits:

- Secure, reliable, and excellent performance

The instances enable customers exclusively use computing, memory, and I/O resources without any virtualization performance overheads or feature loss. In this way, it can meet your business needs for high performance, stability, data security and compliance.

- Highly isolated container deployment

If you deploy container services on ECS bare metal instances, no any virtualization performance overhead will occur. By using the features of multiple ENIs, high bandwidth, and high PPS of ECS bare metal instances, we can implement high-density scheduling of containers, get a higher ratio of income to expenditure.

- Support for RDMA network with high-bandwidth and low-latency

Remote Direct Memory Access (RDMA) network has obvious advantages in high bandwidth, low latency and CPU load reduction, but it has shortcomings in network virtualization. ECS bare metal instances can be connected to RDMA networks to provide high-bandwidth and low-latency network connections for high-performance computing in the cloud.

- AnyStack on ECS bare metal instances

ECS bare metal instance has a complete processor feature same as physical machines. It can adapt to multiple hypervisors, such as VMware, KVM, Xen, and Hyper-V, enabling you to deploy Apsara stack on public clouds. This can meet the needs to quickly and smoothly migrate services to the cloud without modifying the legacy system, provides a cloud environment management experience with consistent online and offline architecture, and meets the needs to deploy hybrid cloud and multi-cloud.

4.1.1.2. Container

Elastic Container Instance (ECI)

A container is a lightweight software packaging and virtualization technology. Container images that are created and packaged can run on virtual machines, physical servers, or public cloud hosts without any modification.

An ECI can provide secure serverless container services. You do not need to manage underlying servers, and do not need to plan capacity during operation. Instead, you only need to provide a packaged Docker image to run containers, and pay only for the resources consumed during the container operation.

By using an ECI, you can directly run containers and pods on Alibaba Cloud without purchasing and managing any ECS instances, eliminating the need to maintain and manage underlying ECS instances. This allows you to focus on business development without having to care about complex infrastructure maintenance.

An ECI can provide vm-level security and resource isolation capabilities. It is deeply optimized for container operating environments, and provides faster startup speeds and operational efficiency than virtual machines.

You can quickly deploy containerized applications in the ECI console, add ECI to your existing business systems by using the ECI SDK, or connect to Kubernetes by using Virtual Kubelet. The scalable capabilities of an ECI make it easy to handle traffic surges.

ACK

Alibaba Cloud Container Service for Kubernetes (ACK) is one of the service platforms in the world that have first passed the Cloud Native Computing Foundation (CNCF) Kubernetes conformance certification. It allows you to create and manage Kubernetes clusters in an automated manner. You can also manage the entire lifecycle of containerized applications.

ACK provides managed Kubernetes clusters. The Kubernetes Master is managed by Alibaba Cloud. You only need to manage worker nodes deployed on Elastic Compute Service (ECS) instances. Managed Kubernetes clusters are highly-available, easy to use, and data-secure. This allows you to focus more on your business development rather than infrastructure management. ACK supports all types of ECS instances, such as virtual machines, Bare Metal Instances, and ECS instances with GPU capabilities. You can select appropriate instance types based on your workloads to optimize performance and reduce computing costs.

As your business demand and Kubernetes cluster grow, multiple correlated services may be deployed in the same Kubernetes cluster. Different types of services may require different types of ECS instances or cluster configurations. Therefore, ACK provides node pools to meet this requirement. A node pool manages a group of ECS instances and has independent resource specifications, billing methods, operating systems, security groups, and automatic scaling policies. It significantly improves the efficiency of running multiple services in one cluster. A node pool automatically scales ECS instances to match the resource needs of applications based on Elastic Scaling Service (ESS) and the elastic scaling policy of the ACK cluster. In addition, ACK provides the Serverless Kubernetes (ASK) cluster service, which is based on Elastic Container Instances (ECIs). ASK allows you to deploy Kubernetes applications without node management or capacity sizing. This provides extreme flexibility for your workloads and free you from node operations and maintenance.

For containerized applications running in an ACK cluster, you can define an automatic scaling policy based on monitoring metrics. Containers or computing resources are scaled out based on the automatic scaling policy to meet the growing need for resources.

ACK has integrated with and optimized the infrastructure capabilities of Alibaba Cloud, and is therefore ready for use. The Terway network plug-in allows you to assign Elastic Network Interfaces (ENIs) to containers. This enables container instances and applications running on virtual machines in the same Virtual Private Cloud (VPC) network to communicate with each other without causing network performance degradation. With Terway, you can easily assign ENIs to a large number of containers for high-density computing in a high speed network. ACK Terway has enhanced the network performance of cloud-native applications to a higher level. ACK also integrates with storage services of Alibaba Cloud, such as Block Storage, Network Attached Storage (NAS) , and Object Storage Service (OSS). You can configure and maintain storage volumes attached to Kubernetes pods in an automated manner. This helps you meet the diverse needs of stateful applications.

ACK has integrated with a variety of Alibaba Cloud services. For example, when an application service is released, it can be automatically bound to a Server Load Balancer (SLB) instance. Users can access the application service through the SLB instance. ACK provides an automated solution for you to maintain and monitor ACK clusters in multiple dimensions. For example, the Application Real-Time Monitoring Service (ARMS)-Prometheus and CloudMonitor services are used to collect and display the monitoring data of cluster resources and containerized applications. Logs of containerized applications can be collected, analyzed, and processed through Alibaba Cloud Log Service. In addition, ACK provides various types of statistical reports for users, such as audit logs, which allow you to track the usage of cluster resources.

Before you use ACK, we recommend that you have completed the following tasks: networking, node pool division, ECS instance type selection, operating system selection, storage sizing, elastic scaling policy setting, logging and monitoring system building, and version control. When you deploy a workload, we recommend that you select the workload type and service release method based on your business demand. During workload processing, we recommend that you conduct inspections and service upgrades on a regular basis, and minimize resources costs. All of the above-mentioned ACK capabilities help you better utilize Alibaba Cloud services.

For more information, visit https://www.aliyun.com/product/kubernetes.

Application Hosting Platform EDAS

In the cloud-native era, the PaaS platform will be integrated into the infrastructure and become a part of the cloud. EDAS is a PaaS that provides enterprises with application hosting and microservice management. It empowers enterprises to release, run, and manage their cloud businesses efficiently.

- Multiple choices of underlying servers

With EDAS, you can select ECS instances and container services for Kubernetes clusters based on your hosting needs. With the evolution of integration from ECS clusters to Kubernetes clusters, EDAS seamlessly integrates container service for Kubernetes and provides a new experience of lightweight O&M and application lifecycle management for Kubernetes clusters. In addition to hosting services of Kubernetes clusters, it also supports one-click, multi-availability-zone deployment of application instances, application version management, tracking changes, and extreme elasticity at the application layer. EDAS provides you a one-stop PaaS service.

- Multiple publishing modes

EDAS allows you to use the console, APIs, CLI, and SDK for deployment. It also supports continuous application integration through Yunxiao and Jenkins and supports the automatic deployment of applications through Cloud Toolkit.

- Multiple deployment methods

Multiple deployment methods are supported, such as WAR packages, JAR packages, and images.

- Application management

EDAS provides you with full lifecycle management services from creation to the launch of an application, including publishing, starting, stopping, scaling out, scaling in, and deleting. With Alibaba's rich O&M experience in ultra-large-scale clusters, EDAS allows you to easily operate and maintain those applications that are running on thousands of instances.

- Auto scaling

Auto Scaling is a service to automatically adjust computing resources based on your volume of user requests. When the demand for computing resources increases, Auto Scaling automatically adds ECS instances to serve additional user requests or removes instances in the case of decreased user requests.

- Monitoring and management integration

Alibaba Cloud upgraded the security protection mechanism to integrate application monitoring and management. You can check and measure, mark in grey, roll back the applications, perform automatic monitoring, intelligent diagnosis, and report generation in more dimensions.

- Canary release

Canary allows you to set the grey traffic and the number of releases in a batch to gradually scale up your traffic based on the situation, greatly reducing the risk of deployment.

- Full-link traffic control

Supports multiple grey control at all stages of a business system, identifies the grey traffic based on specified rules, and guides to the deployment group that corresponds to the downstream applications to achieve fast and flexible multi-application, multi-grey control. Resource costs are greatly reduced.

Key Resources: https://www.aliyun.com/product/edas

Function Compute

Function Compute is an event-driven, fully managed computing service. When you use Function Compute, you only need to design and upload your code without having to purchase and manage any infrastructure. Function Compute provides compute resources to run scalable tasks reliably. It also provides features such as log query, performance monitoring, and alarms. With the integration of various events in Function Compute, you can build scalable, reliable, and secure applications and services, or complete a set of multimedia data processing back-end services in a few days. When the event source triggers an event, it automatically invokes the associated functions to process the event. For example, function processing is automatically triggered when you create a new object or delete some objects in the Object Storage Service (OSS), or when the API Gateway receives an HTTP request.

- Cold start performance

Cold start is a classic FaaS problem. Function Compute has made many optimizations to cold start. After optimizing the code package loading, container startup speed, and whole link, it has significantly reduced the time to cold start a function. A simple code package successfully starts in 500 ms and a typical 10 MB code package starts in one second.

- Warm start performance

In continuous call scenarios, Function Compute caches the hotspot information and schedules the load balance to achieve an average system latency of about 10 ms.

- Speed to scale up

Based on the powerful infrastructure provided by Alibaba Cloud, Function Compute can scale up to 500 servers per minute.

- TPS and instances

Function Compute is designed with full consideration for system scalability. It performs automatic load balance through scheduling. Theoretically, it supports an unlimited number of TPS and instances. The TPS of the current system reaches up to 10 W/s at peak time.

Key Resources: https://www.aliyun.com/product/fc

4.1.2. Storage

As the largest storage service provider in China, Alibaba Cloud is committed to helping enterprises break through the boundary between data storage and circulation. Alibaba Cloud wants to meet the diversified needs of different enterprises and provide core data value for enterprises’ digital transformation.

After more than a decade of technology development and accumulation, Alibaba Cloud has provided a complete storage portfolio of enterprise-level, high-performance products, covering the public cloud, private cloud, and hybrid cloud. Alibaba Cloud offers the most complete storage products of any cloud computing vendor in the world. The Alibaba Cloud Storage family includes Object Storage Service (OSS), Enhanced Block Storage (EBS), Network Attached Storage (NAS), Cloud Paralleled File System (CPFS), TableStore, and Hybrid Backup Recovery (HBR).

4.1.2.1. Performance Overview

Different business scenarios have varied I/O load characteristics and have various requirements for storage products. When you select your storage products, we recommend that you take into account the input/output operations per second (IOPS), throughput, latency, shared access or not, sequential or random read/write, data persistence, and other factors to meet the data access requirements of upper-layer applications.

In terms of performance, you need to focus on the following core metrics: IOPS, throughput, and latency. The following sections describe the features of block storage, file storage, and object storage based on these metrics. We recommend that you fully understand these metrics and use them as the basis for choosing a proper storage product.

Category	Type	IOPS	Throughput	Latency	Recommended business scenarios
Block storage	Cloud disk or local disk	Up to million IOPS per disk	Maximum throughput per disk: 4GB	100 microseconds Level	For OLTP businesses, low-latency and high random access are required.
File storage	Capacity NAS, performance NAS, and CPFS	Millions of IOPS	Hundreds of GB level	Millisecond level	Business scenarios requiring shared storage access, such as containers, high-performance computing, AI training, genetics, EDA, and autonomous driving
Object storage	Standard, low frequency, and archive	Scalable to millions of QPS	Scalable to TB level	Millisecond level	Scenarios such as data lakes, mobile apps, big websites, image sharing, and hot audio and video playback.

Metrics description

IOPS

Input/output operations per second (IOPS) measures the number of write/read operations that can be performed each second. IOPS directly affects the performance of transaction-intensive applications, such as databases.

Throughput

The throughput measures the size of data transferred per second, in GB/s or MB/s. Throughput directly affects the performance of applications that require a large number of read/write operations, such as Hadoop offline computing applications.

Latency

Latency is the time required to process an I/O request, in the unit of s, ms, or us. High latency can lead to performance degradation or service suspension.

4.1.2.2. Key products

Block storage

Block storage is a high-performance and low-latency block storage service provided by Alibaba Cloud for ECS, ECS bare metal instance, and containers. It supports random read and write operations, including cloud-based disks that can be created based on the distributed storage architecture, and local disks that are based on the local disks of physical servers. Block storage is similar to a hard disk. You can format a block storage device and create a file system on it to meet the data storage needs in a general business scenario.

Performance metrics

The metrics for measuring storage performance include IOPS, throughput, and latency. Some block storage devices also have requirements on the capacity. For example, those ESSDs with different performance levels have varied capacity ranges.

Category	Enhanced SSDs		
Performance Level (PL)	PL3	PL2	PL1
Single disk capacity	1,261 – 32,768 GiB	461 – 32,768 GiB	20 – 32,768 GiB
Max IOPS	1,000,000	100,000	50,000.
Max throughput	4,000 MB/s	750 MB/s	350 MB/s
Single-channel random write latency	0.2 ms		

Relations with instance performance and configuration

The enterprise-level ECS instance families feature the isolation capability of storage I/O performance. Dedicated storage bandwidths are assigned to ECS instances and disks to avoid the storage I/O preemption among different ECS instances. The new generation of enterprise-level instance families ensures stable and consistent storage I/O performance of applications, especially at peak time.

If your business application is I/O-sensitive and requires consistent storage I/O performance, we recommend that you select a new generation instance family with isolation capability of storage I/O performance:

- Large and medium-sized databases, such as Oracle, MySQL, SQL Server, PostgreSQL, Cassandra, and MongoDB databases.
- Enterprise-level applications, such as ERP and CRM.

The storage I/O performance of an ECS instance and NAS varies with different instance families. The storage I/O performance of an instance depends on the type of instance you choose. A higher instance type provides stronger storage I/O performance (IOPS and throughput).

After you create an ECS instance and choose an attached NAS, the final storage I/O performance is:

Scenario 1: If the total storage performance required by the attached disks exceeds the performance capability that the instance type can deliver, the final storage I/O performance is the value of the instance type.

Scenario 2: If the total storage performance required by the attached disks does not exceed the performance capability that the instance type can deliver, the final storage I/O performance is the value of the attached disks.

For example: We take the ecs.g6.8xlarge instance type as an example, which supports up to 60,000 IOPS. If you attach a 1,600 GIB ESSD PL2 disk to the instance and its IOPS is 81,800, the maximum storage IOPS will be 60,000 instead of 81,800.

After you learn the relations between instance storage performance and disk storage performance, you can choose the proper instance types and block storage devices based on the performance data that meets your business needs and avoid application performance bottlenecks caused by improper configurations.

Apsara File Storage NAS

Alibaba Cloud Apsara File Storage NAS is a file storage service for compute nodes such as ECS instances, E-HPC, and container services. It is a shared, scalable, reliable, and high-performance distributed file system.

NAS provides a wide variety of features, such as scalable capacity, shared access, support from many standard protocols, secure and compliant, encrypted, and flexible in data access, data transmission, and data backup. NAS can be mounted to any type of computing products, such as ECS, ACK, EHPC, and bare metal instances. By using the network connection, you can mount the storage to a compute node at different locations across VPCs or regions.

Combined with different product performance and its market positioning, NAS is divided into multiple types based on its capacity and performance.

Capacity NAS

Capacity NAS uses SATA HDDs as storage media, providing high-performance storage at relatively low costs. It applies to shared file storage services that require high throughput, scalability on demands, and cost-efficiency. It provides better cost benefits for services that require infrequent reads/writes and have much margin on latency response.

Performance Specifications:

- Capacity: 10 PB
- Latency: milliseconds
- IOPS: up to 15,000 (4,000 random read/write)
- Throughput: linear scale up to 10 GB/s

Performance NAS

Performance NAS uses SSDs as storage media, providing high throughput and IPOS, and low latency performance for workloads. It applies to shared file storage services that require higher throughput, scalability on demands, and low read/write latency. It provides performance benefits for services that require frequent read/write operations and quick response.

Performance Specifications:

- Capacity: 1 PB
- Latency: milliseconds
- IOPS: maximum 30 KB (4 KB random I/O read/write on hard disk)
- Throughput: linear scale up to 20 GB/s

Extreme NAS

Extreme NAS is a high-performance shared file storage system based on a new generation of network architecture and all-flash storage. This fully managed cloud storage service is integrated with Alibaba Cloud’s compute services to fully deliver the optimal computing capability of the public cloud.

Performance Specifications:

- Capacity: up to 256 TB
- Latency: 100 microseconds with optimized OPS performance for small files
- IOPS:10 - 200 K
- Throughput: the initial bandwidth is 150 Mbit/s and it can be scaled up to 1,200 Mbit/s.

Cloud Paralleled File Storage (CPFS)

CPFS is a fully managed and scalable parallel file system that meets the requirements for high-performance computing. CPFS provides a unified namespace that allows simultaneous access from hundreds of clients. It provides a throughput of hundreds of Gbit/s, an IOPS of millions, and latency of sub-milliseconds.

CPFS can provide high-performance and cost-effective computing and storage for scenarios such as genetic computing, petroleum exploration, meteorological analysis, machine learning, big data analysis, and media file processing.

CPFS can provide a peak bandwidth of hundreds of GB, an IOPS of millions, and latency of sub-milliseconds. The specific bandwidth and IOPS varies depending on the file system that you purchase. For a CPFS file system of 50 TB, it can provide a bandwidth of 5 GB and about 350,000 IOPS.

After a file system is created, the capacity and performance will be a fixed value that you cannot directly change. However, you can upgrade your current file system. The performance improves when you select a higher specification.

Object Storage Service (OSS)

Alibaba Cloud OSS is a cloud storage service provided by Alibaba Cloud that is secure, low-cost, highly reliable, and can store massive amounts of data. The data persistence is no less than 99.9999999999% and the service level availability (or business continuity) is no less than 99.995%.

Using RESTful API interfaces provided by OSS, you can store and access any type of data anytime, anywhere from any web application.

You can use API and SDK interfaces provided by Alibaba Cloud or OSS migration tools to migrate massive amounts of data into or out of Alibaba Cloud OSS. You can choose the Standard type to store mobile apps, big websites, images, and audio and video files. You can also choose Infrequent Access (IA) storage and Archive storage to store infrequently accessed data for a long time.

Performance metrics

The default bandwidth and QPS for upload or download by an account in a region are:

- Bandwidth

10 Gbit/s for regions in China Mainland and 5 Gbit/s for other regions. If the threshold is reached, your request may be restricted.

- QPS: 10,000 times/s.

If the threshold is reached, extra requests will be rejected.

If your offline big data processing requires higher bandwidth (10 to 100 Gbit/s) or more QPS, please contact Alibaba Cloud after-sales support.

Transfer acceleration

Alibaba Cloud OSS provides a transfer acceleration service to help enterprises deploy their business globally and improve the user's upload and download experience. OSS transfer acceleration is designed to accelerate uploading and downloading data to/from cloud storage on the Internet. By using a smart scheduling system, optimized route selection and protocol stacks, and customized transfer algorithms, it can be an end-to-end acceleration solution.

OSS uses data centers distributed globally to implement transfer acceleration. When a data transfer request is sent, it is resolved and routed over an optimal network path and protocol to the data center where your bucket resides.

OSS transfer acceleration applies to accelerate access and improve user experience.

- Accelerates remote data transfer

Some customers may experience poor upload and download experiences due to long transmission distance, such as global forums and top online collaborative office platforms. In this case, OSS transfer acceleration can be used to allow users from different regions to transfer data over an optimal network. This service can accelerate data transfers and improve access experience for users across different regions.

- Upload and download large files by GB or TB

To upload or download large files over the Internet at a long distance, you can choose this service to accelerate your transfer. Transfer acceleration is based on the Internet transfer route selection and protocol stack optimization, so it can greatly reduce the timeout proportion during data transmission. Uploading with sharding allows you to re-transfer if there are any errors. You can integrate sharding uploads with transfer acceleration to upload and download large files from a long distance.

This allows an acceleration of downloading non-static and non-hotspot data. For example, on album apps, games, e-commerce, comments on social media apps, enterprise portals, and financial apps, the user's download experience may directly affect product competitiveness and customer retention. OSS transfer acceleration is a service designed to accelerate OSS uploads and downloads. You can enable transfer acceleration to maximize bandwidth utilization to accelerate data transfer.

Performance documentation

Refer to the following documents to learn more about product details about Alibaba Cloud's storage performance.

- Block Storage Performance: https://www.alibabacloud.com/help/doc-detail/25382.htm

- Storage I/O performance of instances: https://www.alibabacloud.com/help/doc-detail/147898.htm
- Performance tests on Block Storage: https://www.alibabacloud.com/help/doc-detail/147897.htm
- Apsara File Storage NAS Performance: https://www.alibabacloud.com/help/doc-detail/61136.htm
- Apsara File Storage NAS Extreme: https://www.alibabacloud.com/help/doc-detail/124577.htm
- Performance testing for Apsara File Storage NAS: https://www.alibabacloud.com/help/doc-detail/95501.htm
- Object Storage Service Limits: https://www.alibabacloud.com/help/doc-detail/54464.htm
- Transfer acceleration: https://www.alibabacloud.com/help/doc-detail/131312.htm

Product Documentation

Refer to the following documents to learn more about Alibaba Cloud's storage products.

- Object Storage OSS: https://www.alibabacloud.com/help/product/31815.htm
- NAS: https://www.alibabacloud.com/help/product/27516.htm
- EBS: https://www.alibabacloud.com/help/doc-detail/63136.htm
- Hybrid Cloud Array Hybrid Cloud Storage Array: https://www.alibabacloud.com/help/product/53942.htm
- Hybrid Backup Recovery: https://www.alibabacloud.com/help/product/60939.htm

4.1.3. Databases

Alibaba Cloud provides a wide portfolio of cloud database solutions to meet your data storage, processing, analysis, and management needs, promote your business development, and enhance your enterprise value. Our database systems provide comprehensive support for all mainstream, open-source, and commercial database solutions.

Regardless of your enterprise scale, the Alibaba Cloud database can boost your business development. The classic database solutions like RDS and PolarDB support the entire Alibaba system during the Double 11 holiday (a shopping day similar to the USA's Black Friday) every year.

4.1.3.1. Ecosystem

ApsaraDB for RDS is a stable, reliable, and scalable online database service. It supports more than 90% of the world's mainstream open source and commercial databases (such as MySQL, SQL Server, and Redis) in operation and maintenance. It provides ApsaraDB for PolarDB with more than six times the performance of a normal open source database. In addition, it has features like, disaster recovery, backup, recovery, monitoring, and migration.

- ApsaraDB for RDS provides a comprehensive solution for hosting relational databases. It supports ACID and SQL standards to quickly meet complex business scenarios. It features extreme performance metrics, including a maximum of 137,000 QPS and 100,000 concurrent requests.
- ApsaraDB for NoSQL supports databases such as caches, documents, and column-based storage that can easily handle business spikes. It features ultimate performance metrics: up to tens of millions of QPS and PB-level data storage.
- ApsaraDB for Data Warehouse supports HTAP databases that can process online transactions, analysis for massive amounts of data, and its relevant tools. It features PB-level horizontal scalability. You can scale up PB-level expansion in 10 ms.
- Data Transmission Service (DTS) is an integrated database service that supports data migration and synchronization between various data stores. Database Backup Service (DBS) can allow you to back up databases from Alibaba Cloud, on-premises IDCs, hybrid cloud, or third-party clouds. Data Management Service (DMS) integrates online database development and visualized tools.

4.1.3.2. Multiple choices

ApsaraDB provides a variety of products to choose from based on your workload.

- ApsaraDB for PolarDB

This one-master-multiple-slave architecture allows all read/write and read-only nodes on an instance to access a data replica at the same time. This can significantly reduce your storage costs. It supports master/slave switchover with zero data loss. This resolves the issue of data inconsistency between read-only and read/write nodes caused by asynchronous replication. It takes only a few minutes to expand read-only replicas, backup, and restore the data.

- Lindorm

It uses a high-availability architecture with master/slave mode and can detect high availability (HA) in real-time. When a core node fails, the service can failover to another region in seconds. Each core node can handle up to 100,000 QPS and provide up to 8 TB of storage space. Flexible scaling of disk and nodes allows you to easily scale up to thousands of instances, supporting 10 million QPS and several petabytes of storage space. It is fully compatible with open source HBase and can be fused with Spark, Phoenix, and Solr.

- ApsaraDB for Cassandra

It uses a masterless architecture and is always online. Your applications will not encounter any performance jitter when a single node in a cluster fails. It can meet some strict scenarios that have strict requirements on service uptime. Initial configuration is two nodes, which is a low threshold for configuration. It supports up to 500 nodes in a cluster, provides PB-level storage, and tens of millions of OPS. It features enterprise-level capabilities such as multi-DC disaster recovery, backup and restoration, security, monitoring, and online scale up/down.

- ApsaraDB for Redis

It provides different architectures, such as single node, master/slave hot-standby, read/write splitting, and distributed cluster architecture. Single node applies to cache-only scenarios. The master-slave architecture supports automatic failover. Read/write splitting can be used in scenarios with more reads than writes. Distributed clusters can be elastically scaled with one click and it theoretically has limitless performance. The read/write splitting architecture applies to scenarios with more reads than writes, providing a write performance of up to 100,000 QPS, and read performance of 600,000 QPS, breaking through the performance bottlenecks for hotkey reads.

- AnalyticDB for MySQL, a cloud-native data warehouse

It uses a new generation, ultra-large-scale MPP+DAG fusion engine, and adopts hybrid row-column storage technology, automatic indexing, and an intelligent optimizer. It quickly discovers your data value. It can perform instant multi-dimensional analysis and perspective on hundreds of billions of data entries in just a few minutes. It supports multiple storage modes including HDD, SSD, and block storage. It supports the elastic separation between computing resources and data resources and can scale from 100 GB up to dozens of PB. This makes it easier to build the most cost-effective resource allocation. Meanwhile, it allows you to write and update large amounts of data in real-time and perform extract-transform-load (ETL) operations. This is the best option for building an enterprise-level cloud data warehouse.

4.1.3.3. Key database services

ApsaraDB for PolarDB, ApsaraDB for Redis, and AnalyticDB for MySQL are critical data services provided by Alibaba Cloud for database solutions that provide relational, NoSQL, and data warehouses, respectively. These services have a large number of configuration options, allowing you to further optimize your storage solution.

For more information about Alibaba Cloud database best practices, see the following document: Alibaba Cloud Database

4.1.4. Networking

As China's biggest networking solutions provider, Alibaba Cloud enables users to explore a range of high-performance, elastic, cost-effective, and fully-automated networking options that integrate seamlessly across global networks. Alibaba Cloud has been at the forefront of cloud networking technology for the past decade, developing and applying technological advances to power businesses with private, public, and hybrid cloud high-performance networking solutions. Alibaba Cloud has established the world's most comprehensive product portfolio in networking that boosts connectivity for on-premises, cross-region, and access networks. Some of the featured Alibaba Cloud networking products include Virtual Private Cloud (VPC), Elastic IP Address (EIP), Network Address Translation (NAT) Gateway, Virtual Private Network (VPN) Gateway, Server Load Balancer (SLB), Express Connect, Cloud Enterprise Network (CEN), and Smart Access Gateway (SAG).

4.1.4.1. Performance Overview

Alibaba Cloud networking products can be classified into two categories in terms of their applications: products that assist in network topology management, such as VPC and CEN, and products used for network traffic processing, such as SLB and NAT Gateway. Therefore, the performances of these networking products are measured with different sets of criteria. One is based on capacity or specifications. You can refer to the sections below for related information about each product.

The other approach assesses the forwarding or processing capability of a product. The following are some key metrics:

- Packets per Second (PPS): the number of data packets that can be forwarded or processed per second (packet-size dependent).
- Bytes per Second (Bps): the number of bytes forwarded or processed per second.
- Concurrent connections: the number of simultaneous connections.
- Connections per Second (CPS): the number of new connections per second.
- Queries per Second (QPS): the number of requests forwarded or processed per second.

4.1.4.2. Key Alibaba Cloud Services

VPC

Product overview

VPC allows you to define a virtual network within Alibaba Cloud. You can provision your own logically-isolated section of Alibaba Cloud, similar to implementing an independent network that would operate in an on-premises data center. VPC provides you with fine-grained control over your virtual infrastructure, allowing you to specify your own IP address range and configure route tables and network gateways. You can launch Alibaba Cloud resources within a VPC, such as Elastic Compute Service (ECS) instances, ApsaraDB for RDS instances, and SLB instances.

Performance metrics

Performance category	VPC	
	Standard VPC instances	Enterprise-level VPC instances (contact support for this feature)
Maximum number of route tables per VPC	10	10
Maximum number of route entries per route table	48	100,000
Maximum number of VSwitches per VPC	24	100
Maximum number of virtual machines (VMs) per VPC	60,000	100,000
Maximum number of network access control lists (network ACLs) per VPC	24	100
Maximum number of rules per network ACL	The limit for ingress rules is 20, and the limit for egress rules is 20.	The limit for ingress rules is 100, and the limit for egress rules is 100.
Maximum internal network bandwidth	500 Gbit/s per zone 100 Gbit/s across zones (we suggest removing this metric)	2 Tbit/s per zone 200 Gbit/s across zones (we suggest removing this metric)

Reference documentation: https://help.aliyun.com/product/27706.html

Elastic IP Address

Product overview

EIPs are public IP addresses that are independent of any instance. Currently, you can associate an EIP with an ECS instance within a VPC, an SLB instance within a VPC, a secondary Elastic Network Interface (ENI), a NAT gateway, or a High-Availability Virtual IP Address (HaVip).

Performance metrics

Peak bandwidth

Pay-by-data-transfer EIP	Pay-by-bandwidth EIP
Each pay-by-data-transfer EIP supports a peak bandwidth of 200 Mbit/s. This value signifies the maximum allowable bandwidth and is for reference only. It is not a service-level guarantee. In the event of resource contention, the peak bandwidth may be limited.	Each pay-by-bandwidth EIP supports a peak bandwidth of 500 Mbit/s. This is a service-level guarantee. In the event of resource contention, the peak bandwidth is guaranteed.
The cumulative sum of the peak bandwidth of all pay-by-data-transfer EIPs in a region cannot exceed 5 Gbit/s. If your service requires a guaranteed or higher peak bandwidth, you must use pay-by-bandwidth EIPs instead.	The cumulative sum of the peak bandwidth of all pay-by-bandwidth EIPs in a region cannot exceed 50 Gbit/s. You can contact your sales representative to request a quota increase.

Instances

A maximum of 20 EIPs can be created per account. You can request a quota increases to adjust the limit.

Reference documentation: https://help.aliyun.com/product/61789.html

Server Load Balancer

Product overview

SLB is a load balancing service that distributes traffic among multiple ECS instances. By spreading the workload evenly, SLB improves application responsiveness. You can also use SLB to eliminate single points of failure (SPOFs) to ensure high availability for your applications.

Performance metrics

Performance category	SLB	Maximum allowed quota after increase
Maximum throughput (bandwidth)	5G	40G
Maximum number of concurrent connections	5 million	50 million
QPS	100,000	300,000
Maximum connections per second	500,000	500,000

Reference documentation: https://help.aliyun.com/product/27537.html

NAT Gateway

Product overview

A NAT gateway is an enterprise-class public network gateway deployed in a VPC. It enables cross-zone disaster recovery and supports source NAT (SNAT) and destination NAT (DNAT) functions. You can use SNAT rules as a more flexible approach to allow traffic from VSwitches or ECS instances to go out to the public network. DNAT rules can be created to translate incoming traffic to servers with port mapping or IP mapping. NAT Gateway can also be used with Internet Shared Bandwidth for enhanced performance and flexibility.

Performance metrics

The key performance metrics of NAT Gateway include Bps, concurrent connections, and CPS. For NAT 1.0, metric values vary by instance specification. For NAT 2.0, maximum throughput is added as a new performance metric and the values of other metrics are several times higher than those of NAT 1.0.

NAT1.0

Performance metrics	**Small**	**Medium**	**Large**	**Super Large**
Maximum number of concurrent connections	10000	50000	200000	1000000
Maximum CPS	1000	5000	10000	30000

NAT2.0

Performance metrics	**NAT Gateway**	**Enhanced NAT Gateway**
Maximum number of concurrent connections	500w	2000w
Maximum CPS	100w	250w
Maximum throughput	12G	100G

Reference documentation: https://help.aliyun.com/product/44413.html

VPN Gateway

Product overview

VPN Gateway builds encrypted tunnels as secure connections between on-premises networks, remote offices, client devices, and the Alibaba Cloud network over the public network. It supports both IPsec and SSL VPN connections, which allows you to establish reliable data transmission with the most suitable solution.

Performance metrics

The key performance metrics of VPN Gateway include maximum throughput, PPS, and maximum number of SSL VPN connections.

Performance category	IPSEC-VPN	SSL-VPN
Maximum throughput	1G	1G
Maximum number of connections	10 peering connections. You can submit a ticket to raise the limit to a maximum of 100.	Default value: 50. You can submit a ticket to raise the limit to a maximum of 1,000.

Reference documentation: https://help.aliyun.com/product/65234.html

Cloud Enterprise Network

Product overview

CEN allows you to create a global network that ensures optimal connectivity and communication for your business with dynamic routing and fast convergence. By using CEN, you can establish secure, private, and enterprise-class interconnectivity between VPCs or between VPCs and on-premises data centers.

Performance metrics

Item	**Limit**
Maximum number of CEN instances per account	5
Maximum number of transit routers per CEN instance	1
Maximum number of route tables per transit router	20
Maximum number of route entries per transit router	2,000
Maximum processing capability per transit router	200 Gbps
Maximum bandwidth for VPC connections within mainland China and between other areas	10 Gbps

Reference documentation: https://help.aliyun.com/document_detail/59870.html

Express Connect

Product overview

Express Connect helps you build high-speed, stable, and secure connections between on-premises and cloud networks across internal networks. Different from Internet service provider (ISP) connections, the physical connections of Express Connect feature improved stability and are free of the risk of data interception.

You can use a leased line to establish a physical connection between your on-premises data center and an access point of Alibaba Cloud, with one end connected to the gateway device at the data center, and the other connected to the Virtual Border Router (VBR) associated with the physical connection. After completing the installation and required configuration, you can use physical connections for reliable and faster communication with minimal latency.

Benefits

The benefits of Express Connect are as follows:

- High-speed interconnection

  Taking advantage of the network virtualization technology of Alibaba Cloud, Express Connect creates direct communication channels between different network environments as data travels through private networks and other than the public network. With Express Connect, no matter how long the distance is, low-latency and high bandwidth are always guaranteed.

- Stability and reliability

  Built on the state-of-the-art infrastructure of Alibaba Cloud, Express Connect ensures stable and reliable communication between networks.

- Security

  Express Connect implements cross-network communication at the network virtualization layer, where all data is transmitted through the tenant-isolated infrastructure of Alibaba Cloud. This eliminates the risk of data theft during transmission.

- Pay-as-you-go billing

  A range of bandwidth specifications are available for you to choose from based on your specific needs.

Performance metrics

Alibaba Cloud provides a range of access ports with up to 100GbE for a single port. If you require a higher specification, you can contact the product designer of networking products. Note that such changes would require multiple SLBs and a dedicated Express Connect access cluster.

Global Accelerator

Product overview

Global Accelerator (GA) is a service that accelerates your latency-sensitive applications on a global scale. By leveraging the high-quality bandwidth and reliable transmission network of Alibaba Cloud, GA delivers a highly available and high-performance architecture where incoming traffic can be accepted from the location closest to end users. GA can also be deployed across regions.

Performance metrics

Performance category	GA
Maximum throughput (bandwidth)	5G
Maximum number of concurrent connections	1 million

Reference documentation: https://www.aliyun.com/product/ga

Smart Access Gateway

Product overview

SAG is a cloud native Software-Defined Wide Area Networking (SD-WAN) solution developed by Alibaba Cloud. Enterprises can utilize SAG as an intelligent gateway to access cloud resources with greater reliability and security.

Benefits

SAG provides the following benefits:

- Intelligence: With highly automated configurations and Zero Touch Provisioning (ZTP), SAG automatically adapts to the fast convergence in the network topology.
- Reliability: SAG adopts nearby public network access within a city. Additionally, it enables access by multiple hosts to Alibaba Cloud through the device-level or link-level active/standby setting.
    - Device-level disaster recovery: Dual-device active/standby failover is implemented so that the traffic is immediately distributed to the standby device when the active device fails.

    - Link-level disaster recovery: All SAG devices use encapsulated dual-link access. The optimal link is automatically detected and designated as the active link. Traffic is distributed to the standby link when the active link fails.
    - Access point disaster recovery: Multiple access points are assigned to each SAG instance. If an access point fails, the system automatically switches over to another access point.
- Security: SAG optimizes security by encrypting the traffic within the hybrid cloud and all data transmitted over the public network.
    - Data encryption: Both the IKE and IPsec protocols are used to encrypt the transmitted data to guarantee data security.
    - Anti-replay: The data source is authenticated to prevent replay attacks.
    - Anti-tampering: Multiple authentication methods are used for verification.
- Centralized management
    - The Alibaba Cloud console is the central platform for managing and configuring SAG devices.

Performance metrics

- An SAG-100WM device can be placed on a table or in an electrical instrument box. The WAN ports support 4G and broadband connections. The LAN ports support wired Ethernet and Wi-Fi connections. The maximum bandwidth of encrypted transmission over a private network is 50 Mbit/s (512 bytes). The SAG-100WM model is recommended for quick access from small and medium-sized branches.
- A SAG-1000 device can be placed in a rack. The WAN port supports an assembly of leased line, broadband, and 4G connections. The LAN port supports wired Ethernet connections. The maximum bandwidth of encrypted transmission over a private network is 500 Mbit/s (512 bytes). The SAG-1000 model is recommended for access from large-sized branches and on-premises data centers.
- SAG-VCPE can be installed on servers, edge computing instances, or virtual machines. The current version supports a maximum bandwidth of 500 Mbit/s (512 bytes) for encrypted transmission over a private network. SAG-VCPE is recommended for accessing Alibaba Cloud with your own device or from another cloud. You can contact the product designer if you need a higher bandwidth.

Performance-related documentation

The following articles provide additional details regarding the performance metrics of each Alibaba Cloud networking product.

- VPC: https://help.aliyun.com/document_detail/27750.html
- EIP: https://help.aliyun.com/document_detail/54479.html
- SLB:
    - https://help.aliyun.com/document_detail/32459.html
    - https://help.aliyun.com/document_detail/85966.html
    - https://help.aliyun.com/document_detail/106192.html
- NAT Gateway: https://help.aliyun.com/document_detail/32382.html

- VPN Gateway: https://help.aliyun.com/document_detail/65242.html
- CEN: https://help.aliyun.com/document_detail/64647.html
- Express Connect: https://help.aliyun.com/document_detail/44849.html
- GA: https://help.aliyun.com/document_detail/153192.html
- SAG: https://help.aliyun.com/document_detail/69234.html

4.1.4.3. Product documentation

The following articles provide detailed introductions to Alibaba Cloud networking products.

- VPC: https://help.aliyun.com/product/27706.html
- EIP: https://help.aliyun.com/product/61789.html
- SLB: https://help.aliyun.com/product/27537.html
- NAT Gateway: https://help.aliyun.com/product/44413.html
- VPN Gateway: https://help.aliyun.com/product/65234.html
- CEN: https://help.aliyun.com/document_detail/59870.html
- Express Connect: https://help.aliyun.com/product/27782.html
- GA: https://help.aliyun.com/document_detail/153189.htm
- SAG: https://help.aliyun.com/document_detail/69227.html

4.2. Measurement

When you first build your scheme, you can estimate the performance baseline based on your familiar product and process as a template. As time goes by, demands and technologies are constantly evolving. A certain measure is required to evaluate your scheme performance, establish a baseline, and record the performance efficiency of your scheme.

You must establish a performance measurement process that includes:

- Clearly defined metrics

It must establish metrics and monitoring mechanisms to collect key performance metrics. We recommend that you integrate both technical and business metrics. For websites or mobile applications, the most important indicators are response time to request and error rate. System-level indicators include the number of threads, garbage collection rates, and waiting status.

- Detailed historical performance baselines

Based on each measurement, you will set up detailed performance definition metrics.

- Automated performance test cases

It should be possible to automatically trigger performance test cases after each change.

You should build a series of test cases to ensure that you can track performance changes over time. For some test cases that need to run for a long time, you can process the change asynchronously, or postpone it to the evening.

- Stress testing

You should create a series of test scripts to run in idempotence, disorder, and in order. The purpose is to check your system performance under a high workload. You can use the Performance Testing Service (PTS) to generate results that clearly show how many workloads are running currently.

- Performance data visualization

Each key indicator should be able to make relevant roles understand. In this way, each role can understand the significant performance changes. This performance data should also contain business indicators to ensure that the target system completes the process as designed.

- Visualization

Use visualization technologies to identify performance issues, hotspots, waiting states, and locations of low utilization. The system can overlay performance metrics on the corresponding architecture chart and call graphs or code to identify problems faster.

4.2.1. Performance test

A stress test can measure the overall performance of a system in different loads and production environments based on your business scenarios. Stress tests will be performed using combined or simplified versions of production data, removing sensitive or identity information.

You can choose traffic playback or manual orchestration of user scenarios to practice the overall system performance. During the drill, make sure that your system can distinguish between drill data and real requests. At the same time, you must have preset key performance indicators and be able to measure these key performance indicators in real-time. You can compare the two factors to ensure that the system performance changes are compared. The Alibaba Cloud performance testing tool (PTS) can evaluate the system performance metrics of infrastructure resources. You can also define your key performance metrics by using the architecture awareness component of the Application High Availability Service (AHAS). You can also use AHAS to configure protection measures to avoid system crashes when key metrics exceed the warning threshold.

You can make full use of the inherent advantages of Alibaba Cloud in the stress testing field, including building traffic that reflects users' geographical locations, orchestrating stress testing scenarios superior script adaptation capabilities, comprehensive customization and parameterization capabilities, rich stress testing flow control, original stress testing mode and second-level control, complete capacity evaluation and inflection point identification, and supporting second-level monitoring. You can discover potential bottlenecks at an extremely low cost before having an impact on real business scenarios.

When writing key business test cases, you should include specific performance requirements for various scenarios. In this case, a performance baseline for the business scenario must be created to support the relevant test scripts. This ensures that you can track key business scenarios throughout the process of performance changes.

Stress tests often affect online businesses. We recommend that you select a proper write or read model based on your business requirements. We also recommend that you use protective measures to mitigate DDoS attacks when traffic exceeds system traffic.

4.2.1.1. Critical Alibaba Cloud services

The critical service at the stress testing layer is PTS, which provides a variety of stress testing tools and collects key metrics that can explain the architecture performance. With PTS, you can also establish a performance baseline.

4.2.1.2. Related resources

Application high availability service (AHAS) can quickly locate performance bottlenecks and automatically protect workloads that exceed the threshold. For more information, see the following website: https://www.aliyun.com/product/ahas

4.2.2. Bottleneck locating

Due to the diversity of open source software and cloud services, the heterogeneity of development languages, and the organizational and capability differences of IT teams in enterprises, standardization has become more complicated. You can use architecture awareness to easily measure structural changes and locate the causes of performance changes.

The Application High Availability Service (AHAS) collects data from the operating system, standard third-party interfaces, and detects process-level call relationships. It uses the feature library algorithm to identify the technical components used by the processes. Finally, the application architecture is displayed visually in three areas: server, container, and process. When the structure or performance changes, architecture awareness can quickly locate the relative changes.

4.2.2.1. Key Alibaba Cloud services

The key to identifying bottlenecks in visualization is the Alibaba Cloud service AHAS. This provides standard Alibaba Cloud APIs to quickly identify application architecture changes.

4.2.3. Monitoring

Alibaba Cloud provides two monitoring systems for different scenarios. CloudMonitor provides monitoring for Alibaba Cloud infrastructure products and services, such as ECS, RDS, and SLB. ARMS provide monitoring of customers' server applications, frontend applications, and mobile applications.

4.2.3.1. CloudMonitor

During the construction of the architecture, you can use the functions of cloud monitoring to discover and solve potential problems. By monitoring the metrics and events of infrastructure and services, alarms can be triggered when the threshold is reached or a specific event occurs. These alarms can also be used to automatically trigger some automated O & M actions.

CloudMonitor can send alert notifications to you by phone, text message, email, or DingTalk. It also supports a callback API to customize alerts. In addition, the CloudMonitor alarm service integrates with log service, MNS, Function Compute, ESS, and O & M orchestration services. These services can be used to perform O & M operations to solve performance problems.

CloudMonitor provides the following monitoring services:

- **Host monitoring**

It provides monitoring capabilities against servers, covering the Alibaba Cloud ECS and the customer's servers on the cloud. After you install the host monitoring agent of CloudMonitor, it collects basic monitoring information of the servers such as CPU, memory, disk, network, and other basic indicators of some columns. By monitoring the thresholds of these metrics, you can detect the performance problems of the server as soon as possible. You can also monitor processes in host monitoring to quickly discover the top five processes of CPU consumption.

- **Cloud service monitoring**

It provides monitoring capabilities for major cloud services on Alibaba Cloud, such as computing, storage, and network. It is turned on by default out-of-the-box. Monitoring based on indicator thresholds can identify dependent service performance bottlenecks as soon as possible.

- **Site monitoring**

It simulates and sends user requests to your site from nationwide terminal nodes to test and monitor the network quality provided by network operators all over the country. The site monitoring function allows you to monitor the performance of the current service from the outside.

- **Log monitoring**

This allows you to monitor services based on existing logs. You can collect and analyze local logs to monitor the services and performance.

- **Event monitoring**

Event Monitoring covers cloud service faults, O&M events, and user-business exceptions. It provides summary statistics by service, level, name, and application group, and supports the alerting feature. It provides automated O & M services based on these events.

CloudMonitor can collect and track various metrics, events, collect and monitor log files, and set alarms. CloudMonitor can be used to monitor various Alibaba Cloud resources, such as ECS instances, ApsaraDB for RDS instances, and ApsaraDB for Redis instances. You can also use CoudMonitor to monitor custom metrics generated by your applications and services and log files from the application. You can use log monitoring after connecting to the log service.

4.2.3.2. Real-time application monitoring

Real-time application monitoring is divided into two sub-products: application monitoring and browser monitoring. Application monitoring collects application data to provide you with application observability, browser monitoring helps you better measure the user experience of your app by collecting relevant performance data from your browser.

By installing the ARMS agent, the application real-time monitoring service comprehensively monitors applications, helps you quickly sort out the call relationships, locate slow API errors, and analyze the call stack and memory structure. As a result, the efficiency of online problem diagnosis is greatly improved.

Browser monitoring collects the access performance data of all real users online, including page loading time, runtime exceptions, API call status, and time consumption. It helps you understand your personal user experience and quickly analyze the causes of reduced user experience, such as JS errors, CDN fluctuations, and backend latency.

By using application monitoring and browser monitoring together with our pre-built diagnosis model, we can quickly locate things that cause user experience problems and effectively reduce the time required to fix the problems.

Key Resources: https://www.aliyun.com/product/arms

4.3. Trade-offs

4.3.1. Application level

You can use the application-level cache or in-memory mode to implement these features at the code level. At the architecture level, we recommend using the highly reliable, dual-node, hot standby architecture of ApsaraDB for Redis. To meet your business needs, we recommend using seamlessly scalable cluster architecture with high read/write performance and flexible capacity configuration. After the request is cached, the execution time is significantly reduced. In this way, you will be able to scale out based on the buffer layer and help high-strength use components to reduce the load level.

The read/write splitting feature supported by each specification adopts a chain replication architecture. This architecture allows the overall instance performance to increase linearly by increasing the number of read-only instances, making full use of the physical resources of each read-only node.

The enterprise-level, performance-enhanced instance mainly optimizes in two aspects: multi-threaded performance enhancement and multi-module integration.

- An instance with enhanced performance of Redis separates tasks in each stage and uses multiple threads to process the corresponding tasks in parallel. Apsaradb for Redis and its community instances of the same specification have three times higher read and write performance. This removes the limit on the performance of frequent read/write hot data.
- The ApsaraDB for Redis Enterprise Edition performance enhancement series integrates multiple customized Redis modules, including TairString, TairHash, TairGIS, TairBloom, and TairDoc, to extend the applicability of Redis in many ways. It simplifies business development in complex scenarios so you can focus on business innovation.

You can also use circuit breaking degradation to protect the system when performance degradation occurs. For example, slow SQL statements may cause severe performance issues in the system. You must use effective methods to detect performance issues and take measures to prevent performance deterioration.

Many scenarios have an impact on performance:

- The system you built has encountered more external requests than the system can handle. These requests are accumulated at the application entry point and waiting for processing. This leads to longer request processing time, more threads in the application process, higher host CPU usage, and higher system load. Causing a sharp decline in performance.
- Dependent components are unstable. When the dependent components are unstable, the performance of the constructed system can also be severely affected.

You can use AHAS to detect performance problems promptly and quickly stop losses to prevent rapid performance deterioration. AHAS provides real-time traffic analysis that allows you to detect system performance problems within seconds. AHAS provides various three-dimensional methods for you to quickly take action by using application-side stops.

You can use AHAS to make a trade-off between performance and traffic.

4.3.2. Database level

Databases provide enhanced performance through single-write-multiple-read or distributed middleware, which enables a database to exceed the capacity of a single database. This can better meet the actual needs of high-strength, read-only database workloads.

ApsaraDB for PolarDB adopts a "one primary node and multiple secondary nodes" architecture in a comprehensive managed service mode. All read/write and read-only nodes of the same instance access and store the same data copy. MaxCompute supports up to 100 TB of storage space. You can scale out a maximum of 16 nodes. Each node supports up to 88 VCPUs. The serverless distributed storage space is automatically scaled based on the data volume. This uses the computing and storage separation architecture to greatly improve resource utilization and performance. ApsaraDB for PolarDB is up to six times faster than traditional MySQL database engines in handling large amounts of concurrent queries. Each node can handle more than one million queries per second. It takes less than five minutes to scale out the number of compute nodes. These features enable you to handle workload spikes with ease.

Distributed Relational Database Service (DRDS) is designed to solve the scalability problem of standalone relational databases. DRDS is lightweight, flexible, stable, efficient, supports horizontal or vertical parallelism, and parallel computing, to meet the scalability challenges of high concurrency, storage capacity, and online computing. Based on the stable RDS for MySQL instance, the database and table sharding function is used to bring the DRDS database into optimal stability. The peak TPS can reach 87 million times per second. The DRDS database supports 550,000 orders per second. These combined features withstood the test of the Double 11 holiday in 2019.

4.3.3. Geographical level

Alibaba Cloud Content Delivery Network (CDN) is a distributed network built over the transport network and contains edge node clusters deployed in different regions. CDN offloads traffic from origins to prevent network congestion. You can use CDN to accelerate website content delivery in different regions or scenarios.

Alibaba Cloud CDN caches the resources from origins to the accelerating nodes that are distributed across the globe. When a user requests access to these resources, the system does not need to reroute the requests to the origin. Instead, the system automatically obtains the resources cached on the CDN node closest to the user. For more information about how to access Alibaba Cloud CDN, see the quick start manuals.

Currently, some CDN nodes can be accessed through IPv6.

Alibaba Cloud has over 2,800 nodes worldwide. Mainland China has more than 2,300 nodes, covering 31 provincial regions, with a large number of nodes located in first-tier cities and provincial capitals. More than 500 nodes are deployed across 70 countries and regions outside China (Hong Kong, Macao, and Taiwan.)

Each node in Alibaba Cloud CDN supports 10 Gbit/s optical NICs, 40 TB to 1.5 PB storage, 40 to 200 Gbit/s operational bandwidth, and 130 Tbit/s reservable bandwidth.

The widely distributed high-performance CDN nodes help to accelerate data transmission and cope with emergencies.

Users handled the traffic peak during the Double 11 holiday with the help of national acceleration nodes, the intelligent elastic scheduling system, and security protection capabilities. It supports a peak rate of over 100 million QPS to ensure that hundreds of millions of buyers around the world can quickly browse high-definition pictures and videos and order smoothly.

4.3.4. Asynchronous cache

To reduce coupling and improve performance between systems in your architecture design, you can often introduce message queue products into your application architecture design. Alibaba Cloud provides a wide range of message-oriented middleware options, including RocketMQ, AMQP, Kafka, MQTT, and MNS. It supports various standard messaging protocols in the industry.

Message queue for Apache RocketMQ

RocketMQ is a distributed message middleware with low latency, high concurrency, high availability, and high reliability built by Alibaba Cloud based on Apache RocketMQ. It provides asynchronous decoupling and load shifting for distributed application systems and supports features for Internet applications, such as massive message accumulation, high throughput, and reliable retry.

Product benefits

- **Load shifting**

Large events, such as flash sales, red packet snatching, and enterprise success, may cause high traffic pulses. The system may become overloaded or crash due to a lack of corresponding protection or too many failed requests. This affects users' experience. Message queue for Apache RocketMQ provides load shifting to solve this problem.

- **Asynchronous decoupling**

As the core system of e-commerce, the transaction system can attract the attention of hundreds of downstream business systems, including the logistics, shopping cart, points, and stream computing analysis. When each transaction order is created, the overall service system is large and complex. RocketMQ can implement asynchronous communication and application decoupling, ensuring the continuity of services at the primary site.

- **Sequential sending and receiving**

Several scenarios need to ensure the sequence in daily life, such as the time-first principle of securities trading, order creation, payment, refund, and other processes in the trading system, as well as the handling of boarding messages of passengers on flights. Ordered messages in the message queue for RocketMQ are generated in the First-In-First-Out (FIFO) order.

- **Consistency of Distributed Transactions**

Final data consistency must be ensured in transaction systems and payment packets. Distributed Transactions of the RocketMQ version are introduced in large numbers to implement decoupling between systems and ensure final data consistency.

Message Queue for AMQP

Alibaba Cloud has developed AMQP, which is fully compatible with the RabbitMQ open source ecosystem and multi-language clients. It provides distributed, high throughput, low-latency, and scalable cloud message services. It is always ready to use, eliminating the need for deployment and O&M, and enabling rapid cloud migration. Alibaba Cloud provides a fully managed service that is professional, reliable, and secure.

Product benefits

- Upper performance of platform: Support millions of queues, linear scaling capability, and no concurrency limit.
- Scale-out of single queue: solves the performance bottleneck of RabbitMQ single queue and supports horizontal scaling of single queue, without concurrency limit.
- Auto Scaling: container service for Kubernetes has excellent scalability and linear performance. It supports auto scaling based on business requirements and is transparent to users.
- Massive message accumulation: Despite the accumulation of massive messages, HybridDB for MySQL maintains high performance without affecting the normal services of the cluster. Message producers are isolated from consumers, which supports a large number of concurrent message producers and allows stable consumption.

Message Queue for MQTT

Message Queue for MQTT is a lightweight, message-oriented middleware (MOM) launched by Alibaba Cloud for mobile Internet and IoT scenarios. Based on the features of message transmission in mobile Internet and IoT scenarios, it supports mainstream communication protocols such as STOMP and GB. Message Queue for MQTT supports native TCP persistent connections, SSL encryption, Websocket, and other transmission modes at the data layer, including C/C++, Java, iOS, Android, and others.

Performance benefits

The MQTT edition needs to handle tasks such as mobile connection access, connection management, and data forwarding. It is equivalent to a connection gateway with unlimited scalability capabilities, backend data persistence, and message storage that can be used with other Alibaba Cloud MQ products, such as the traditional MOM (RocketMQ or Kafka). Message Queue for MQTT is designed with a distributed architecture. Without single point failure and infinite scalability between components, the system ensures that capacity is adjusted according to your online usage and is completely transparent to users.

Message Queue for Kafka

Message Queue for Kafka is a distributed, high throughput, and scalable message queue service provided by Alibaba Cloud. It is widely used in big data fields, such as log collection, monitoring data aggregation, streaming data processing, and online and offline analysis. It has become an indispensable part of the big data ecosystem.

Product benefits

- High Reliability: Messages are persistently stored in message queues with data reliability of 99.999999%;
- High Availability: Kubernetes clusters are deployed across zones and based on open-source architectures. The service availability reaches 99.9% MB.
- Massive message accumulation: In the case of massive message accumulation, the message queue for the Kafka cluster can always maintain a high throughput.
- Tens of thousands of topics: Supports tens of thousands of topics with highly concurrent reads/writes and maintains a high throughput for message queue for Kafka clusters.
- Elastic Computing: You can scale up as needed based on the scale of your business. This service does not affect upper-layer business applications.
- Cluster scaling: Brokers can be scaled out across zones and data centers.
- Partition capacity expansion: Supports the fast capacity expansion of tens of thousands of topics and unlimited queues.

Key Resources: https://www.aliyun.com/product/ons

5. Conclusion

These best practices allow customers to achieve and maintain the performance efficiency of their applications on the cloud. The customer should start with their personal business scenarios and make decisions based on service resource allocation and architecture construction in the cloud. Additionally, the customer needs to periodically measure the preceding metrics, identify performance baselines through automated testing, and select the correct resource types and configurations based on testing and monitoring data. Customers should make good use of the elasticity and out-of-the-box features of cloud resources and evolve the architecture quickly and securely to ensure performance and efficiency. Customers should be more proactive in using the monitoring tools provided by cloud service providers to promptly detect performance issues or bottlenecks based on data orientation and make a further trade-off between the architecture design and relevant resource configuration to ensure the performance of off-premises applications.

Alibaba Cloud aims to help you build an efficient architecture design. At the same time, Alibaba Cloud aims to help you realize the value of your business partnership. It is truly a win-win situation. We recommend that customers adopt the architecture design principles and best practices mentioned in this white paper to efficiently use services and resources on the cloud to meet their performance and efficiency requirements.