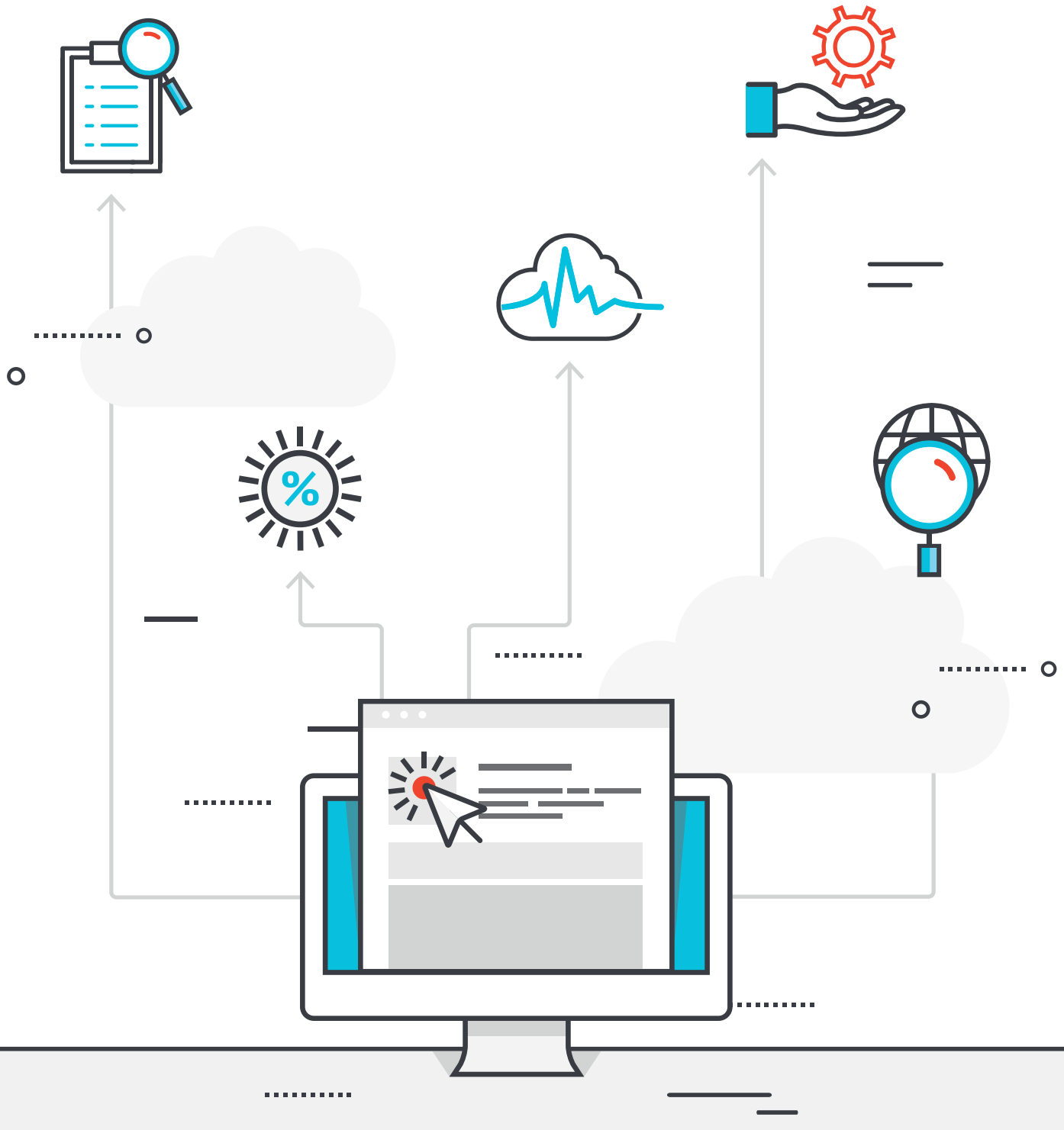
Alibaba Cloud

Capacity Planning and Evaluation for Greater Business Productivity



Contents

01 Introduction

With "to be on the cloud" emerging as the latest technological requirement and trend, the usage and optimization of IT resources have become imperative for businesses and their respective CTOs.

To begin with, most organizations thought cloud migration was the solution to all their problems. However, managing cloud resources has emerged as a primary challenge that organizations today grapple with. Capacity planning and evaluation seek to mitigate this problem.

Capacity planning refers to the ability to forecast future use of cloud infrastructure, which is essential for business service requirements. It incorporates the business volume, service level and application system performance for unified planning and management. Furthermore, it establishes relationship models between these three aspects to help enterprises effectively manage the cloud infrastructure input costs, and improve the service output capacity of application systems.

This whitepaper provides answers to questions on the topic of capacity planning and evaluation. Business leaders, CTOs, Vice Presidents, and technology leads will find this whitepaper helpful in understanding the importance of capacity planning and evaluation along with the various solutions provided by Alibaba Cloud.

02 Industry Pain Points and Business Risks

Cloud infrastructure utilization and database management are critical for modern organizations. However, while cloud infrastructure comes with its own advantages and groundbreaking capability, it also brings its own pain points to organizations and their leaders. Not addressing these pain points has proven to be destructive for organizations. Capacity planning and evaluation seeks to address the following problems:

Evaluating Existing Cloud Infrastructure with Business Needs

Organizations often face questions whether their cloud infrastructure setup matches business needs, if it is insufficient or has it been over-deployed?

Utilizing Current Cloud Infrastructure Resources

Over a decade has passed since the introduction of cloud computing to the IT arena. Since then organizations have jumped at the opportunity to implement on the cloud. However, this has resulted in under or over utilization of IT resources, which can prove disastrous to small and medium-sized businesses.

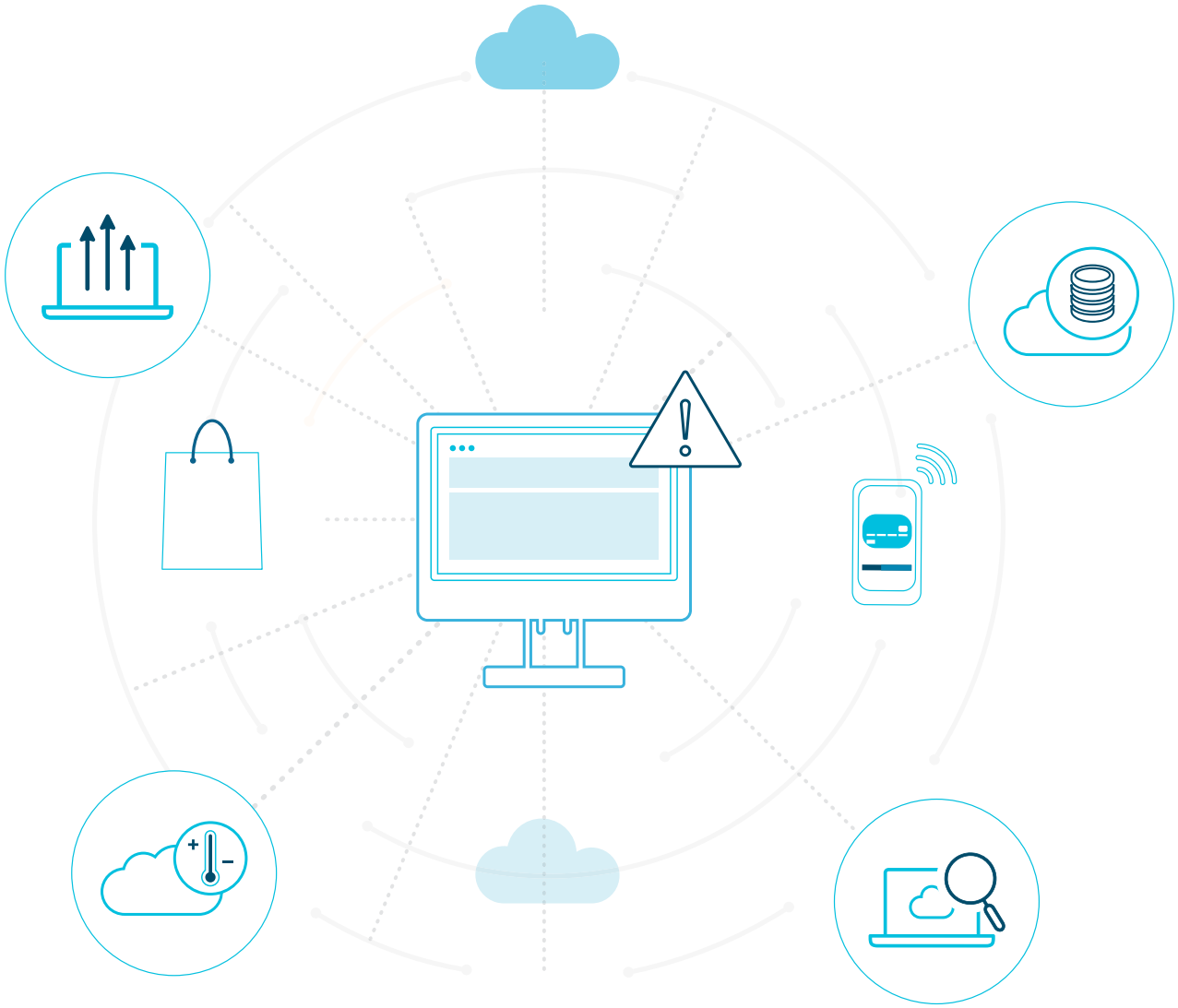
Business Productivity

Attaining profitable business productivity is a common organizational goal for organizations, and is even more critical for the survival of small businesses. An improperly planned and incorrectly utilized cloud infrastructure can severely hinder business productivity.

Cloud Infrastructure Resources Forecasting

Organizations are constantly striving to increase their revenue and customer reach. Growth in business volume requires cloud infrastructure capable of matching spikes in website traffic or database expansion.

Businesses thrive on the income their products generate. However, when the company's products start demanding a large chunk of capital, the product transforms into a business risk and potential liability instead of a revenue generating resource. There are several aspects a business owner has to consider in order to avoid capital gains turning to capital losses. One crucial consideration is managing business infrastructure and planning with a particular set of rules. Capacity planning and evaluation seeks to address this challenge.



03 Significance of Capacity Planning on the Cloud

Every organization, irrespective of its size, category, and expertise has a common goal of effectively utilizing and implementing its IT resources. With cloud being one of the most critical resources of an organization, its capacity planning becomes even more consequential. Below are important points regarding capacity planning for enterprise cloud users.

Reduction of Procurement Costs

Capacity planning for existing systems can determine the optimal cloud infrastructure resource allocations and facilitate investment budgeting.

Enhancing Application Reliability

Capacity planning can accurately predict the time of excessive resourcing or capacity loads to reduce the probability of system downtime.

Greater Application Availability

Through short-term, uninterrupted capacity planning, service quality requirements and response time gaps can be monitored promptly, allowing precautionary measures to be taken to avoid compromised service quality.

Identifying and Re-purposing Underutilized IT Infrastructure

Organizations often stock underutilized cloud infrastructure in one way or the other. Instances, where organizations are unaware of their cloud infrastructure, are not unheard of. Capacity planning helps identify and repurpose underutilized infrastructure.

Migrating Workloads to New IT Infrastructure

Capacity planning helps organizations and their leaders decide the possibility of a successful migration to new IT infrastructure. Capacity planning can address concerns including the quantity of data resources to migrate and what IT infrastructure to migrate the workload to.

Optimal Migration to the Cloud Can Reduce Downtime

With abilities to reduce downtime to the minimum, organizations can use cloud resources to the fullest and facilitate an effective and optimal cloud migration. Capacity planning not only helps organizations identify cloud infrastructure that has the potential to be more efficiently utilized but also facilitates improved customer experiences. It seeks to enhance the overall experience and presence of the determinants of a business environment - particularly the organization, the customers and the cloud infrastructure of the organization.

04 Methods of Capacity Planning

There are two capacity planning methods available to organizations - resource limit analysis and resource factor analysis.

4.1 Resource Limit Analysis

Resource limit analysis studies the resources that will become a bottleneck for the system under a load. The analysis steps are as follows:

- Measure the requested frequency of the cloud product or service and monitor changes in the requested frequency over a given period
- Measure actual usage of the cloud product or service and application software, and monitor the changes in resource utilization over time
- Present the requests to the cloud server by resource usage
- Infer the limits of requests to the cloud product or service based on each resource

First, identify the cloud product or service type and the request type served by the cloud product or service. Examples could be an HTTP request to a web service built on Elastic Compute Service (ECS), a query request to the cache database service cache built on ECS, or a query request to the Relational Database System (RDS) database service data.

The next step is to determine the specific system resources the request will require. For online systems, measuring the current rate of requests corresponding to the resource utilization is possible. First, determine which resource will reach 100% utilization or the alert threshold first, and then check the request rate at that time.

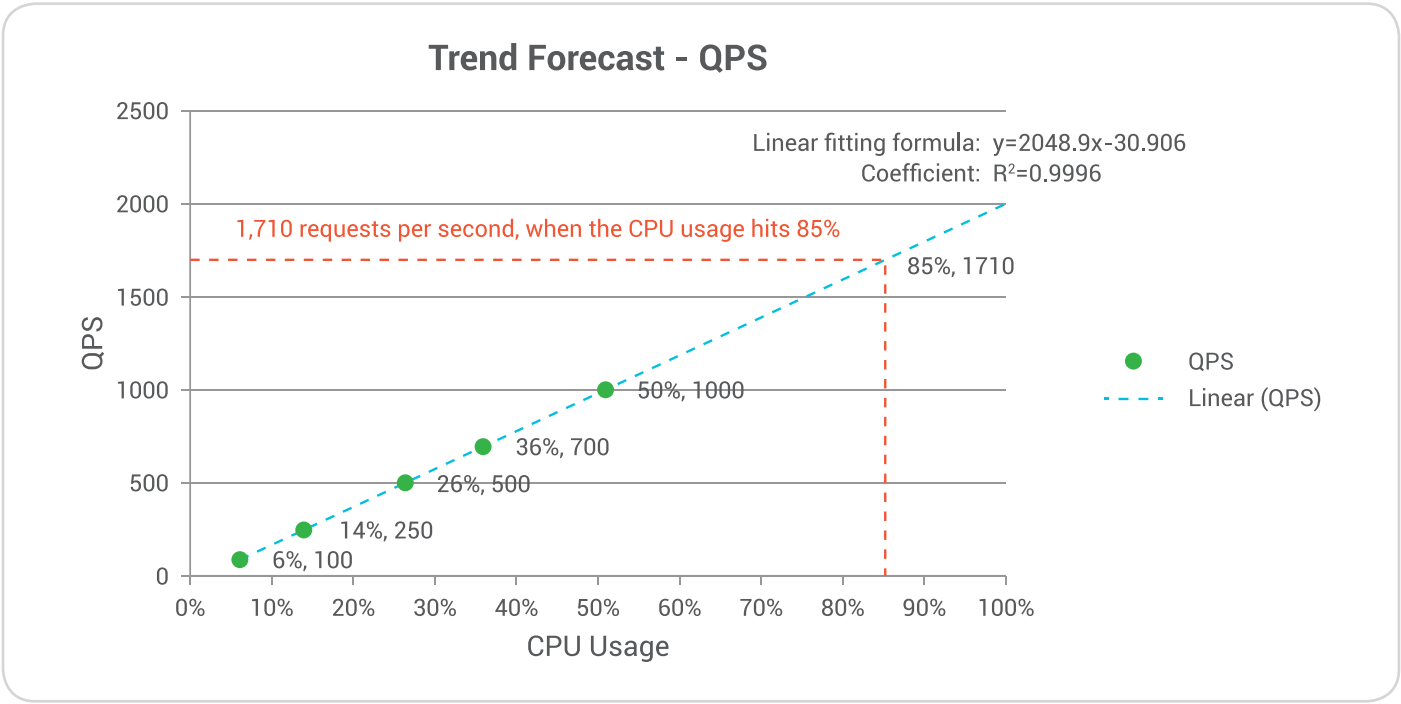
For offline systems, you can use the load pressure testing tool to simulate the request to be initiated in the test environment while measuring the resource usage. You can increase the workload to identify the resource limit or alert threshold through measurement.

An example below elaborates on the steps of resource analysis mentioned above:

The current system executes 1,000 requests per second. The busiest resources are 16 CPUs, with their average utilization rates being 50%. When the CPU usage reaches 85%, it is estimated to become the average workload. Calculate the CPU consumption per second by requests and number of requests supported when the CPU usage reaches 85%:

- Every request CPU% = total CPU%/total number of requests = 16 CPU * 50%/1000 = 0.8% CPU consumed per request
- Maximum number of requests per second = 85% * 16 CPU/0.8% CPU consumed per request = 1360/0.8 = 1700 requests per second
- When the CPU usage reaches 85%, the number of requests is forecasted to reach 1,700 per second. This is a rough forecast, and you may encounter other constraints before reaching the request rate.

The above practice uses one data point: the application's throughput of 1,000 requests per second and the server CPU utilization of 40%. If the monitoring continues for a longer duration, you can collect a vast number of different throughput and utilization data values to improve the accuracy of the forecast. The following resource limit analysis chart is a method for inferring the maximum throughput of an application based on multiple measured data values.



This graph describes how server CPU utilization increases as website traffic increases. A close analysis will help identify the maximum number of requests a server can handle. This tells us when we need a new server when the traffic is increasing on the website.

Resource limit planning becomes imperative to ensure that the organization's website remains available in the long run. Businesses strive to reach as many potential customers as possible which is reciprocated by a substantial increase in website traffic. Heavy loading of the website may result in choking of services, and this could be disastrous to overall business operations. Hence, resource limit analysis plays a prominent role in ensuring long-term website service availability.

4.2 Factor Analysis

Achieving the desired performance of a cloud service or newly deployed system requires the adjustment of several constraints. These adjustments include the number of CPU cores, memory size, cloud disk type, and storage capacity. Generally, we invest the minimum cost to achieve the performance required for the application system.

Through a combination of tests and evaluations, we can determine the most cost-effective combination. However, this can quickly get out of control, because you will need to perform 256 tests and evaluations for two possible factors out of eight.

The solution is to test and evaluate a finite set of a combination and carry out factor analysis based on the Cannikin Law as shown below:

- Test and evaluate the performance when all factors are at their maximum settings.
- Alter the factors one by one and assess the performance (the result should be that the change to each factor leads to performance reductions).
- Based on the evaluation results, generate statistics on the percentage of performance reductions arising from alterations in each factor and the cost savings of the cloud product resources.
- Use the highest performance as the starting point, choose cost-saving factors, and ensure that the combination with performance reduction can still meet the required business performance requirements.
- Re-measure the changed configuration and confirm the delivered performance.

For an application system with eight factors, perform only ten tests using this method.

Factor Analysis helps organizations to determine the most cost effective combination to ensure that cloud resources are effectively planned.

05 Capacity Evaluation

Capacity is the resource ceiling pre-allocated to a specific application, such as traffic and bandwidth. Similar to traffic systems, network and interfaces connect information systems. Smooth traffic flow is dependent on road width, traffic lights, and temporary control measures. Similarly, the proper functioning of an application system is subject to the influences of capacity configuration. Evaluating the capacity of your cloud infrastructure or any IT resource is called capacity evaluation.

5.1 Capacity Indicator

As the name suggests, capacity indicator measures the processing capacity of the system, while any apparent constraints that may come up serve as the signal to stop the pressure test. For example, for CPU-intensive systems, we often choose transactions per second (TPS) as a system capacity indicator to measure the system's processing power, and the selection of constraints will focus on whether the CPU usage will become the bottleneck first. Storage-oriented systems use data traffic (MB/S) as the capacity indicator. The TPS of storage-oriented systems depends on the business data size, so the traffic is more suitable for identifying the system's processing capacity. The constraints will focus on the network traffic or IO, which will become the first bottleneck.

Outlined below are commonly used capacity indicators.

System Processing Capability

System processing capability refers to the information processing capability of the system using the system cloud service, resource platform and application software platform. System processing capability uses the number of transactions processed per second as a measurement. There are two methods to understand transactions: first, the process of a trade from the business personnel perspective; second, the process of a transaction application and response from the system perspective. System processing capability uses:

- **HPS (Hits per Second):** The number of hits per second. Unit: hits/second.
- **TPS (Transaction per Second):** The number of transactions handled per second by the system. Unit: transactions/second.
- **QPS (Query per Second):** The number of queries handled per second by the system. Unit: queries/second.

Error Rate

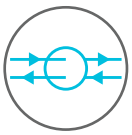
The error rate refers to the probability of failed transactions when the system is under load. The error rate = (number of failed transactions/total number of transactions) * 100%. Timeouts cause a better and more stable error rate of systems. Different systems have different requirements for the error rate. The error rate should not exceed 0.3%, meaning the success rate should not be less than 99.7%.

CPU Utilization

CPU indicators mainly refer to the CPU utilization, including user mode (user), system mode (sys), wait mode (wait), and idle mode (idle). CPU utilization should be less than or equal to 75%, the industry alarm value range. The CPU sys% should be less than or equal to 30%, and the CPU wait% should be less than or equal to 5%.

Memory Utilization

Modern operating systems are equipped with the cache in the memory to maximize the use of memory, so memory utilization of 100% does not mean that there are bottlenecks in the memory. The bottlenecks of systems are mainly dependent on the SWAP (swap with virtual memory) space utilization. In typical cases, SWAP space utilization should be less than 70% or too many swaps will cause lower system performance.



Disk Throughput

Disk throughput is the amount of data that passes through the disk per unit of time in the absence of disk failures. Disk indicators include the megabytes of data read and written per second, disk busy rate, the number of disk queues, average service time, average wait time, and space utilization. Among them, the disk busy rate is an essential reference that directly reflects whether the disk contains any bottlenecks. Under normal circumstances, the disk busy rate is less than 70%.

Network Throughput

Network throughput refers to the data size that passes through the network, per unit of time, in the absence of network failures. The unit is bytes/second. The network throughput metric is used to measure the system's requirement for network equipment or link transmission capability. When the network throughput indicators approach the maximum transmission capacity of the network equipment or link, you need to consider upgrading the network equipment.

Business Traffic Modeling

This indicator analyzes the amount of website traffic over a period. Further, it also considers the system resource consumption conditions, such as the consumption of CPU, memory, I/O, and other resources in the peak period, as well as the transaction response time, transaction throughput, transaction success rate, and other indicators of the application. Business traffic modeling can be used in various modes of approach depending on whether an organization has an online system or an offline one.

A good cloud provider can help you know more about the functionality and effectiveness of these indicators.

5.2 Capacity Evaluation Tools

Several open-source tools are available for capacity evaluation. Below is a brief discussion of these tools.

5.2.1 Enterprise-Level Performance Testing Tool – LoadRunner

LoadRunner is a load testing tool that predicts system behavior and performance. LoadRunner can test the entire enterprise architecture by identifying and looking for problems by simulating the concurrent loads from a large number of users and with real-time performance monitoring. The tool allows enterprises to minimize the test time, optimize performance, and accelerate the release cycle of application systems. Additionally, LoadRunner supports a broad range of protocols and technologies to provide customized performance testing solutions for your special environment.

Virtual User Generator

LoadRunner's Virtual User Generator helps create virtual users and the system load. The system load can generate virtual users to simulate the business operations of real users in the form of a virtual user. It first records the business process (i.e., placing an order or booking a ticket) and then converts it into a test script. With virtual users, a vast number of users can be generated to access the tested system at the same time, on Windows, UNIX, or Linux machines.

Controller

After creating the virtual users, set the load scheme (the number of business process combinations and virtual users). With LoadRunner's Controller, you can quickly organize a multi-user test scheme. The Controller's synchronization point feature provides an interactive environment where you can establish continuous and cyclic loads and manage load testing schemes. Moreover, you can use its agenda planning service to define when users access the system to generate the load, which automates the testing process. Additionally, organizations can use the Controller to limit the load scheme in which all users perform the same action at the same time (e.g., logging into an inventory application to simulate the peak load scene). Additionally, it can monitor the performance of various components in the system architecture, including servers, databases, and network devices to help customers determine the system configuration.

Analysis

Once the test is complete, LoadRunner collects all testing data and provides advanced analysis and reporting tools to quickly locate performance issues and trace the cause. With LoadRunner's web transaction details monitor, you can get an idea of the time required to download all the images, frames, and text to each web page. Additionally, the web transaction detail monitor breaks down the end-to-end response time used for the client, the network, and the server, making it easy to identify the problem and locate the components that go wrong. For example, you can break down the network latency to determine the time it takes for Doman Name Servers (DNS) resolution, connection to a server, or Secure Sockets Layer (SSL) authentication. Using the LoadRunner analysis tool, you can quickly find the location and cause of the error and make appropriate adjustments.

LoadRunner integrates with a real-time monitor that observes the running performance of the application system at all times during the load test process. These performance monitors display the transaction performance data (e.g., response time) in real-time and real-time performance of other system components, including application servers, web servers, network devices, and databases. This allows you to evaluate the running performance of these system components from both the client and the server during the testing process to discover problems more quickly.

5.2.2 On-cloud Pressure Test Platform – Performance Test Services

Performance Test Service (PTS) is an on-cloud performance testing platform that centrally manages test machines, test scripts, test scenarios, test tasks, and test results. PTS is designed to swiftly scale and enable dynamic configuration of domain names to meet the ever-growing demand for cluster pressure tests.

Organizations can use PTS to perform an overall evaluation of their system's performance in the Alibaba Cloud environment. This will help identify system performance bottlenecks for optimization purposes and provide an understanding of the system performance indicators for future expansion.

Alibaba Cloud users will need to purchase test machines (ECS, RDS) to use PTS. Furthermore, PTS determines which servers are available for the user's account during the pressure test. The pressure test traffic will only apply to the server under the user's account. If you do not purchase a server, you will not be able to use PTS.

During operation, PTS will generate pressure testing traffic through the pressure generators. If you have more requirements on the pressure traffic or region, you can dynamically scale and globally deploy the PTS pressure generators.

06 Alibaba Cloud Capacity Planning and Evaluation Solution

Being a global cloud solution provider, Alibaba Cloud caters to all the needs of an organization. Capacity planning is one of the areas where Alibaba Cloud's tools and services excel in providing service to customers.

6.1 Auto Scaling

Auto Scaling is a management function that automatically adjusts elastic computing resources, based on your business' needs and policies. You can adjust elastic computing resources automatically based on your organizational needs, seamlessly increase ECS instances to cope with traffic peaks and surges. Additionally, it automatically reduces the number of ECS instances when business demands drop, saving on costs.

Alibaba Cloud's Auto Scaling service comes with the following features:

Dynamic Scaling Mode

Based on CloudMonitor performance indicators (e.g., CPU and memory usage) to automatically increase or reduce ECS instances.

Timed Scaling Mode

Configures periodic tasks to increase or reduce ECS instances in a timed manner. When the periodic demands change, you can set up dynamic scaling modes at the same time to tackle unexpected changes.

Fixed Number Mode

Through the "minimal number of instances" attribute, you can maintain an appropriate number of ECS instances to achieve healthy operation and ensure real-time availability in daily scenarios.

Automatic Configuration of the Server Load Balancer and RDS

When ECS instances are increased or reduced, Auto Scaling will add ECS instances to the Server Load Balancer or remove ECS instances, and automatically add the IP addresses of the corresponding ECS instances to the RDS access whitelist, or alternatively remove IP addresses.

Auto Scaling is a capacity-planning technique that minimizes the error from unknown traffic demand. Additionally, it eliminates inefficiency by requisitioning resources when needed and decommissioning them when not. Furthermore, it ensures that cloud resources are utilized as per business requirements, thus ensuring smooth functionality and at a reasonable cost.

6.2 CloudMonitor

CloudMonitor provides users with a comprehensive understanding of the usage, performance, and running states of their Alibaba Cloud resources. Organizations can use this service to collect monitoring metrics for Alibaba Cloud resources, detect Internet service availability, and set metric alarms. It can monitor ECS, RDS, Server Load Balancer and other types of Alibaba Cloud resources. Cloud Monitor can also monitor Internet application availability via universal network protocols, such as HTTP and Internet Control Message Protocol (ICMP). The alarm service then enables you to respond quickly, ensuring the smooth operation of your applications. Currently, CloudMonitor provides three types of services: **site monitoring**, **cloud service monitoring**, and **custom monitoring**.

CloudMonitor provides an extremely rich array of application scenarios, that include:

Cloud Service Monitoring

CloudMonitor allows you to check the running statuses of your products, as well as various metrics on its product page. You can also set alarm rules for the monitored items.

Daily Management Scenario

When performing day-to-day management of Alibaba Cloud products, you can conveniently view the running statuses of monitored products by directly logging onto the CloudMonitor console.

Timely Handling of Exceptions

CloudMonitor sends an alarm message when the monitored data reaches the alarm threshold configured in the alarm rules. In this way, you can receive timely exception notifications and check the cause of the exception.

Timely Scaling Scenario

Once you set the alarm's rules for various monitored items such as bandwidth, number of connections, and disk space utilization, you can readily understand the current status of the cloud services and receive alarm notifications promptly when the business volume grows for service scaling.

Site Monitoring

The site monitoring service provides monitoring settings for eight protocols, allowing you to detect the availability, response time, and packet loss rate of the site. This gives access to the complete picture of the availability of the site.

Custom Monitoring

Custom monitoring is designed as a supplement to cloud service monitoring. If the CloudMonitor does not provide a monitored item you need, you can create a new monitored item, and report acquired metric data to CloudMonitor. It will then display metric charts and provide alarm functionality for the newly monitored item.

CloudMonitor helps businesses to keep an eye on each and every resource running in their account. This facilitates the evaluation of the resource consumption and maximizes return on investment (ROI).

6.3 Alibaba Cloud CDN

The Alibaba Cloud CDN (Content Delivery Network) is built to automatically handle spikes in traffic, reduce load speed on your origin site, and support storage capacity of up to 1.5 petabytes.

With its low latency and high data transfer rate, Alibaba Cloud CDN directs user requests to the most suitable node based on network congestion to retrieve content in the most efficient manner. Alibaba Cloud CDN has a high-performance cache system built with powerful clusters of servers to maximize throughput. It relies on an intelligent object heat algorithm and hierarchical 'hot' cache resources for precise resource acceleration.

Alibaba Cloud CDN provides the following benefits:

Fast Load Speed

Alibaba Cloud CDN improves the availability of servers or nodes with high-speed read/write storage using SSDs. It accelerates content distribution with page optimization and smart compression technologies.

- The page optimization features remove spaces, line breaks, TABs, annotations, and other redundant page content to reduce page size.
- Smart compression reduces the size of the content and accelerates content distribution. Additionally, CDN increases the response time by collapsing multiple JavaScript/Cascading Style Sheets (CSS) files into a single request.

Advanced Scheduling

CDN reduces complexity by supporting millions of domain names with a single machine. It also ensures high availability with multi-level scheduling policies. Moreover, Alibaba Cloud CDN is simple and easy to use. It coordinates security systems, refreshes systems and content management systems using multi-system interaction techniques. It also enhances user experience with a real-time data scheduling option and advanced traffic predictions. You can seamlessly integrate Alibaba Cloud CDN with other products to improve data transfer speed and easy content distribution.

Alibaba Cloud CDN not only helps to deliver content faster to users but also reduces resource consumption at the back-end, hence minimizing infrastructure costs.

6.4 Server Load Balancer

Server Load Balancer is a ready-to-use service that integrates with ECS to manage varying traffic levels without manual intervention. Server Load Balancer is configured above ECS instances to receive incoming traffic first. It then distributes incoming traffic across multiple ECS instances, detects unhealthy or unsafe instances and routes traffic to healthy and safe instances only. Importantly, Server Load Balancer also ensures high availability of applications by eliminating any single-point-of-failure and protects from SYN flood and Distributed Denial of Service (DDoS) attacks.

The benefits of Server Load Balancer are as follows:

High Availability

Server Load Balancer distributes incoming traffic to healthy ECS instances within and across availability zones. It facilitates local disaster recovery by using multi-zone deployment model for certain regions. Additionally, it supports global load balancing and cross-region disaster recovery when used in combination with DNS.

Flexibility

Server Load Balancer facilitates multiple traffic scheduling algorithms to distribute traffic evenly. Furthermore, it supports Weighted Round Robin (WRR) and least connections scheduling algorithms. Additionally, it increases server load balancing capabilities and distributes traffic evenly by configuring ECS instance weights.

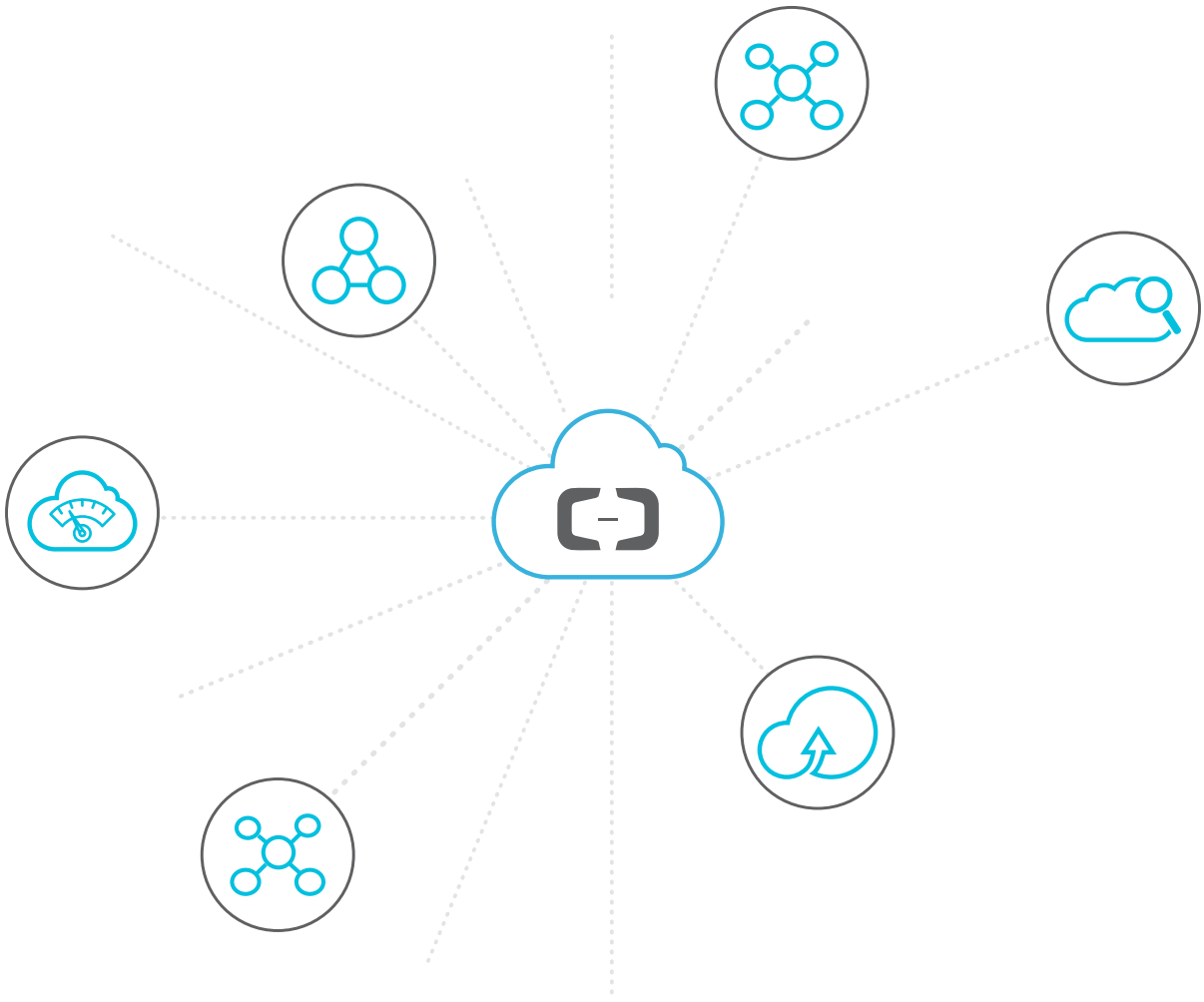
Easy-to-use Console

Like all Alibaba Cloud products, Server Load Balancer is an easy to use console that provides various management methods that allow users to create, modify, and manage their Server Load Balancer.

Server Load Balancer ensures organizations' IT infrastructure to stay within the evaluated limits of capacity planning by distributing traffic evenly across the servers. Furthermore, it assists Auto Scaling to scale up and scale down resources when required.

6.5 Object Storage Service

Alibaba Cloud Object Storage Service (OSS) is an easy-to-use service that enables you to store, backup, and archive large amounts of data on the cloud. It acts as an encrypted central repository where organizations can securely access files from around the globe. OSS guarantees up to 99.9% availability and is a perfect fit for global teams and international project management. An important feature of Alibaba Cloud's OSS is that it comes with an infinite amount of data storage.



07 Conclusion

Cloud usage calls for effective capacity planning and evaluation. In fact, capacity planning and evaluation is an aspect that no organization can afford to overlook. Its impact affects not just the proper utilization of resources but can determine the very existence of an organization.

Best use of capacity planning and evaluation can be made when organizations estimate server workload, setup application performance objectives and select the correct instance type.

Keeping these important characteristics in mind, Alibaba Cloud products and services aim to provide capacity planning and evaluation solutions to organizations across the table.

Get in touch with our experts to ensure that your business is linked with the latest techniques in cloud infrastructure planning and evaluation.

Alibaba Cloud