Monitor and Management

Chapter 6

Outline

1. Cloud monitor mechanisms
2. Service-level agreement (SLA)
3. Cloud management mechanism

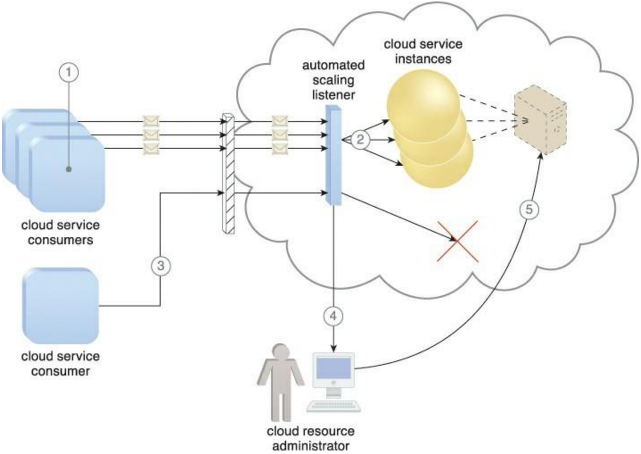
Cloud Usage Monitor

- Responsible for collecting and processing IT resource usage data
- Used to measure network traffic and message metrics
    - actively monitoring a virtual server and detects an increase in usage
- Continuous monitoring to ensure new and modified files are scanned in real-time

Automated Scaling Listener

- The automated scaling listener mechanism is a service agent that monitors and tracks communications between cloud service consumers and cloud services for automatic/dynamic scaling purposes
- Deployed within the cloud, typically near the firewall, from where they automatically track workload status information
- Workloads can be determined by the volume of cloud consumer-generated requests or via backend processing demands triggered by certain types of requests
- Auto-scaling settings controlled by cloud consumers determine the runtime behaviour of automated scaling listener agents, which run on the hypervisor that monitors the resource usage of the virtual servers

When workloads exceed current thresholds or fall below-allocated resources



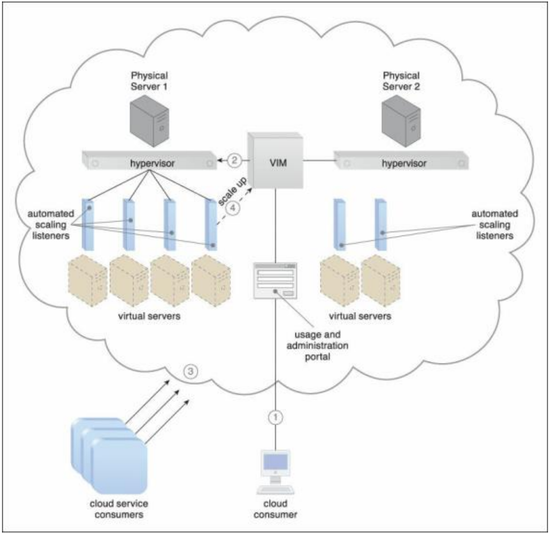
When workloads exceed current thresholds or fall below-allocated resources

1. Three cloud service consumers attempt to access one cloud service simultaneously
2. The automated scaling listener scales out and initiates the creation of three redundant instances of the service
3. A fourth cloud service consumer attempts to use the cloud service
4. Programmed to allow up to only three instances of the cloud service, the automated scaling listener rejects the fourth attempt and notifies the cloud consumer that the requested workload limit has been exceeded
5. The cloud resource administrator accesses the remote administration environment to adjust the provisioning setup and increase the redundant instance limit

Example: Auto scaling

One cloud consumer has it set up so that whenever resource usage exceeds 80% of a virtual server's capacity for 60 consecutive seconds, the automated scaling listener triggers the scaling-up process by sending the virtual infrastructure manager (VIM) platform a scale-up command. Conversely, the automated scaling listener also commands the VIM to scale down whenever resource usage dips 15% below capacity for 60 consecutive seconds

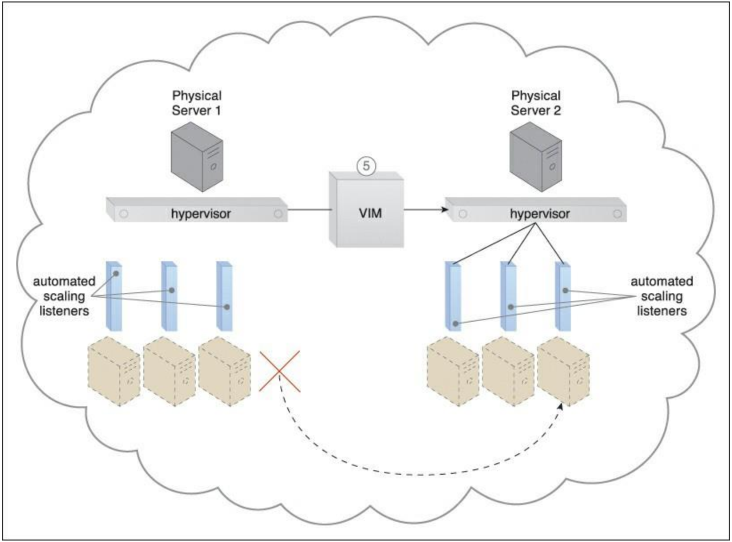
Example: Auto scaling



Example: Auto scaling

1. A cloud consumer creates and starts a virtual server with 8 virtual processor cores and 16 GB of virtual RAM
2. The VIM creates the virtual server at the cloud service consumer's request and allocates it to Physical Server 1 to join 3 other active virtual servers
3. Cloud consumer demand causes the virtual server usage to increase by over 80% of the CPU capacity for 60 consecutive seconds
4. The automated scaling listener running at the hypervisor detects the need to scale up and commands the VIM accordingly
5. The VIM determines that scaling up the virtual server on Physical Server 1 is not possible and proceeds to live migrate it to Physical Server 2

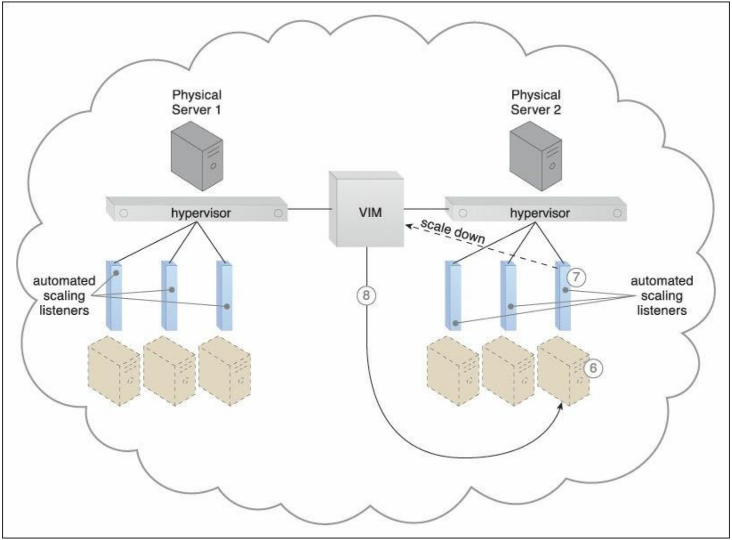
Example: Auto scaling



Example: Auto scaling

6. The virtual server's CPU/RAM usage remains below 15% capacity for 60 consecutive seconds
7. The automated scaling listener detects the need to scale down and commands the VIM
8. The automated scaling listener scales down the virtual server (8) while it remains active on Physical Server 2

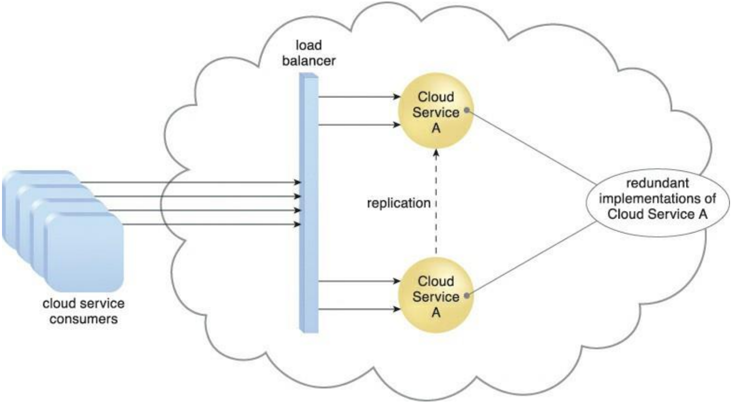
Example: Auto scaling



Load Balancer

- Process of distributing workloads across computing resources in a cloud computing environment and carefully balancing the network traffic accessing those resources
- A common approach to horizontal scaling is to balance a workload across two or more IT resources to increase performance and capacity beyond what a single IT resource can provide
- General objectives: optimizing IT resource usage, avoiding overloads, and maximizing throughput
- Typically located on the communication path between the IT resources generating the workload and the IT resources performing the workload processing

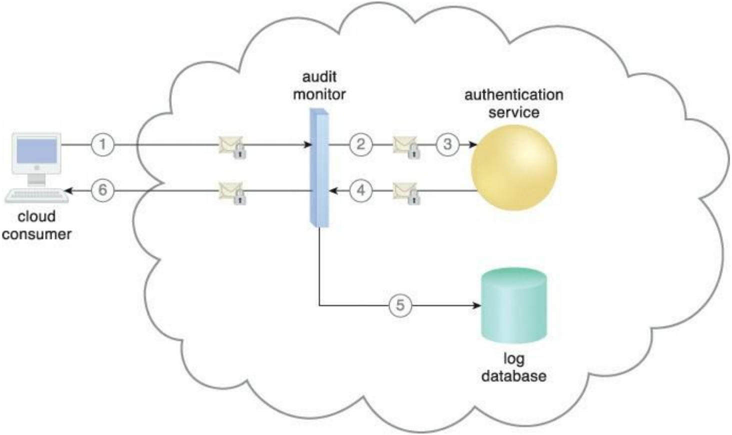
Load Balancer



Audit Monitor

- Used to collect audit tracking data for networks and IT resources in support of (or dictated by) regulatory and contractual obligations
- Intercepts “login” requests and stores the requestor’s security credentials, as well as both, failed and successful login attempts, in a log database for future audit reporting purposes

Audit Monitor



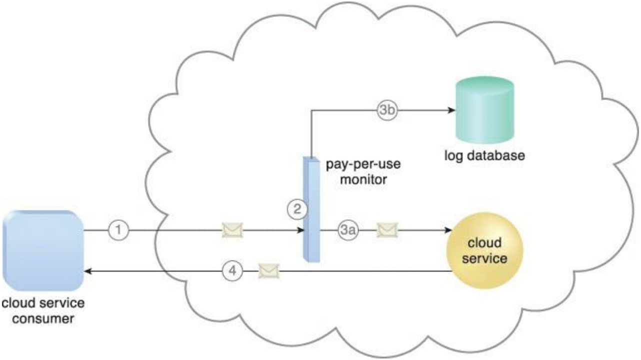
Audit Monitor

1. A cloud service consumer requests access to a cloud service by sending a login request message with security credentials
2. The audit monitor intercepts the message (HTTP Header) and forwards it to the authentication service
3. The authentication service processes the security credentials
4. A response message is generated for the cloud service consumer, in addition to the results from the login attempt
5. The audit monitor intercepts the response message and stores the entire collected login event details in the log database, as per the organization's audit policy requirements
6. Access has been granted, and a response is sent back to the cloud service consumer

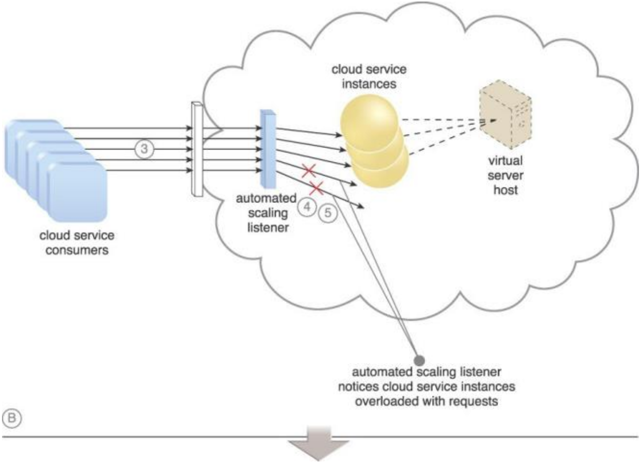
Pay-Per-Use Monitor

- Measures cloud-based IT resource usage in accordance with predefined pricing parameters and generates usage logs for fee calculations and billing purposes
- The data collected by the pay-per-use monitor is processed by a billing management system that calculates the payment fees
  - The pay-per-use monitor stores the value timestamp in the log database
- Designed as a monitoring agent that transparently intercepts and analyzes runtime communication with a cloud service

Pay-Per-Use Monitor



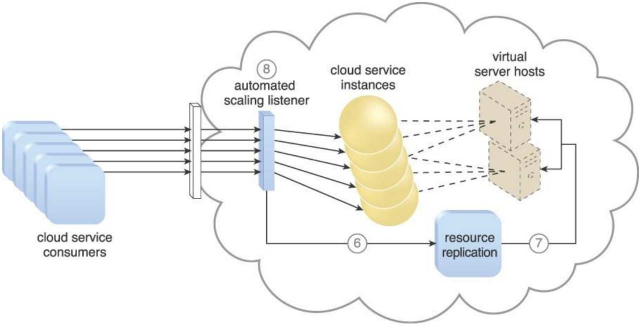
How can IT resources be scaled automatically in response to fluctuating demand?


cloud service instances
virtual server host
3
automated scaling listener
4
5
cloud service consumers
automated scaling listener notices cloud service instances overloaded with requests
B

How can IT resources be scaled automatically in response to fluctuating demand?

- The automated scaling listener monitors the cloud service to determine if predefined capacity thresholds are being exceeded.
- The number of service requests coming from cloud service consumers further increases
- If the cloud service implementation is deemed eligible for additional scaling, the automated scaling listener initiates the scaling process
- The automated scaling listener sends a signal to the resource replication mechanism
- The resource replication mechanism then creates more instances of the cloud service and accommodated increased workload

How can IT resources be scaled automatically in response to fluctuating demand?



How can IT resources be scaled automatically in response to fluctuating demand?

Mechanisms

- Automated Scaling Listener
- Cloud Usage Monitor
- Pay-Per-Use Monitor
- Resource Replication
- Virtual Server

Service-level agreement (SLA)

- An agreement between a cloud service provider and a customer that ensures a minimum level of service is maintained
    - It's the blueprint that governs the relationship between the customer and the service provider
    - Agree upon some rules to create a trusted bond between them.
    - Also reflects the provider's quality of service and its infrastructure
- It guarantees levels of reliability, availability and responsiveness to systems and applications
    - specifies who governs when there is a service interruption; and describes penalties if service levels are not met
- Different providers and customers have their own service level agreements
- There are mainly four aspects to keep in mind when you make a Service Level Agreement
    - Efficiency of systems
    - Speed
    - Volume and quality of work
    - Responsiveness

Types of Service Level Agreements

1. **Customer-Based SLA:**
   - agreement includes the details of all relevant services that a client will be provided while leveraging only a single contract
   - used for single customers
2. **Service-Based SLA:**
   - one similar type of service for all its customers as the service is limited to the unchanging standard
     - same service is provided to all the end-users signing Service-Based SLA
   - simple for vendors
3. **Multi-Level SLA:**
   - agreement is customised according to the need of the customer or company.
   - allows the user to create an arrangement that contains many conditions under the same deal
   - Multi-Level is further divided into 3 types:
     1. **Corporate Level:**
     2. **Customer Level:**
     3. **Service Level:**

Multi-Level SLA

1. **Corporate Level:** This SLA does not need frequent updates as it rarely changes. It involves a brief discussion of all aspects of the agreement and applying it to all the customers using it.
2. **Customer Level:** This agreement discusses the service issues associated with the number of retries allowed and other factors.
3. **Service Level:** In this agreement, all aspects attributed to a particular service regarding a customer group are included.

Customer SLA: example

Company A is an e-commerce retailer that depends heavily on its cloud-based platform for various business operations, such as order processing, inventory management, and customer communication. To ensure that the platform always operates efficiently and is available to customers, the company has negotiated a customer-based SLA with its cloud service provider.

- The SLA includes the following:
  1. **Availability:** The cloud provider guarantees 99.99% availability of the e-commerce platform during business hours, which are defined as 8:00 am to 8:00 pm local time. The provider also guarantees 99.99% availability during non-business hours.
  2. **Performance:** The cloud provider guarantees a maximum response time of 500 milliseconds for all transactions on the e-commerce platform. The provider also guarantees a minimum throughput of 1,000 transactions per second.

Customer SLA: example

3. **Support:** The cloud provider agrees to provide 24/7 technical support via phone, email, or chat. The provider also guarantees a response time of no more than 2 hours for all support requests.
4. **Security:** The cloud provider guarantees that the e-commerce platform will comply with all industry standards and regulations for security

Service SLA: example

Company A is a small consulting firm that provides financial and accounting services to its clients. The company relies heavily on a cloud-based file-sharing and collaboration platform to share sensitive financial data with its clients. To ensure that the platform is always available and secure, the company negotiates a service-based SLA with its cloud service provider.

- The SLA includes the following:
  1. **Availability:** The cloud provider guarantees 99.99% availability of the file sharing and collaboration platform during business and non-business hours.
  2. **Security:** The provider guarantees that the platform will be encrypted both at rest and in transit, and that regular vulnerability assessments and penetration testing will be conducted.
  3. **Support:** The cloud provider agrees to provide 24/7 technical support and guarantees a maximum response time of 2 hours for all support requests related to the file sharing and collaboration platform.

Cloud management mechanism

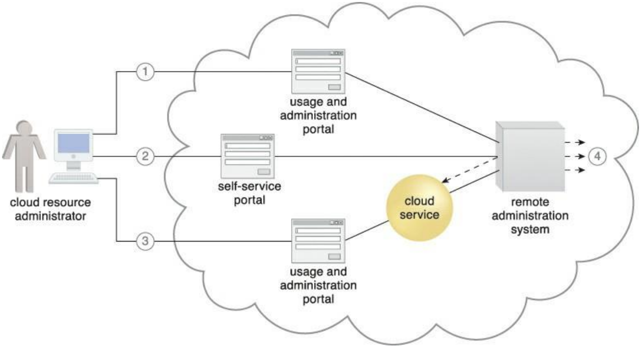
- Process of managing, maintaining, and optimizing a computer environment for achieving better performance, availability of resources, security, and to fulfill the requirements of the users
- It helps to manage the cloud resources and services and enables perfect communication between the users and cloud vendors
- It ensures the smooth functioning of the cloud and maintains the security of the cloud applications.
- It enables real-time monitoring of the applications and data which eliminates the chances of virus and malware attacks.

1. Remote Administration System
2. Resource Management System
3. SLA Management System
4. Billing Management System

Remote Administration System

- establish a portal for access to administration and management features of various underlying systems, including the resource management, SLA management, and billing management systems
- tools and APIs provided by a remote administration system are generally used by the cloud provider to develop and customize online portals that provide cloud consumers with a variety of administrative controls.
- two primary types of portals that are created with the remote administration system
  1. Usage and Administration Portal: portal that centralizes management controls to different cloud-based IT resources and can further provide IT resource usage reports
  2. Self-Service Portal: essentially a shopping portal that allows cloud consumers to search an up-to-date list of cloud services and IT resources that are available from a cloud provider (usually for lease).

Remote Administration System


1
2
3
4
usage and administration portal
self-service portal
usage and administration portal
cloud service
remote administration system
cloud resource administrator

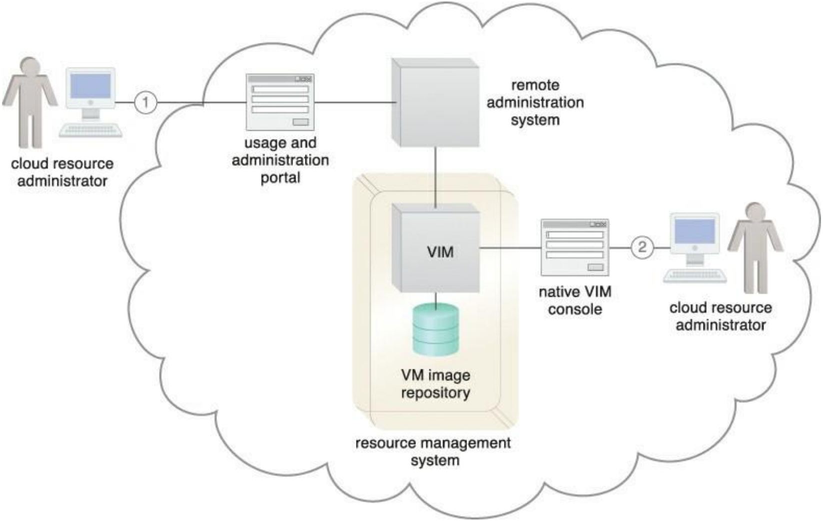
Resource management system

- mechanism helps coordinate IT resources in response to management actions performed by both cloud consumers and cloud providers
- core to this system is the virtual infrastructure manager (VIM) that coordinates the server hardware
    - virtual server instances can be created from the most expedient underlying physical server
- A VIM is a commercial product that can be used to manage a range of virtual IT resources across multiple physical servers

Resource management system

- Tasks that are typically automated and implemented through the resource management system include:
  - managing virtual IT resource templates that are used to create pre-built instances, such as virtual server images
  - allocating and releasing virtual IT resources into the available physical infrastructure in response to the starting, pausing, resuming, and termination of virtual IT resource instances
  - coordinating IT resources in relation to the involvement of other mechanisms, such as resource replication, load balancer, and failover system
  - enforcing usage and security policies throughout the lifecycle of cloud service instances
  - monitoring operating conditions of IT resources

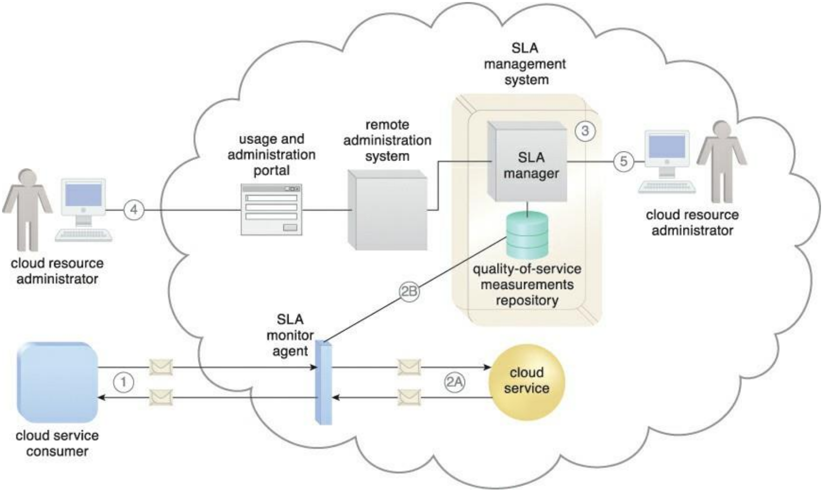
Resource management system


cloud resource administrator
1
usage and administration portal
remote administration system
VIM
VM image repository
resource management system
native VIM console
2
cloud resource administrator

SLA Management System

- generally includes a repository used to store and retrieve collected SLA data based on pre-defined metrics and reporting parameters
- SLA monitor mechanisms to collect the SLA data that can then be made available in near-real time to usage and administration portals to provide on-going feedback regarding active cloud services
- The metrics monitored for individual cloud services are aligned with the SLA guarantees in corresponding cloud provisioning contracts.

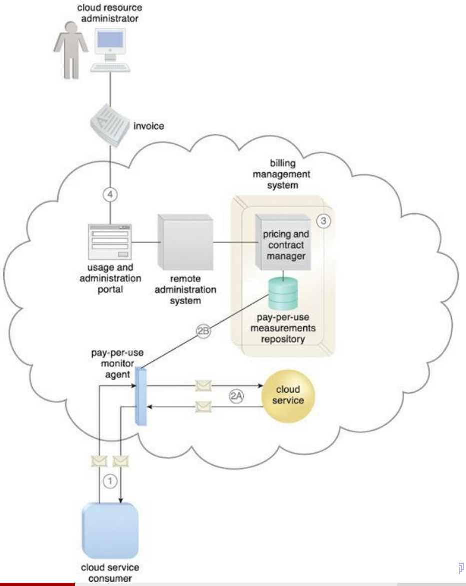
SLA Management System



Billing Management System

- mechanism is dedicated to the collection and processing of usage data as it pertains to cloud provider accounting and cloud consumer billing
- system relies on pay-per-use monitors to gather runtime usage data that is stored in a repository that the system components then draw from for billing, reporting, and invoicing purposes
- allows for the definition of different pricing policies, as well as custom pricing models on a per-cloud consumer and/or per IT resource basis
- Pricing models can vary from the traditional pay-per-use models, to flat-rate or pay-per-allocation modes, or combinations.

Billing Management System


cloud resource administrator
invoice
4
billing management system
3
usage and administration portal
remote administration system
pricing and contract manager
pay-per-use measurements repository
2B
pay-per-use monitor agent
2A
cloud service
1
cloud service consumer