Module III- Computing Mechanisms

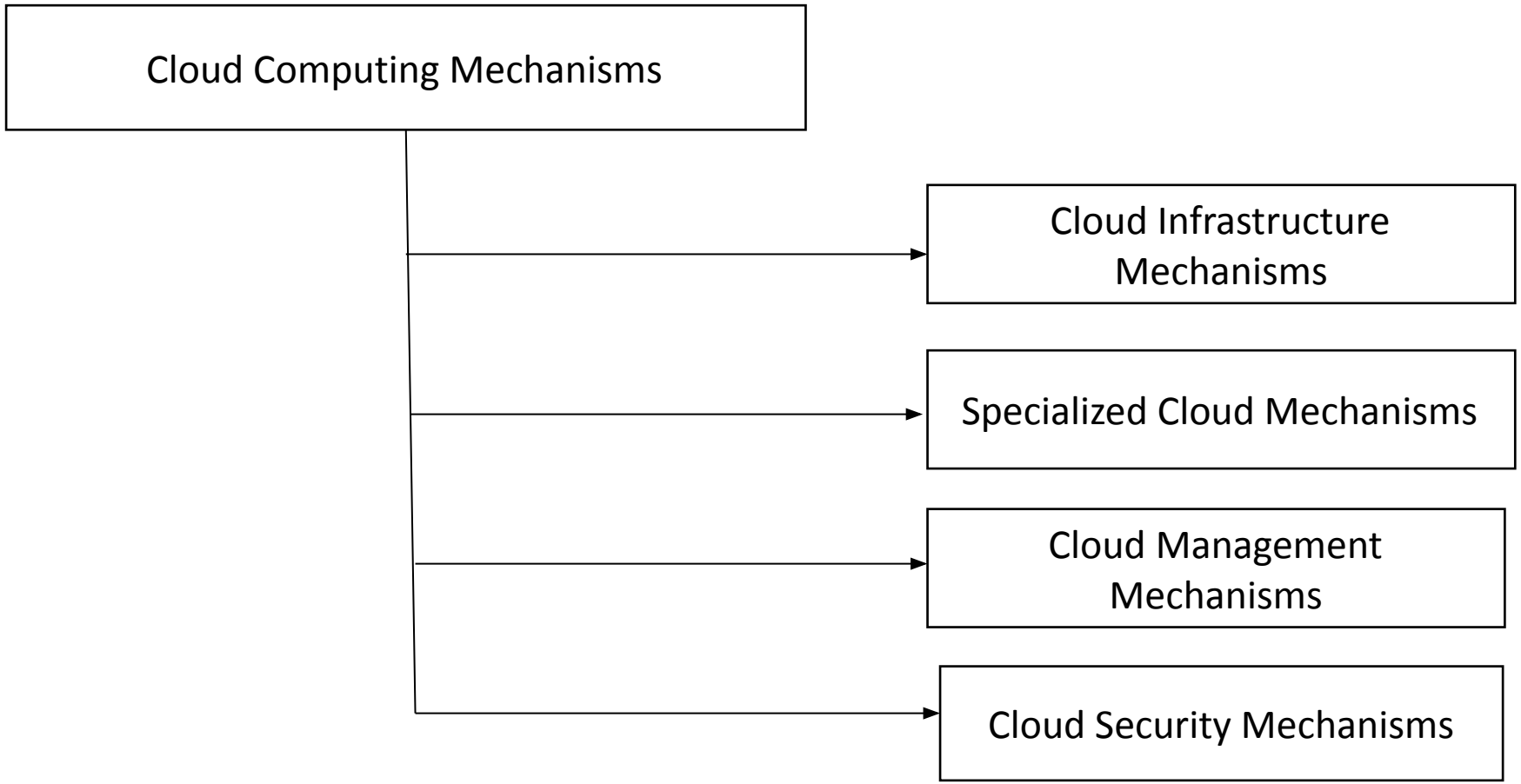
Dr. Madhukrishna Priyadarsini

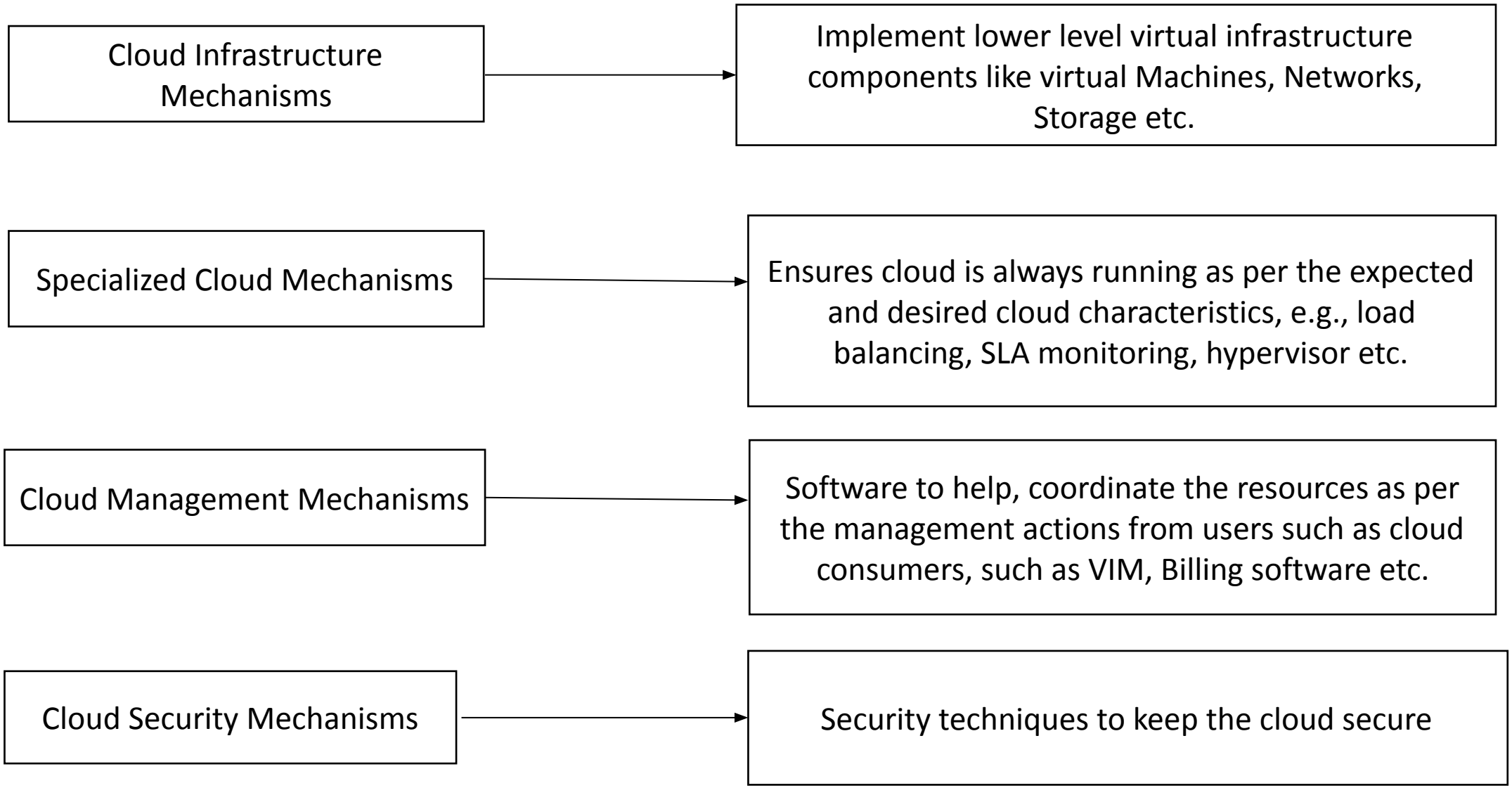
Assistant Professor

NIT Trichy

Introduction

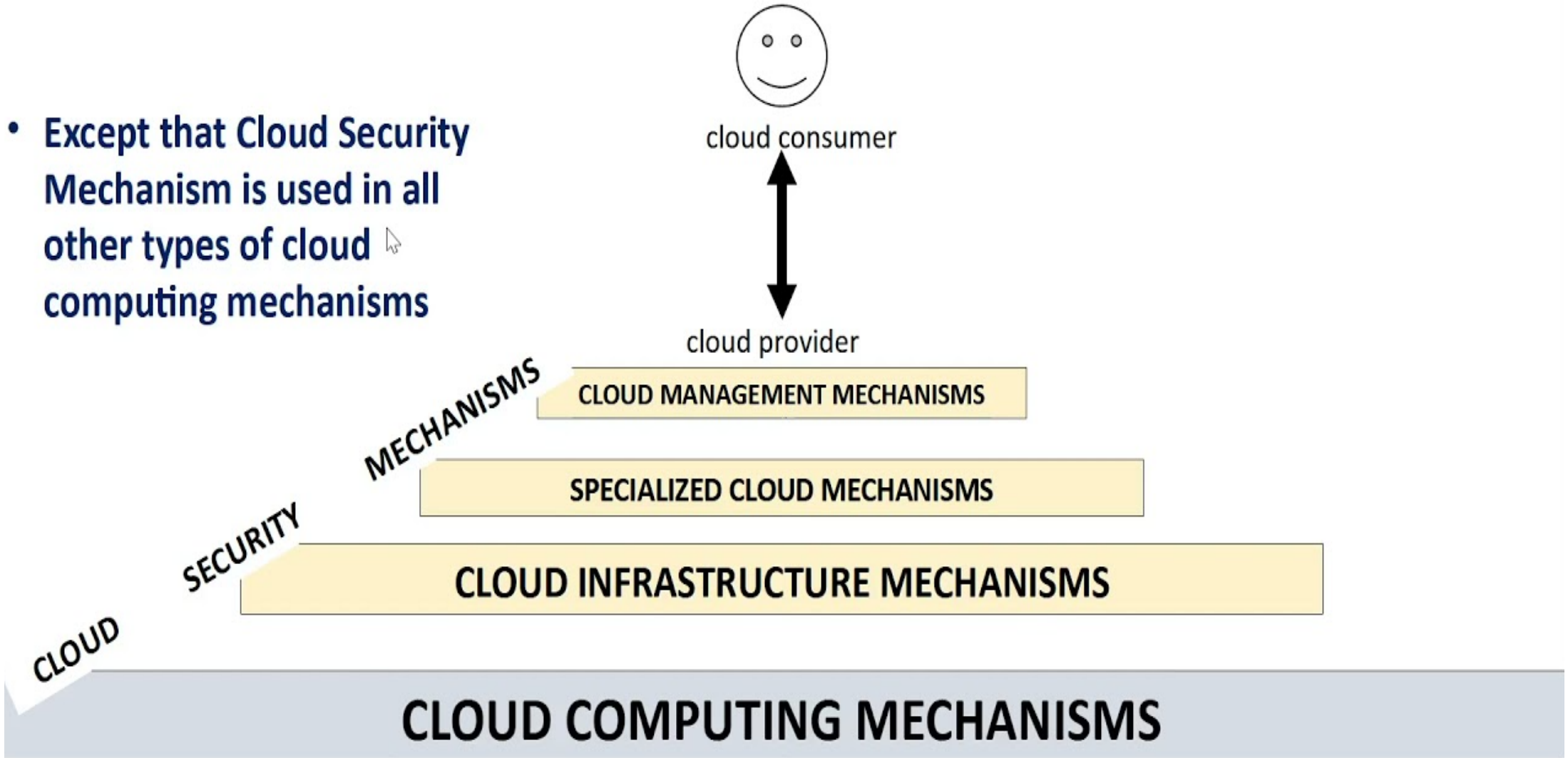
- Mechanism refers to techniques that are used to solve something.
- Cloud computing mechanism consists of all such techniques that are used to enable cloud computing.

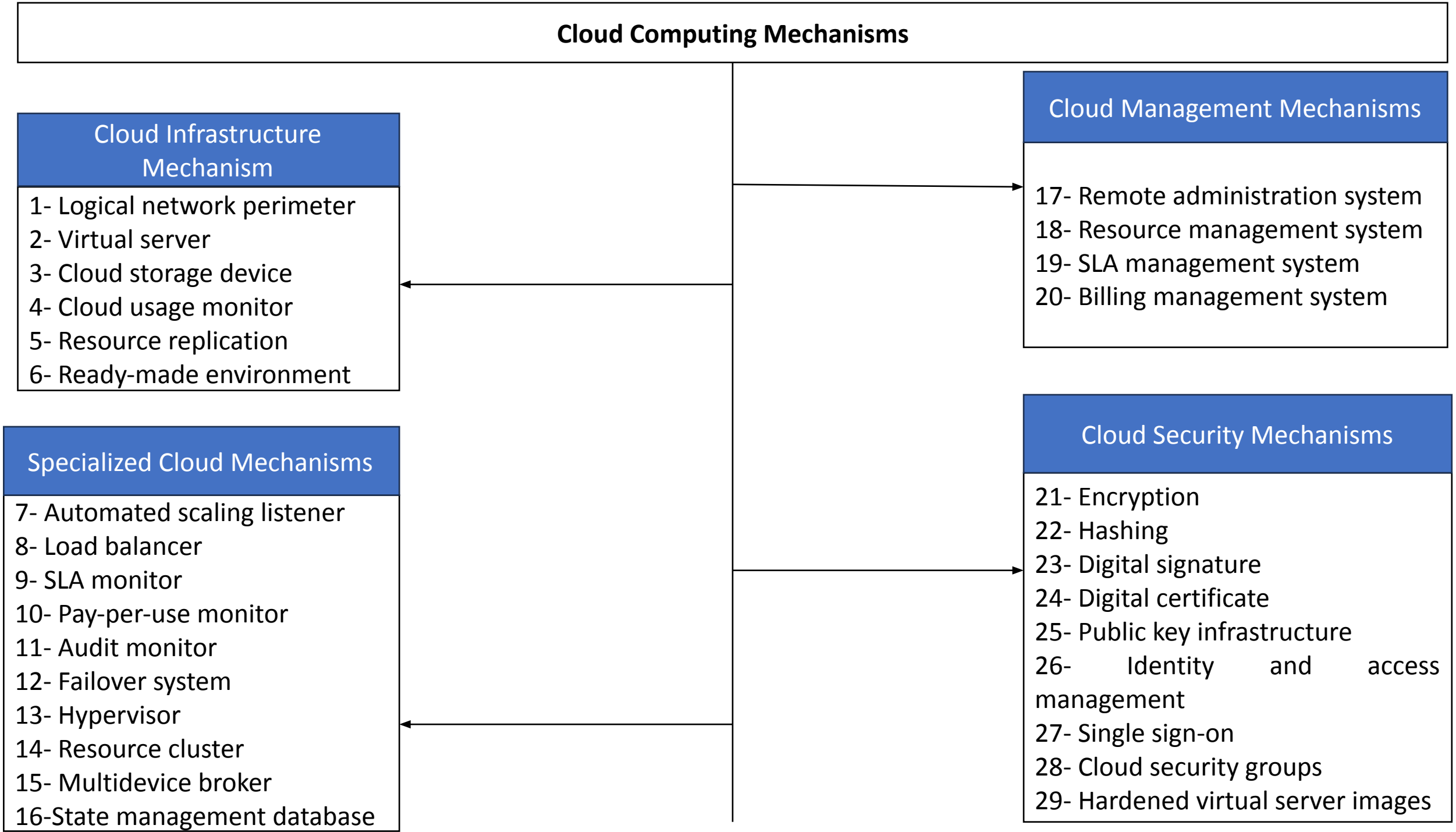



Cloud Infrastructure Mechanisms
Implement lower level virtual infrastructure components like virtual Machines, Networks, Storage etc.
Specialized Cloud Mechanisms
Ensures cloud is always running as per the expected and desired cloud characteristics, e.g., load balancing, SLA monitoring, hypervisor etc.
Cloud Management Mechanisms
Software to help, coordinate the resources as per the management actions from users such as cloud consumers, such as VIM, Billing software etc.
Cloud Security Mechanisms
Security techniques to keep the cloud secure

- Note that each mechanism is supporting the underlying mechanism

- Except that Cloud Security Mechanism is used in all other types of cloud computing mechanisms




Cloud Computing Mechanisms
Cloud Infrastructure Mechanism
1- Logical network perimeter
2- Virtual server
3- Cloud storage device
4- Cloud usage monitor
5- Resource replication
6- Ready-made environment
Cloud Management Mechanisms
17- Remote administration system
18- Resource management system
19- SLA management system
20- Billing management system
Specialized Cloud Mechanisms
7- Automated scaling listener
8- Load balancer
9- SLA monitor
10- Pay-per-use monitor
11- Audit monitor
12- Failover system
13- Hypervisor
14- Resource cluster
15- Multidevice broker
16-State management database
Cloud Security Mechanisms
21- Encryption
22- Hashing
23- Digital signature
24- Digital certificate
25- Public key infrastructure
26- Identity and access management
27- Single sign-on
28- Cloud security groups
29- Hardened virtual server images

Cloud Infrastructure Mechanisms

Cloud infrastructure mechanisms are

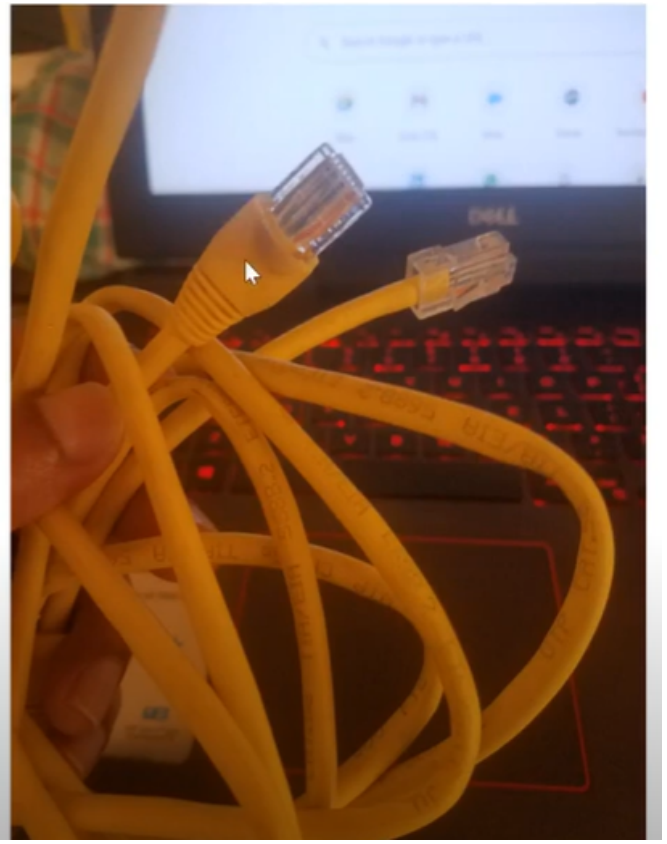
1. Foundational building blocks of cloud environment
2. Establish primary artifacts to form the basis of fundamental cloud technology architecture
3. Infrastructure core components

There are six cloud infrastructure core components/mechanisms that are common to cloud platform:

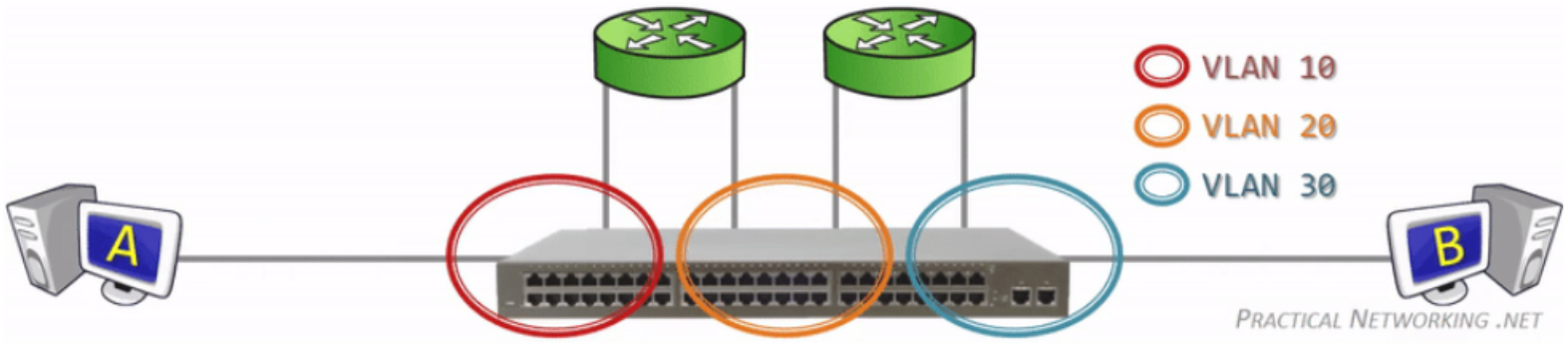
- Logical network perimeter → Techniques to implement **networks** in the cloud
- Virtual server → Techniques to implement **machines** in cloud
- Cloud storage device → Techniques to implement **storage** in cloud
- Cloud usage monitor → Techniques to **monitor usage** data of cloud resources
- Resource replication → Techniques to **replicate/duplicate** resources such as networks, machines, software etc.
- Ready-made environment → Techniques to provide a ready-made platform solution to do something

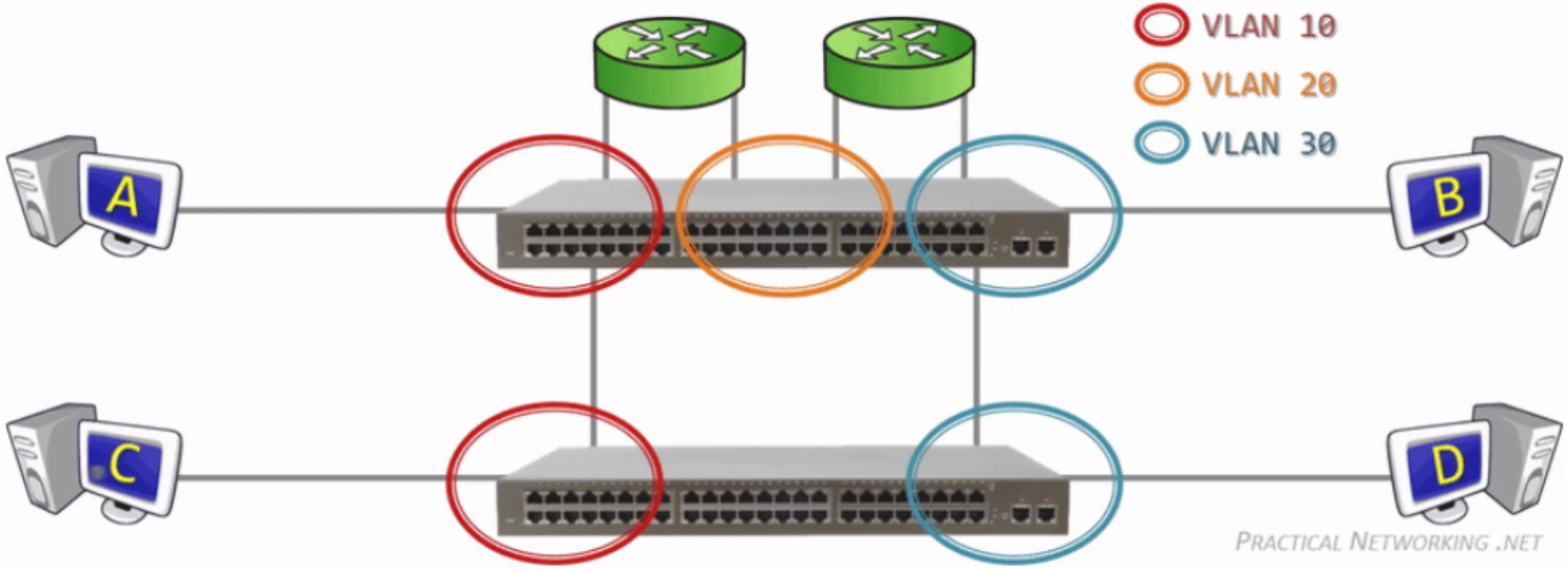
1. Logical Network Perimeter

- Logical network perimeter is isolation of a network environment from the rest of communication networks.
- Logical network perimeter establishes a network boundary that can encompass and isolate a group of related cloud-based IT resources.



VLANs


VLAN 10
VLAN 20
VLAN 30
A
B
PRACTICAL NETWORKING .NET


VLAN 10
VLAN 20
VLAN 30
A
B
C
D
PRACTICAL NETWORKING .NET

Why to use logical network perimeter?

The logical network perimeter mechanism can be implemented to:

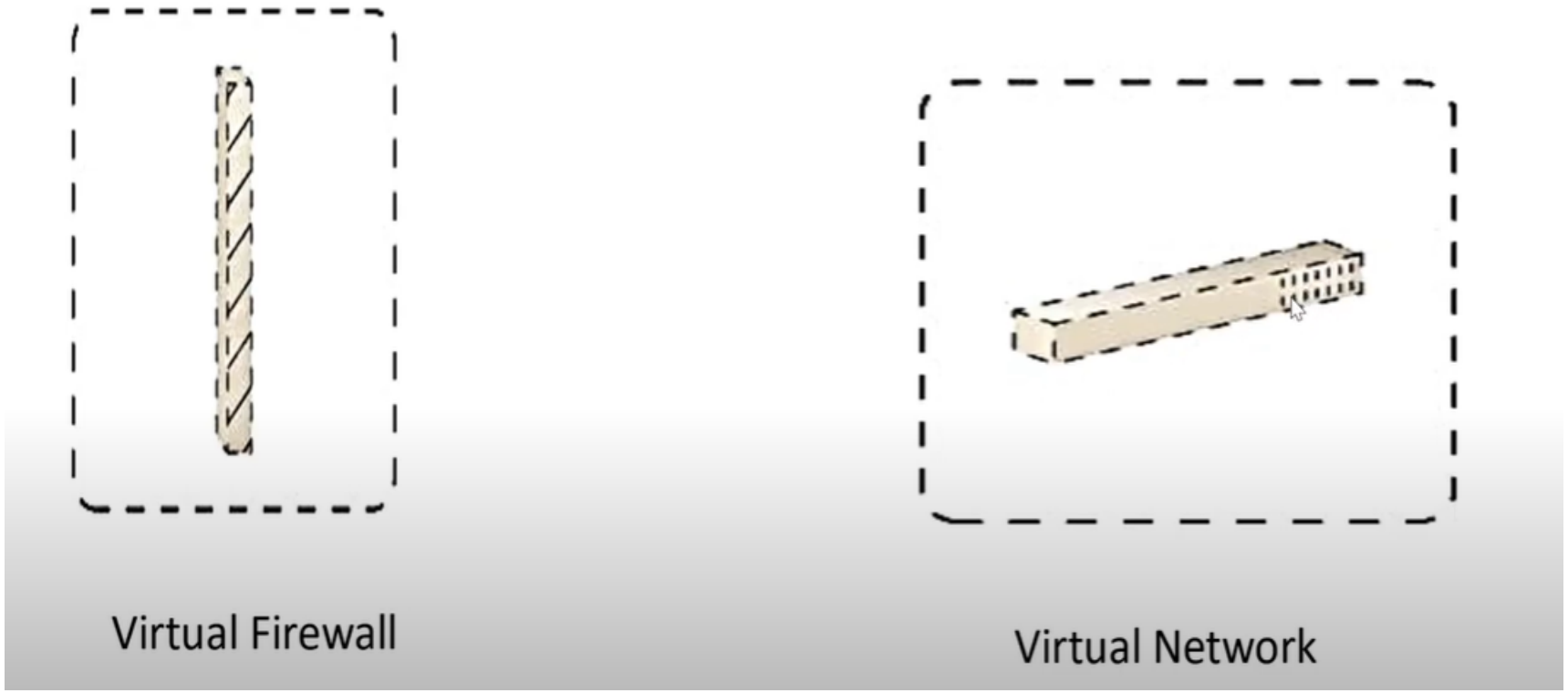
- Isolate IT resources in a cloud from non-authorized users
- Isolate IT resources in a cloud from non-users
- Isolate IT resources in a cloud from cloud consumers
- Control the bandwidth that is available to isolate IT resources

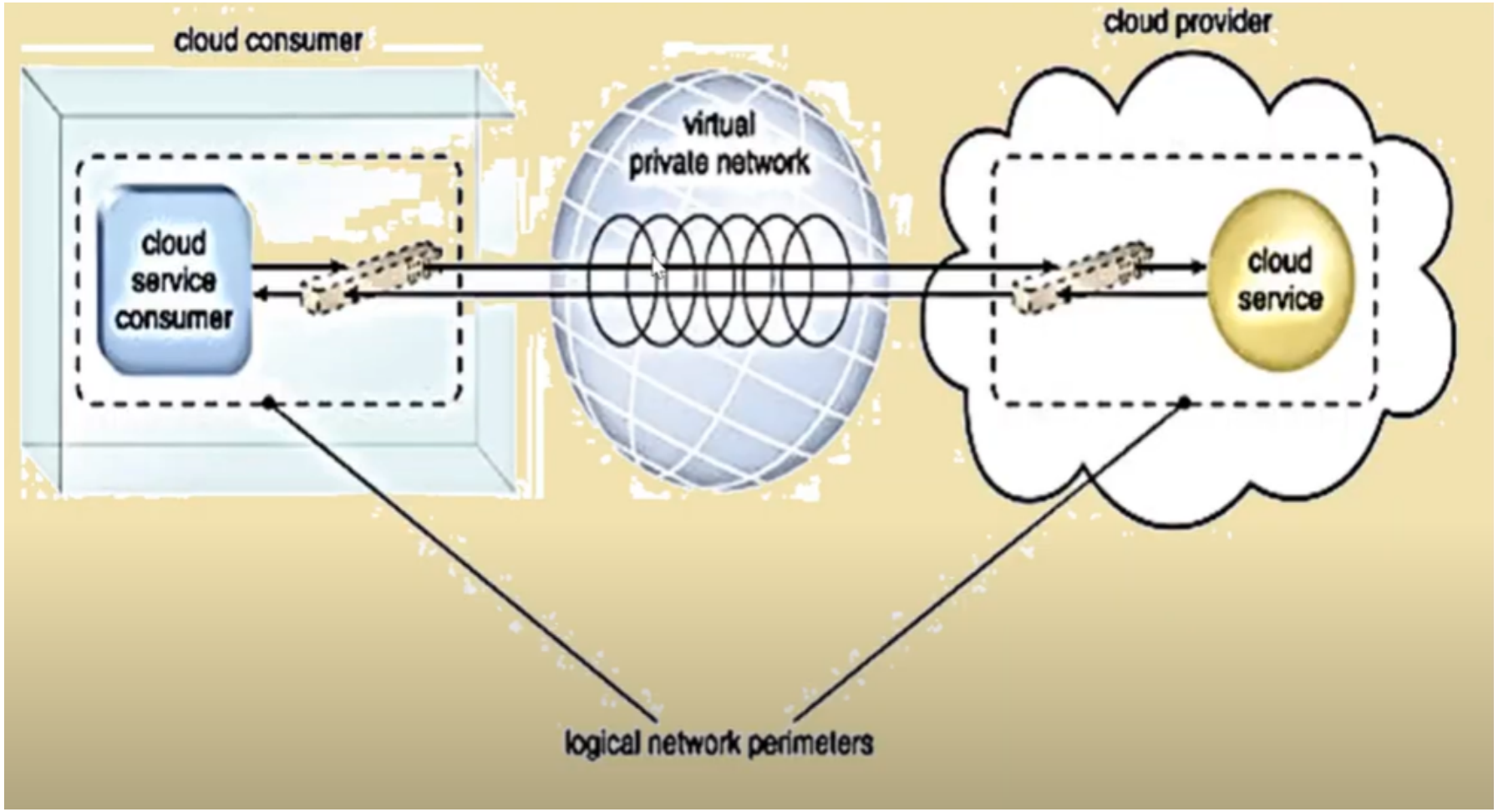
How is logical network perimeter implemented?

- Logical network perimeters are typically established via network devices that supply and control the connectivity of a data center, and are commonly deployed as virtualized IT environments that include:

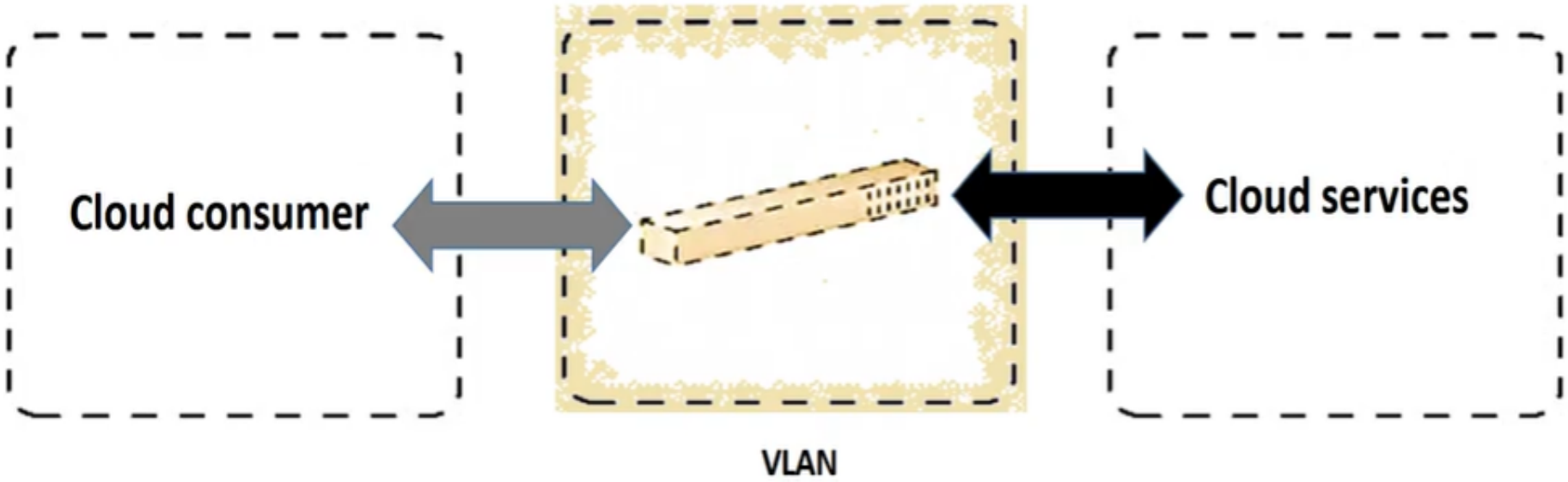
1. Virtual Firewall-> AN IT resource that actively filters network traffic to and from the isolated network such as virtual network while controlling its interactions with the internet.
2. Virtual Network-> Usually acquired through VLANs, this IT resource isolates the network environment within the data center infrastructure.

Symbols used in logical network perimeter





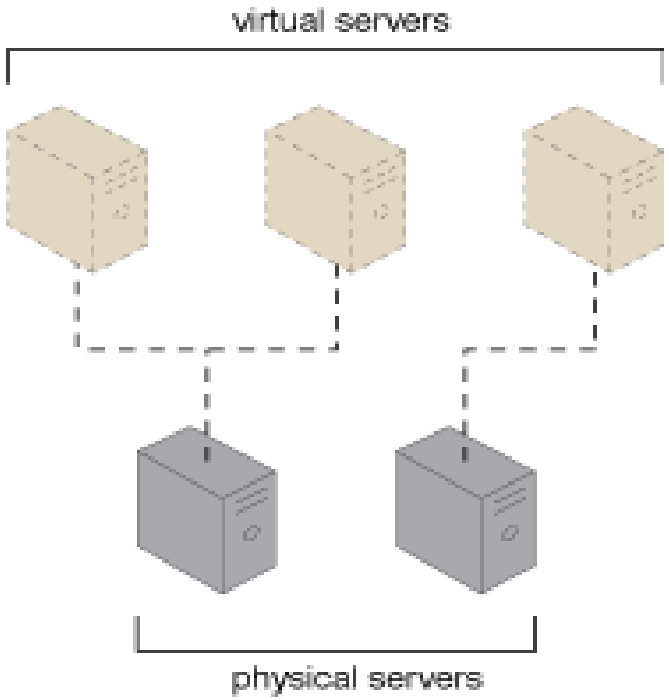
(Two logical network perimeters surround the cloud consumer and cloud provider environments)



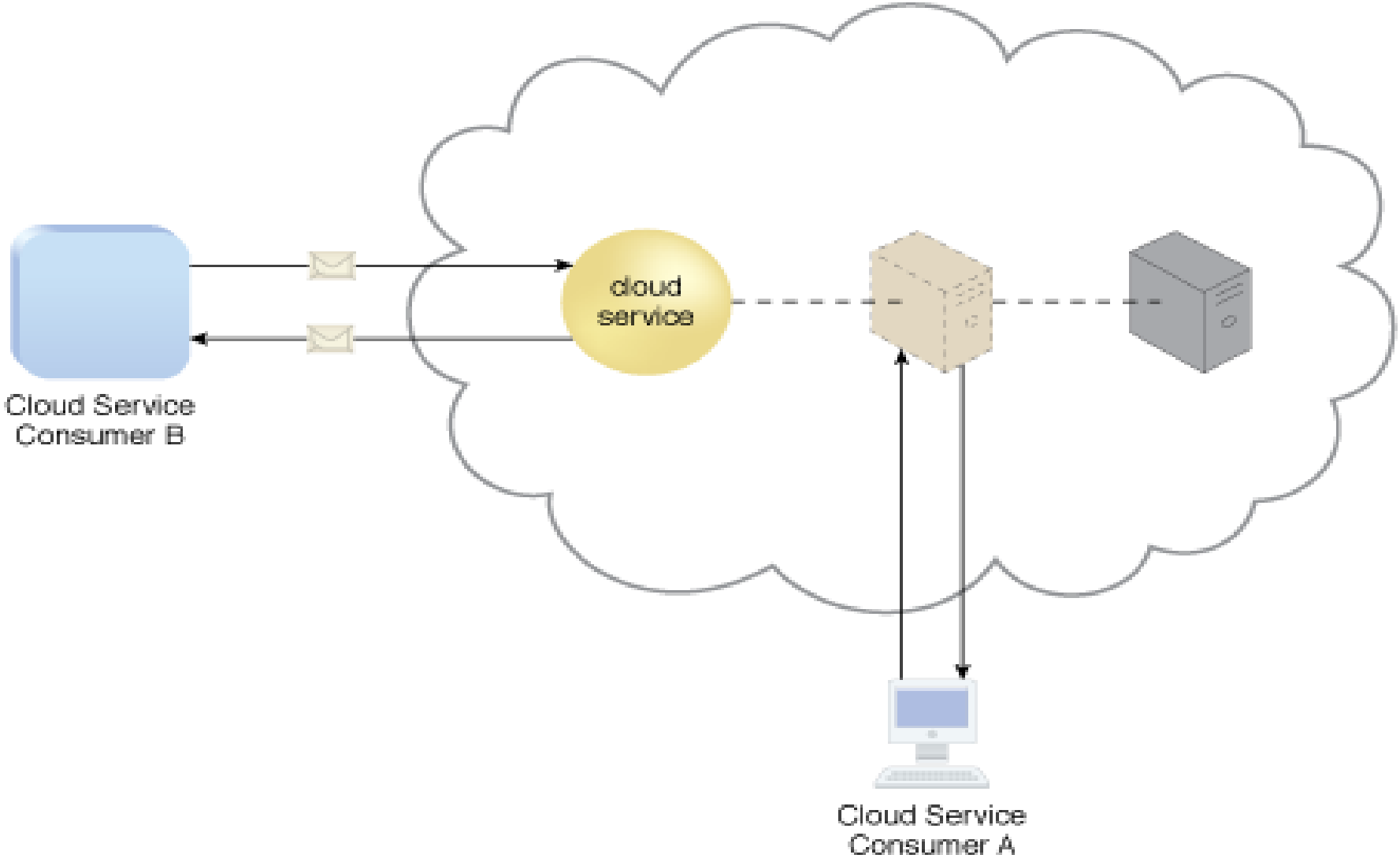
VLAN (virtual network) does network segmentation to keep cloud consumer in a different network whereas cloud services in a different network

2. Virtual Server

- A virtual server is a virtualization software that emulates a physical server
- Virtual servers are used by the cloud providers to share the same physical server with multiple cloud consumers by providing cloud consumers with individual virtual server instances.
- Virtual server is also referred to as virtual machine or virtual instance.

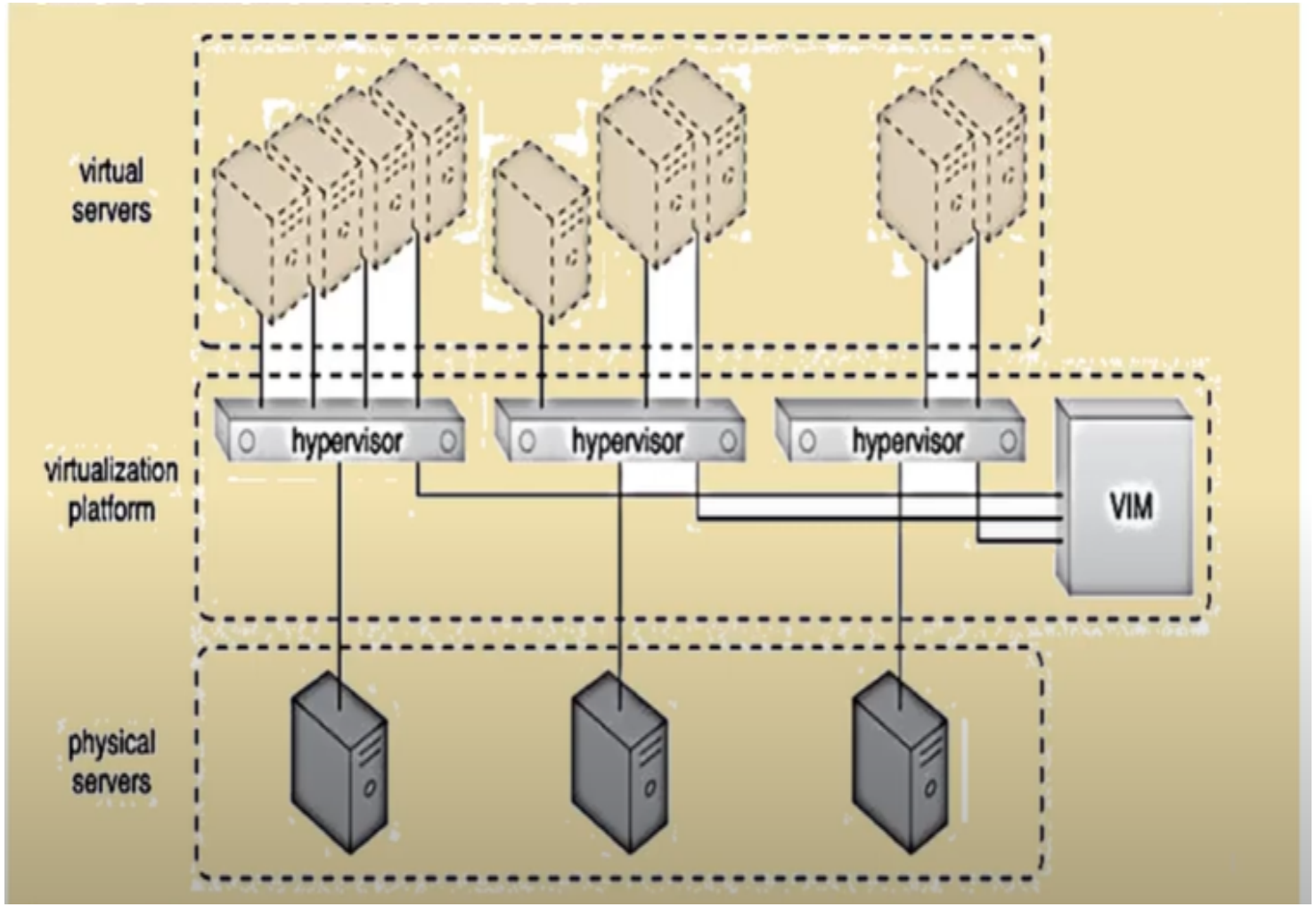


- As a commodity mechanism, the virtual server represents the most foundational building block of cloud environments.
- Virtual server (an IT resource) can be used to host
  - Numerous IT resources (e.g., CPU, storage, memory)
  - Cloud based solutions (such as Java platforms)
  - Cloud computing mechanisms (a service agent program that monitors the cloud)
- Virtual servers are created or instantiated from image files.
- Such a resource allocation process technique can be completed rapidly and on-demand.
- There may be more than one cloud consumers on the same physical server but each using a different virtual server instance.
- Cloud consumers that install or lease virtual servers can customize their environments independently from other cloud consumers that may be using virtual servers hosted by the same underlying physical server.



(A virtual server that hosts a cloud service being accessed by cloud consumer B, whereas the cloud consumer A directly accesses the virtual server for administrative purpose)

- Hypervisor software create and manage one or more virtual server software whereas virtual infrastructure management software (VIM) manages one or more hypervisor.



3. Cloud Storage Device

- The cloud storage device represents storage devices that are represented specifically for cloud based provisioning.
- Instances of these storage devices can be **virtualized**, similar to how physical servers can spawn virtual server images.
- Cloud storage devices are commonly able to provide **fixed increment capacity allocation** in support of the pay-per-use mechanism.
- Cloud storage devices are exposed for remote access via cloud storage services.
- Confidentiality, security, integrity, legal and regulatory requirements are important to on these storage as it contains cloud consumer's data.
- Storages are WAN-based and not LAN-based and hence performance of large databases, network reliability, latency requirement has to be met too.
- Cloud storage device mechanisms are accessible via cloud based API (also referred as storage interface) .

- Cloud storage device mechanism provide common logical units of data storage, such as;

  - **Files**-> Collection of data are grouped into files that are located in folders. Example: in AWS, one can mount file system to store files.

  - **Blocks-**> the lowest level of storage and closest to the hardware, a block is the smallest unit of data that is still individually accessible. Example: in AWS, one can allocate Elastic Block Storage (EBS) for blocked data.

  - **Datasets**-> sets of data are organized into table-based, delimited or record format. Example: in AWS, one can create Relational Database System (RDS) to store data in table format.

  - **Objects**-> data and its associated meta-data are organized as web-based resources. Example: in AWS, one can create Simple Storage Service (S3) objects which can be directly accessed via a web URL.

4. Cloud Usage Monitor

- The cloud usage monitor is a lightweight and autonomous software program responsible for collecting and processing IT resource usage data.

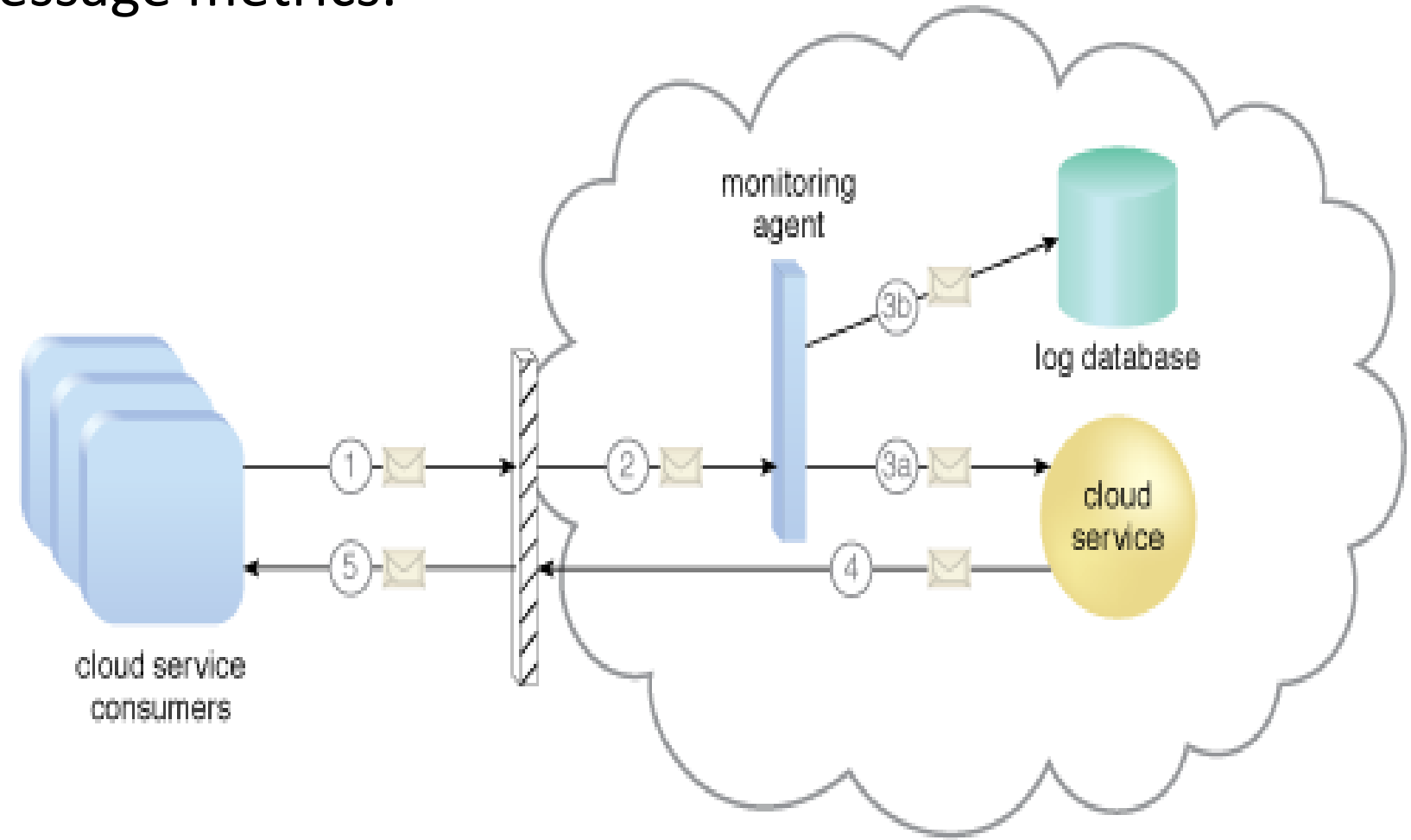
- The IT resources such as virtual server status (CPU, memory, storage), networks (bandwidth, latency, throughput) are examples of usage data that needs to be collected and processed by cloud usage monitor.

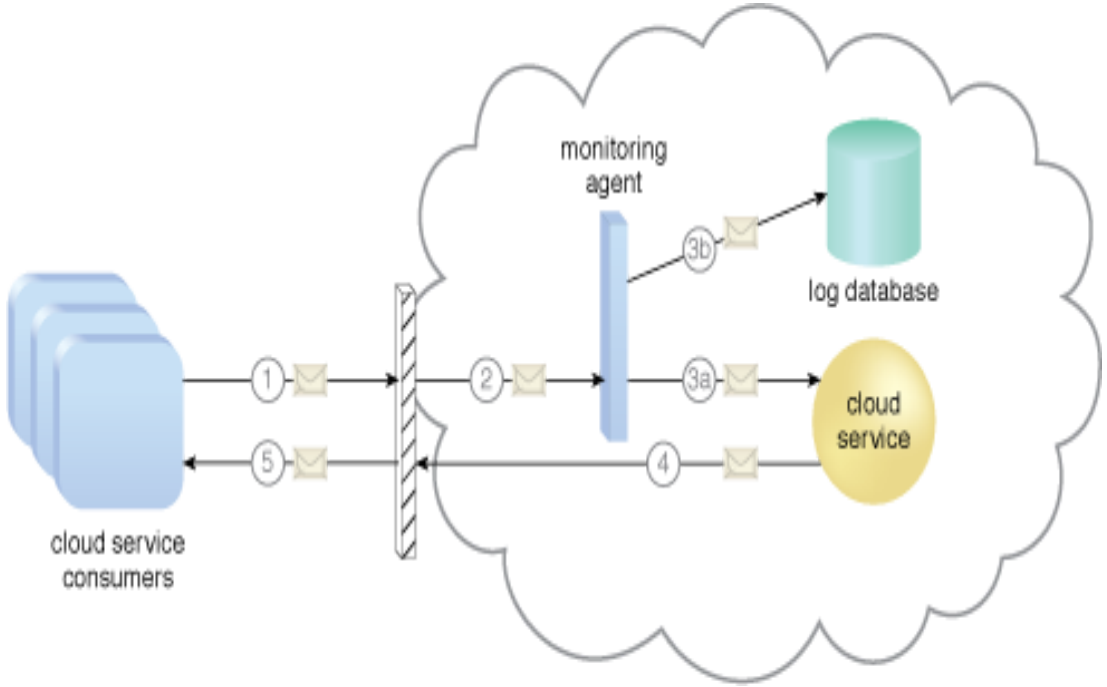
- The monitor collects these data and stores it in a separate log-database for post processing and reporting purposes.

- There are 3 agent-based implementation formats of cloud usage monitor
  - Monitoring agent
  - Resource agent
  - Polling agent

Monitoring agent-> is an intermediary, event-driven program that exists as a service agent and resides along existing communication paths to transparently monitor and analyze.

- This type of cloud usage monitor is commonly used to measure network traffic and message metrics.

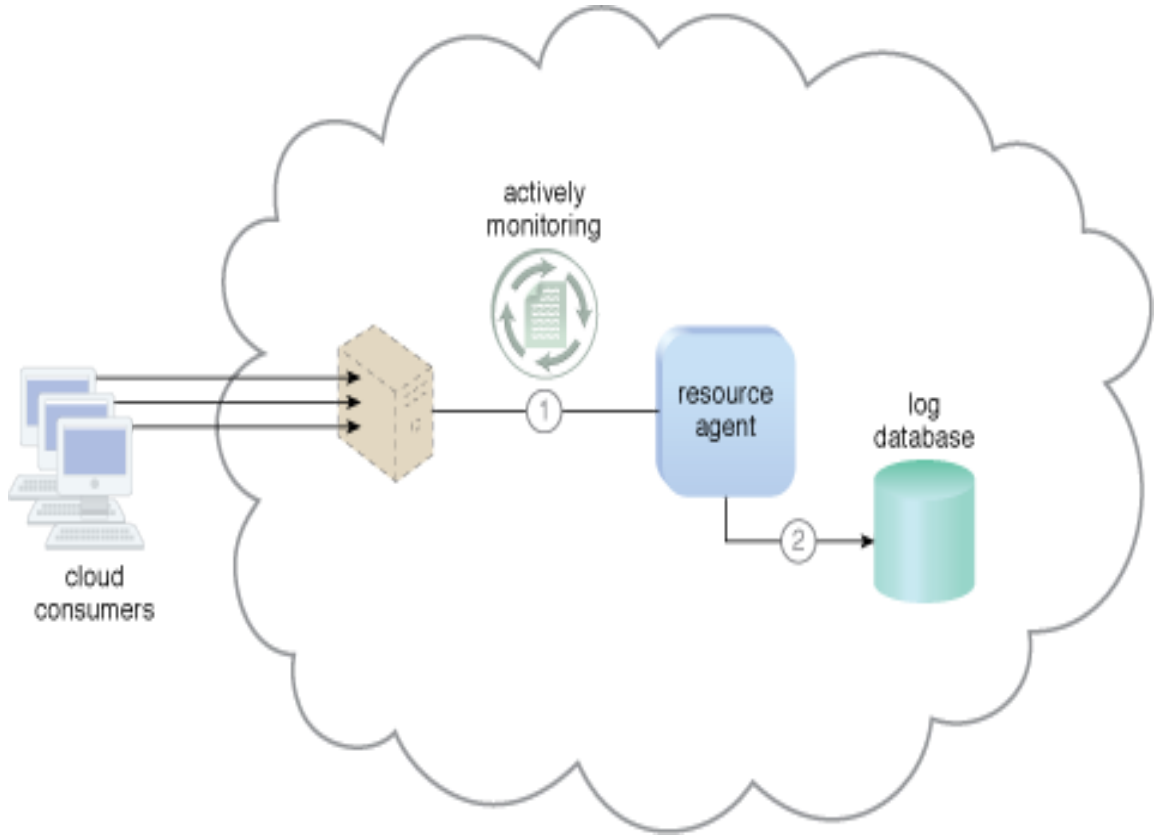




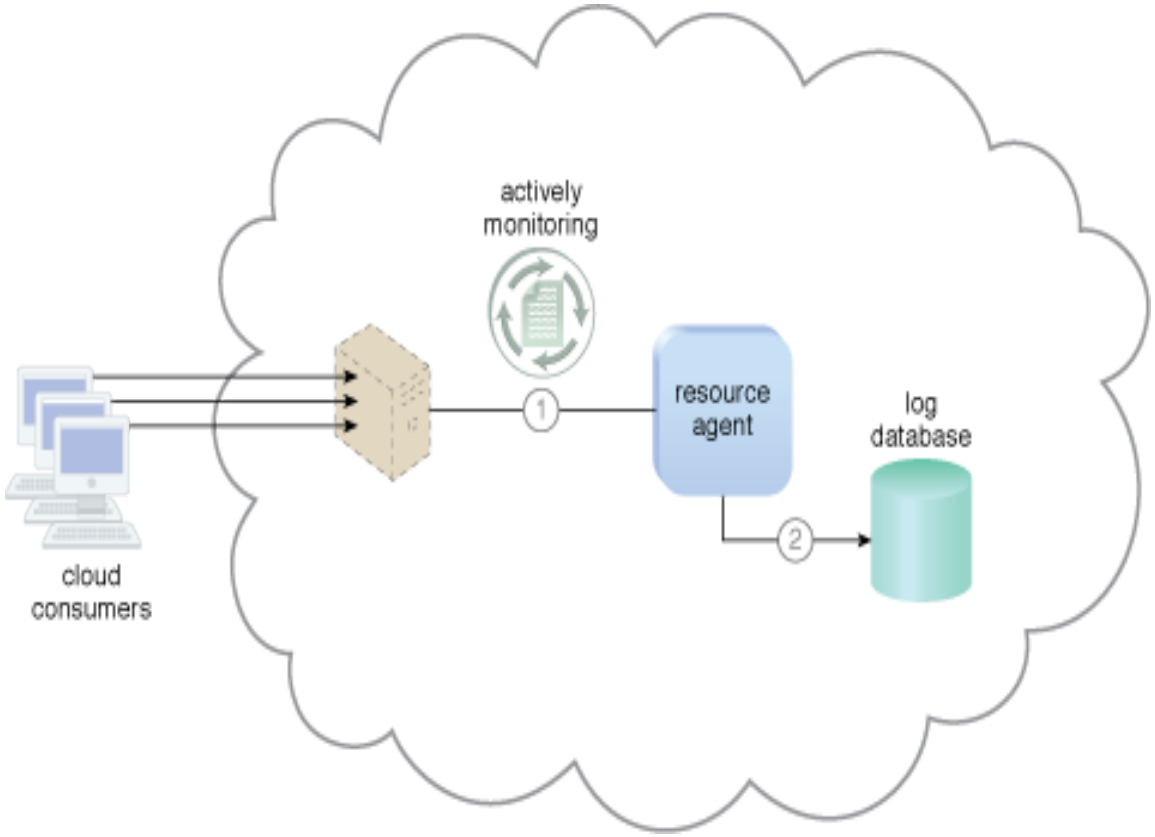
- A cloud service consumer sends a request message to cloud service (1)
- The monitoring agent intercepts the message to collect relevant usage data (2)
- Before allowing it to continue cloud service (3(a))
- The monitoring agent stores the collected usage data in a log data base (3(b))
- The cloud service replies with a response message (4)
- The response is sent back to the cloud service consumer without being interpreted by the monitoring agent (5)

Resource Agent-> is a processing module that collects usage data by having event-driven interactions with specialized resource software.

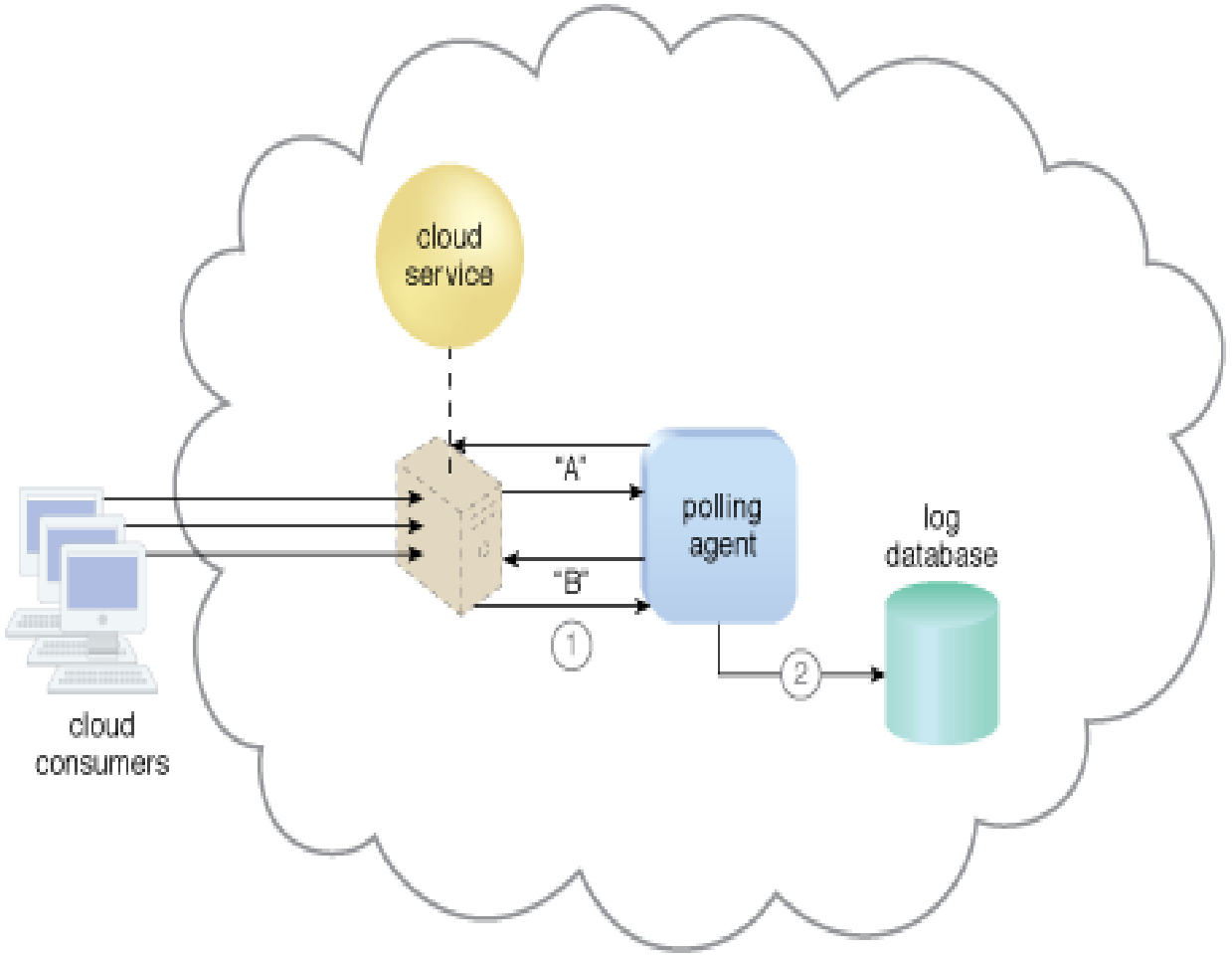
- The module is used to monitor usage metrics based on pre-defined, observable events at the resource software level, such as initiating, resuming, suspending, and vertical scaling



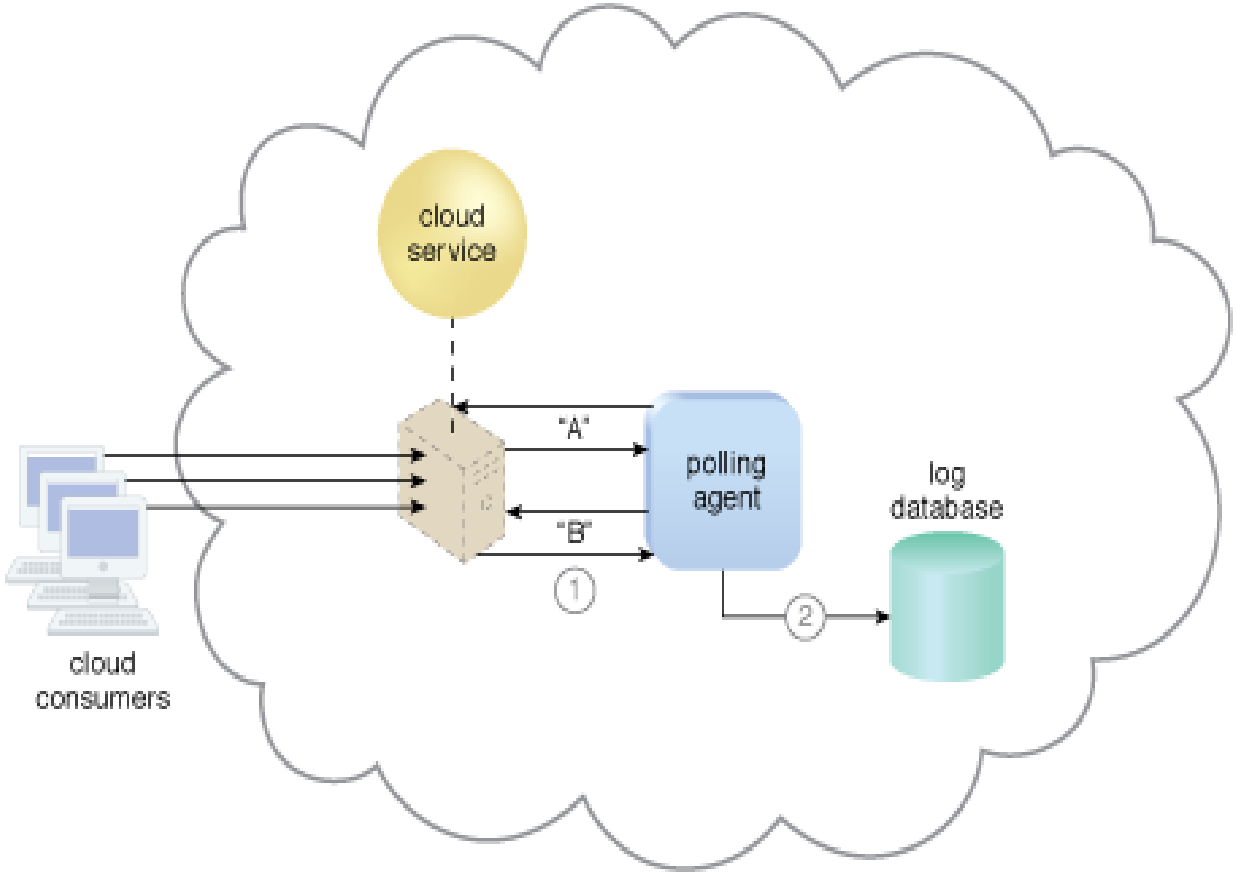
- The resource agent is actively monitoring a virtual server and detects an increase in usage (1)
- The resource agent receives a notification from the underlying resource management program that virtual server is being scaled up and stores the collected usages data in a log database, as per its monitoring metrics (2)



Pooling Agent-> is a processing module that collects cloud service usage data by polling IT resources.



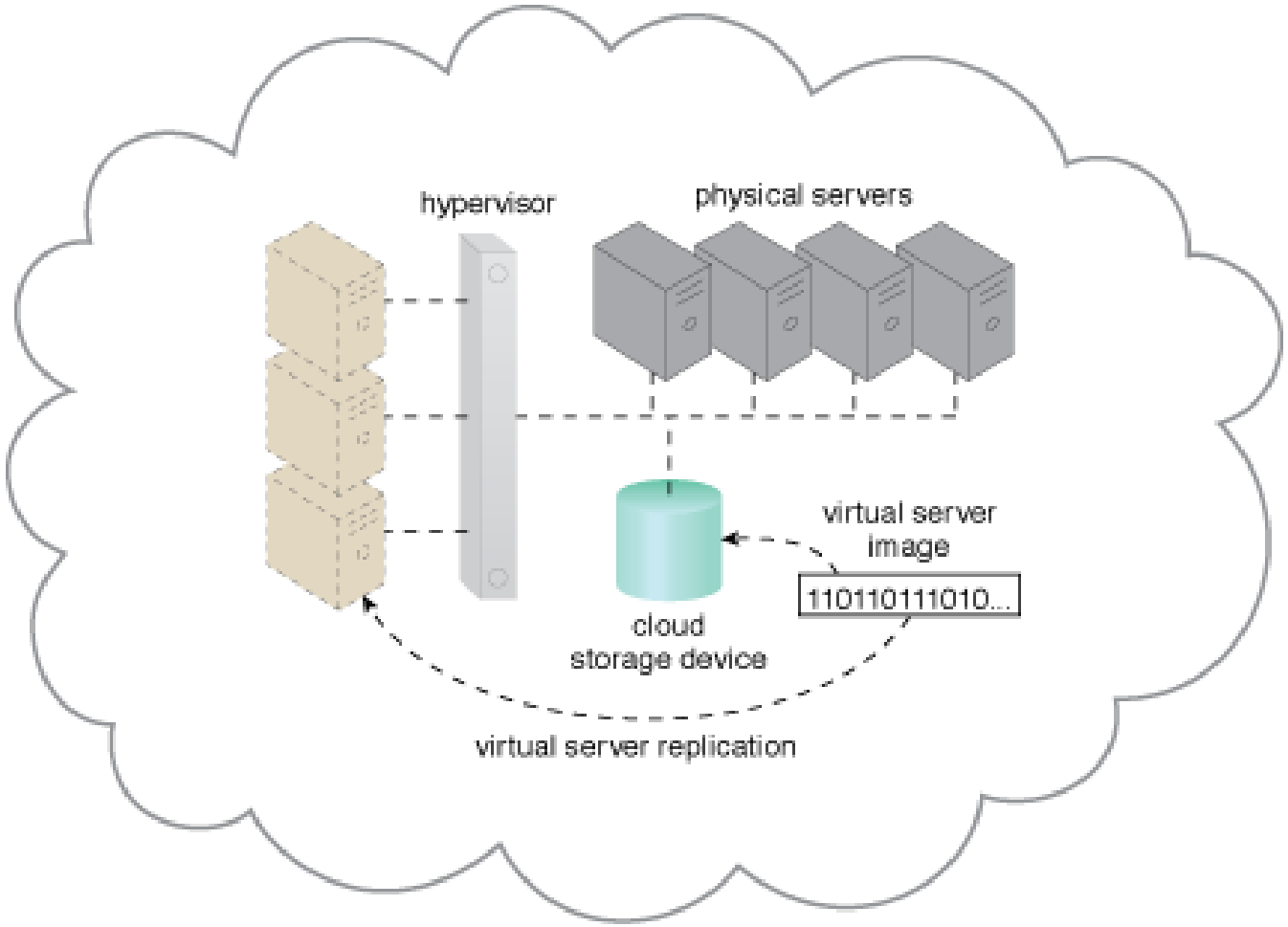
- Polling agent monitors the status of a cloud service hosted by a virtual server by sending periodic polling request messages and receiving polling response messages that report usage status “A” after a number of polling cycles until it receives a usage status of “B” (1)
- Upon which the polling agent records the new usage status in the log database (2)



- Difference between resource agent and polling agent is the direction of request and response.
- In resource agent mechanism, the resource event is sent to the resource agent by a software (eg. A hypervisor or VIM sends information about virtual server). The agent takes a note of the event and logs it into the database.
- In polling agent mechanism, the polling agent polls the resource for the status change. Initially it receives the status A , and while the polling continues the status changes to B, and the polling agent records the status change in the log database. Example: virtual machine has moved from “pending” to “starting” to “started”.
- There is one time event notification in case of resource agent.
- There is continuous polling in case of polling agent.

5. Resource Replication

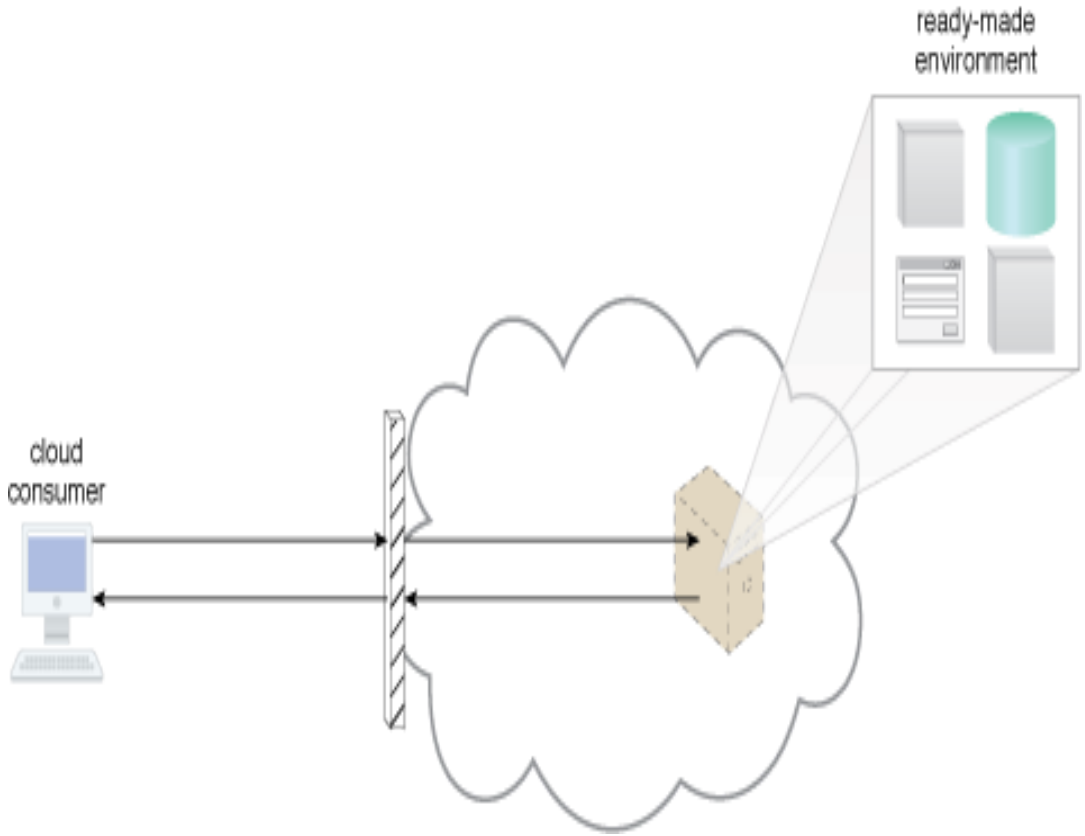
- Resource replication can be taught of as multiplying the existing infrastructure such as a virtual machine.



- Resource replication is defined as creation of multiple instances of the same IT resource, replication is typically performed when an IT resource’s availability and performance need to be enhanced.
- The hypervisor replicates several instances of a virtual server using stored virtual server image.
- Such replication is independent of the location where the resource needs to be replicated.
- Such replication technique makes a resource available at any place.
- In case of resource failure, the resource can be replicated anywhere.

6. Ready-made Environment

- Ready-made environment can be thought of as some platform that is provided ready-made. It will already be installed and start with default configuration.



- A ready-made environment can be a Java platform, a Hadoop platform, a web server like Apache and database MySQL may come pre-installed, pre-configured etc.
- A cloud consumer accesses a ready-made environment hosted on a virtual server.
- The ready-made environment mechanism is a defining component of the PaaS cloud delivery model, that represents a pre-defined cloud-based platform comprised of a set of already-installed IT resources, ready to be used, and customized by a cloud consumer.
- These environments are utilized by cloud consumers to remotely develop and deploy their own services and applications within a cloud.
- Typical ready-made environments include pre-installed IT resources such as database, middleware, development tools, and governance tools.
- A ready-made environment is generally equipped with a complete software development kit (SDK) that provides cloud consumers with programmatic access to the development technology that comprise their preferred programming stacks.

Specialized Cloud Mechanisms

- A typical cloud technology architecture contains numerous moving parts to address distinct usage requirements of IT resources and solutions.
- Specific cloud mechanism contains a specific runtime function in support of one or more cloud characteristics.

The cloud characteristics that specialized cloud mechanisms tries to support are

- On demand usage
  - Ability of the cloud provider to provide any service at any time.
- Ubiquitous access
  - Ability of the cloud provider to provide any service anywhere on any device.
- Multi-tenancy (resource pooling)
  - Ability of the cloud provider to manage multiple cloud consumers by sharing limited physical resource.
- Elasticity
  - Ability of the cloud provider to let consumers grow and shrink its resource as per the application needs
- Measured usage
  - Ability of the cloud provider to measure the quantity and quality of usages of any particular resource.
- Resiliency
  - Ability of the cloud provider to ensure that applications can heal itself and be made available again irrespective of any attack or disaster.

There are 10 specialized cloud mechanisms that are used in cloud platforms

7- Automated scaling listener → Program which listens to resources configured for scaling needs

8- Load balancer → Program which distributes the load across multiple IT resources

9- SLA monitor → Program which monitors for SLA specific data

10- Pay-per-use monitor → Program which tracks the billing usage and stops the resource usage if required

11- Audit monitor → Program which monitors the cloud ensure regulations and contract obligations

12- Failover system → A system that ensures duplicated redundant system is available in case of system failures

13- Hypervisor → A software that manages virtual servers

14- Resource cluster → A software that manages a cluster of similar dispersed resources

15- Multidevice broker → A broker software that is tuned to interface with any user devices eg. Mobile, desktop

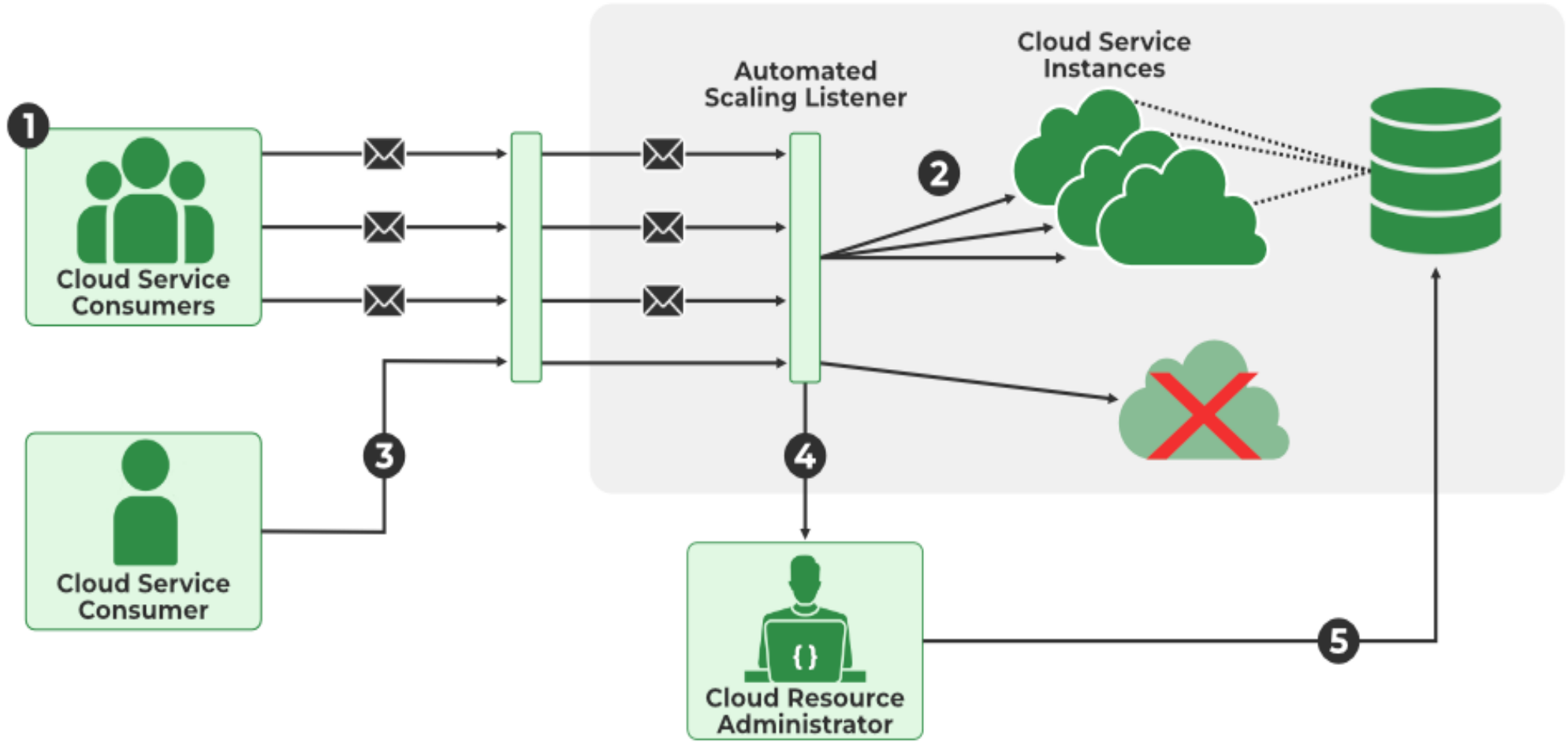
16-State management database → A shared database between similar resources that stores service data temporarily to enable restoration of services or functionality on any redundant resource in case of any failure or need

7. Automated Scaling Listener

- What scaling means?
  - Scaling refers to grow or shrink something as needed. For eg., we need more virtual servers to manage a lot of web requests, we scale out. Sometimes, there is hardly any web requests, we do not need more virtual servers, we scale in.
- The automated scaling listener mechanism is a service agent that monitors and tracks communication between cloud service consumers and cloud services for dynamic scaling purposes.
- Automated scaling listeners are deployed within the cloud, typically near the firewall, from where they automatically track workload status information.
  - This is because cloud consumers first have to cross the firewall before accessing any services. So, this is the best place to have your program to listen, to know the quantity of requests, and scale out and scale in quickly.

- Workload can be determined by the
  - Volume of cloud consumer generated requests or
  - Via back-end processing demands triggered by certain types of requests
  - Note that, even though the request may be little, the workload may be huge. For example; a small amount of incoming data can result in a large amount of processing.

- Automated scaling listeners can provide different types of responses to workload fluctuation conditions such as;
  - Automatically scaling IT resources out or in based on parameters previously by the cloud consumer. (In AWS, this happens via auto scaling)
  - Automatic notification to the cloud consumer when workloads exceed current thresholds or fall below allocated resources. The cloud consumer can choose to configure auto scale or manually scale his or her resources.


Automated
Scaling Listener
Cloud Service
Instances
1
Cloud Service
Consumers
2
3
4
Cloud Service
Consumer
Cloud Resource
Administrator
5

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

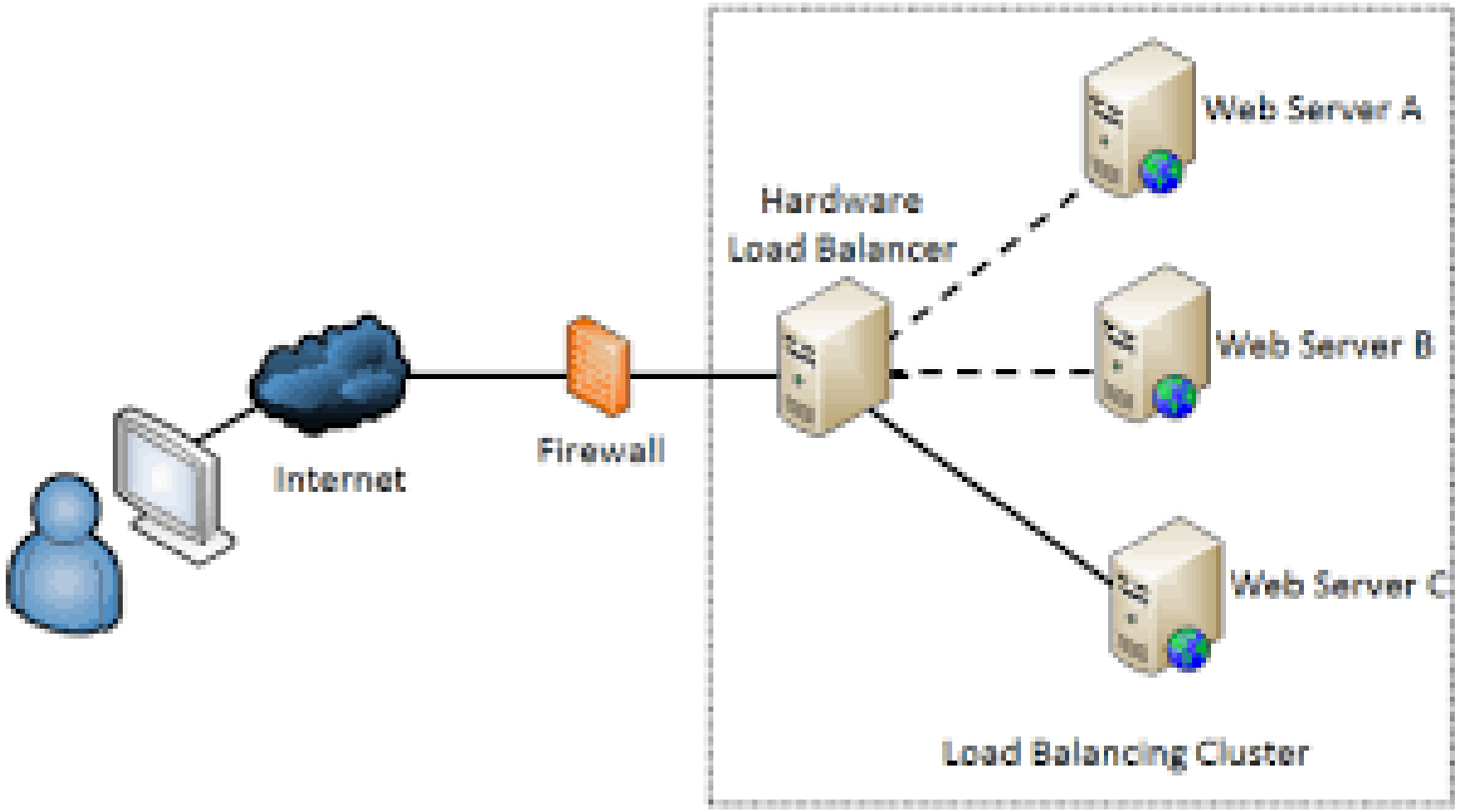
- **Horizontal scaling Vs. vertical scaling**

- Scaling as we know that something grows or shrinks. Now, the question is in which direction it scales?

- **Horizontal Scaling->** Scales out or scales in. It tries to create another resource of the same type. Example: when virtual servers of similar configuration is added to manage the load, is called that it is horizontally scaling out.

- **Vertical Scaling->** Scales up or scales down. It tries to increase or decrease the capacity of the resource by replacing it with that of a scaled requirement. Example; cloud consumer have registered for storage of 1 GB, then after a year he/she vertically scales up the storage to 5 GB.

8. Load Balancer



Beyond simple division of labor algorithms, load balancer can perform a range of specialized runtime workload distribution functions that include

- Asymmetric Distribution-> Larger workloads are issued to IT resources with higher processing capacities.

Example: upload of bank reconciliation will be balanced moreover download of banking statements.

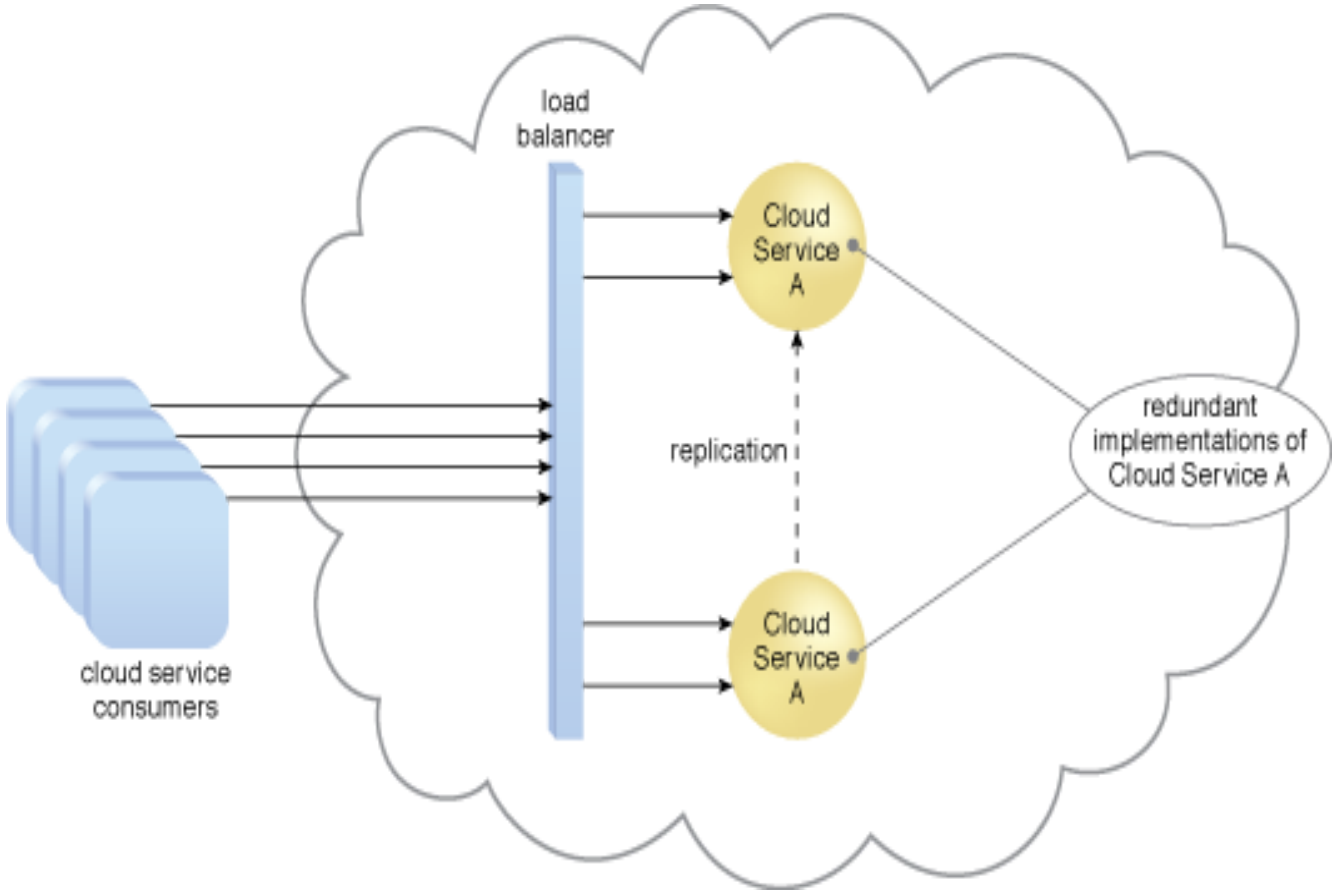
- Workload Prioritization-> Workloads are scheduled, queued, discarded, and distributed according to their priority levels.

Example: financial payments may be balanced more over checking balance on website.

- Content-Aware Distribution-> Requests are distributed to different IT resources detected by the request content.

Example: videos may be balanced moreover text data.

- A load balancer is implemented as a service agent, transparently distributes incoming workload request messages across two redundant cloud service implementations, which in turn maximizes performance for the cloud service consumer.



- The load balancer mechanism can exist as a
  - Multi-layer network switch
  - Dedicated hardware appliance
  - Dedicated software-based system (common in server operating system)
  - Service agent (usually controlled by cloud management software)

- The load balancer is typically located on the communication path between the IT resources generating the workload and the IT resources performing the workload processing.

- The mechanism can be designed as a transparent agent that remains hidden from the cloud service consumers, or as a proxy component that abstracts the IT resources performing their workload.

9. SLA Monitor

- What is SLA (Service Level Monitor)?

- Example: when you get a car, you can not see how its designed internally or what quality it is, but you may agree to buy a car based on a mileage it can offer. Your mileage is something that you can measure or through which you can test the product, and the SLA with the manufacturer could be “Car should give 50km/hr mileage”.

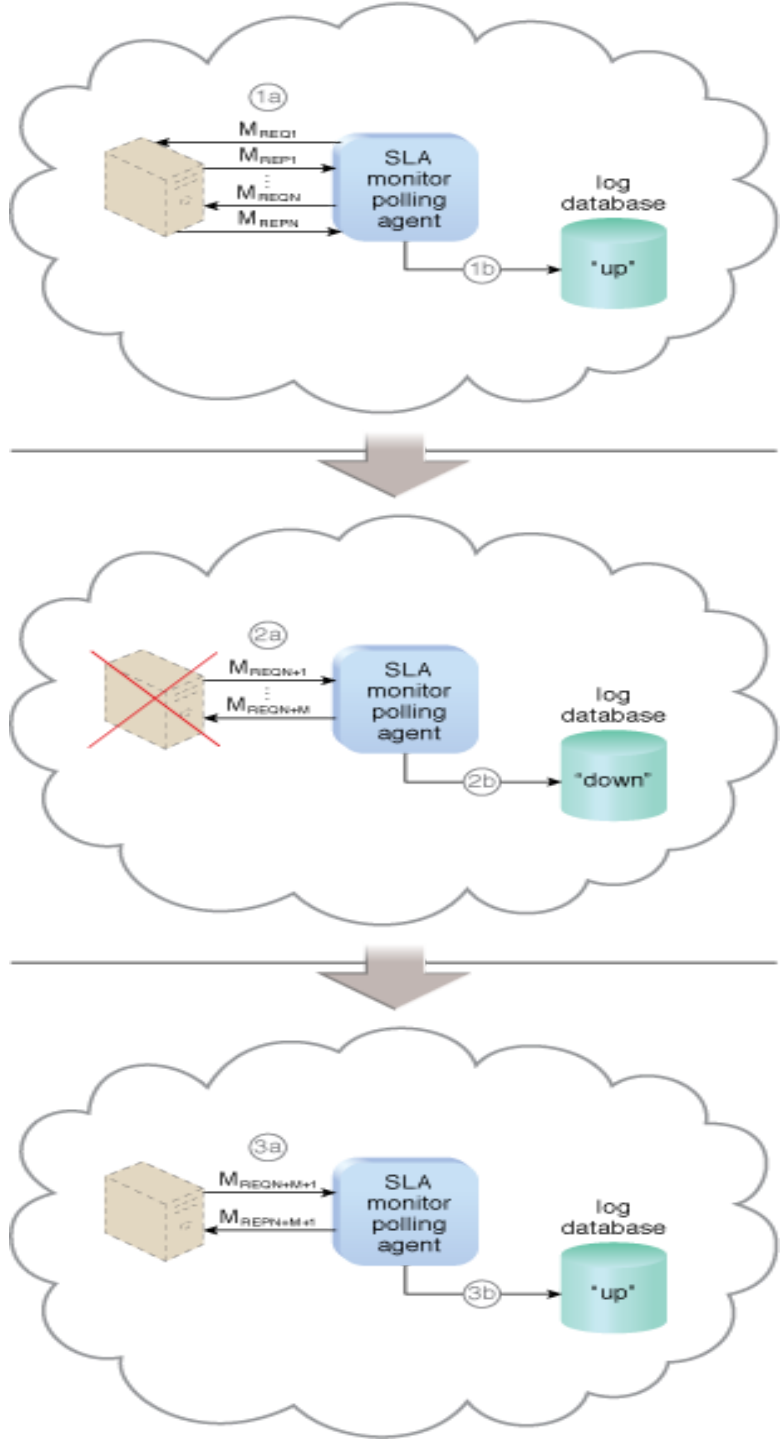
- An SLA provides details of various measurable characteristics related to IT outcomes such as, up time, security characteristics, and other specific QoS features, including availability, reliability, and performance. Since the implementation of the service is hidden from the cloud consumer, an SLA becomes a critical specification.

- SLA monitor essentially monitors if the services are meeting the SLAs between the cloud consumer and cloud provider.

- The SLA monitor mechanism is specifically used to observe the run time mechanism of the cloud services to ensure that they are fulfilling the contractual QoS requirements that are published in SLA.

- The data collected by the SLA monitor is processed by an SLA management system to be aggregated in to a SLA reporting metrics.

- The system can proactively repair the failover cloud services when exception conditions occur, such as when the SLA monitor reports a cloud service as “down”.


1a
M_{REQ1}
M_{REP1}
M_{REQN}
M_{REPN}
SLA monitor polling agent
log database
1b
"up"
2a
M_{REQN+1}
M_{REQN+M}
SLA monitor polling agent
log database
2b
"down"
3a
$M_{REQN+M+1}$
$M_{REPN+M+1}$
SLA monitor polling agent
log database
3b
"up"

- The SLA monitor polls the cloud service by sending over polling request messages (M to M). The monitor receives polling response messages (M to M) that report the service was “up” at each polling cycle 1(a).
- The SLA monitor stores the “up” time – time period of all polling cycles 1 to N in the log database 1(b).
- The SLA monitor polls the cloud service that sends polling request messages (M to M). The polling response messages are not received 2(a).
- The response messages continue to time out, so the SLA monitor stores the “down” time- time period of all polling cycles N+1 to N+M in the log database 2(b).
- The SLA monitor sends a polling request message (M) and receives the polling response message (M) 3(a).
- The SLA monitor stores the “up” time in the log database.

10. Pay-per-use Monitor

- The pay-per-use monitor measures cloud-based IT resource usage in accordance with pre-defined pricing parameters and generates usage logs for free calculations and billing purposes.

- The pricing may be based on
  - Request/response message quantity
  - Transmitted data volume
  - Bandwidth consumption

- The data collected by the pay-per use monitor is processed by a billing management system that calculates the payment fees.

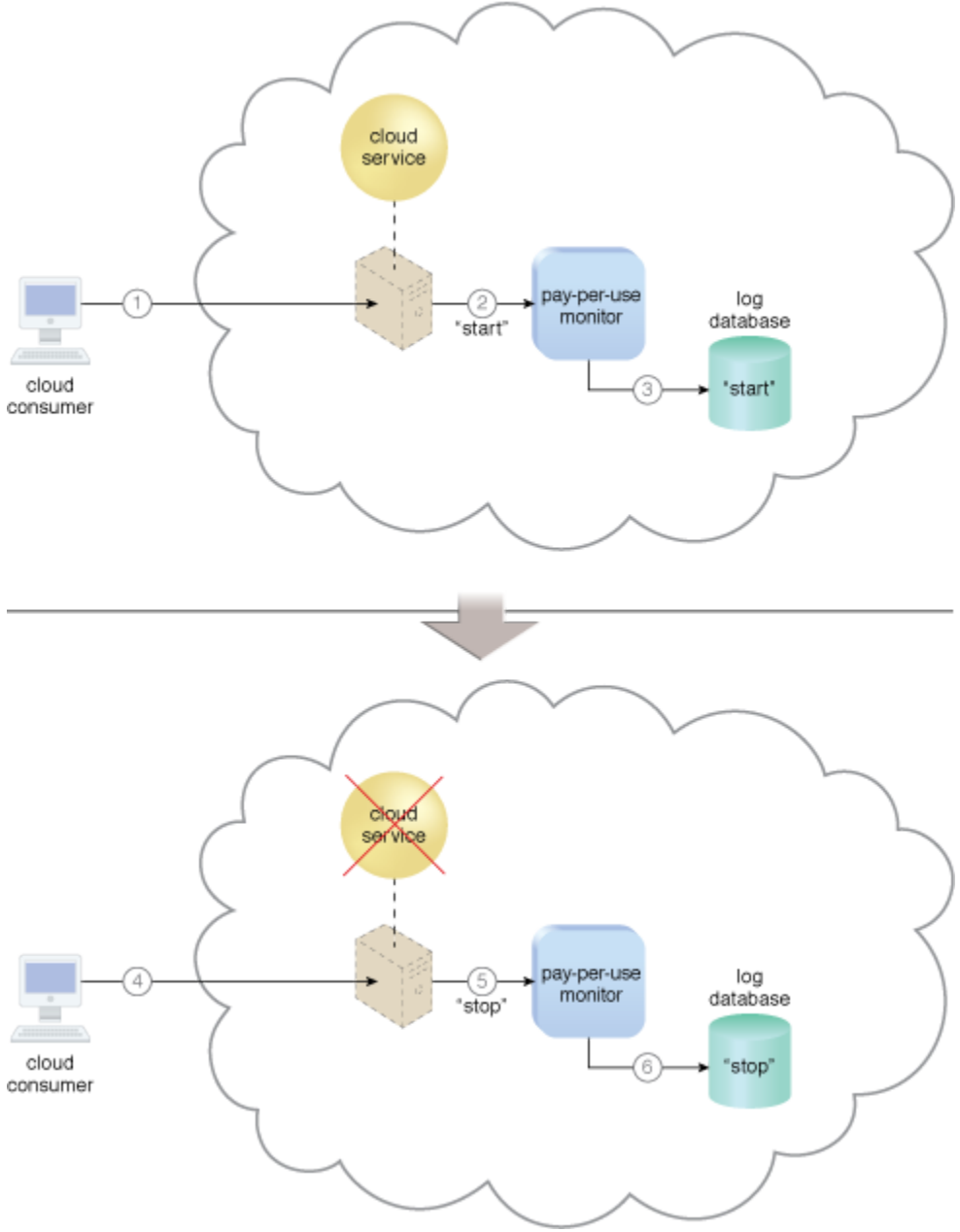


Figure shows pay-per-use monitor implemented as a resource-agent used to determine the usage period of a virtual server

- A cloud consumer requests the creation of a new instance of a cloud service (1)
- The IT resource is instantiated and the pay-per-use monitor receives a “start” event notification from the resource software (2)
- The pay-per-use monitor stores the value timestamp in the log-database (3)
- The cloud consumer later requests that the cloud service instance be stopped (4)
- The pay-per-use monitor receives a “stop” event notification from the resource software (5)
- And stores the value timestamp in the log database (6)

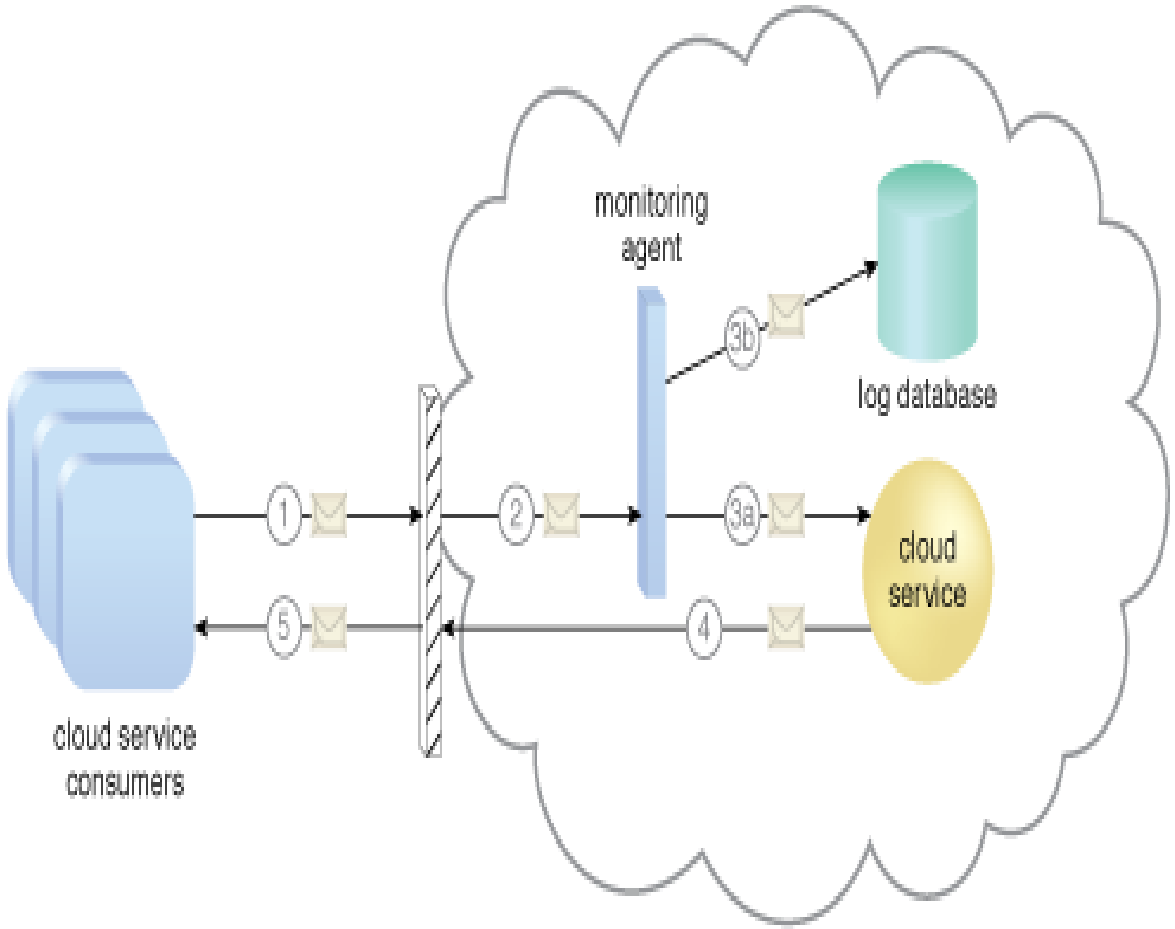
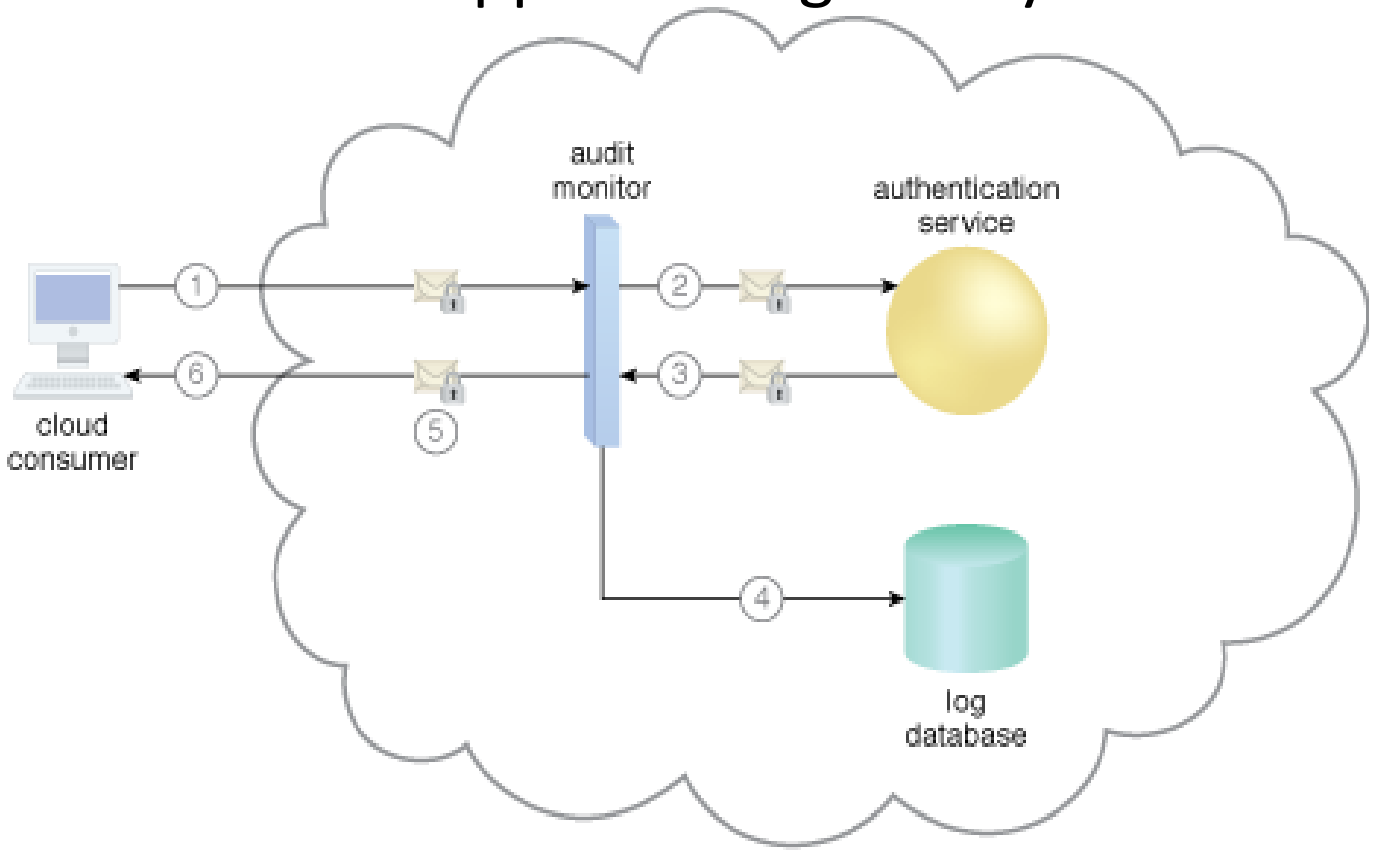


Figure illustrates a pay-per-use monitor designed as a monitoring agent that transparently intercepts and analyzes run time communication with a cloud service

- A cloud service consumer sends a request message to the cloud service (1)
- The pay-per-use monitor intercepts the message (2)
- Forwards it to the cloud service 3(a)
- And stores the usage information in accordance with its monitoring metrics 3(b)
- The cloud service forwards the response messages back to the cloud service consumer (4)

11. Audit Monitor

The audit monitor mechanism is used to collect audit tracking data for networks and IT resources in support of regulatory or contractual obligations



The figure depicts as an audit monitor implemented as a monitoring agent that 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

- A cloud service consumer requests access to a cloud service by sending a login request message with security credentials (1)
- The audit monitor intercepts the message (2)
- And forwards it to the authentication service (3)
- The authentication service processes the security credentials. A response message is generated for the cloud service consumer, in addition to the results in the login attempt (4)
- 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 (5)
- Access has been granted, and a response is sent back to the cloud service consumer (6)

12. Failover System

Example: we usually carry two pens for our written exams. Its because when the first pen fails for any reason, we can switch to redundant one. In this case, the person is the failover system who ensures that the pen is always available for the exam.

- The failover system mechanism is used to increase the reliability and availability of IT resources by using established clustering technology to provide redundant implementations.

- A failover system is configured to automatically switch over to a redundant or standby IT resource instance whenever the currently active IT resource becomes unavailable.

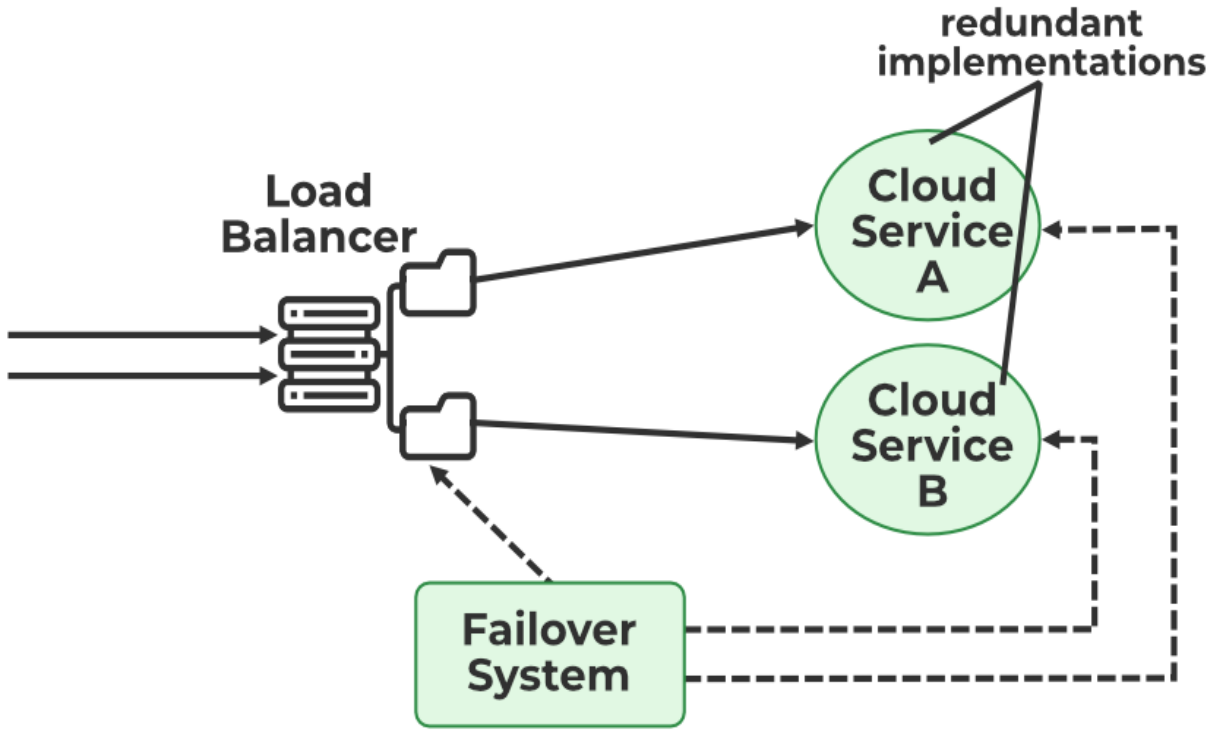
- Failover systems are commonly used for mission critical programs and reusable services that can introduce a single point of failure for multiple applications.

- A failover system can span more than one geographical region, so that each location hosts one or more redundant implementations of the same IT resource.

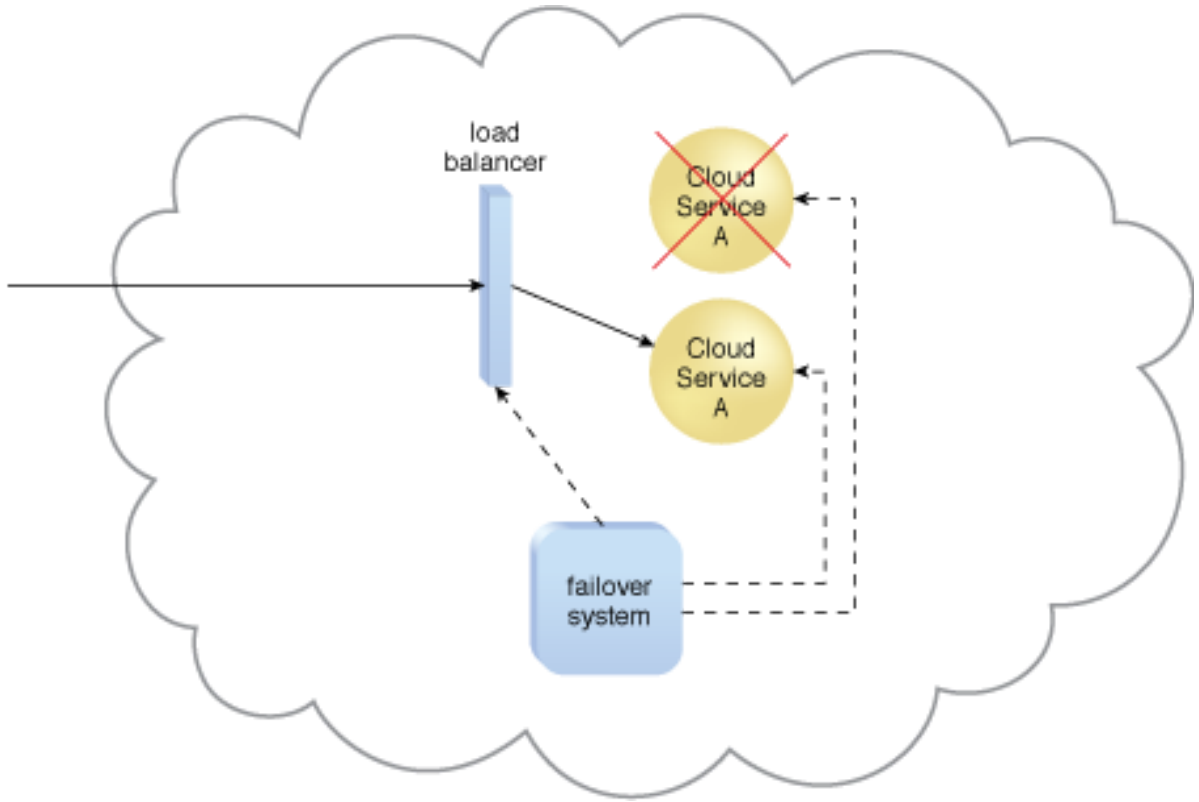
- The resource replication mechanism is sometimes utilized by the failover system to provide redundant IT resource instances, which are actively monitored for the detection of errors and unavailability conditions.

- Failover systems come in two basic configuration types
  - Active-Active
  - Active-Passive

- Active-Active Failover System
  - The redundant system is active
  - In active-active configuration, the redundant implementation of the IT resource actively serve the workload synchronously
  - Load balancing among active instances is required
  - When a failure is detected, the failed instance is removed from the load balancing scheduler
  - Which ever IT resource remains operational whenever a failure is detected takes over processing



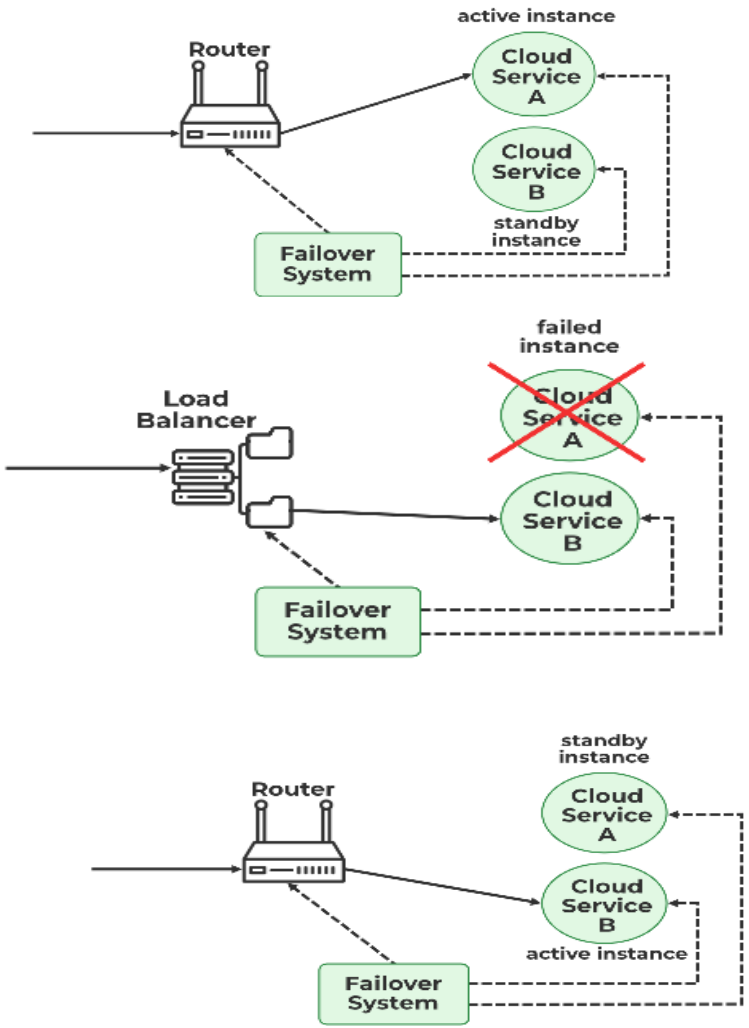
(Failover system monitors the operational status of cloud service A)



When a failure is detected in one cloud service A implementation, the failover system commands the load balancer to switch over the workload to the redundant cloud service A implementation

When a failed cloud service A is recovered or replicated in to an operational cloud service, the failover system commands the load balancer to distribute the workload again

- # Active-Passive Failover System
  - In an active-passive configuration, a standby or inactive implementation is activated to take over the processing from the IT resource that become unavailable, and the corresponding workload is redirected to the instance taking over the operation.
  - The example of carrying two pens to the exam hall is an example of active-passive failover system.



- The failover system monitors the operational status of cloud service A
- The cloud service A implementation acting as the active instance is receiving cloud service consumer request

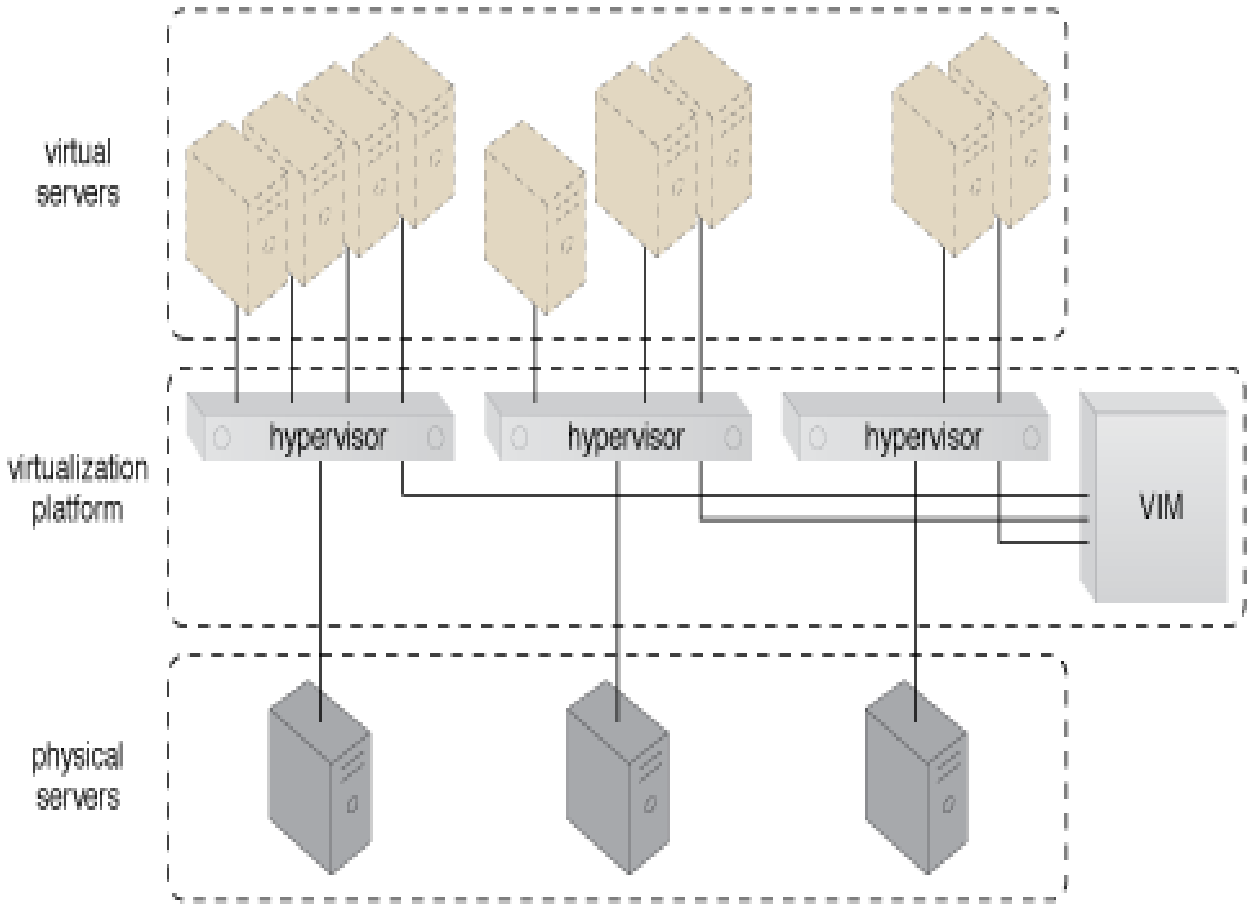
- The cloud service A implementation acting as the active instance encounters a failure that is detected by the failover system, which subsequently activates the inactive cloud service B implementation and redirects the workload towards it
- The newly invoked cloud service B implementation now assumes the role of active instance

- The failed cloud service A is recovered or replicated an operational cloud service, and is now positioned as standby instance, while the previously invoked cloud service A continues to serve as the active instance

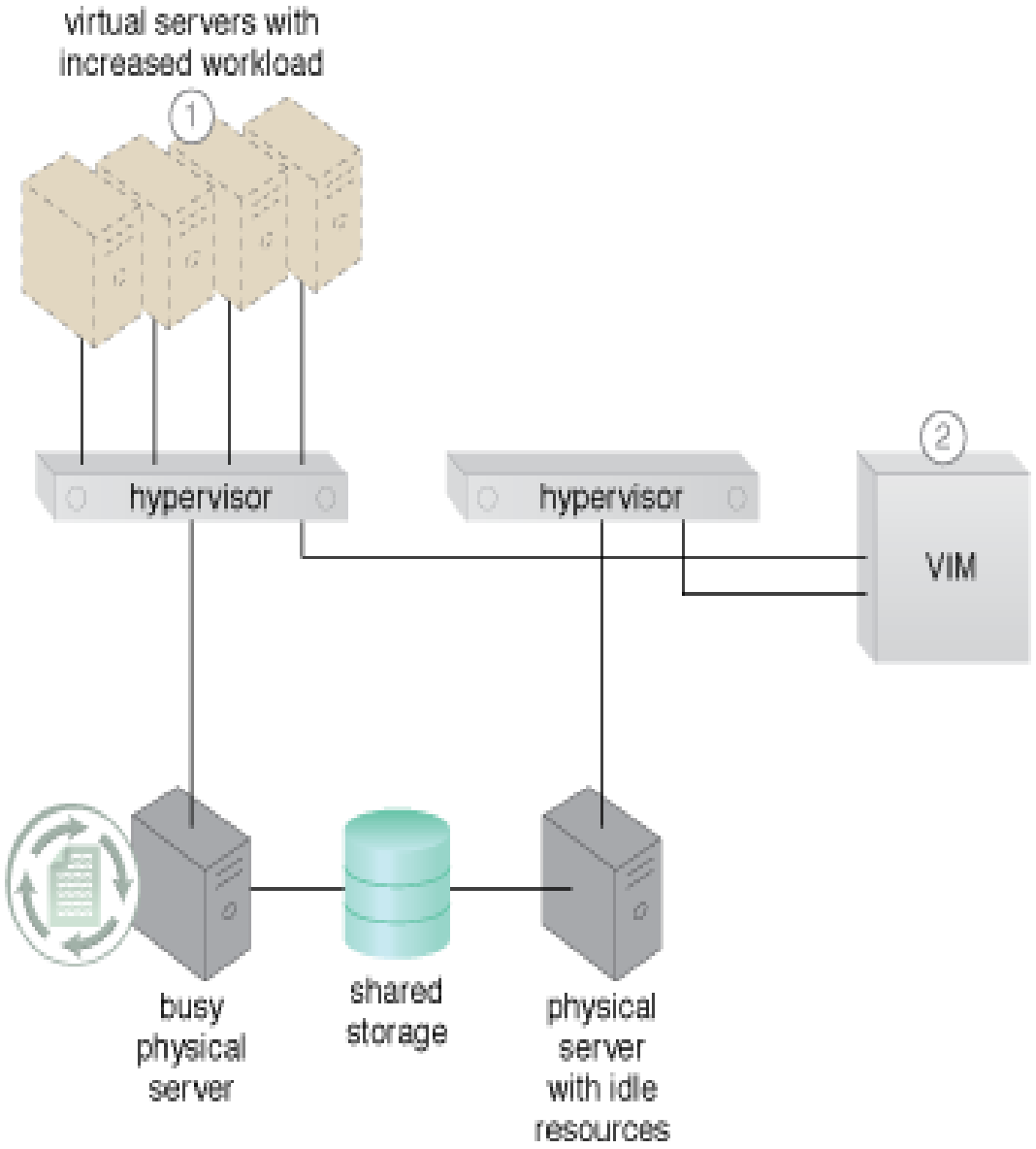
13. Hypervisor

- The hypervisor mechanism is a fundamental part of virtualization infrastructure that is primarily used to generated virtual server instances of a physical server.
- A hypervisor is generally limited to one physical server and therefore create only virtual images of that server.
- Similarly, a hypervisor can only assign virtual servers it generates to the resource pools that reside on the same underlying physical server.
- A hypervisor has limited virtual server management features, such as increasing the virtual server’s capacity or shutting it down.

- A VIM provides a range of features for administering multiple hypervisors across physical servers.

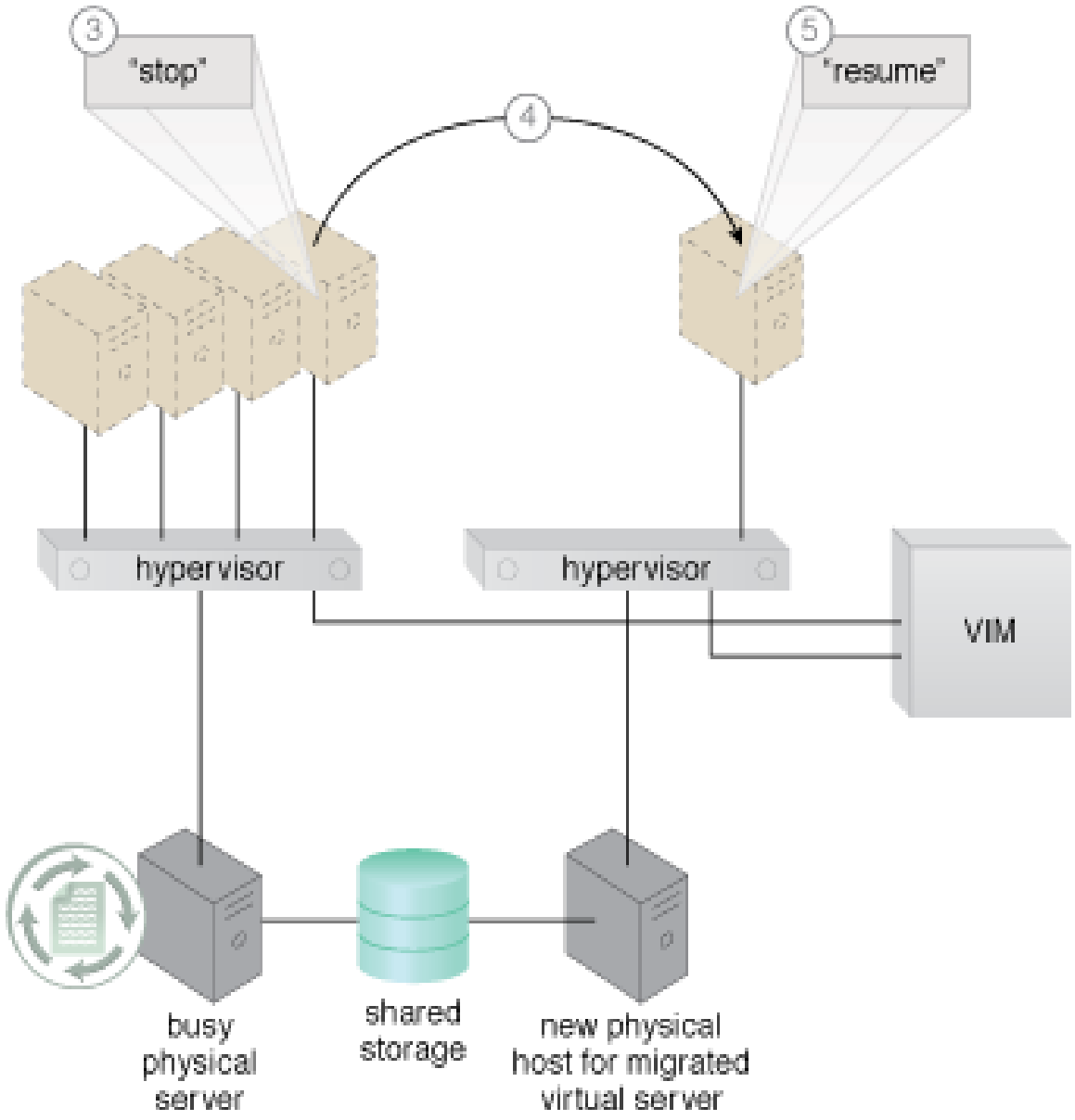


- Virtual servers are created by individual hypervisor on individual physical servers.
- All three hypervisors are jointly controlled by the same VIM.
- Hypervisor software can be installed directly in bare metal servers and provides features for controlling, sharing, and scheduling the usages of hardware resources such as process power, memory, and I/O. These can appear to each virtual server's OS as dedicated resources.



- A virtual server capable of auto scaling experiences an increase in its workload (1)

- The VIM decides that the virtual server can not scale up because its underlying physical server host is being used by other virtual servers (2)



- The VIM commands the hypervisor on the busy physical server to suspend execution of the virtual server (3)

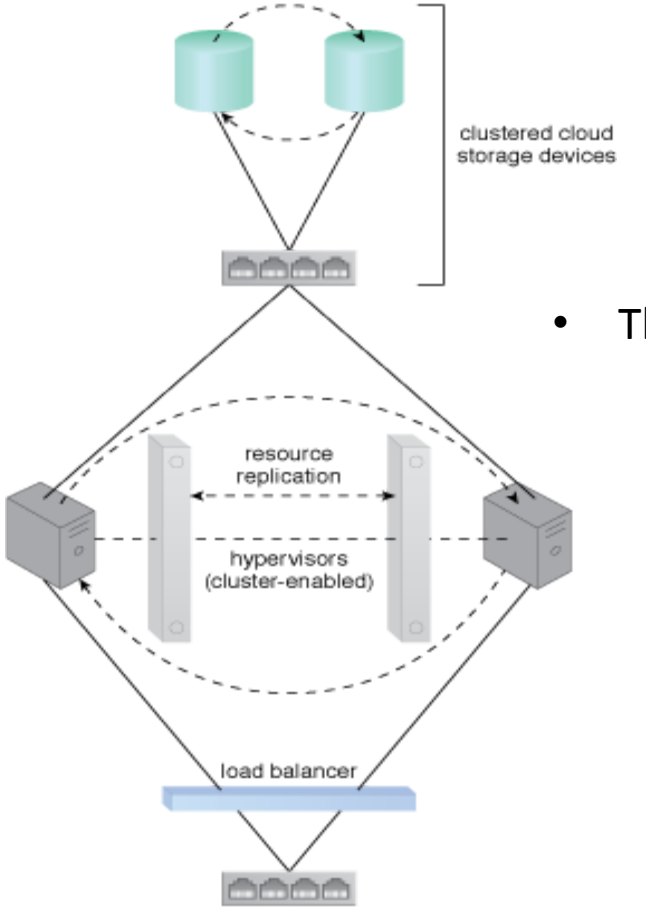
- The VIM then commands the instantiation of the virtual server on the idle physical server. State information (such as dirty memory pages and processor registers) is synchronized via a shared cloud storage device (4)

- The VIM commands the hypervisor at the new physical server to resume the virtual server processing (5)

14. Resource Cluster

- **What is cluster?**
- A cluster can be thought of as one unit consisting of similar components. For example; the Indian team is a cluster unit of top players from India.
- Clusters serve one functionality. For example; the Indian team cluster function is to win for India.
- Cluster helps to distribute the load between all units which serve the same goal, which is to win. Example; the players belonging to the clusters are the resources.
- In IT world or in Cloud, we make clusters of databases, clusters of virtual web machines, or clusters that serve parallel computation etc.

- The cloud based IT resources that are geographically diverse can be logically combined into groups to improve their allocation and use.
- The resource cluster mechanism is used to group multiple IT resource instances so that they can be operated as a single IT resource. This increases the combined computing capacity, load balancing, and availability of the clustered IT resources.



- The curved dashed lines are used to indicate that IT resources are clustered.

- Resource cluster architectures rely on high speed dedicated network connections, or cluster nodes between IT resource instances to communicate about workload distribution, task scheduling, data sharing, and system synchronization.

- A **cluster management platform** that is running as distributed middleware in all of the cluster nodes is usually responsible for these activities.

- This platform implements a coordination function that allows distributed IT resource to appear as one IT resource, and also executes IT resources inside the cluster.

- Common resource cluster types include:

1. Server Cluster → Cluster of server machines
2. Database Cluster → Cluster of databases
3. Large dataset Cluster → Cluster of datasets

Server Cluster:

- Physical or virtual servers are clustered to increase performance and availability.
- Hypervisors running on different physical servers can be configured to share virtual server execution state (such as memory pages and processor register state) in order to establish clustered virtual servers.
- In such configurations which usually requires physical servers to have access to shared storage, virtual servers are able to live migrate from one to another.
- In this process, the virtualization platform suspends the execution of a given virtual server at one physical server and resumes it on another physical server. The process is transparent to the virtual server OS and can be used to increase scalability by live migrating a virtual server that is running on an overloaded physical server to another physical server that has suitable capacity.

Database Cluster:

- Database clusters are designed to improve data availability, this high-availability resource cluster has a synchronization feature that maintains the consistency of data being stored at different storage devices used in the cluster.
- The redundant capacity is usually based on an active-active or active-passive fail over system committed to maintaining the synchronization conditions.

Large Dataset Cluster:

- Data partitioning and distribution is implemented so that the target datasets can be efficiently partitioned without compromising data integrity or computing accuracy.
- Each cluster node processes workloads without communicating with other nodes as much as in other cluster types.

- Many resource clusters require cluster nodes to have almost identical computing capacity and characteristics in order to simplify the design of and maintain consistency within the resource cluster architecture.
- The cluster nodes in high availability cluster architectures need to access and share common storage IT resources.
- This can require two layers of communication between the nodes- one for accessing the storage device and another to execute IT resource orchestration.

Two types of resource clustering

a. Load balanced cluster → Cluster of servers working together under a load balancer to balance load

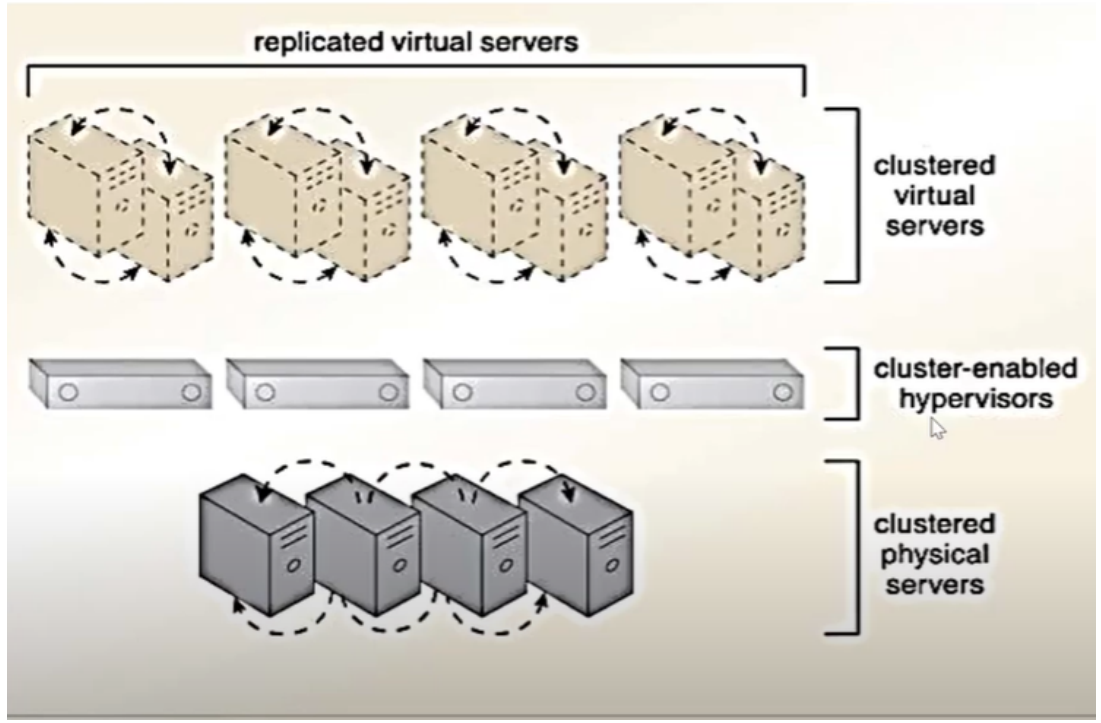
b. High availability cluster → Cluster of servers working together under a failover system to ensure, if one fails then the other one is available

Load balanced cluster:

- This resource cluster specializes in distributing workloads among cluster nodes to increase IT resource capacity while preserving the centralization of IT resource management.
- It usually implements a load balancer mechanism that is either embedded within the cluster management platform or setup as a separate IT resource.

High availability cluster:

- It maintains system availability in the event of multiple node failures, and has redundant implementations of most or all of the IT resources.
- It implements a failover system mechanism that monitors failure conditions and automatically redirects the workload away from any failed nodes.
- The provisioning of clustered IT resources can be considerably more expensive than the provisioning of individual IT resources that have an equivalent computing capacity.



HA virtualization cluster

- An HA virtualization cluster of physical servers is deployed using a cluster-enabled hypervisor, which guarantees that the physical servers are constantly in sync.
- Every virtual server that is instantiated in the cluster is automatically replicated in at least two physical servers.

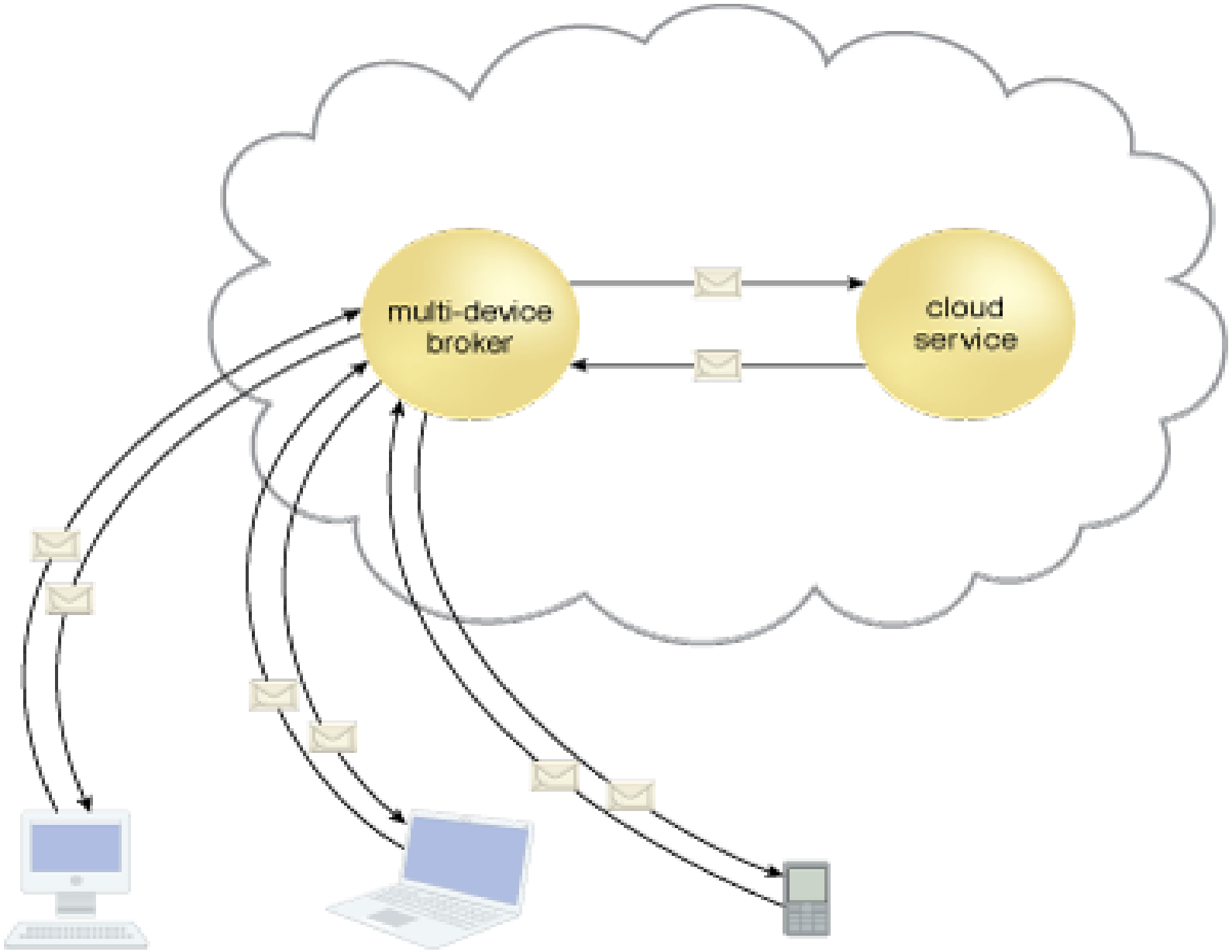
15. Multi-device Broker

- There are many types of devices the cloud consumer works with on Internet. The cloud consumer may be using laptop to access the cloud provider, a Mozilla Firefox browser, a smart phone, a tablet and even there are different variations in every type of device. When the cloud provider web portal interact with the cloud consumers, the multi-device broker ensures that it can display the data as per the device of the cloud consumer.

- An individual cloud service may need to be accessed by a range of cloud service consumers differentiated by their hosting hardware devices and/or communication requirements.

- To overcome incompatibilities between a cloud service and cloud service consumer, mapping logic needs to be created to transform or convert information that is exchanged at run time.

- The multi-device broker mechanism is used to facilitate run time data transformation so as to make a cloud service accessible to a wider range of cloud service consumer programs and devices.


multi-device
broker
cloud
service

- Multi-device brokers commonly exist as gateways or incorporate gateway components to enable the broker server such as;

- XML gateway-> transmits and validates XML data

- Cloud storage gateway-> transforms cloud storage protocols and encodes storage devices to facilitate data transfer and storage

- Mobile device gateway-> transforms the communication protocols used by mobile devices in to protocols that are compatible with a cloud service

The levels at which the transformation logic can be created include:

- Transport protocols
- Messaging protocols
- Storage device protocols
- Data schemas/data models

16. State Management Database

- A state can be thought of as “what is the status of something at a given time?”

- Example: A customer uses a shopping cart program on a browser to select items to purchase. The state of the shopping cart may be stored in browser memory or server memory. But, since the memory is temporary the shopping cart information may be lost when browser gets closed or server resets for some reason.

- A state management database which temporarily stores data of a particular process until done, helps it to recover back to the original state.

- A website may be stored on a cluster of load balancing web servers. If a user logged on to one web server then the other web servers should know the state too. A common state management database is used for all the load balancing web servers serves such a purpose.

- A state management database also helps to keep the state of virtual machine.

- Example; several virtual machines may go through various states of initiating, starting, started, stopping, stopped, terminated. If we store current status of all virtual machines in database, one can re-initialize the virtual machine to the last known state.

- State management databases are mainly used by cloud services, especially those are involved in long-running run time activities.

Cloud Management Mechanisms

- Cloud management mechanisms are techniques that let you do management tasks such as set up, configure, maintain, monitor any cloud IT resources such as virtual machine, storage etc.

- The tasks are also exposed as API which can be called from any program to programmatically manage the task.

There are four cloud management mechanisms

17. Remote administration system → How to administer the cloud remotely?

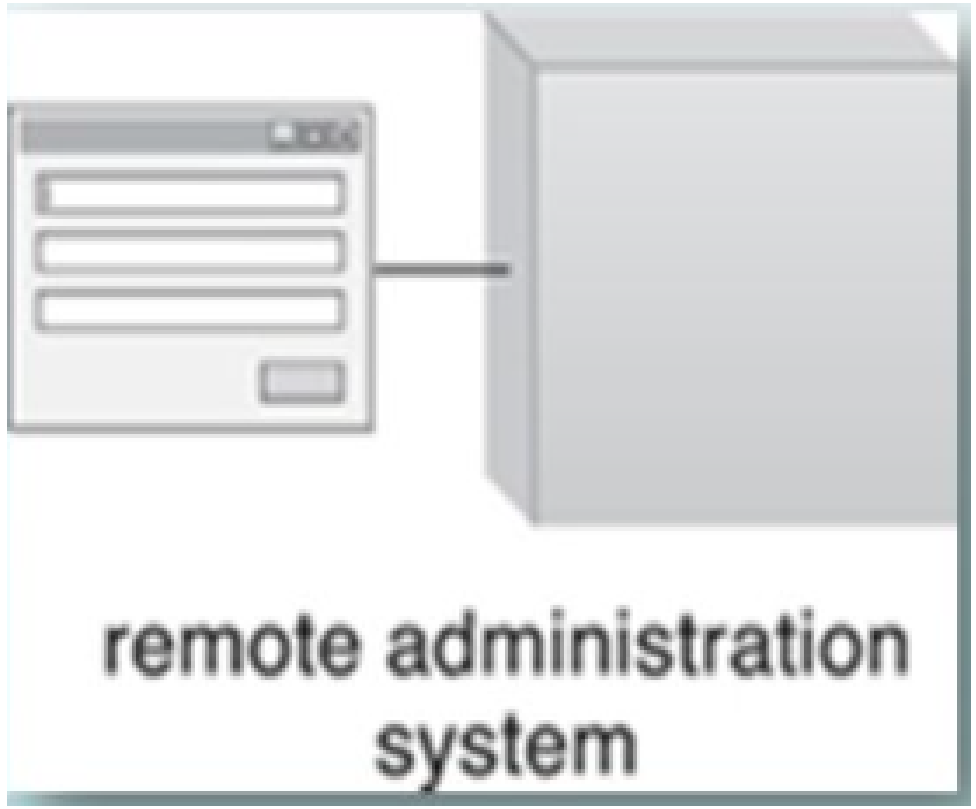
18. Resource management system → How to manage the resources such as physical servers, hypervisors, virtual images?

19. SLA management system → How to manage the SLAs?

20. Billing management system → How to do the billing?

17. Remote Administration System

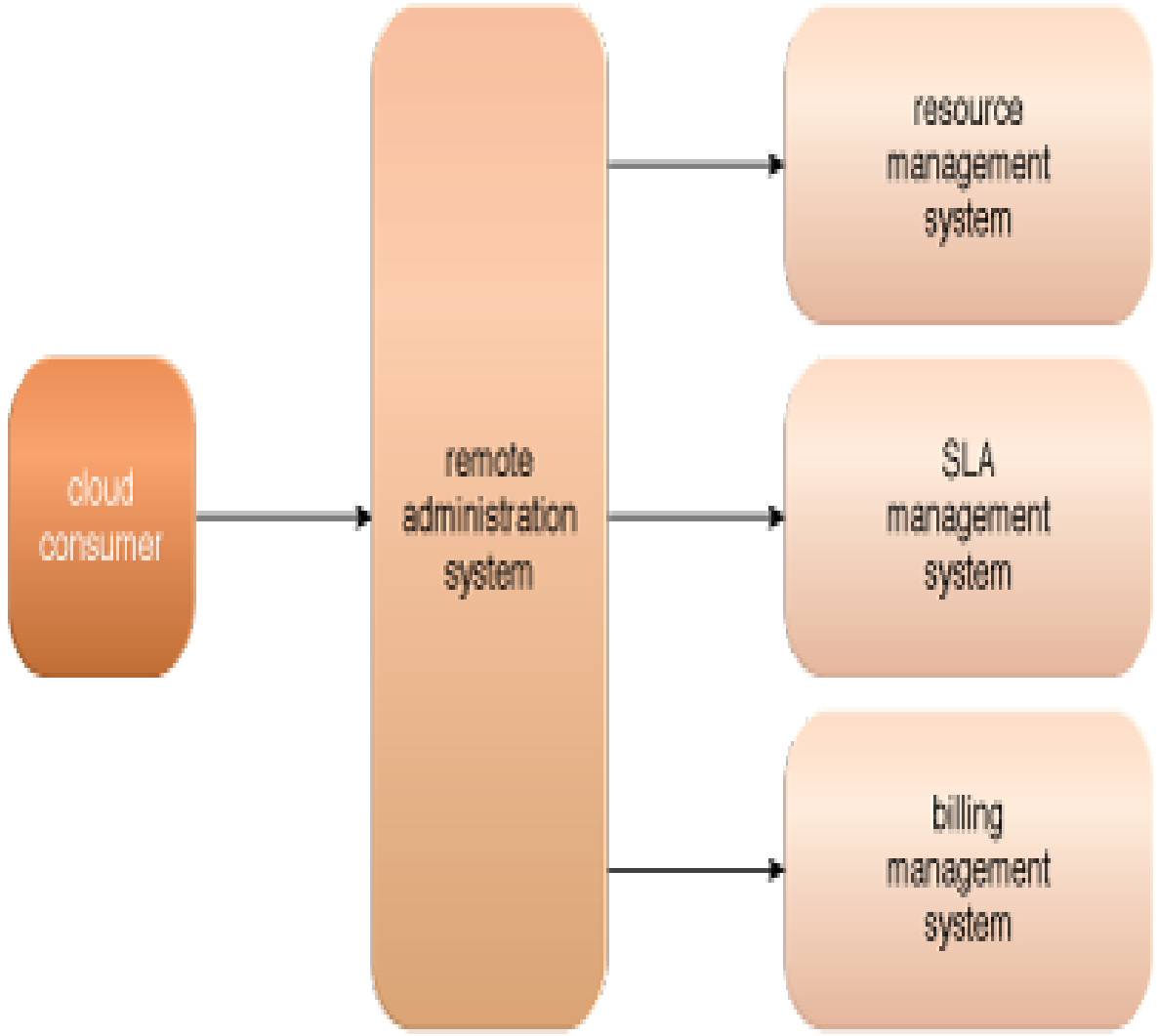
The remote administration system mechanism provides tools and user-interfaces for external cloud resource administrators to configure and administer cloud based IT resources.



- The symbol for the remote administration system.
- The displayed user-interface will typically be labelled to indicate a specific type of portal (web).
- AWS management console is an example of such a portal website to set up and configure cloud IT resources.

The portal can be used to implement other cloud management mechanisms namely

- Resource management
- SLA management
- Billing management



- Remote administration is of two types

Self-service portal

- It is to help shop for cloud services
- This is 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).

Usages and administration portal

- Portal for the portal to use and administer the purchased resources.
- A general purpose portal that centralizes management controls to different cloud based IT resources and can further provide IT resources usage reports.

Tasks that can commonly be performed by cloud consumers via a remote administration console includes:

1. Configuring and setting up cloud services
2. Provisioning and releasing IT resource for an on-demand cloud services

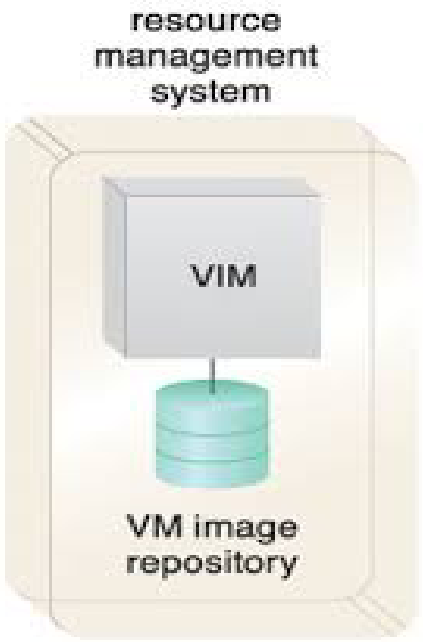
3. Monitoring cloud service status, usages, and performance
4. Monitoring QoS and SLA fulfillment

5. Managing leasing costs and usages fees
6. Managing user accounts, security credentials, authorization, and access control

7. Tracking internal and external access to leased services
8. Planning and assessing IT resource provisioning
9. Capacity planning

18. Resource Management

- The resource management system mechanism helps coordinate IT resource in response to management action performed by both cloud consumers and cloud providers
- Virtual infrastructure management (VIM) is an example of resource management system that is responsible to manage resources such as hypervisors, virtual machine images etc.



- A resource management system encompassing a VIM platform and a virtual machine image repository
- The VIM may have additional repositories, including one dedicated to storing operational data

- Core to resource management system is a virtual infrastructure manager (VIM) that coordinates the server hardware so that 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.

- For example; a VIM can create and manage multiple instances of hypervisor across different physical servers or allocate a virtual server on one physical server to another (or to a resource pool).

Tasks that are typically automated and implemented through the resource management system include:

1. Managing virtual IT resource template that are used to create pre-built instances, such as virtual server images.

2. Allocating and releasing virtual IT resources in to available physical infrastructure in response to the starting, pausing, resuming, and termination of virtual IT resource instances.

3. Coordinating IT resources in relation to the involvement of other mechanisms such as resource replication, load balancer, and failover system.

4. Enforcing usages and security policies through out the lifecycle of cloud service instances.

5. Monitoring operational conditions of IT resources.

- Since VIM is a cloud provider task rather than cloud consumer, the cloud consumer special native console to do the management.

- Just like cloud consumers can remotely access APIs to administer, cloud providers can also remotely access APIs to manage the cloud resources for its various cloud consumers.

19. SLA Management

- The SLA management system mechanism represents a range of commercially available cloud management products that provides features pertaining to the administration, collection, storage, reporting, and runtime notification of SLA data.

SLA management system

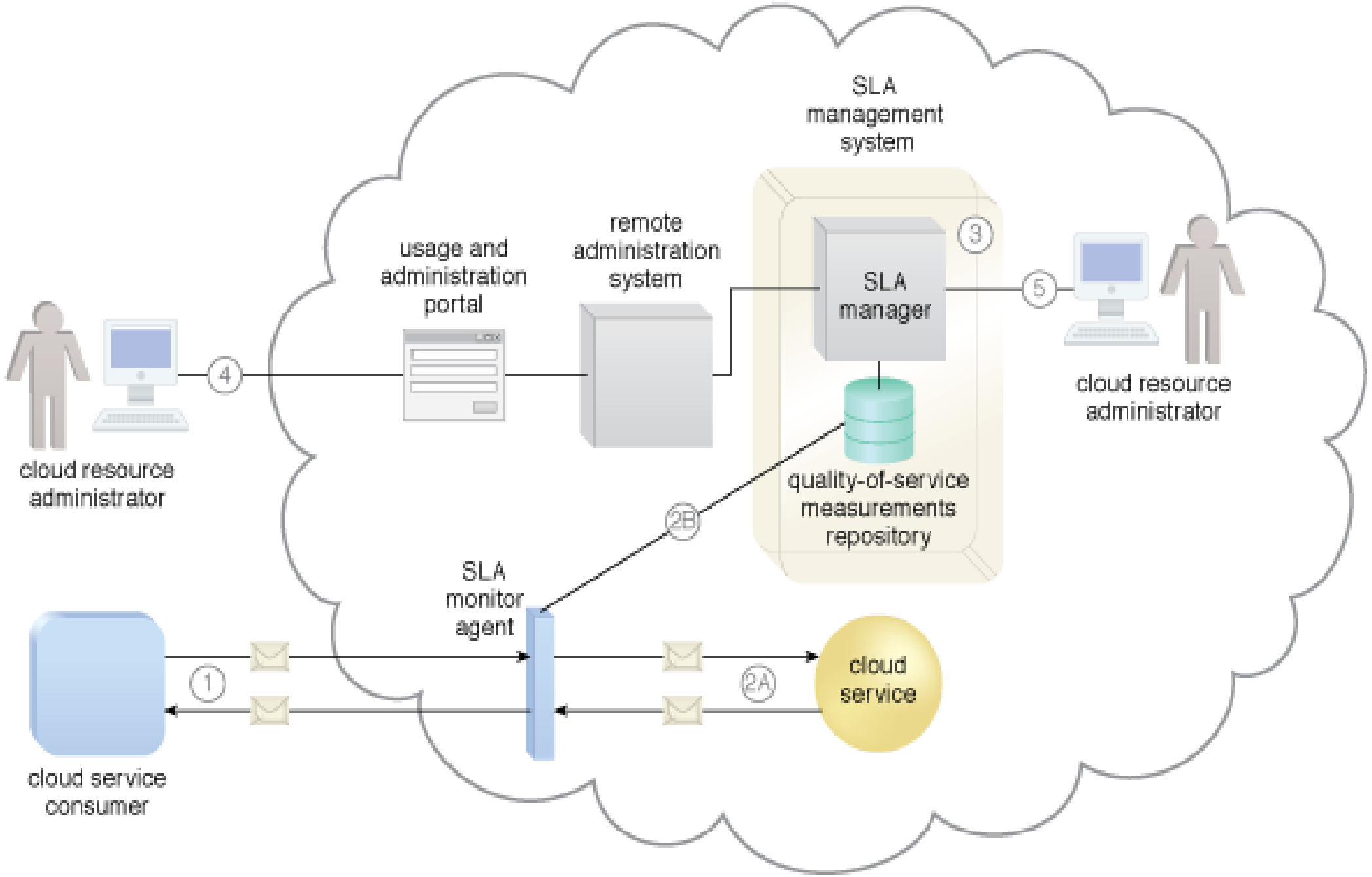
An SLA management system consists of

- SLA Manager-> Manages SLA
- QoS measurements repository-> Defines SLA metrics to be met

- In SLA management system deployment will generally include a repository used to store and retrieve collected SLA data based on predefined metrics and reporting parameters.

- It will further rely on one or more SLA monitor mechanisms to collect the SLA data that can 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.

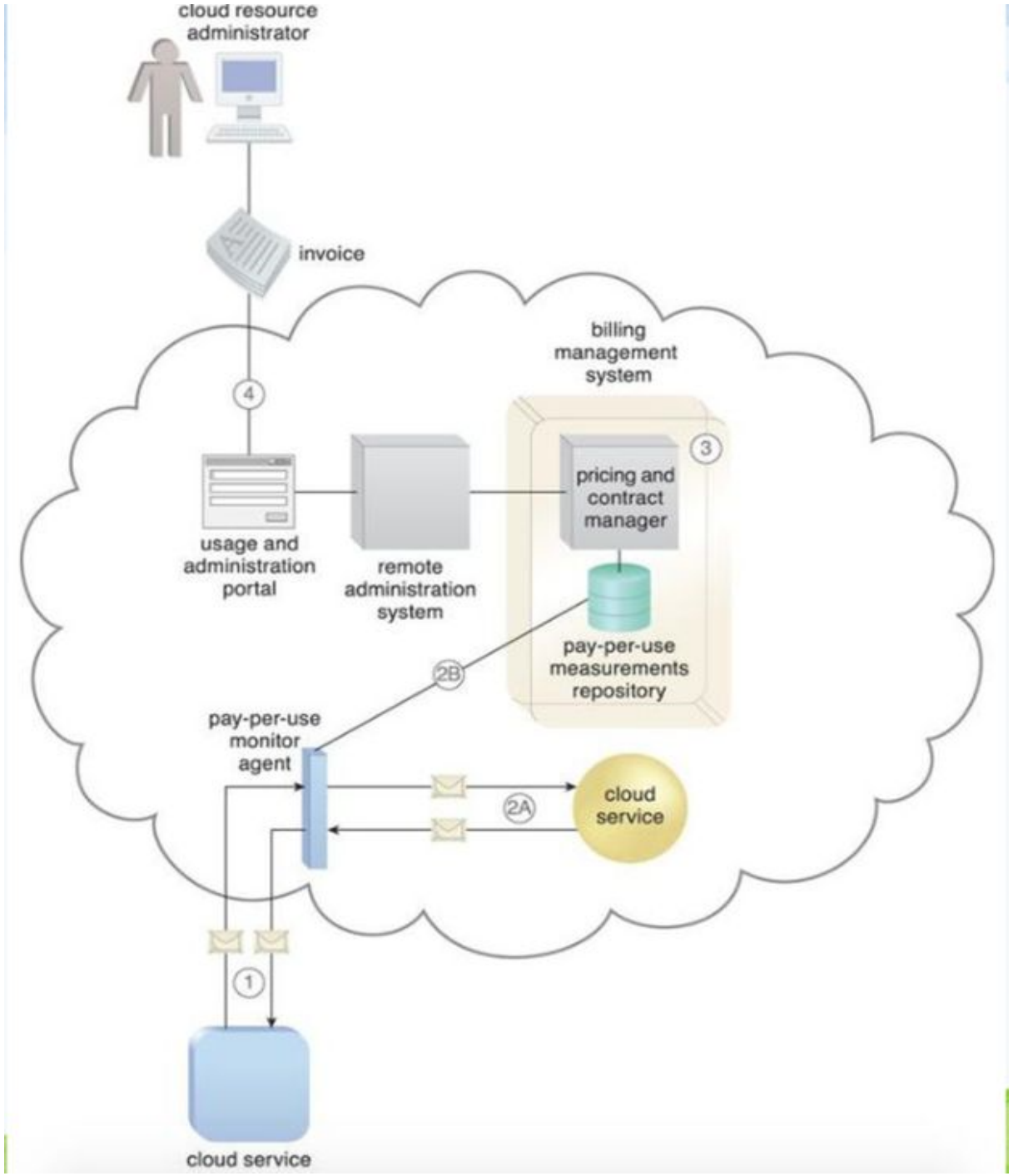

SLA
management
system
usage and
administration
portal
remote
administration
system
SLA
manager
3
5
cloud resource
administrator
cloud resource
administrator
4
quality-of-service
measurements
repository
2B
SLA
monitor
agent
1
2A
cloud
service
cloud service
consumer

- A cloud service consumer interacts with a cloud service (1)
- A SLA monitor intercepts the exchanged messages, evaluates the interaction, and collects relevant run time data in relation to QoS guarantees defined in the cloud service's SLA (2A)
- The data collected is stored in a repository (2B)
- That is part of the SLA management system (3)
- Queries can be issued and reports can be generated for an external cloud resource administrator via a usage and administration portal (4)
- Or for an internal cloud resource administration via the SLA management system's native user-interface (5)

20. Billing Management System

- The 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.
- Specifically, the bill management system relies on pay-per-use monitors to gather run time usage data that is stored in a repository that the system components then draw from for billing, reporting , and invoicing process.
- A billing management system comprised of a

1. Pricing and contract manager-> software to manage the pricing
2. Pay-per-use measurements repository-> pricing agreements with cloud consumer



- A cloud service consumer exchanges messages with a cloud service (1)
- A pay-per-use monitor keeps track of the usages and collects data relevant to billing (2A)
- Which is forwarded to the repository that is part of the billing management system (2B)
- The system periodically calculates the consolidated cloud service usage fees and generated and invoice for the cloud consumer (3)
- The invoice may be provided to the cloud consumer through the usage and administration portal (4)

- The billing management system 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 model
  - Flat-rate
  - Pay-per-allocation modes
  - Or combinations thereof

- Billing arrangements be based on

- Pre-usage payments-> ×
- **Post-usage payments-**> The post-usage payment type can include pre-defined limits or it can be set up (with the mutual agreement of the cloud consumer) to allow for unlimited usage (and consequently no limit on subsequent billing).

- When the limits are established they are usually in the form of usage quotas
- When quotas are exceeded the billing management system can block further usage request by the cloud consumer.

Overview of Cloud Security Mechanisms

Types of security in cloud:

1. Network security
2. Application security
3. Data security

Cloud computing security basic terms and concepts:

1. Threat agents (anonymous attacker, malicious service agent, trusted attacker, malicious insider)
2. Cloud security threats (traffic eavesdropping, malicious intermediary, DoS, insufficient authorization, virtualization attack)
3. Cloud security mechanisms
4. Encryption
5. Hashing
6. Digital signature
7. Public-key Infrastructure (PKI)
8. Identity and Access Management (IAM) (authentication, authorization, user management, central user repository)
9. Single Sign-on (SSO)
10. Cloud based security groups
11. Hardened virtual server images (remove redundant programs, closing unnecessary ports, disabling unused services, internal root accounts, and guest access)