Chapter 1

Cloud computing is a specialized form of distributed computing that introduces utilization models for remotely provisioning scalable and measured resources.

Cost Reduction (cloud computing doesn't require those)

- Technical personnel
- Upgrades and patches that introduce additional testing and deployment cycles
- Utility bills and capital expense investments for power and cooling
- Security and access control measures that need to be maintained and enforced to protect infrastructure resources
- Administrative and accounts staff that may be required to keep track of licenses and support arrangements.

Clustering

A cluster is a group of independent IT resources that are interconnected and work as a single system.

Grid Computing (or “computational grid”)

provides a platform in which computing resources are organized into one or more logical pools.

Note// sometimes referred to as a “super virtual computer”.

Virtualization

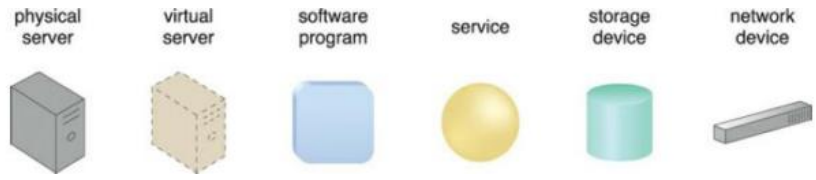
represents a technology platform used for the creation of virtual instances of IT resources.

Cloud

refers to a distinct IT environment that is designed for the purpose of remotely provisioning **scalable** and **measured** IT resources.

IT Resource

is a physical or virtual IT-related artifact that can be either software-based, such as a virtual server or a custom software program, or hardware-based, such as a physical server or a network device.



The party that provides cloud-based IT resources is the **cloud provider**.

The party that uses cloud-based IT resources is the **cloud consumer**.

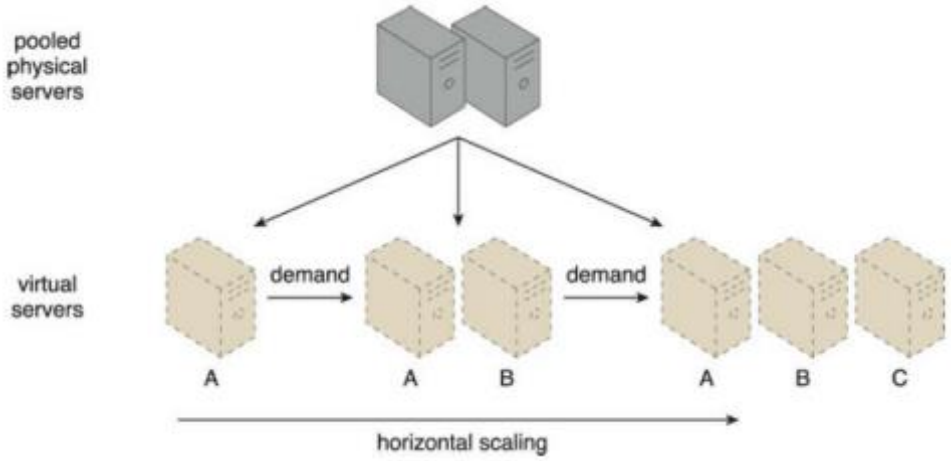
Scaling

represents the ability of the IT resource to handle increased or decreased usage demands.

1. Horizontal Scaling

The allocating or releasing of IT resources that are of the same type is referred to as horizontal scaling.

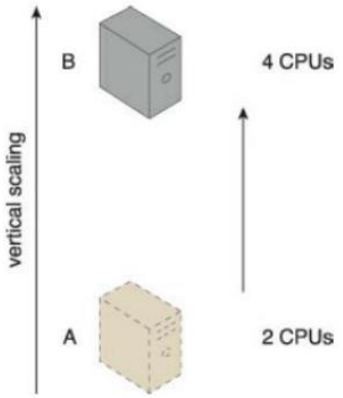
- The **horizontal allocation** of resources is referred to as **scaling out.**
- And the **horizontal releasing** of resources is referred to as **scaling in.**



2. Vertical Scaling

When an existing IT resource is replaced by another with higher or lower capacity

- **Replacing** an IT resource **with another that has a higher capacity** is referred to as **scaling up.**
- And **replacing** an IT resource **with another that has a lower capacity** is considered **scaling down.**



Horizontal Scaling VS Vertical Scaling

Horizontal Scaling	Vertical Scaling
less expensive (through commodity hardware components)	more expensive (specialized servers)
IT resources instantly available	IT resources normally instantly available
resource replication and automated scaling	additional setup is normally needed
additional IT resources needed	no additional IT resources needed
not limited by hardware capacity	limited by maximum hardware capacity

Goals and benefits of cloud computing (Advantages)

1. On-demand access to PAY-AS-YOU-GO computing resources.
2. The ability to add or remove IT resources.
3. Applications are not locked into devices or locations.
4. Cost Reduction.
5. Scalability.
6. Increased availability and reliability

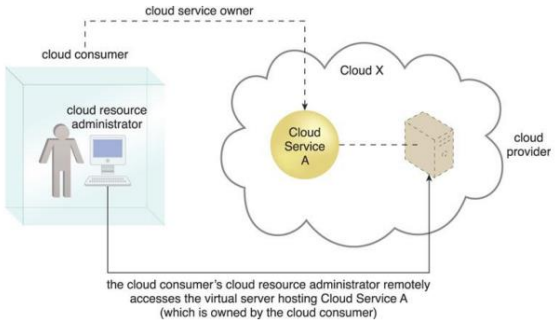
Risks and challenges (Disadvantages)

1. Responsibility over data security becomes shared with the cloud provider.
2. Unreliable cloud provider may not maintain the guarantees of SLA's.
3. longer geographic distance between the cloud consumer and provider.
4. limited portability between cloud providers.

Chapter 2

ROLES in cloud

- **Cloud Service Owner** The person or organization that legally owns a cloud service.
- **Cloud Resource Administrator** is the person or organization responsible for administering a cloud-based IT resource (including cloud services)



Note//A cloud resource administrator can be with a cloud consumer organization and administer remotely accessible IT resources that belong to the cloud consumer.

- **Cloud Auditor** A third-party (often accredited) that conducts independent assessments of cloud environments assumes the role of the cloud auditor.
- **Cloud Broker** A third-party that acts as an intermediary between the cloud consumer and the cloud provider.
  This role is assumed by a party that assumes the responsibility of managing and negotiating the usage of cloud services between cloud consumers and cloud providers.
- **Cloud Carrier** The party responsible for providing the wire-level connectivity between cloud consumers and cloud providers

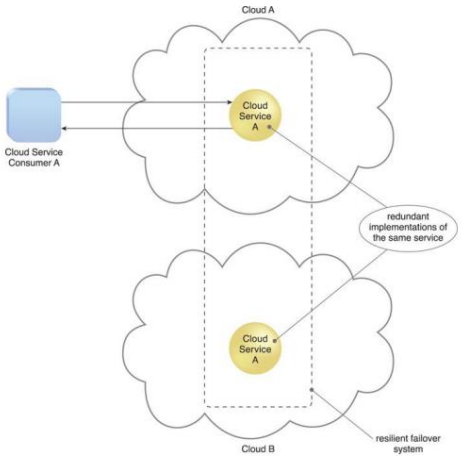
BOUNDARIES in cloud

- **Organizational Boundary**
  represent the boundary of an organizational set of IT assets and IT resources.
- **Trust Boundary**
  represent the extent to which IT resources are trusted.

Cloud characteristics

1. **On-demand usage**
   a delivery model in which computing resources are made available to the user as needed.
   A cloud consumer can unilaterally access cloud-based IT resources giving the cloud consumer the freedom to self-provision these IT resources.
2. **Ubiquitous access**
   represents the ability for a cloud service to be widely accessible.
3. **Multitenancy (and resource pooling)**
   software program that enables an instance of the program to serve different consumers each is isolated from the other.
   multiple customers of a cloud vendor are using the same computing resources.
4. **Elasticity**
   is the automated ability of a cloud to transparently scale IT resources.
5. **Measured usage**
   represents the ability of a cloud platform to keep track of the usage of its IT resources
6. **Resiliency**
   is a form of failover that distributes redundant implementations of IT resources across physical locations.

Resilient system



A resilient system in which Cloud B hosts a redundant implementation of Cloud Service A to provide failover in case Cloud Service A on Cloud A becomes unavailable.

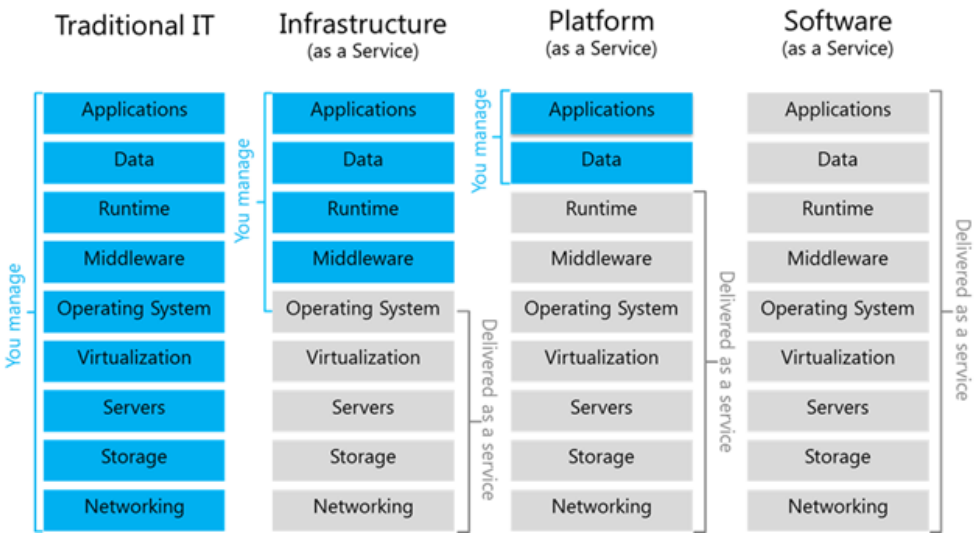
Cloud Delivery Model

A cloud delivery model represents a specific, pre-packaged combination of IT resources offered by a cloud provider.

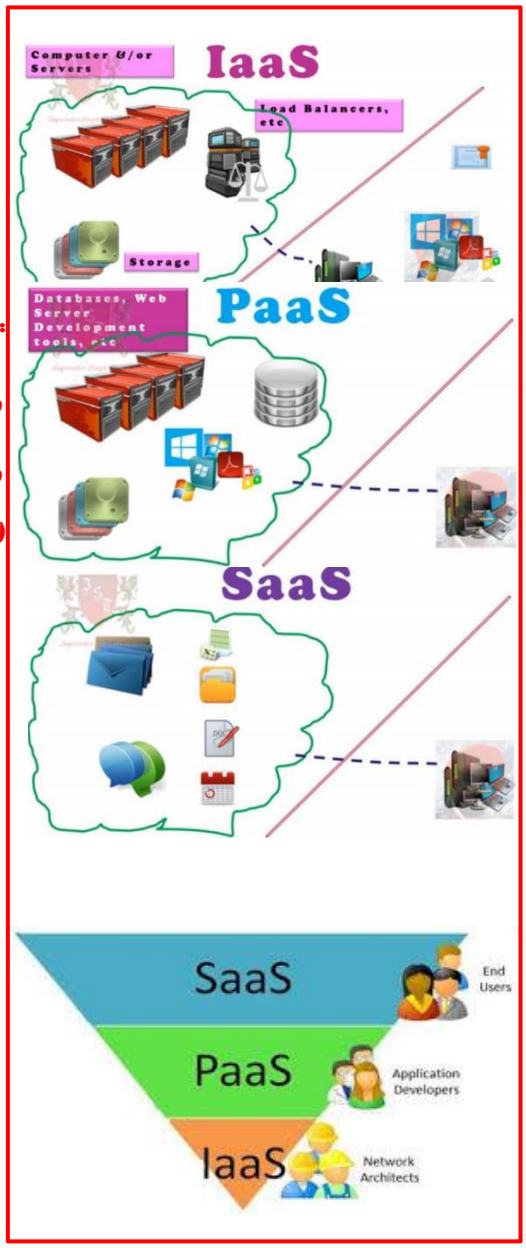
1. **Infrastructure-as-a-Service (IaaS)**
   represents a self-contained IT environment comprised of infrastructure-centric IT resources.
   This environment can include hardware, network, connectivity, operating systems, and other “raw” IT resources.

2. **Platform-as-a-Service (PaaS)**
   represents a pre-defined “ready- to-use” environment typically comprised of already deployed and configured IT resources.

3. **Software-as-a-Service (SaaS)**
   A software program positioned as a shared cloud service and made available as a “product” or generic utility.
   used to make a reusable cloud service widely available (often commercially) to a range of cloud consumers.



هذي الصور للتوضيح فقط

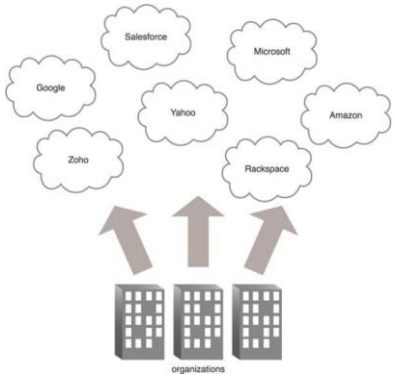


Cloud Deployment Models

represents a specific type of cloud environment, primarily distinguished by ownership, size, and access.

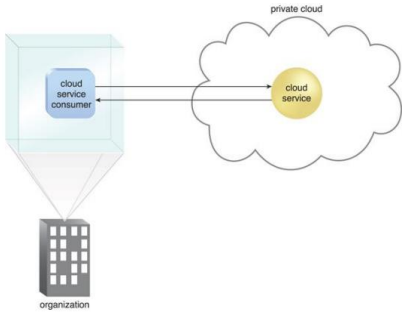
1. Public Clouds

is a publicly accessible cloud environment owned by a third-party cloud provider.



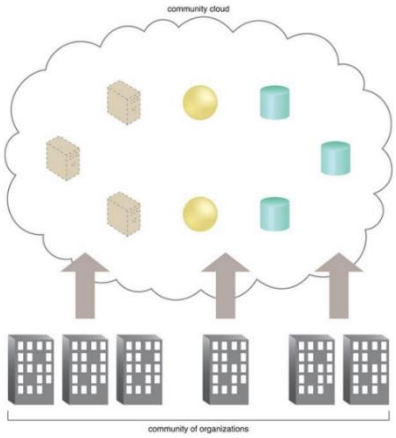
2. Private Clouds

A private cloud is owned by a single organization.



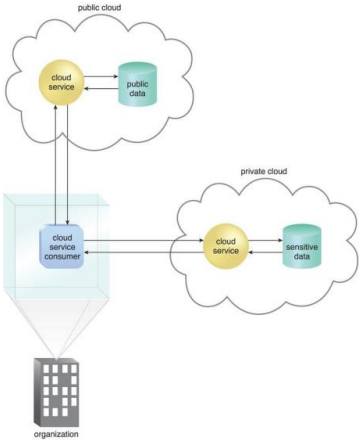
3. Community Clouds

is similar to a public cloud except that its access is limited to a specific community of cloud consumers.



4. Hybrid Clouds

is a cloud environment comprised of two or more different cloud deployment models.
Such as public and private cloud models



Chapter 3

Note// The **physical** IT resource layer refers to the **facility infrastructure** that houses computing/networking systems

Virtualization

is the process of converting a physical IT resource into a virtual IT resource.

Note// Data centers consist of both physical and virtualized IT resources.

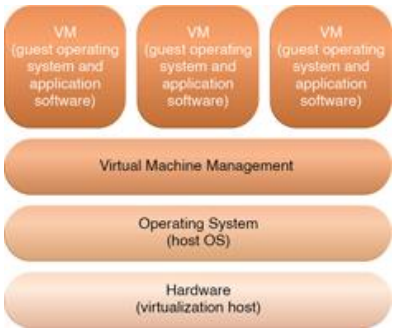
types of IT resources can be virtualized.

- **Servers** – A physical server can be abstracted into a virtual server.
- **Storage** – A physical storage device can be abstracted into a virtual storage device or a virtual disk.
- **Network** – Physical routers and switches can be abstracted into logical network fabrics, such as VLANs.
- **Power** – A physical UPS and power distribution units can be abstracted into what are commonly referred to as virtual UPSs.

Operating system-based virtualization

is the installation of virtualization software in a pre-existing operating system.

Note// VM is first installed into a full **host operating system**.

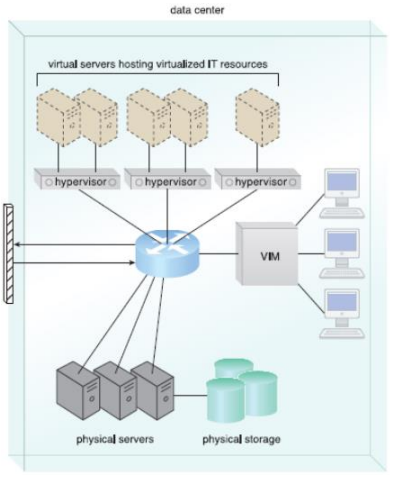


Data center technology

1. **Virtualization**
   Data centers consist of both physical and virtualized IT resources.
2. **Standardization and Modularity**
   Data centers are built upon standardized commodity hardware and designed with modular architectures.
3. **Automation**
   Data centers have specialized platforms that automate tasks like provisioning, configuration, patching, and monitoring without supervision.
4. **Remote Operation and Management**
   Most of the operational and administrative tasks of IT resources in data centers are commanded through the network's remote consoles and management systems.

DATA CENTER

The common components of a data center working together to provide virtualized IT resources supported by physical IT resources.



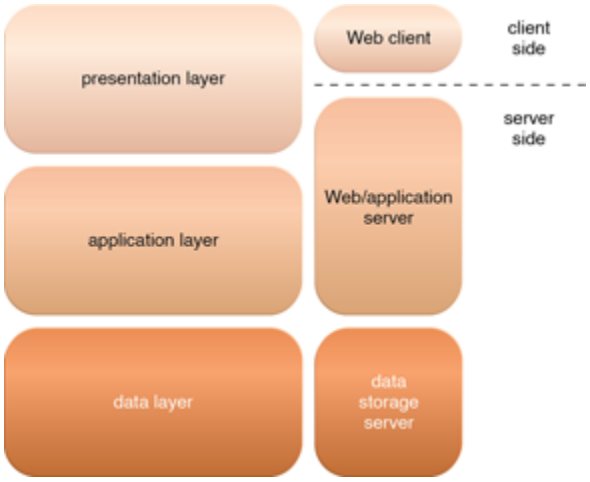
Web technology

is generally used as both the implementation medium and the management interface for cloud services.

Web application

is based on the basic three-tier model.

1. first tier is called the **presentation layer** which represents the userinterface.
2. middle tier is the **application layer** that implements application logic.
3. third tier is the **data layer** that is comprised of persistent data stores.



Cloud usage monitor

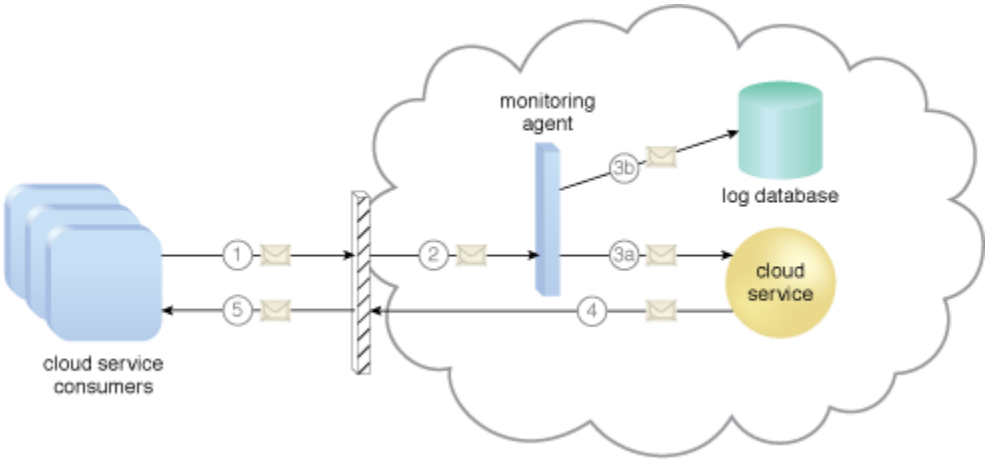
mechanism is a lightweight and autonomous software program responsible for collecting and processing IT resource usage data.

a method of reviewing, observing, and managing the operational workflow in a cloud-based IT infrastructure.

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 dataflows.

a collectd-based daemon that gathers system and application metrics from virtual machine instances and sends them to Monitoring.

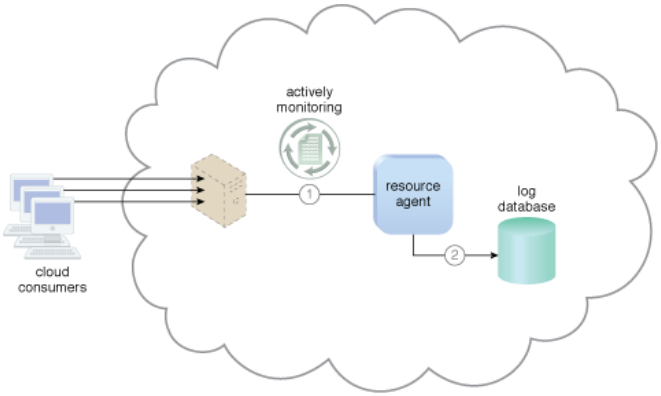


1. A cloud service consumer sends a request message to a cloud service.
2. The monitoring agent intercepts the message to collect relevant usage data.
3.
    - (a) before allowing it to continue to the cloud service.
    - (b) The monitoring agent stores the collected usage data in a log database.
4.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

Resource agent

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

Note// This module is used to **monitor usage metrics** based on predefined, observable events at the resource software level, such as **initiating**, **suspending**, **resuming**, and **vertical scaling**.

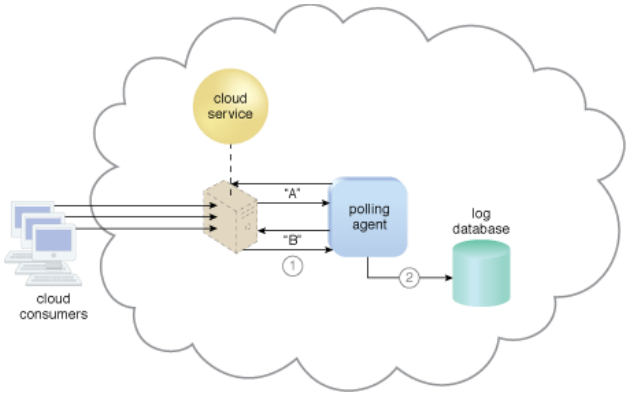


1. The resource agent is actively monitoring a virtual server and detects an increase in usage.
2. The resource agent receives a notification from the underlying resource management program that the virtual server is being scaled up and stores the collected usage data in a log database, as per its monitoring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

Polling agent

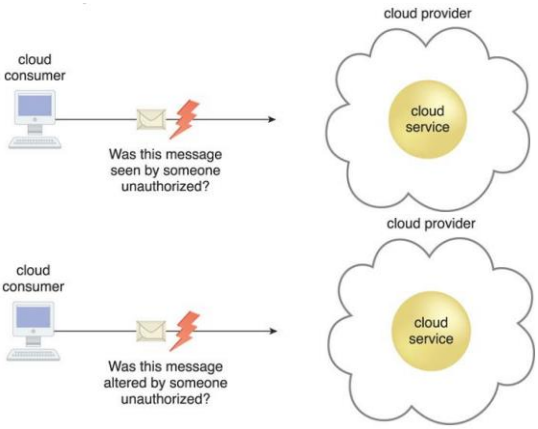
is a processing module that collects cloud service usage data by polling IT resources.

Note// This type of cloud service monitor is commonly used to **periodically monitor IT resource status**, such as **uptime** and **downtime.**



1. A 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”
2. upon which the polling agent records the new usage status in the log database

Chapter 4



- **Confidentiality**
  It is the characteristic of something being made accessible only to authorized parties.

- **Integrity**
  It is the characteristic of not having been changed by an unauthorized party.

- **Authenticity**
  is the characteristic of something having been provided by an authorized source.

- **Availability**
  is the characteristic of being accessible and usable during a specified time period.

- **Threat**
  is a potential security violation that can challenge defenses in an attempt to breach privacy and/or cause harm.

- **Vulnerability**
  is a weakness that can be exploited either because it is protected by insufficient security controls, or because existing security controls are overcome by an attack.

- **Risk**
  Risk is the possibility of loss or harm arising from performing an activity.
  Two metrics that can be used to determine risk for an IT resource are
  1- The probability of a threat occurring to exploit vulnerabilities in the IT resource.
  2- The expectation of loss upon the IT resource being compromised.

- **Security Controls**
  are counter measures used to prevent or respond to security threats and to reduce or avoid risk.

- **Security Policies**
  establishes a set of security rules and regulations.
  Note// Often, security policies will further define how these rules and regulations are implemented and enforced.

Threat agents

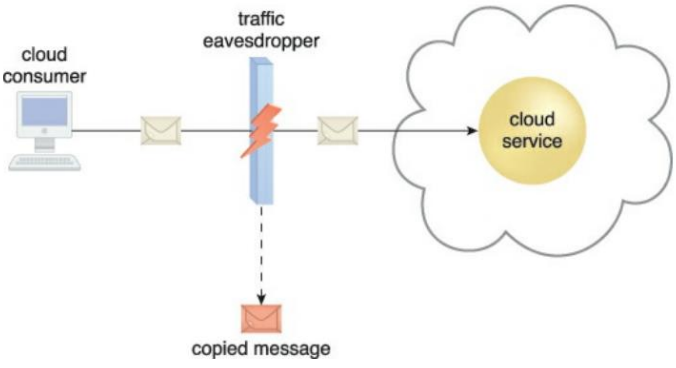
is an entity that poses a threat because it is capable of carrying out an attack.

Note// Cloud security threats can originate either **internally** or **externally**, from **humans** or **software** programs.

1. **Anonymous attacker**
   is a non-trusted cloud service consumer without permissions in the cloud.
   Note// It typically exists as an external software program that launches network-level attacks through public networks.
   Note// anonymous attackers often resort to committing acts like **bypassing user accounts** or **stealing user credentials**.
2. **Malicious service agent**
   is able to intercept and forward the network traffic that flows within a cloud.
   Note// It typically exists as a service agent (or a program pretending to be a service agent) with compromised or malicious logic. It may also exist as an external program able to remotely intercept and potentially corrupt message contents.
3. **Trusted attacker** // also known as **malicious tenants.**
   an authorized cloud service consumer with legitimate credentials that it uses to exploit access to cloud-based IT resources.
   Note// **hacking of weak authentication** processes, the **breaking of encryption**, the **spamming of e-mail** accounts, or to launch common attacks, such as **denial of service** campaigns.
4. **Malicious insider**
   are human threat agents acting on behalf of or in relation to the cloud provider.
   Note// They are typically current or former **employees** or **third parties** with access to the cloud provider's premises.
   Note// the malicious insider may have administrative privileges for accessing cloud consumer IT resources.

Traffic Eavesdropping

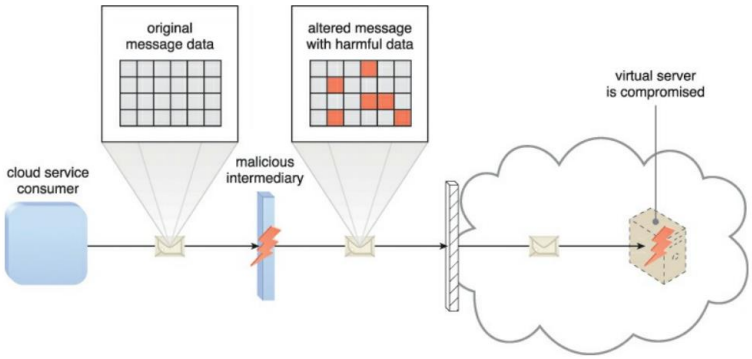
occurs when the data that being transferred to or within a cloud is passively intercepted by a malicious service agent for illegitimate information-gathering purposes.



An externally positioned malicious service agent carries out a traffic eavesdropping attack by intercepting a message sent by the cloud service consumer to the cloud service. The service agent makes an unauthorized copy of the message before it is sent along its original path to the cloud service.

Malicious Intermediary threat

arises when messages are intercepted and altered by a malicious service agent.

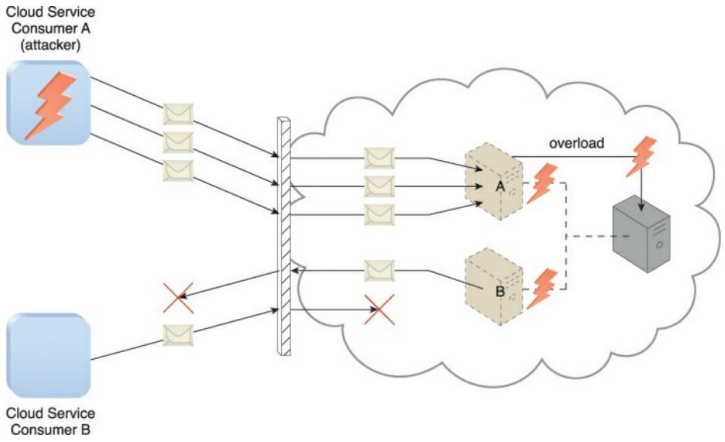


The malicious service agent intercepts and modifies a message sent by a cloud service consumer to a cloud service (not shown) being hosted on a virtual server. Because harmful data is packaged into the message, the virtual server is compromised.

Denial of Service (DoS) attack

Is used to overload IT resources to the point where they cannot function properly.

- The workload on cloud services is artificially increased with imitation messages or repeated communication requests.
- The network is overloaded with traffic to reduce its responsiveness and cripple its performance.
- Multiple cloud service requests are sent.

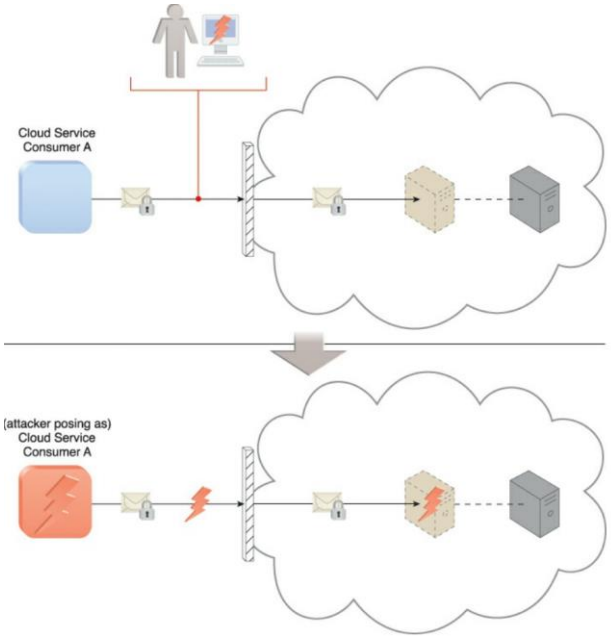


Cloud Service Consumer A sends multiple messages to a cloud service (not shown) hosted on Virtual Server A. This overloads the capacity of the underlying physical server, which causes outages with Virtual Servers A and B. As a result, legitimate cloud service consumers, such as Cloud Service Consumer B, become unable to communicate with any cloud services hosted on Virtual Servers A and B.

Insufficient Authorization attack

occurs when access is granted to an attacker erroneously or too broadly, resulting in the attacker getting access to IT resources that are normally protected.

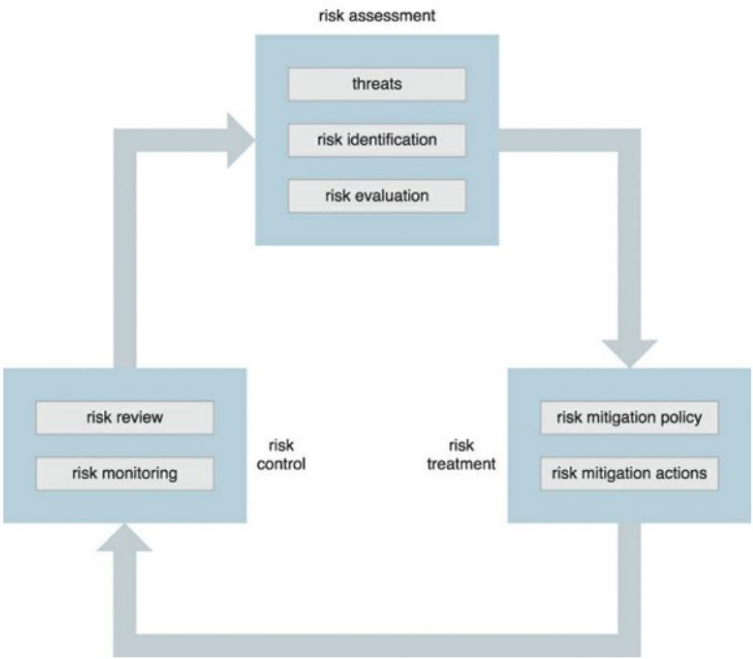
Note// This is often a result of the attacker gaining direct access to IT resources, A variation of this attack, known as **weak authentication**.



An attacker has cracked a weak password used by Cloud Service Consumer A. As a result, a malicious cloud service consumer (owned by the attacker) is designed to pose as Cloud Service Consumer A in order to gain access to the cloudbased virtual server

Risk Management

1. **Risk Assessment**
   the cloud environment is analyzed to identify potential vulnerabilities and shortcomings that threats can exploit.
   Note// The cloud provider can be asked to produce statistics and other information about past attacks (successful and unsuccessful) carried out in its cloud.
2. **Risk Treatment**
   Mitigation policies and plans are designed during the risk treatment.
3. **Risk Control**
   The risk control stage is related to risk monitoring, a three-step process that is comprised of surveying related events, reviewing these events to determine the effectiveness of previous assessments and treatments, and identifying any policy adjustment needs.

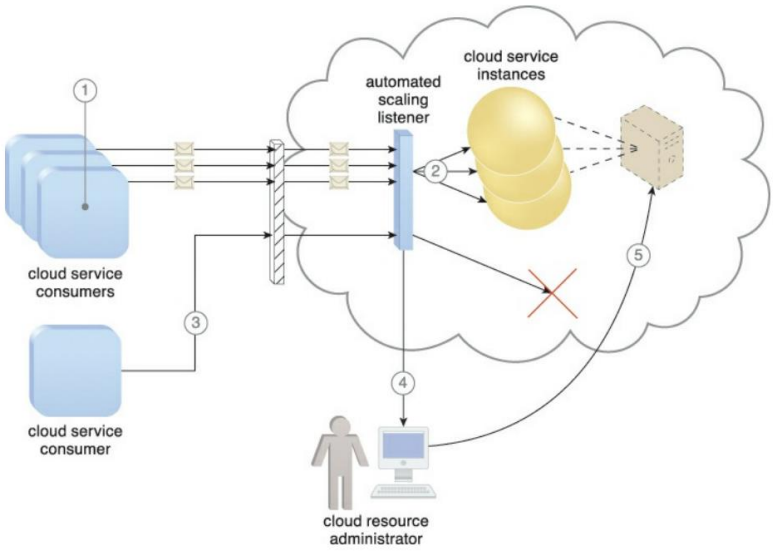


Chapter 5

Automated Scaling Listener

is a service agent that monitors and tracks communications between cloud service consumers and cloud services **for dynamic scaling purposes**.

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 consumer's cloud resource administrator accesses the remote administration environment to adjust the provisioning setup and increase the redundant instance limit.

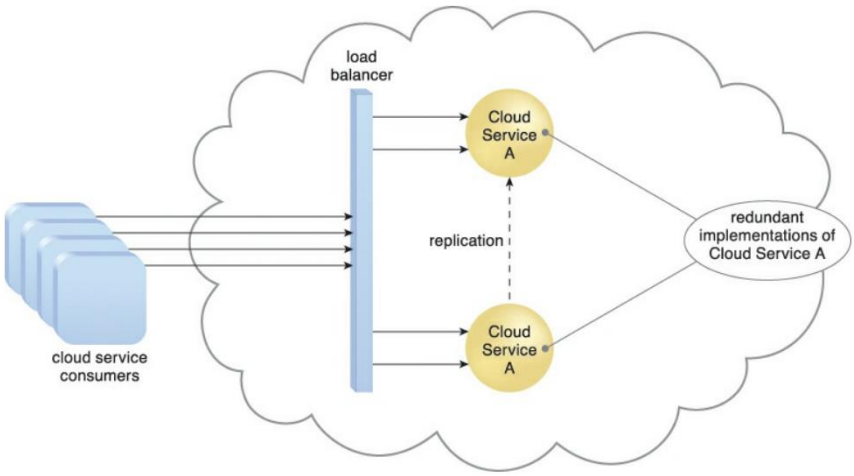


Load Balancer

Used to balance a workload across two or more IT resources to increase performance and capacity.

Note// The load balancer mechanism is a **runtime agent** with logic fundamentally based on this premise.

- **Asymmetric Distribution**
  larger workloads are issued to IT resources with higher processing capacities.
- **Workload Prioritization**
  workloads are scheduled, queued, discarded, and distributed workloads according to their priority levels.
- **Content-Aware Distribution**
  requests are distributed to different IT resources as dictated by the request content

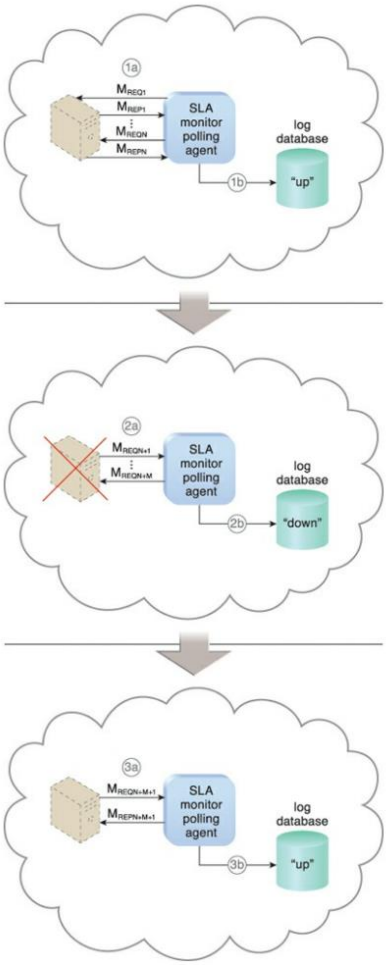


A load balancer **implemented as a service agent** transparently distributes incoming workload request messages across two redundant cloud service implementations, which in turn maximizes performance for the cloud service consumers.

SLA Monitor

Note// SLA **(Service Level Agreement)**

is used to specifically observe the runtime performance of cloud services to ensure that they are fulfilling the contractual QoS requirements that are published in SLAs.



1. The SLA monitor polls the cloud service by sending over polling request messages (MREQ1to MREQN). The monitor receives polling response messages (MREP1 to MREPN) that report that the service was “up” at each polling cycle (1a). The SLA monitor stores the “up” time—time period of all polling cycles 1 to N—in the log database (1b).

2. The SLA monitor polls the cloud service that sends polling request messages (MREQN+1 to MREQN+M). Polling response messages are not received (2a). 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 (2b).

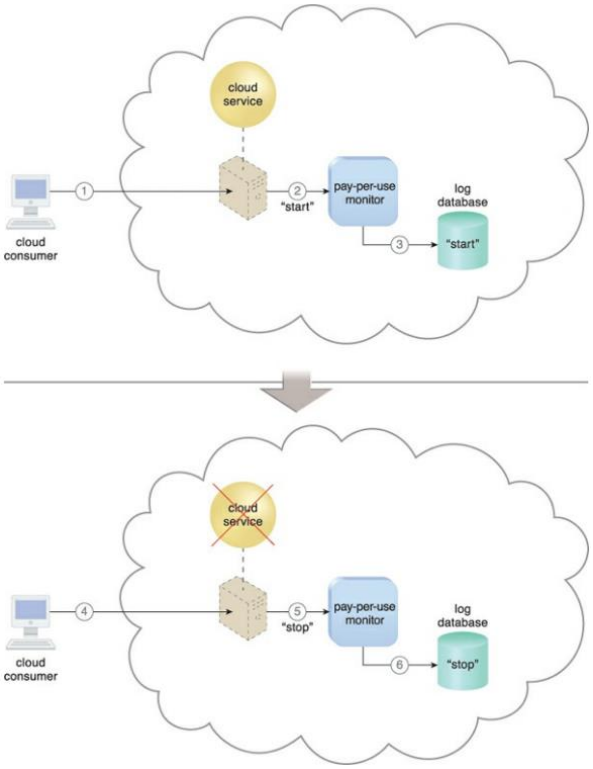
3. The SLA monitor sends a polling request message (MREQN+M+1) and receives the polling response message (MREPN+M+1) (3a). The SLA monitor stores the “up” time in the log database (3b).

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.

monitoring variables are

- request/response message quantity
- transmitted data volume
- bandwidth consumption



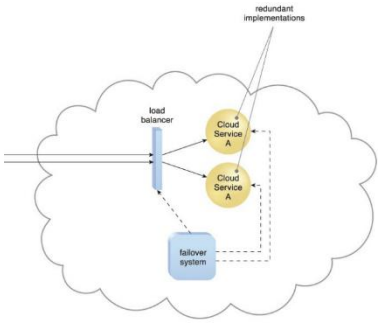
1. A cloud consumer requests the creation of a new instance of a cloud service (1).
2. The IT resource is instantiated and the pay-per-use monitor receives a “start” event notification from the resource software (2).
3. The pay-per-use monitor stores the value timestamp in the log database (3).
4. The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

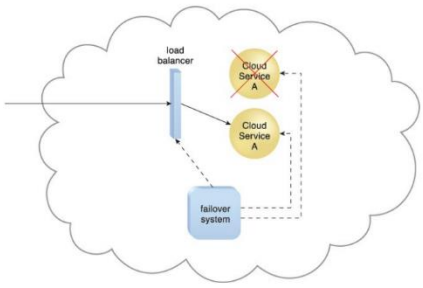
Failover System

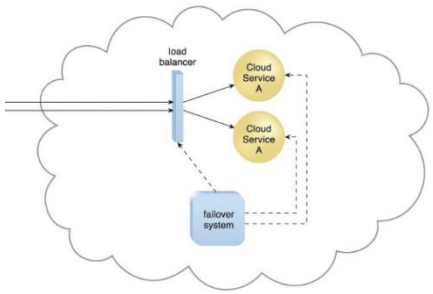
is used to increase the reliability and availability of IT resources by using established clustering technology to provide redundant implementations.

Note// 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.

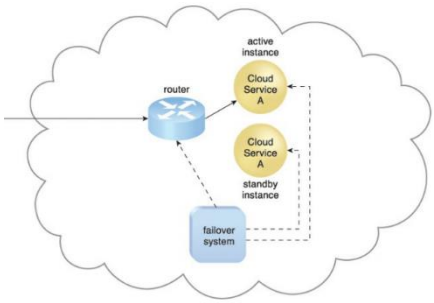
- **Active-Active**
  redundant implementations of the IT resource actively serve the workload synchronously.
  Note// When a failure is detected, the failed instance is removed from the load balancing scheduler.

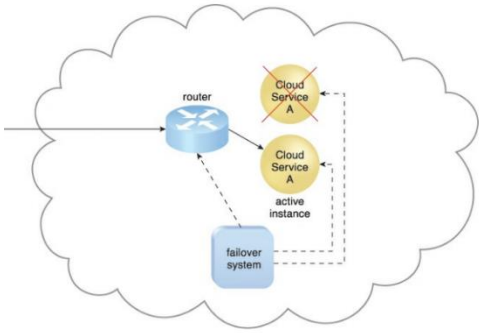


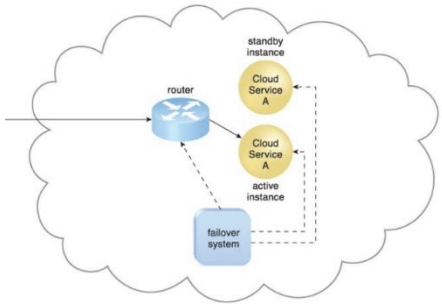




- **Active-Passive**
  a standby or inactive implementation is activated to take over the processing from the IT resource that becomes unavailable.
  Note// and the corresponding workload is redirected to the instance taking over the operation.





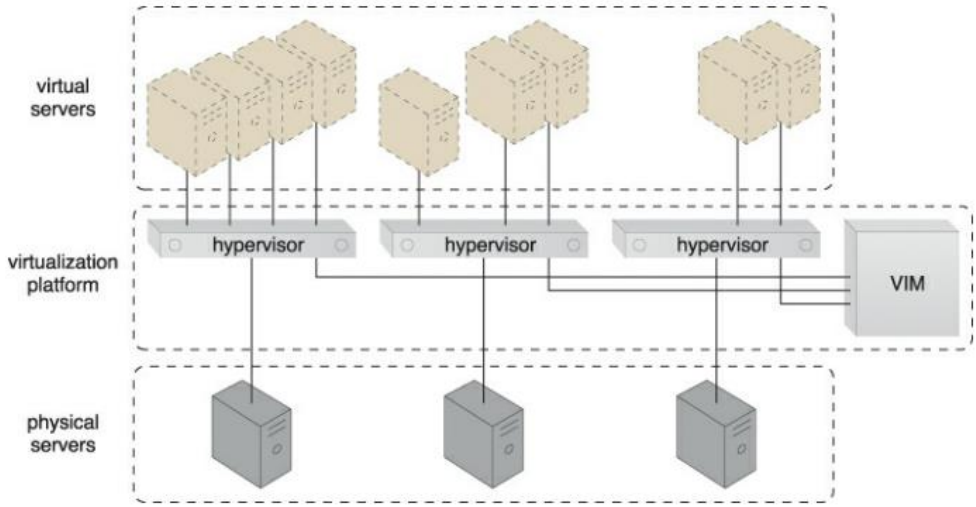


Hypervisor

is a fundamental part of virtualization infrastructure that is primarily used to generate virtual server instances of a physical server.

Note// hypervisor is generally limited to one physical server and can therefore only create virtual images of that server

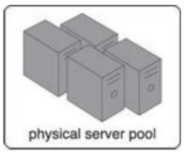
Virtual servers are created via individual **hypervisor on individual physical servers**. All three hypervisors are jointly controlled by the **same VIM**.



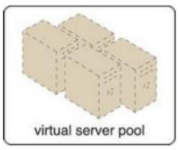
Resource Pooling Architecture

is based on the use of one or more resource pools, in which identical IT resources are grouped and maintained by a system that automatically ensures that they remain synchronized.

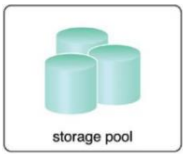
1. **Physical server pools**
   are composed of networked and synchronized physical servers.
   Physical server pools are composed of networked servers that have been installed with operating systems and other necessary programs and/or applications and are ready for immediate use.



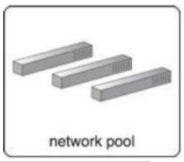
2. **Virtual server pools**
   are composed of networked and synchronized virtual servers.
   Virtual server pools are usually configured using one of several available templates chosen by the cloud consumer during provisioning. For example, a cloud consumer can set up a pool of mid-tier Windows servers with 4 GB of RAM or a pool of low tier Ubuntu servers with 2 GB of RAM.



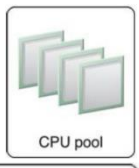
3. **Storage pools**
   are composed of networked and synchronized storage devices.
   Storage pools, or cloud storage device pools, consist of file-based or block-based storage structures that contain empty and/or filled cloud storage devices.



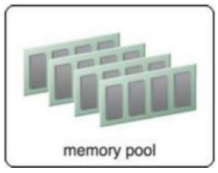
4. **Network pools**
   are composed of networked and synchronized network devices.
   Network pools (or interconnect pools) are composed of different preconfigured network connectivity devices. For example, a pool of virtual firewall devices or physical network switches can be created for redundant connectivity, load balancing, or link aggregation.



5. **CPU pools**
   are composed of networked and synchronized CPUs.
   CPU pools are ready to be allocated to virtual servers, and are typically broken down into individual processing cores.

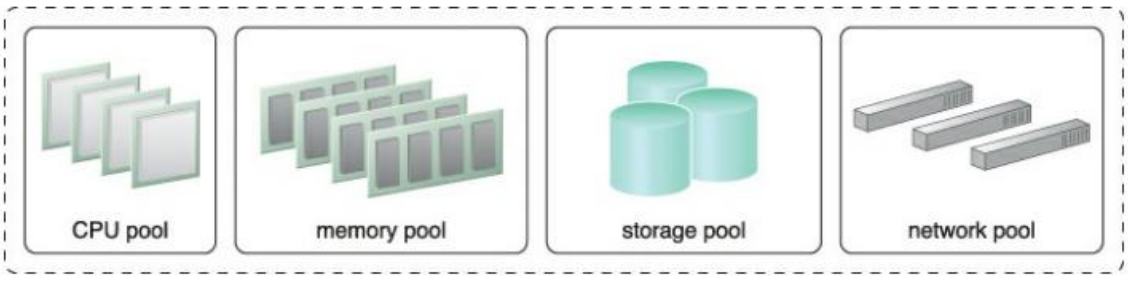


6. **Pools of physical RAM**
   are composed of networked and synchronized RAMs.
   Pools of physical RAM can be used in newly provisioned physical servers or to vertically scale physical servers.



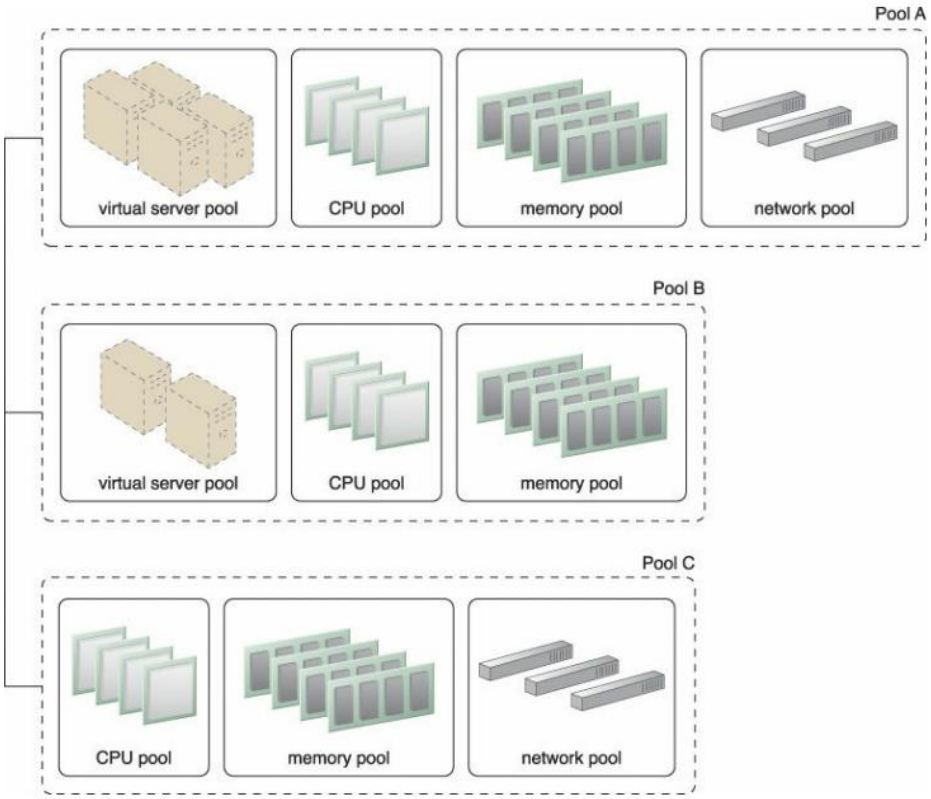
Dedicated pools

can be created for each type of IT resource and individual pools can be grouped into a larger pool



Multiple pools

Pools B and C are sibling pools that are taken from the larger Pool A, which has been allocated to a cloud consumer. This is an alternative to taking the IT resources for Pool B and Pool C from a general reserve of IT resources that is shared throughout the cloud.



Dynamic Scalability Architecture

is an architectural model based on a system of predefined scaling conditions that trigger the dynamic allocation of IT resources from resource pools.

- **Dynamic Horizontal Scaling** – IT resource instances are scaled out and in to handle fluctuating workloads.
- **Dynamic Vertical Scaling** – IT resource instances are scaled up and down when there is a need to adjust the processing capacity of a single IT resource.
- **Dynamic Relocation** – The IT resource is relocated to a host with more capacity.