Chapter 7: Cloud Infrastructure Mechanisms

- are **foundational** building blocks of cloud environments that establish **primary** artifacts to form the **basis** of **fundamental** cloud technology architecture.
- **Mechanisms**:
    - Logical Network Perimeter
    - Virtual Server
    - Cloud Storage Device
    - Cloud Usage Monitor
    - Resource Replication
    - Ready-Made Environment

7.1 Logical Network Perimeter

- the **isolation** of a network environment from **the rest of** a communications network
- establishes **a virtual network** boundary that can **encompass and isolate** a group of related cloud-based IT resources that may be physically distributed
- This mechanism can be implemented to:
    - isolate IT resources in a cloud from:
        - **non-authorized users**
        - **non-users**
        - **cloud consumers**
    - **control the bandwidth** that is available to isolated IT resources
- **Isolation mechanisms:** isolated via network devices that supply and control the connectivity of a data center and are commonly deployed as virtualized IT environments that include:
    - **Virtual Firewall** – An IT resource that actively **filters** network traffic to and from the isolated network while controlling its interactions with the Internet.
    - **Virtual Network** – Usually acquired through **VLAN**s, this IT resource isolates the network environment within the data center infrastructure.
- The **cloud consumer'**s IT environment and **the cloud IT resources** are connected via so called VPN implemented by point-to-point encryptuon of the data packets between two endpoints
-

7.2 Virtual Server

Virtualization Mechanisms:

- A **core** technology
- Allows **multiple cloud consumers** to share the same physical server
- Instant **VM** (Virtual Machine) **creation** by copying template VM image file (on-demand resource provisioning)

- On-line **scaling up/down** (by allocating more or less **cores**) or **out/in** (by adding/removing VM instances)
- On-line **server migration** by replicating VM image file
- Seamless service **failover** by reinstating the same VM image file
- Effective **load balancing** by even provisioning and real-time online migration
- Easy administration and self-provisioning

7.3 Cloud Storage Device

Definition & concerns:

- storage devices that are designed specifically for cloud-based provisioning
- Possibly **virtualized** in or **distributed** in general
- Usually upper-bounded due to capacity allocation in support of the **pay-per-use** mechanism
- Open to remote access via cloud storage services (via **RESTful APIs**)
- Main concern: the **security, integrity** and **confidentiality** of data
- **Legal** and **regulatory** issues for relocating data across geographical or national boundaries
- Performance issues as well due to remote and/or **large data** access

Cloud storage levels:

- **Files**: collections of data are grouped into files that are located in folders
- **Blocks**: the lowest level of storage and the closest to the hardware - a block is the smallest unit of data that is still individually accessible
- **Datasets**: sets of data organised into a table-based, delimited or record format
- **Objects**: data and its associated metadata organised as web-based resources

Each data storage levels associated with a certain type of technical interfaces or APIs:

- Network storage interfaces
- Object storage interfaces
- Database storage interfaces:
    - Relational data storage
    - Non-relational data storage

7.4 Cloud Usage Monitor

Definition:

- A **lightweight** and **autonomous** software program responsible for **collecting** and **processing** IT resource usage data
- **Different formats** are available based on the type of usage metrics and the was usage data needs to be collected

- Agent-based implementation in which usage data are collected and forwarded to **a log database** for post-processing and reporting purpose
- **Monitoring Agent:**
    - An **intermediary**, **event-driven** program that exists as a service agent and resides along existing communication paths, to **transparently** monitor and analyze dataflow
    - Commonly used to measure **network traffic** and **message metrics**
- **Resource Agent:**
    - Processing module that collects usage data by having **event-driven**, interactions with specialized resource software
    - Commonly used to monitor usage metrics based on predefined, observable events, at the resource software level, such as: **initiating**, **suspending**, **resuming** and **vertical scaling**
- **Polling Agent:**
    - A processing module that collects cloud service usage data, by polling IT resources
    - Commonly used to **periodically** monitor IT resource status, such as **uptime** and **downtime**

7.5 Resource Replication

Definition:

- The creation of multiple instances of the same IT resources
- Primarily to enhance the **availability** and the **performance** of IT resources

The nature of virtualization technology:

- VM, configuration, memory status and data are stored in **image files** in a virtualized environment
- Resource replication can then be easily done via **replication** of image files

7.6 Ready-Made Environment

Definition:

- A 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:
    - **database,**
    - **middleware (multitenant apps),**
    - **development tools (SDK),** and
    - **governance.**

Chapter 8: Specialized Cloud Mechanisms

8.1 Automated Scaling Listener

- A service agent that monitors and tracks communications between cloud service consumers and cloud services for dynamic scaling purpose
- Deployed within the cloud, typically near the firewall from where these agents automatically track workload status information

How does it work:

- automatically track workload status information
- Workloads can be determined by:
    - he volume of cloud consumer-generated requests
    - back-end processing demands triggered by certain types of requests

8.2 Load Balancer

- A runtime agent to balance a workload across two or more IT resources to increase performance and capacity beyond what a single IT resource can provide
- An attempt to distribute overall workload as evenly as possible across all available IT resources

Implementation Mechanisms:

- **Round-robin distribution:** a simple division of labor distribution (one after another)
- **Less load first distribution:** assign a new request to one with the smallest current load
- **Asymmetric distribution**: larger workloads are issued to IT resources with higher processing capacities
- **Workload prioritization:** workloads are scheduled, queued, discarded and distributed according to their priority levels
- **Content-aware distribution:** requests are distributed to different IT resources as dictated by the request content

Load balancers come in the form of:

- Multi-layer network switch (layer 4 or higher)
- Dedicated hardware appliance
- Dedicated software-based system (common in server os),
- Service agent

8.3 SLA Monitor

Definition:

- A resource agent that monitors and keeps track of the runtime performance of cloud services to ensure that they are fulfilling the contractual QoS requirements that are published in SLAs

Implementation Mechanisms:

- Periodic polling,
- SLAs reporting metrics as **uptime** and **downtime**,
- Proactively **repair** or **failover** cloud services when exception condition occurs,
- **Health checker** or **heartbeat checker** in traditional High Availability systems.

8.4 Pay-Per-Use Monitor

Definition:

- An **event-driven** or **monitoring** resource agent that measures cloud-based IT resource usage in accordance with predefined pricing parameters and generates usage logs for **fee calculations** and **billing purposes**.

Typical monitoring variables:

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

8.5 Audit Monitor

Definition:

- A **monitoring agent** that collects audit tracking data for networks and IT resources in support of (or dictated by) **regulatory** and **contractual** obligations.

8.6 Failover System

Definition:

- A system to increase the **reliability** and **availability** of IT resources **by using** established clustering technology to provide **redundant implementations**.
- Commonly used for mission-critical programs and reusable services that can introduce a single point of failure for multiple applications.
- Can span across more than one geographical region.

Two basic configurations:

- Active-active:
    - Redundant implementations of the IT resource actively serve the workload synchronously.
    - When a failure is detected, the failed instance is removed from the load balancing scheduler.
    - The remaining IT resource takes over the processing.
- Active-passive:

    - A stand-by or inactive implementation is activated to take over the processing from the IT resource that becomes unavailable and the corresponding workload is redirected to the instance taking over the operation.

8.7 Hypervisor

Definition:

- The hypervisor mechanism is a fundamental part of virtualization infrastructure that is primarily used to generate virtual server instances of a physical server.
- Limited to **one** physical server.

Responsible for:

- Creating VMs,
- Increasing VM capacity,
- Decreasing VM capacity,
- Shutting VM down.

8.8 Resource Cluster

Definition:

- A mechanism to group multiple IT resource instances together for them to act as a **single IT resource.**
- A technology to logically combine multiple physical IT resources to **improve the availability** and **to increase computing capacity**.

Implementation mechanisms:

- Distributed middleware implementation (cluster middleware)
    - Basic role:
        - Workload distribution,
        - Task scheduling,
        - Data sharing,
        - System synchronization,
        - SSI (Single System Image),
        - SAP(Single Access Point), etc.

Types:

- **Server cluster**
    - Clustering physical or virtual servers
        - Increase performance and availability
    - Hypervisors running on different physical servers can be configured to share **virtual execution states:**
        - Memory pages,
        - Processor register states.
    - Possible live VM migration.
- **Database cluster - RDBMS or SQL type:**
    - Designed to improve data **availability** - not performance

    - **Data synchronization** between different storage devices to maintain the **consistency & availability**
    - Redundant capacity required to synchronize data among physically distributed multiple storage
- **Large dataset cluster - NoSQL or bigdata processing**
    - A large dataset partitioned and distributed across multiple storages for **independent processing**

Types of Resource Clusters:

- **Load Balancer 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 set up as a separate IT resource.
- **HA Cluster**
    - A high-availability cluster maintains system availability in the event of multiple node failures, and has redundant implementations of most or all of the clustered IT resources.
    - It implements a failover system mechanism that monitors failure conditions and automatically redirects the workload away from any failed nodes.

8.9 Multidevice Broker

Definition:

- A mechanism to facilitate runtime **data transformation** so as to make a cloud service accessible **to a wider range of cloud server** consumer programs and service
- Commonly implemented as a **gateway** or incorporate **gateway components** 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 into protocols that are compatible with a cloud service

8.10 State Management Database

Definition:

- **A state management database** is a storage device that is used to temporarily persist state data for software programs.
- As an alternative to caching state data in memory, software programs an off-load state data to the database in order to reduce the amount of runtime memory they consume.