# Cloud Computing and Virtualisation / Cloud Computing Technologies

Week 8:
Cloud Architectures

# Outline

- Week 6. Cloud Architectures
  - Fundamental Cloud Architectures
  - Advanced Cloud Architectures
  - Specialised Cloud Architectures


# Chapter 11
# Fundamental Cloud Architectures

## Fundamental Cloud Architectures

1. Workload Distribution Architecture
2. Resource Pooling Architecture (sibling/nested)
3. Dynamic Scalability Architecture (HVR)
4. Elastic Resource Capacity Architecture (scripts)
5. Service Load Balancing Architecture (indep/in-built)
6. Cloud Bursting Architecture (on-premiseto cloud)
7. Elastic Disk Provisioning Architecture (thick/thin)
8. Redundant Storage Architecture

# Fundamental Cloud Architectures

- Workload Distribution Architecture: combination of horizontal scaling and a load balancer allowing the even distribution of workload among IT resources
  - Audit Monitor: type and geographical location of IT resources may require monitoring for legal and regulatory requirements
  - Cloud Usage Monitor: used for runtime workload tracking and data processing
  - Hypervisor: distribution of workload between hypervisors and hosted virtual servers
  - Logical Network Perimeter: isolates cloud consumer network boundaries where workloads are distributed
  - Resource Cluster: Cluster IT resources inactive/active mode commonly used
  - Resource Replication: can generate new instances of virtualized It resources in response to runtime demands



# Fundamental Cloud Architectures

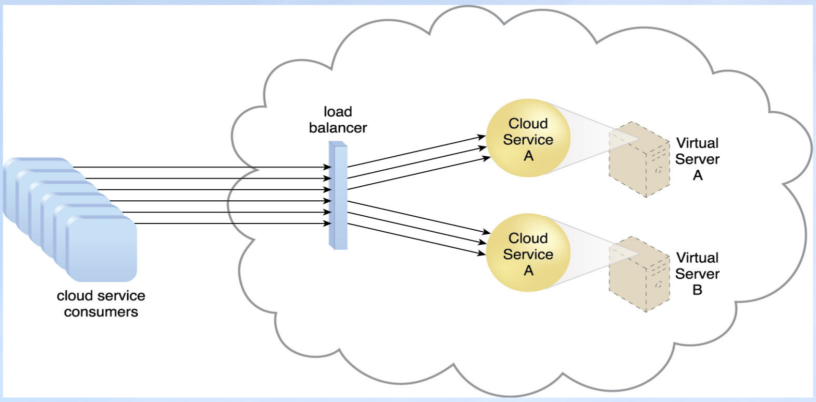



Figure 11.1 A redundant copy of Cloud Service A is implemented on Virtual Server B. The load balancer intercepts cloud service consumer requests and directs them to both Virtual Servers A and B to ensure even workload distribution.

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Fundamental Cloud Architectures

- Resource Pooling Architecture: uses one or more resource pools where identical IT resources are synchronised
    - Examples: (identical) physical server pools, virtual server pools, storage pools, network pools, CPU pools, memory pools, etc.
  - Audit Monitor: monitors pool usage to ensure compliance with privacy and regulation requirements
  - Cloud Usage Monitor: required for runtime tracking and synchronisation
  - Hypervisor: provides virtual servers with access to resource pools in addition to hosting the virtual servers/pools themselves
  - Logical Network Perimeter: logically organises and isolates resource pools
  - Pay-Per-Use Monitor: collects usage and billing information from pools
  - Remote Administration System: interfaces with backend systems to provide administration functionality
  - Resource Management System: provides tools and permission management options for cloud consumers to manage pools
  - Resource Replication: generates new instances of IT resources for pools



# Fundamental Cloud Architectures

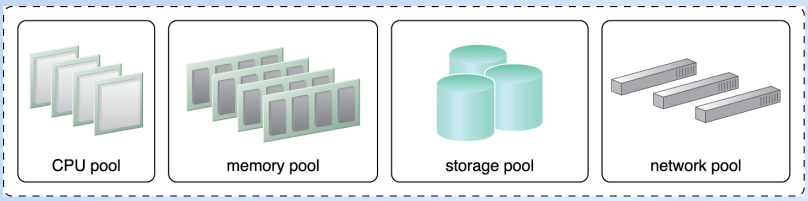
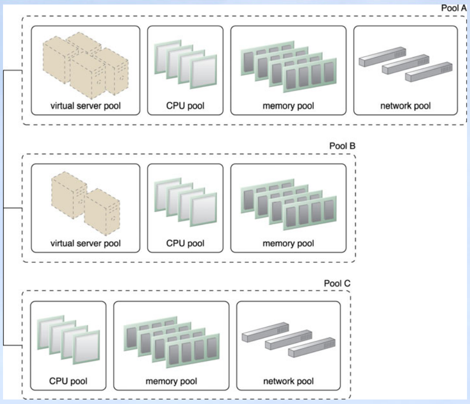



Figure 11.2 A sample resource pool that is comprised of four sub-pools of CPUs, memory, cloud storage devices, and virtual network devices.

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Fundamental Cloud Architectures

Figure 11.3 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.

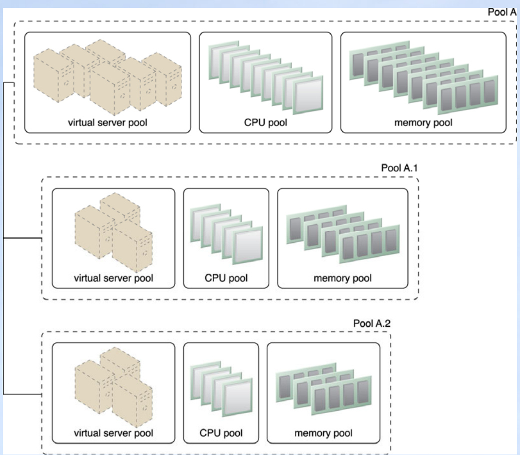



Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini



# Fundamental Cloud Architectures

Figure 11.4 Nested Pools A.1 and Pool A.2 are comprised of the same IT resources as Pool A, but in different quantities. Nested pools are typically used to provision cloud services that need to be rapidly instantiated using the same type of IT resources with the same configuration settings.




Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Fundamental Cloud Architectures

- Dynamic Scalability Architecture: triggers dynamic allocation of IT resources from resource pools based on predefined scaling conditions
    - Dynamic horizontal scaling
    - Dynamic vertical scaling, and
    - Dynamic relocation (relocate resources to a host with more capacity)
  - Cloud Usage Monitor: tracks runtime usage in response to dynamic fluctuations caused by this architecture
  - Hypervisor: invoked to create or remove virtual server instances or to be scaled itself
  - Pay-Per-Use Monitor: collects usage cost information in response to scaling of IT resources



# Fundamental Cloud Architectures

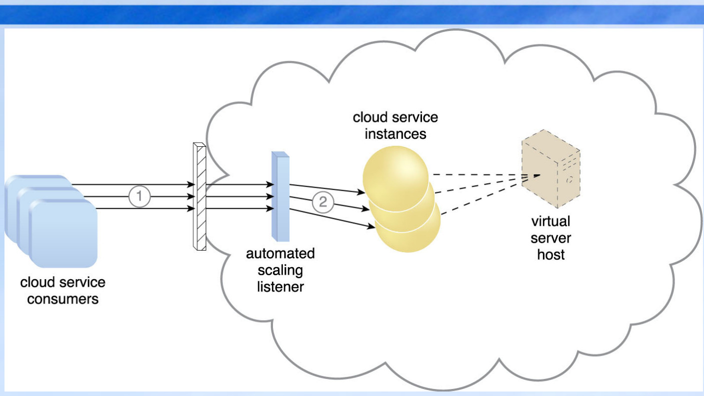



Figure 11.5 Cloud service consumers are sending requests to a cloud service (1). The automated scaling listener monitors the cloud service to determine if predefined capacity thresholds are being exceeded (2).

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Fundamental Cloud Architectures

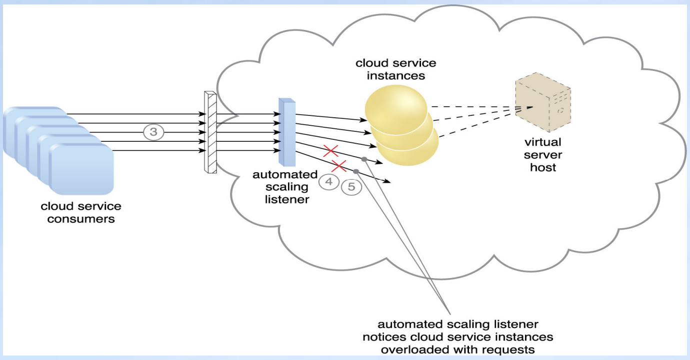



Figure 11.6 The number of requests coming from cloud service consumers increases (3). The workload exceeds the performance thresholds. The automated scaling listener determines the next course of action based on a predefined scaling policy (4). If the cloud service implementation is deemed eligible for additional scaling, the automated scaling listener initiates the scaling process (5).

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini



# Fundamental Cloud Architectures

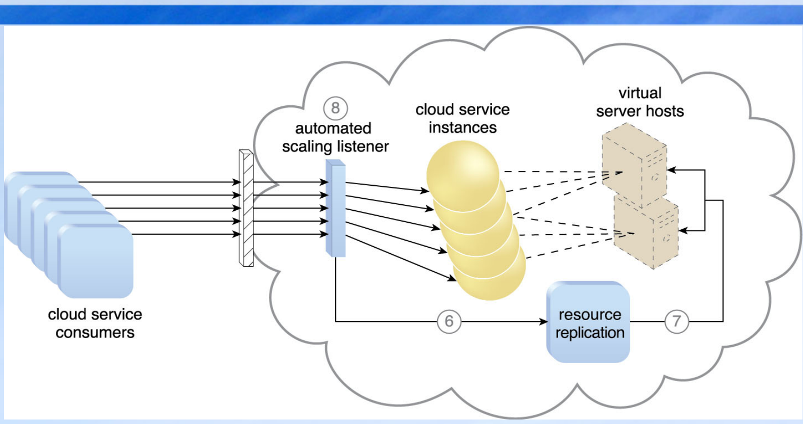



Figure 11.7 The automated scaling listener sends a signal to the resource replication mechanism (6), which creates more instances of the cloud service (7). Now that the increased workload has been accommodated, the automated scaling listener resumes monitoring and detracting and adding IT resources, as required (8).

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

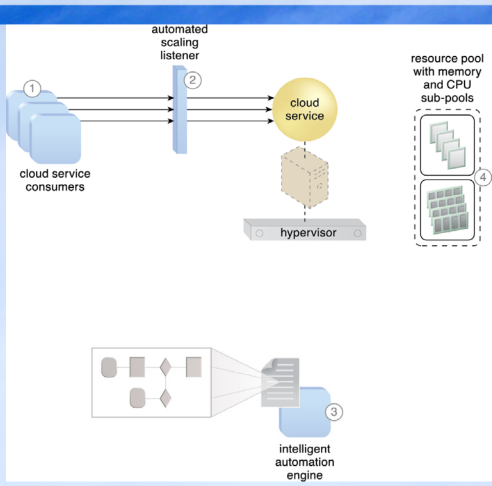
# Fundamental Cloud Architectures

- Elastic Resource Capacity Architecture: allocates and reclaims CPUs and RAM in immediate response to fluctuating requirements of IT resources
  - Cloud Usage Monitor: collect resource usage information on IT resources before, during, and after scaling, to help define the future processing capacity thresholds
  - Pay-Per-Use Monitor: collects usage cost information as it fluctuates with elastic provisioning
  - Resource Replication: generates new instances of scaled IT resources



# Fundamental Cloud Architectures

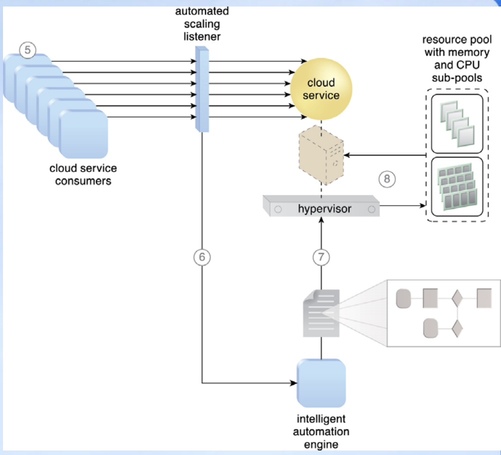
Figure 11.8 Cloud service consumers are actively sending requests to a cloud service (1), which are monitored by an automated scaling listener (2). An intelligent automation engine script is deployed with workflow logic (3) that is capable of notifying the resource pool using allocation requests (4).




Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Fundamental Cloud Architectures

Figure 11.9 Cloud service consumer requests increase (5), causing the automated scaling listener to signal the intelligent automation engine to execute the script (6). The script runs the workflow logic that signals the hypervisor to allocate more IT resources from the resource pools (7). The hypervisor allocates additional CPU and RAM to the virtual server, enabling the increased workload to be handled (8).




Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

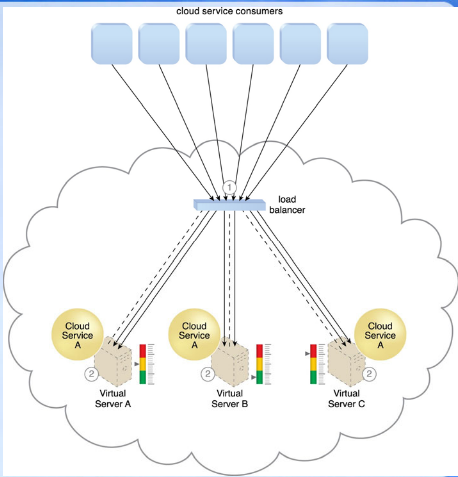


# Fundamental Cloud Architectures

- Service Load Balancing Architecture: variation of workload distribution architecture for scaling cloud services
  - Cloud Usage Monitor: monitors cloud service instances and their respective IT resource consumption levels
  - Resource Cluster: active-active cluster groups are used to help balance workloads across cluster members
  - Resource Replication: generates cloud service implementations in support of load balancing requirements

# Fundamental Cloud Architectures

Figure 11.10 The load balancer intercepts messages sent by cloud service consumers (1) and forwards them to the virtual servers so that the workload processing is horizontally scaled (2).

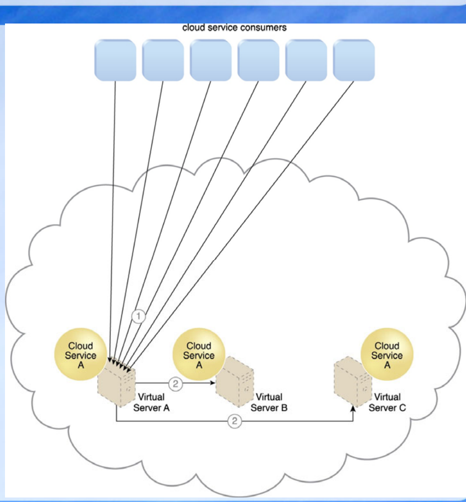



Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini



# Fundamental Cloud Architectures

Figure 11.11 Cloud service consumer requests are sent to Cloud Service A on Virtual Server A (1). The cloud service implementation includes built-in load balancing logic that is capable of distributing requests to the neighboring Cloud Service A implementations on Virtual Servers B and C (2).




Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Fundamental Cloud Architectures

- Cloud Bursting Architecture: a form of dynamic scaling where on-premise IT resources are primarily used but "bursts out" to cloud-based IT resources during high-demand periods
  - Automated Scaling Listener: determines when to start redirecting requests to pre-deployed cloud-based IT resources
  - Resource Replication: maintains synchronicity between on-premise and cloud-based IT resources



# Fundamental Cloud Architectures

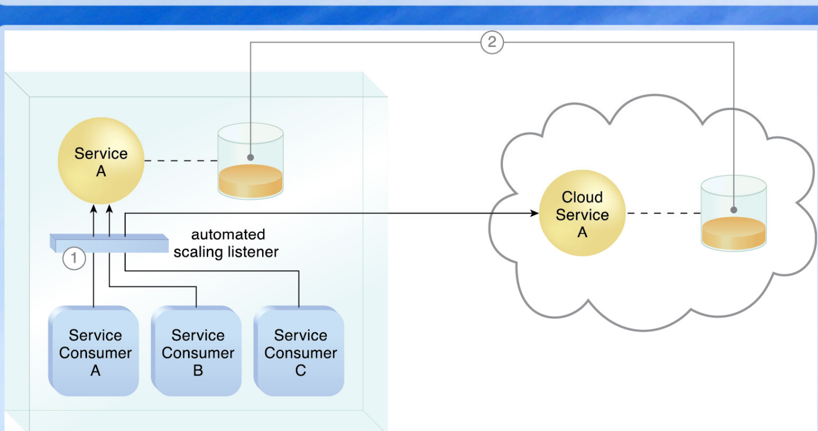



Figure 11.12 An automated scaling listener monitors the usage of on-premise Service A, and redirects Service Consumer C's request to Service A's redundant implementation in the cloud (Cloud Service A) once Service A's usage threshold has been exceeded (1). A resource replication system is used to keep state management databases synchronized (2).

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Fundamental Cloud Architectures

- Elastic Disk Provisioning Architecture: thin provisioning of storage to allow charging based on actual data consumption not maximum capacity
  - Cloud Usage Monitor: tracks and logs fluctuations in storage usage
  - Resource Replication: used if required to convert dynamic thin-disk storage into static thick-disk storage
    - Thin – storage is provisioned but not necessarily allocated from physical resources (can thin provision more storage than is physically available)
    - Thick –physical storage is allocated/reserved to match provisioned



# Fundamental Cloud Architectures

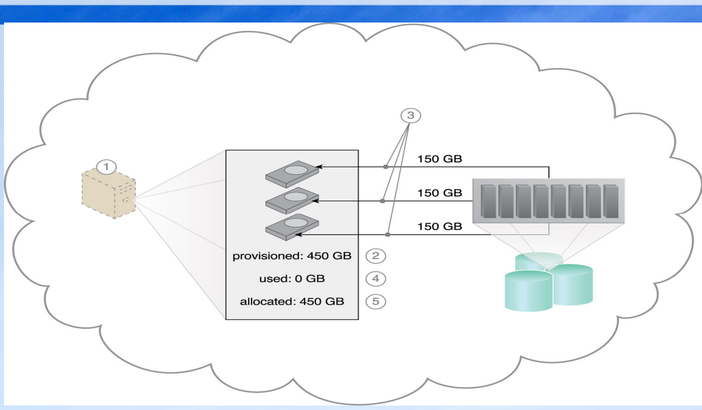



Figure 11.13 The cloud consumer requests a virtual server with three hard disks, each with a capacity of 150 GB (1). The virtual server is provisioned according to the elastic disk provisioning architecture, with a total of 450 GB of disk space (2). The 450 GB is allocated to the virtual server by the cloud provider (3). The cloud consumer has not installed any software yet, meaning the actual used space is currently 0 GB (4). Because the 450 GB are already allocated and reserved for the cloud consumer, it will be charged for 450 GB of disk usage as of the point of allocation (5).

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Fundamental Cloud Architectures

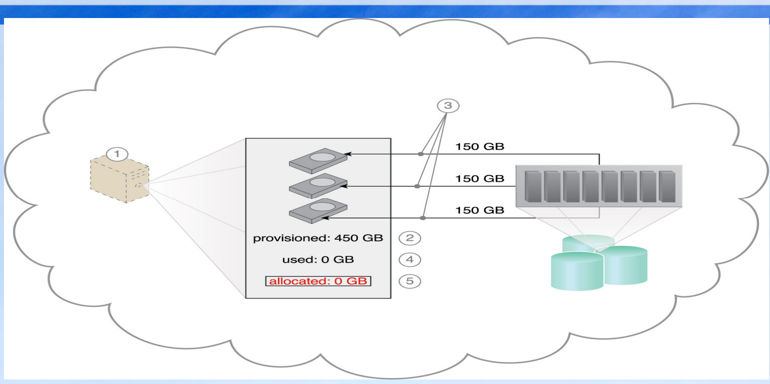



Figure 11.14 The cloud consumer requests a virtual server with three hard disks, each with a capacity of 150 GB (1). The virtual server is provisioned by this architecture with a total of 450 GB of disk space (2). The 450 GB are set as the maximum disk usage that is allowed for this virtual server, although no physical disk space has been reserved or allocated yet (3). The cloud consumer has not installed any software, meaning the actual used space is currently at 0 GB (4). Because the allocated disk space is equal to the actual used space (which is currently at zero), the cloud consumer is not charged for any disk space usage (5).

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini



# Fundamental Cloud Architectures

- Redundant Storage Architecture: use of secondary duplicate cloud storage device as part of a failover system
  - Storage replication (a variation of resource replication) used to synchronously/asynchronously replicate data from primary to secondary device
  - Storage service gateway diverts cloud consumer requests to the secondary device whenever the primary device fails
  - May be stored in a separately geographical region (subject to legal/regulatory requirements)

# Fundamental Cloud Architectures

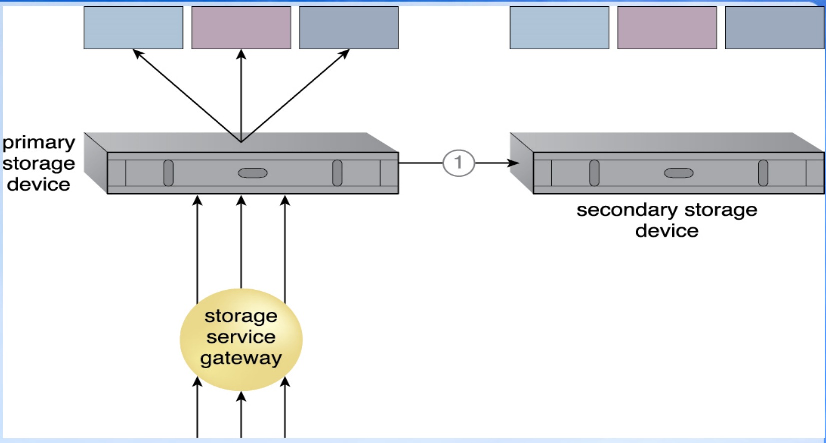



Figure 11.16 The primary cloud storage device is routinely replicated to the secondary cloud storage device (1).

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini



# Fundamental Cloud Architectures

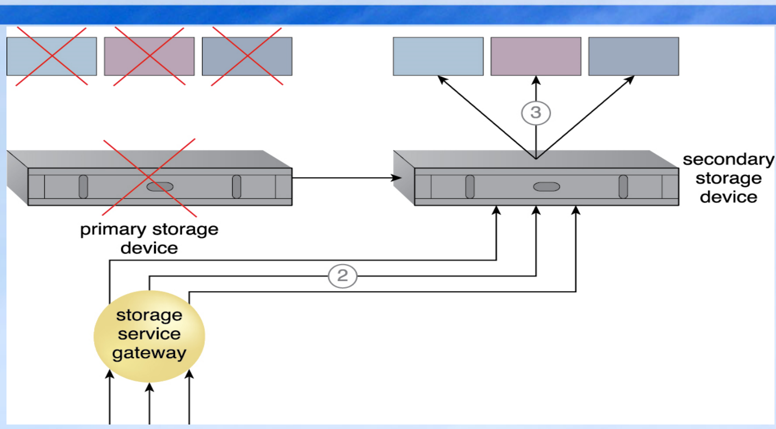



Figure 11.17 The primary storage becomes unavailable and the storage service gateway forwards the cloud consumer requests to the secondary storage device (2). The secondary storage device forwards the requests to the LUNs, allowing cloud consumers to continue to access their data (3).

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Fundamental Cloud Architectures

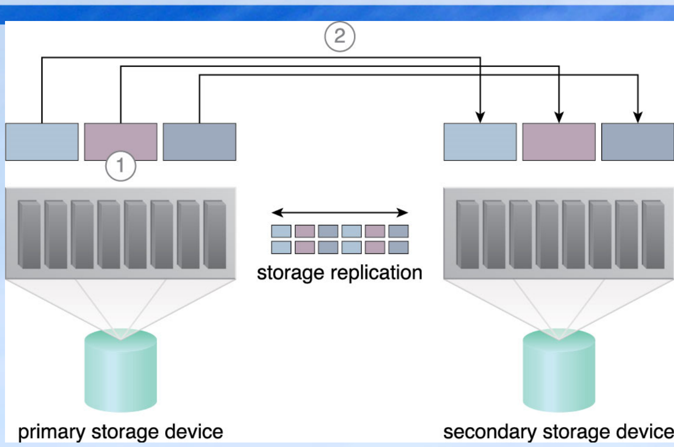



Figure 11.18 Storage replication is used to keep the redundant storage device synchronized with the primary storage device.

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini




# Chapter 12
# Advanced Cloud Architectures

# Advanced Cloud Architectures

- Hypervisor Clustering Architecture: establishes a high-availability cluster of hypervisors across multiple physical servers
  - Failure of a hypervisor can cascade to the virtual servers they host
  - Upon detecting a hypervisor failure, the hosted virtual servers can be moved to another physical server / hypervisor to maintain runtime operations (controlled by the VIM)
  - Logical Network Perimeter: ensure that none of the hypervisors of other cloud consumers are accidentally included in a given cluster
  - Resource Replication: hypervisors in the same cluster exchange status and availability to ensure changes to the cluster are replicated to all hypervisors



# Advanced Cloud Architectures

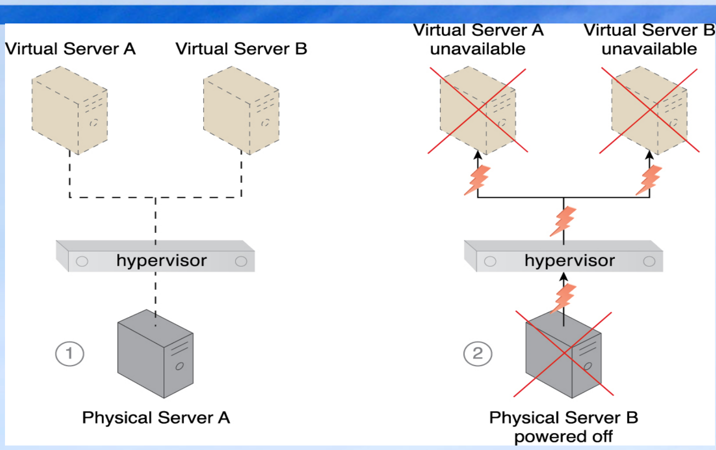
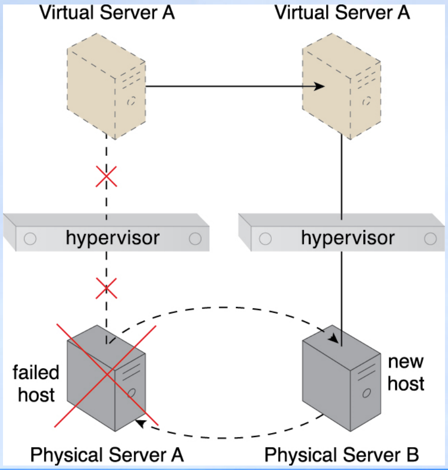



Figure 12.1 Physical Server A is hosting a hypervisor that hosts Virtual Servers A and B (1). When Physical Server A fails, the hypervisor and two virtual servers consequently fail as well (2).

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Advanced Cloud Architectures

Figure 12.2 Physical Server A becomes unavailable and causes its hypervisor to fail. Virtual Server A is migrated to Physical Server B, which has another hypervisor that is part of the cluster to which Physical Server A belongs.

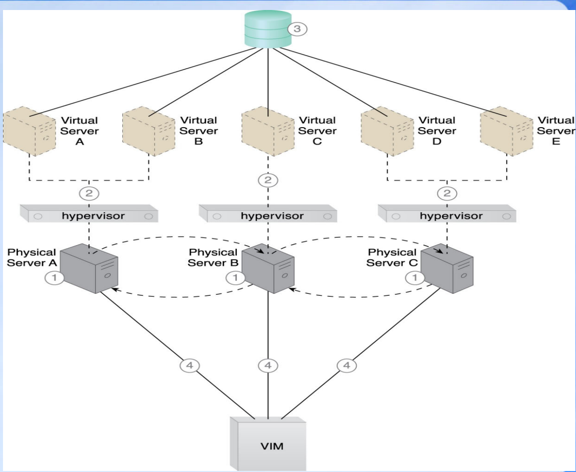



Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini



# Advanced Cloud Architectures

Figure 12.3 Hypervisors are installed on Physical Servers A, B, and C (1). Virtual servers are created by the hypervisors (2). A shared cloud storage device containing virtual server configuration files is positioned in a shared cloud storage device for access by all hypervisors (3). The hypervisor cluster is enabled on the three physical server hosts via a central VIM (4).




Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Advanced Cloud Architectures

- Load Balanced Virtual Server Instances Architecture: dynamically calculates virtual server instances and associated workloads before distributing processing across physical servers
    - Physical server over- and under-utilisation (imbalance) can increase dramatically over time
  - Automated Scaling Listener: dynamically monitors workload on virtual servers and initiates load balancing
  - Load Balancer: distributes workload between hypervisors
  - Logical Network Perimeter: ensures destination of relocated virtual server is compliant with SLA and privacy regulations
  - Resource Replication: virtual server replication may be required to support load balancing



# Advanced Cloud Architectures

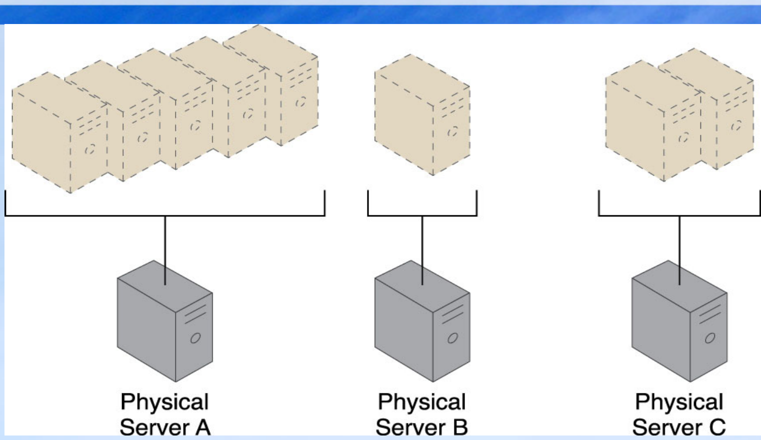
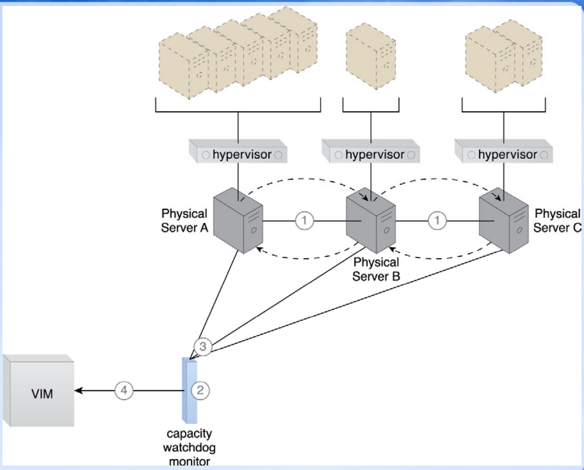



Figure 12.7 Three physical servers have to host different quantities of virtual server instances, leading to both over-utilized and under-utilized servers.

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Advanced Cloud Architectures

Figure 12.9 The hypervisor cluster architecture provides the foundation upon which the load-balanced virtual server architecture is built (1). Policies and thresholds are defined for the capacity watchdog monitor (2), which compares physical server capacities with virtual server processing (3). The capacity watchdog monitor reports an over-utilization to the VIM (4).

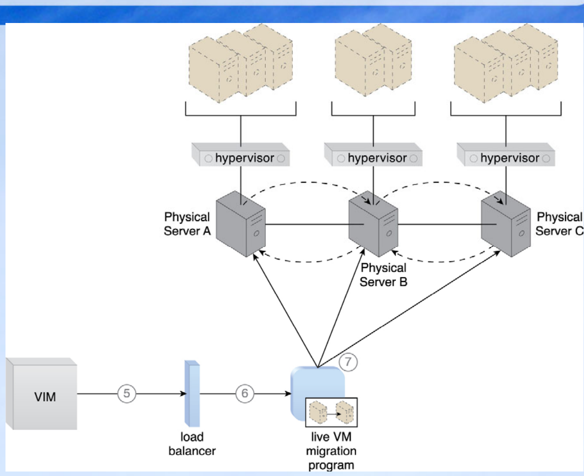



Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini



# Advanced Cloud Architectures

Figure 12.10 The VIM signals the load balancer to redistribute the workload based on pre-defined thresholds (5). The load balancer initiates the live VM migration program to move the virtual servers (6). Live VM migration moves the selected virtual servers from one physical host to another (7).




Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Advanced Cloud Architectures

- Non-Disruptive Service Relocation Architecture: predefined events trigger duplication or migration of a cloud service at runtime to avoid disruption
  - New cloud service implementation is guaranteed to be successfully receiving and responding to cloud service consumer requests before the original service implementation is deactivated or remoted
  - Cloud Usage Monitor: continuously track IT resource usage and system activity
  - Pay-Per-Use Monitor: collects data for service usage cost calculations at both source and destination locations
  - Resource Replication: instantiate the shadow copy of the cloud service at the destination
  - SLA Management System: processes SLA data to obtain cloud service availability assurances
  - SLA Monitor: collects SLA information required by the SLA management system



# Advanced Cloud Architectures

- Zero Downtime Architecture: virtual servers can be moved dynamically to different physical server hosts, in the event that their original physical server host fails
  - Virtual servers are stored on a shared volume to enable other physical hosts to access the relevant files
  - Audit Monitor: checks whether relocation of virtual servers also relocates hosted data to prohibited locations
  - Cloud Usage Monitor: monitor actual IT resource usage to help ensure virtual server capacities are not exceeded
  - Hypervisor: hosts the affected virtual servers
  - Logical Network Perimeter: provide and maintain isolation required to ensure cloud consumers remain within logical boundaries after virtual server relocation
  - Resource Cluster: creates active-active cluster groups to improve availability of virtual server-hosted IT resources
  - Resource Replication: creates new virtual server and cloud service instances upon primary virtual server failure

# Advanced Cloud Architectures

- Cloud Balancing Architecture: load balances IT resources across multiple clouds
    - Improves/increases performance and scalability of IT resources, availability and reliability of IT resources, and load-balancing and IT resource optimization
  - Automated Scaling Listener: redirects cloud service consumer requests to one of several redundant IT resource implementations, based on current scaling and performance requirements
  - Failover System: ensures redundant IT resources are capable of cross-cloud failover, notifying the Automated Scaling Listener of any failures to ensure requests aren't routed to unavailable/unstable IT resources



# Advanced Cloud Architectures

- Resource Reservation Architecture: IT resources are partly or wholly set aside for the exclusive use of a cloud consumer
    - Protection from resource constraints where demand for a shared IT resource can exceed capacity
    - Protection from resource borrowing where one resource pool temporarily "borrows" resources from another pool causing resource constraints
  - Audit Monitor: checks resource reservation is complying with cloud consumer auditing, privacy, and other regulatory requirements
  - Cloud Usage Monitor: oversees the thresholds that trigger allocation of reserved IT resources
  - Hypervisor: applies reservations for different cloud consumers to ensure they are correctly allocated guaranteed IT resources
  - Logical Network Perimeter: establishes necessary boundaries ensuring IT resources are made exclusively available to cloud consumers
  - Resource Replication: stays informed about IT resource consumption limits for each cloud consumer, replicating and provisioning new IT resources instances expediently

# Advanced Cloud Architectures

- Dynamic Failure Detection and Recovery Architecture: monitors and responds pre-defined failure scenarios, escalating failure conditions that cannot be automatically resolved
  - Audit Monitor: tracks whether data recovery is carried out in compliance with legal or policy requirements
  - Failover System: used during initial attempts to recover failed IT resources
  - SLA Management System and SLA Monitor: provides information regarding SLA guarantees



# Advanced Cloud Architectures

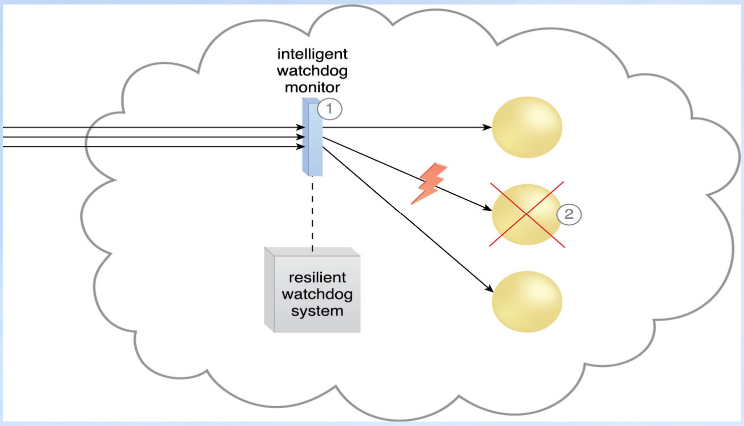



Figure 12.20 The intelligent watchdog monitor keeps track of cloud consumer requests (1) and detects that a cloud service has failed (2).

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Advanced Cloud Architectures

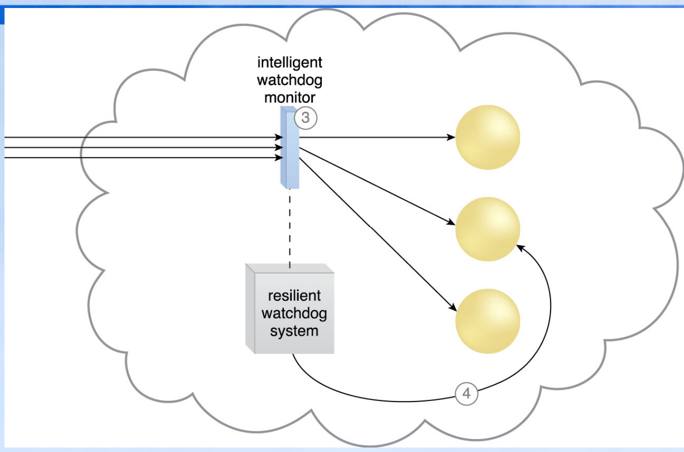



Figure 12.21 The intelligent watchdog monitor notifies the watchdog system (3), which restores the cloud service based on pre-defined policies. The cloud service resumes its runtime operation (4).

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini



# Advanced Cloud Architectures

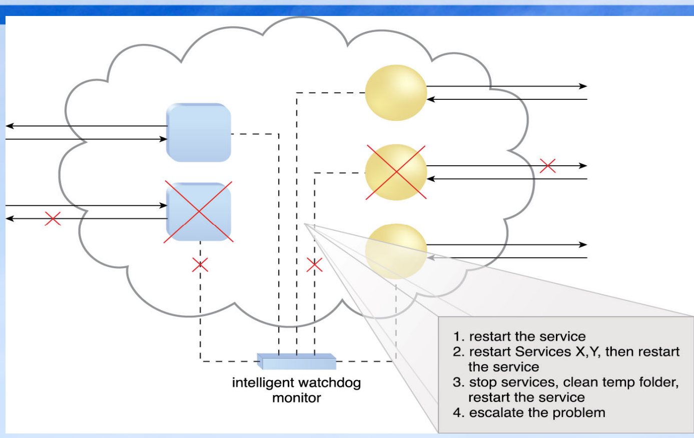



Figure 12.22 In the event of a failure, the intelligent watchdog monitor refers to its pre-defined policies to recover the cloud service step-by-step, escalating the process when a problem proves to be deeper than expected.

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Advanced Cloud Architectures

- Bare-Metal Provisioning Architecture: combines specialised service agents with hardware-based remote access to provide remote control of systems with no operating system installed, i.e., bare-metal
  - Discovery Agent: locates available physical servers for assigning to cloud consumers
  - Deployment Agent: installed in physical server's memory and used for bare-metal provisioning deployment system
  - Discovery Section: scans network to locate physical servers
  - Management Loader: loads management options for cloud consumer
  - Deployment Component: installs operating system on selected physical servers



# Advanced Cloud Architectures

- Bare-Metal Provisioning Architecture(cont):
  - Optional mechanisms include
    - Cloud Storage Device: stores operating system templates and installation files, deployment agents and deployment packages
    - Hypervisor: may be required for provisioning on physical servers
    - Logical Network Perimeter: helps ensure raw physical servers an only be accessed by authorized cloud consumers
    - Resource Replication: deploys a new hypervisor on a physical server to balance hypervisor workload during/after provisioning
    - SLA Management System: ensures availability of bare-metal servers according to SLA stipulations

# Advanced Cloud Architectures

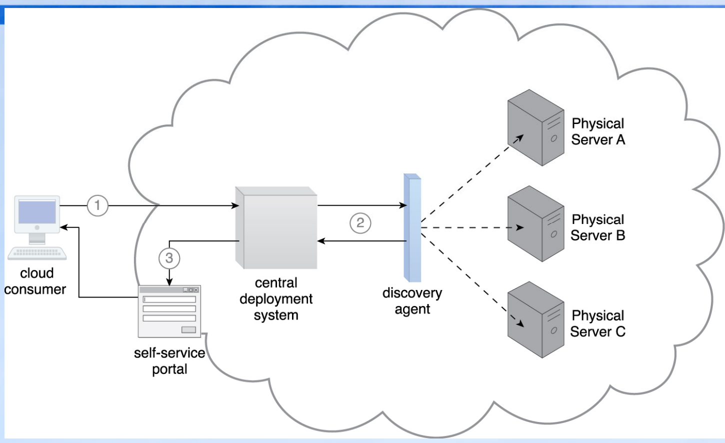



Figure 12.23 The cloud consumer connects to the deployment solution (1) to perform a search using the discovery agent (2). The available physical servers are shown to the cloud consumer (3).

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini



# Advanced Cloud Architectures

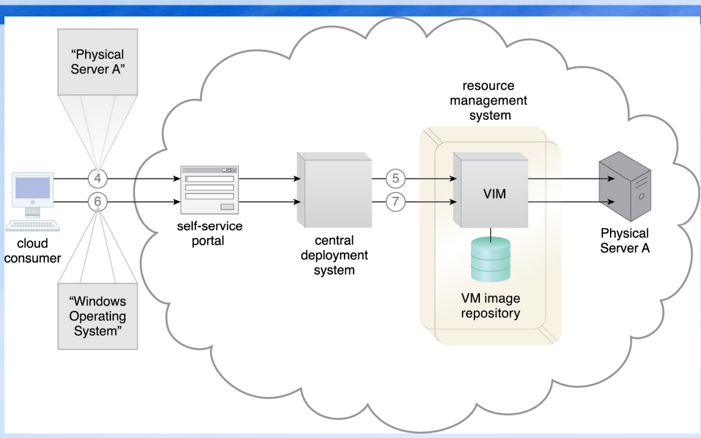



Figure 12.24 The cloud consumer selects a physical server to provision (4). The deployment agent is loaded to the physical server's RAM via the remote management system (5). The cloud consumer selects an operating system and method of configuration via the deployment solution (6). The operating system is installed and the server becomes operational (7).

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Advanced Cloud Architectures

- Rapid Provisioning Architecture: automates provisioning of a wide range of IT resources, either individually or collectively
  - Server Templates: virtual image files that automate creation of new virtual servers
  - Server Images: similar to templates, but used to provision physical servers
  - Application Packages: applications and other software for automated deployment
  - Application Packager: software for creating application packages
  - Custom Scripts: automate administrative tasks
  - Sequence Manager: organises a sequence of automated provisioning tasks
  - Sequence Logger: logs execution of automated provisioning sequences
  - Operating System/Application Configuration Baselines: configurations applied to operating systems and applications automatically after installation to prepare them for use
  - Deployment Data Store: repository of virtual images, templates, scripts, baseline configurations, and other related data.



# Advanced Cloud Architectures

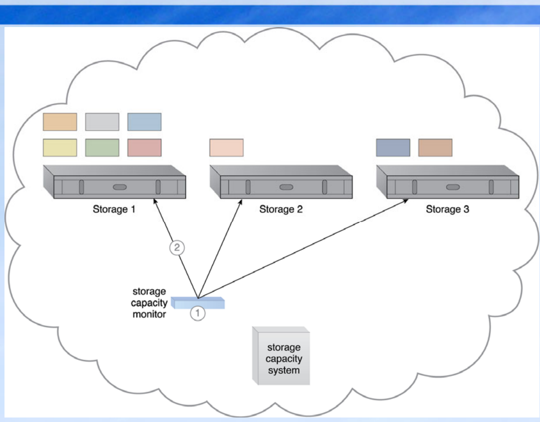



Figure 12.28 The storage capacity system and storage capacity monitor are configured to survey three storage devices in realtime, whose workload and capacity thresholds are pre-defined (1). The storage capacity monitor determines that the workload on Storage 1 is reaching its threshold (2).

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

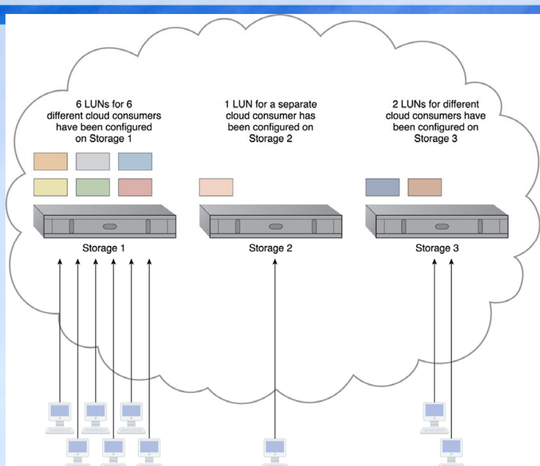
# Advanced Cloud Architectures

- Storage Workload Management Architecture: enables LUNs (virtual disk/partition) to be evenly distributed across available cloud storage devices
  - LUNs can be migrated (moved) between physical storage devices in different locations to balance the I/O load on the physical disks
  - Audit Monitor: checks for compliance with regulatory, privacy, and security requirements for physical data relocation
  - Automated Scaling Listener: monitors workload and responds to fluctuations
  - Cloud Usage Monitor: tracks LUN movements and collects workload distribution statistics
  - Load Balancer: used to horizontally balance workloads across cloud storage devices
  - Logical Network Perimeter: provides isolation of cloud consumer data during/after physical data relocation



# Advanced Cloud Architectures

Figure 12.26 An unbalanced cloud storage architecture has six storage LUNs in Storage 1 for cloud consumers to use, while Storage 2 is hosting one LUN and Storage 3 is hosting two. The majority of the workload ends up with Storage 1, since it is hosting the most LUNs.




Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini

# Advanced Cloud Architectures

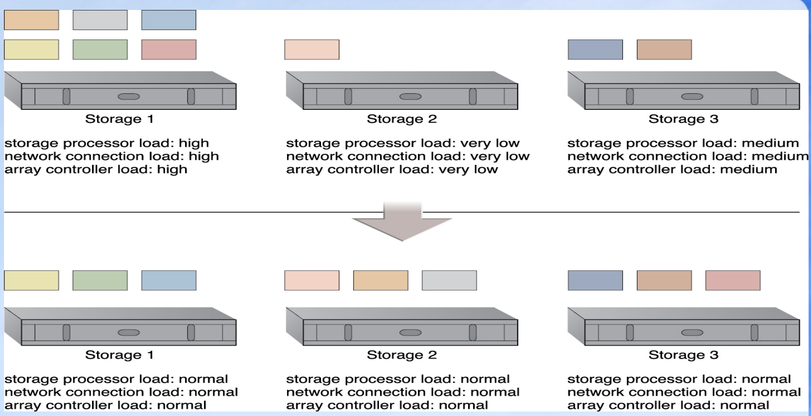



Figure 12.27 LUNs are dynamically distributed across cloud storage devices, resulting in more even distribution of associated types of workloads.

Source: *Cloud Computing* by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini




# Chapter 13
# Specialised Cloud Architectures

## Specialized Cloud Architectures

- Direct I/O Access Architecture: allows virtual servers to bypass the hypervisor to directly access physical hardware, providing improved performance
  - Virtual Server: requires appropriate drivers installed to access the physical I/O card
  - Cloud Usage Monitor: collects usage data
  - Logical Network Perimeter: ensures direct I/O access does not allow a cloud consumer to access the IT resources owned by another cloud consumer
  - Pay-Per-Use Monitor: collects cost information for the allocated physical I/O card
  - Resource Replication: used to replace virtual I/O cards with physical I/O cards



## Specialized Cloud Architectures

- Direct LUN Access Architecture: provides raw block-level access to storage LUNs (virtual disks/partitions) instead of file-based storage
    - Appears to virtual server as unformatted/un-partitioned storage
  - Cloud Usage Monitor: tracks and collects usage information
  - Pay-Per-Use Monitor: collects and classifies usage cost information
  - Resource Replication: relates to how block-based storage is accessed instead of file-based storage

# Specialized Cloud Architectures

- Dynamic Data Normalization Architecture: eliminates redundant data stored on block and file-based storage through use of a de-duplication system
  - Data stored in cloud is usually replicated to ensure reliable storage – no point replicating a copy/replica!



# Specialized Cloud Architectures

- Elastic Network Capacity Architecture: dynamic allocation of additional network bandwidth to avoid bottlenecks
  - Automated Scaling Listener and Intelligent Automation Engine: detect bandwidth use approaching a threshold and dynamically allocate bandwidth when required
  - Cloud Usage Monitor: track elastic network capacity before, during and after scaling
  - Hypervisor: provides access to the physical network
  - Logical Network Perimeter: establishes boundaries to provide allocated network capacity to cloud consumers
  - Pay-Per-Use Monitor: tracks billing-related data
  - Resource Replication: used to add network ports to physical and virtual servers in response to workload demands
  - Virtual Servers: host the IT resources and cloud services that are affected by scaling network capacity

# Specialized Cloud Architectures

- Cross-Storage Device Vertical Tiering Architecture: supports vertical scaling of storage between devices with different storage capacities
    - Example: vertical scaling from disk-based to SSD-based storage for improved performance
  - Automated Scaling Listener: monitors requests and signals storage management program to move LUNs to higher capacity devices once a predefined threshold reached
  - Audit Monitor: ensures data relocation aligns with any legal or data privacy regulations and policies
  - Cloud Usage Monitor: records data transfer and usage at both source and destination storage locations
  - Pay-Per-Use Monitor: collects storage usage information at both source and destination locations, and IT resource usage information for cross-storage tiering functionality



# Specialized Cloud Architectures

- Intra-Storage Device Vertical Data Tiering Architecture: supports vertical scaling within a single cloud storage device for data subject to security, privacy, or various legal constraints
    - Uses a complex storage device supporting different types of hard disks, e.g., SATA, SAS, and SSDs organised into graded tiers
  - Automated Scaling Listener: monitors runtime data processing traffic and moves LUNs to higher or lower grade storage when certain thresholds are met

# Specialized Cloud Architectures

- Load Balanced Virtual Switches Architecture: balance network traffic across multiple uplinks or redundant paths to avoid slow transfers and data loss
  - Cloud Usage Monitor: monitors network traffic and bandwidth usage
  - Hypervisor: hosts virtual servers providing access to virtual switches and external network
  - Load Balancer: distributes network workload across different uplinks
  - Logical Network Perimeter: creates boundaries that protect and limit bandwidth usage for each cloud consumer
  - Resource Replication: generates additional uplinks to the virtual switch
  - Virtual Server: host IT resources that benefit from additional uplinks and bandwidth



# Specialized Cloud Architectures

- Multipath Resource Access Architecture: establishes alternative paths to IT resources to allow cloud consumers to programmatically or manually overcome path failures
  - Cloud Storage Device: requires alternative paths to remain accessible so solutions that rely on data access
  - Hypervisor: requires alternative paths to have redundant links to hosted virtual servers
  - Logical Network Perimeter: maintains cloud consumer privacy even when multiple paths are created
  - Resource Replication: required when a new instance of an IT resource needs to be created to generate the alternative path
  - Virtual Server: host IT resources that have multipath access, hypervisors can provide multipath access to the virtual servers

# Specialized Cloud Architectures

- Persistent Virtual Network Configuration Architecture: centralised network configuration information is replicated to physical server hosts to allow migrated virtual machines to maintain network connectivity
  - Hypervisor: hosts virtual servers with replicated configuration settings
  - Logical Network Perimeter: ensures access to the virtual server and its IT resources is isolated to the rightful cloud consumer throughout migration
  - Resource Replication: replicates virtual switch configurations and network capacity information across hypervisors



# Specialized Cloud Architectures

- Redundant Physical Connection for Virtual Servers Architecture: uses one or more redundant network uplink connections to ensure virtual servers maintain connectivity
  - Failover System: transitions unavailable uplinks to standby uplinks
  - Hypervisor: hosts virtual servers and some virtual switches, providing virtual networks and virtual switches with access to virtual servers
  - Logical Network Perimeter: ensure allocated/defined virtual switches remain isolated for each cloud consumer
  - Resource Replication: replicates current status of active uplinks to standby uplinks to maintain network connectivity

# Specialized Cloud Architectures

- Storage Maintenance Window Architecture: enables cloud service consumers to be automatically and transparently redirected to secondary cloud storage devices to allow for maintenance to primary storage devices
  - Resource Replication: keeps primary and secondary storage devices synchronised



# Summary

- Week 6. Cloud Architectures
  - Fundamental Cloud Architectures
  - Advanced Cloud Architectures
  - Specialised Cloud Architectures