CS 4037
Introduction to Cloud Computing
Lecture 11

Danyal Farhat

FAST School of Computing

NUCES Lahore

Fundamental Cloud Architectures

Lecture's Agenda

- **Workload Distribution Architecture**
- Resource Pooling Architecture
- Dynamic Scalability Architecture
- Elastic Resource Capacity Architecture
- Service Load Balancing Architecture
- Cloud Bursting Architecture
- Elastic Disk Provisioning Architecture

Workload Distribution Architecture

The workload distribution architecture reduces both IT resource overutilization and under-utilization to an extent dependent upon the sophistication of the load balancing algorithms and runtime logic.

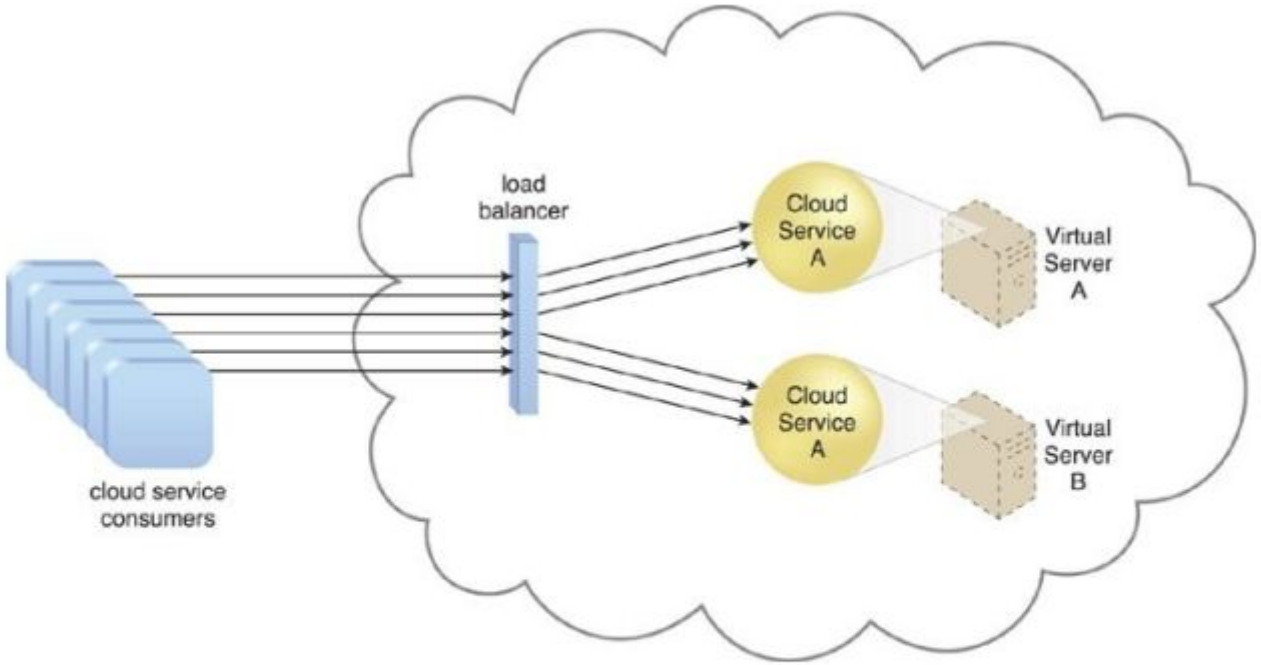


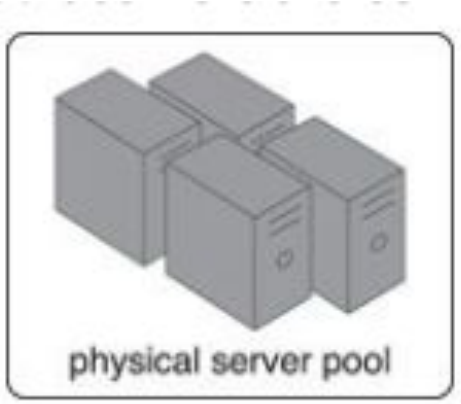
Figure 11.1. A redundant copy of Cloud Service A is implemented on Virtual Server B. The load balancer intercepts cloud service consumer requests and directs them to both Virtual Servers A and B to ensure even workload distribution.

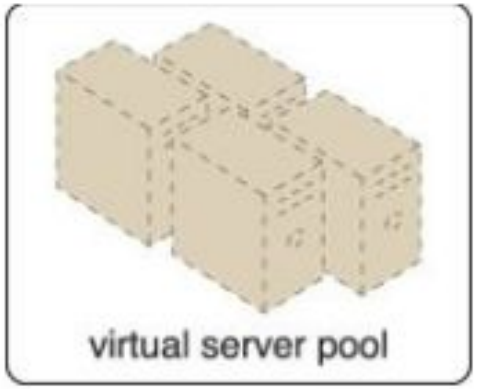
Lecture's Agenda

- Workload Distribution Architecture
- **Resource Pooling Architecture**
- Dynamic Scalability Architecture
- Elastic Resource Capacity Architecture
- Service Load Balancing Architecture
- Cloud Bursting Architecture
- Elastic Disk Provisioning Architecture

Resource Pooling Architecture

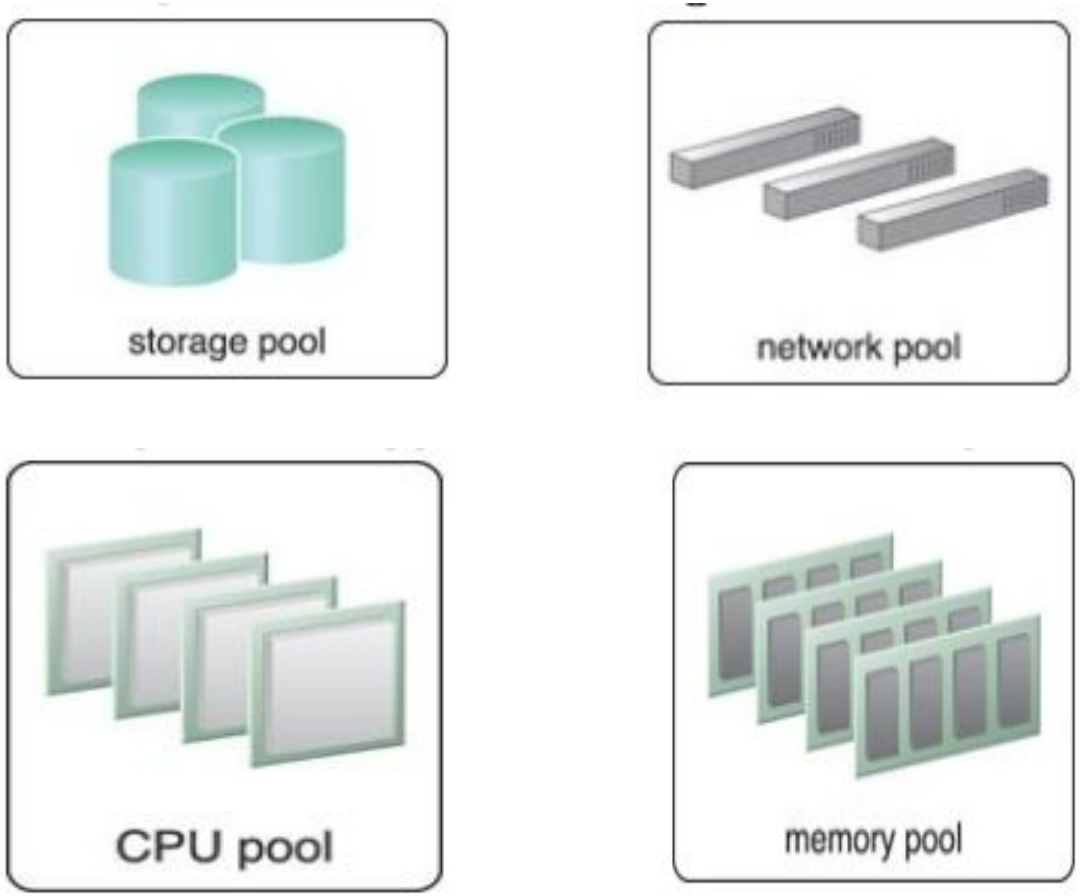
- A resource pooling architecture is based on the use of one or more resource pools, in which identical IT resources are grouped and maintained by a system that automatically ensures that they remain synchronized.
- Common examples of resource pools:





Resource Pooling Architecture (Cont.)

- **Common examples of resource pools:**



Resource Pooling Architecture (Cont.)

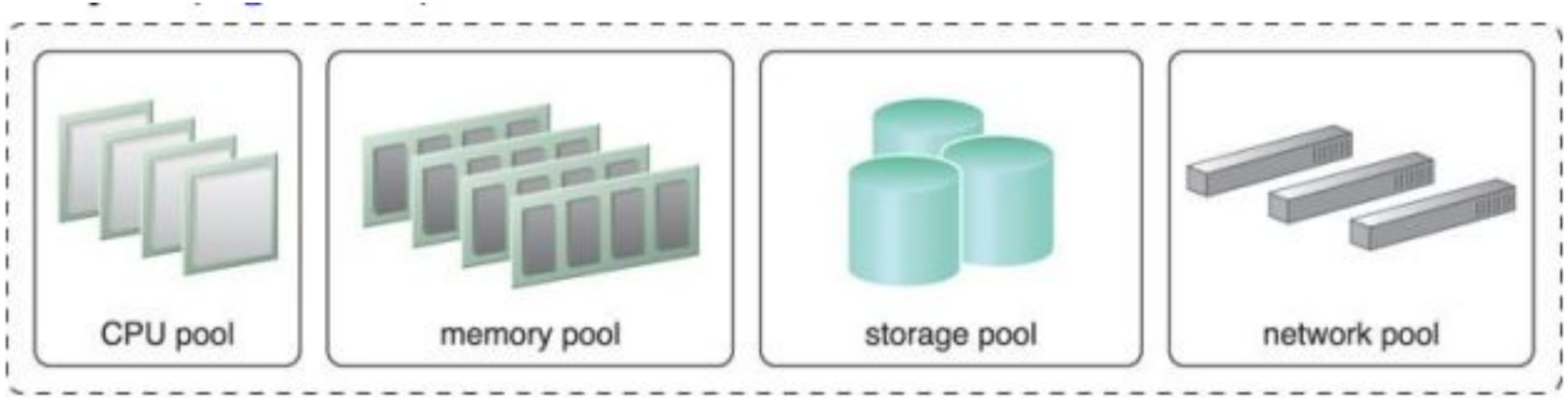


Figure 11.2. A sample resource pool that is comprised of four sub-pools of CPUs, memory, cloud storage devices, and virtual network devices.

Resource Pooling Architecture (Cont.)

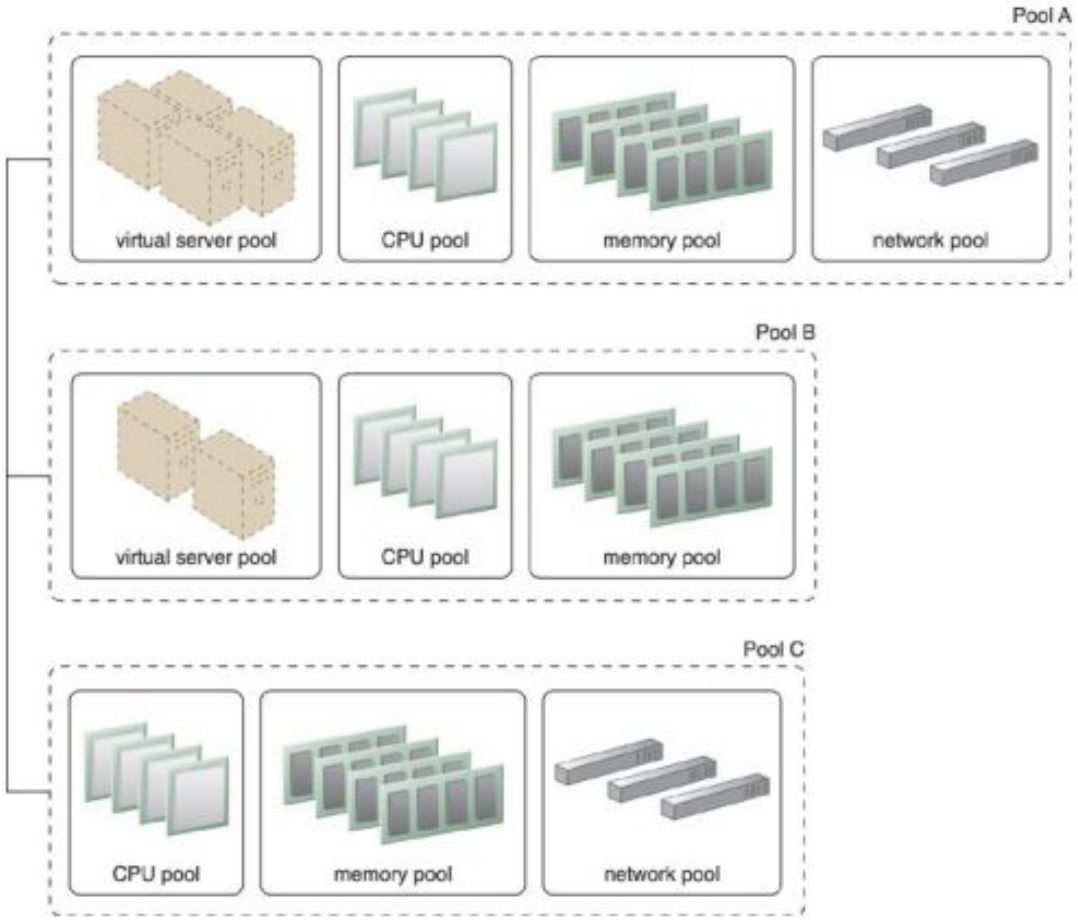


Figure 11.3. Pools B and C are sibling pools that are taken from the larger Pool A, which has been allocated to a cloud consumer. This is an alternative to taking the IT resources for Pool B and Pool C from a general reserve of IT resources that is shared throughout the cloud.

Lecture's Agenda

- Workload Distribution Architecture
- Resource Pooling Architecture
- **Dynamic Scalability Architecture**
- Elastic Resource Capacity Architecture
- Service Load Balancing Architecture
- Cloud Bursting Architecture
- Elastic Disk Provisioning Architecture

Dynamic Scalability Architecture

- The 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 allocation enables variable utilization as dictated by usage demand fluctuations, since unnecessary IT resources are efficiently reclaimed without requiring manual interaction.

Dynamic Scalability Architecture (Cont.)

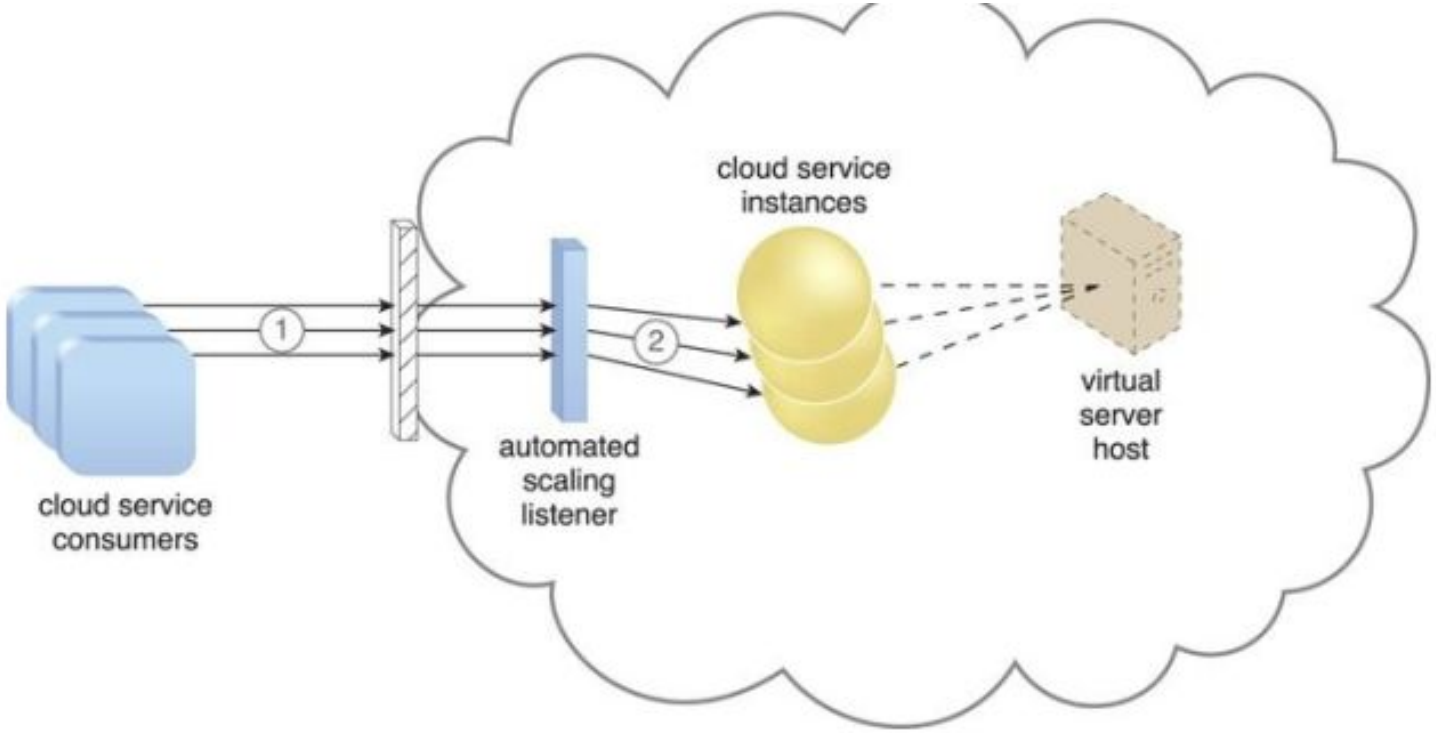


Figure 11.5. Cloud service consumers are sending requests to a cloud service (1). The automated scaling listener monitors the cloud service to determine if predefined capacity thresholds are being exceeded (2).

Dynamic Scalability Architecture (Cont.)

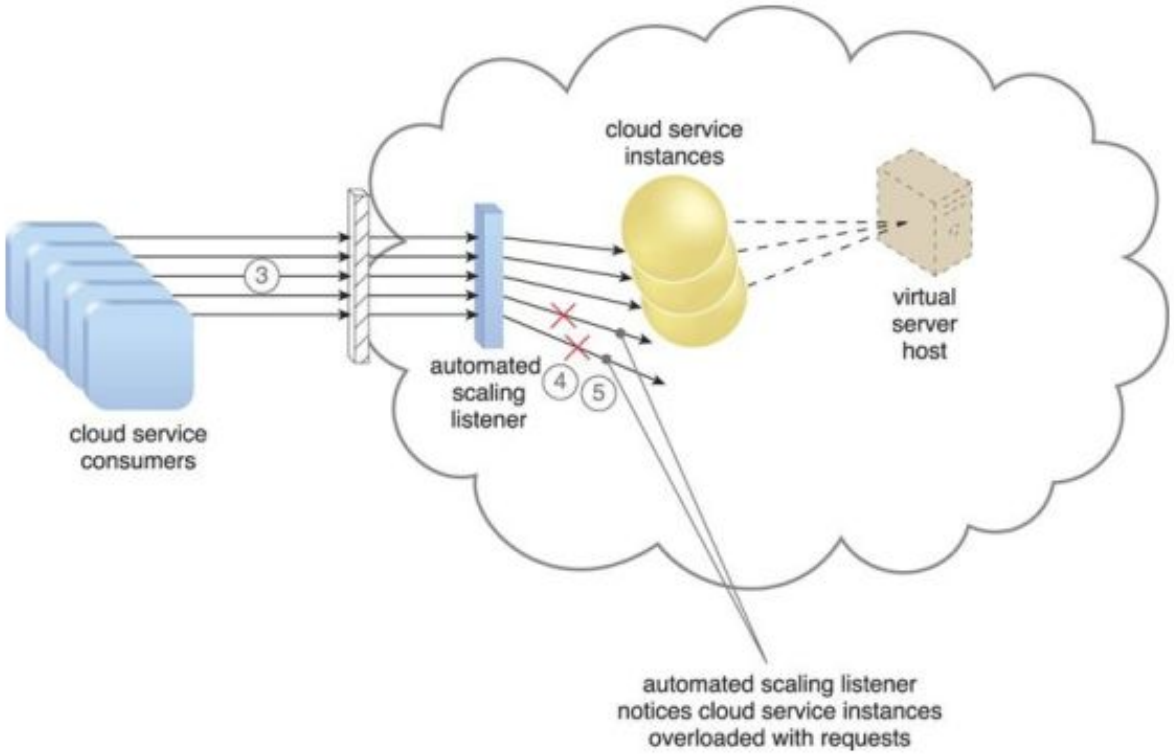


Figure 11.6. The number of requests coming from cloud service consumers increases (3). The workload exceeds the performance thresholds. The automated scaling listener determines the next course of action based on a predefined scaling policy (4). If the cloud service implementation is deemed eligible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

Dynamic Scalability Architecture (Cont.)

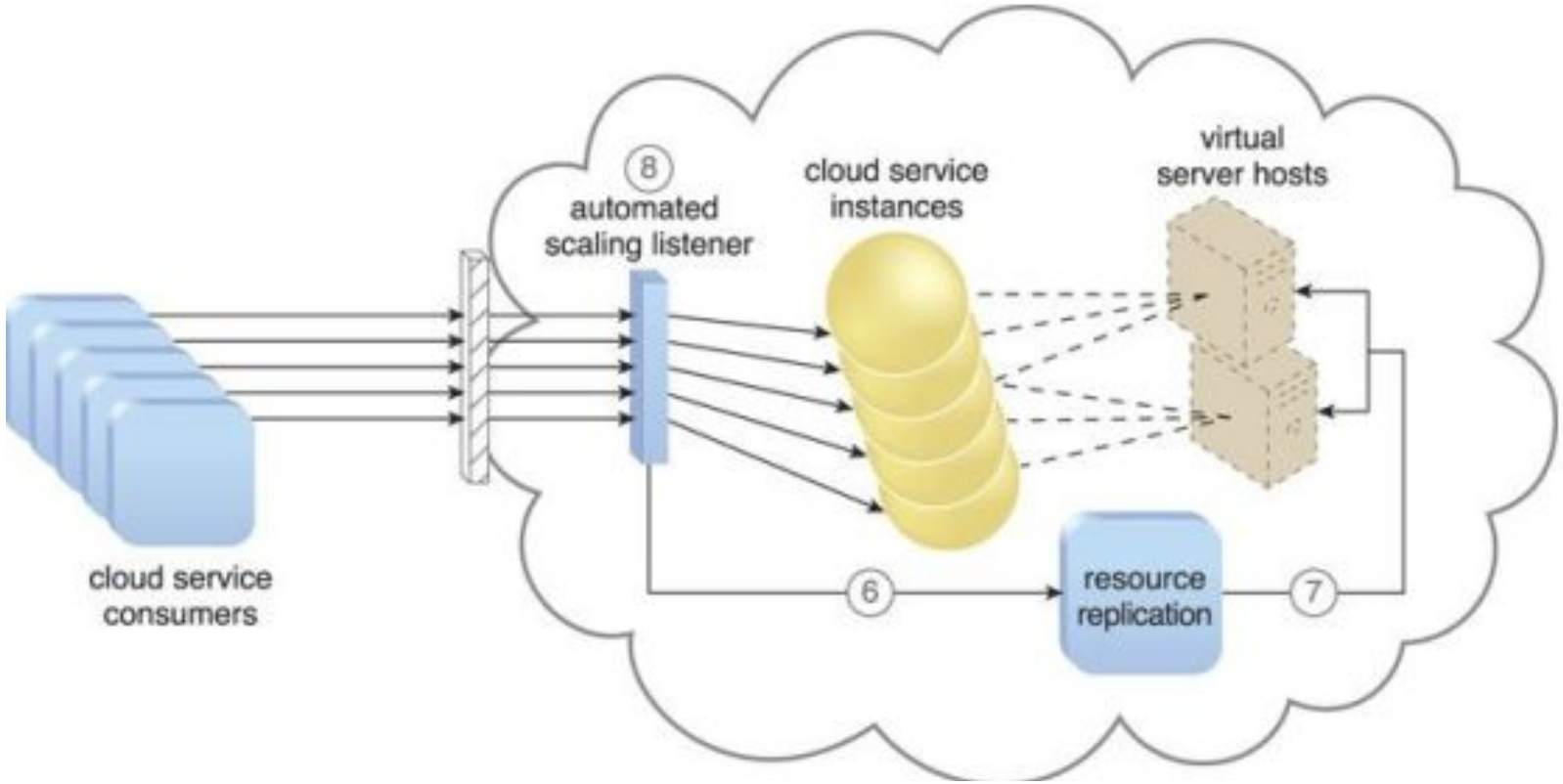


Figure 11.7. The automated scaling listener sends a signal to the resource replication mechanism (6), which creates more instances of the cloud service (7). Now that the increased workload has been accommodated, the automated scaling listener resumes monitoring and detracting and adding IT resources,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

Lecture's Agenda

- Workload Distribution Architecture
- Resource Pooling Architecture
- Dynamic Scalability Architecture
- **Elastic Resource Capacity Architecture**
- Service Load Balancing Architecture
- Cloud Bursting Architecture
- Elastic Disk Provisioning Architecture

Elastic Resource Capacity Architecture

- The elastic resource capacity architecture is primarily related to the dynamic provisioning of virtual servers, using a system that allocates and reclaims CPUs and RAM in immediate response to the fluctuating processing requirements of hosted IT resources.

Elastic Resource Capacity Architecture (Cont.)

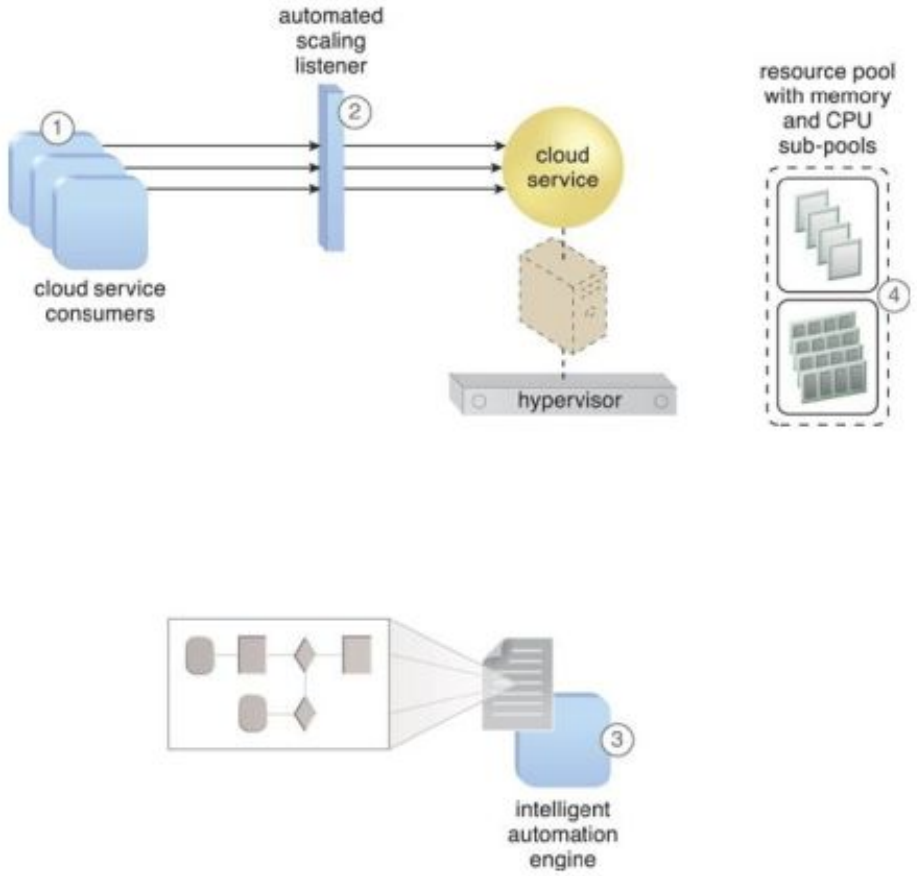


Figure 11.8. Cloud service consumers are actively sending requests to a cloud service (1), which are monitored by an automated scaling listener (2). An intelligent automation engine script is deployed with workflow logic (3) that is capable of notifying the resource pool using allocation requests (4).

Elastic Resource Capacity Architecture (Cont.)

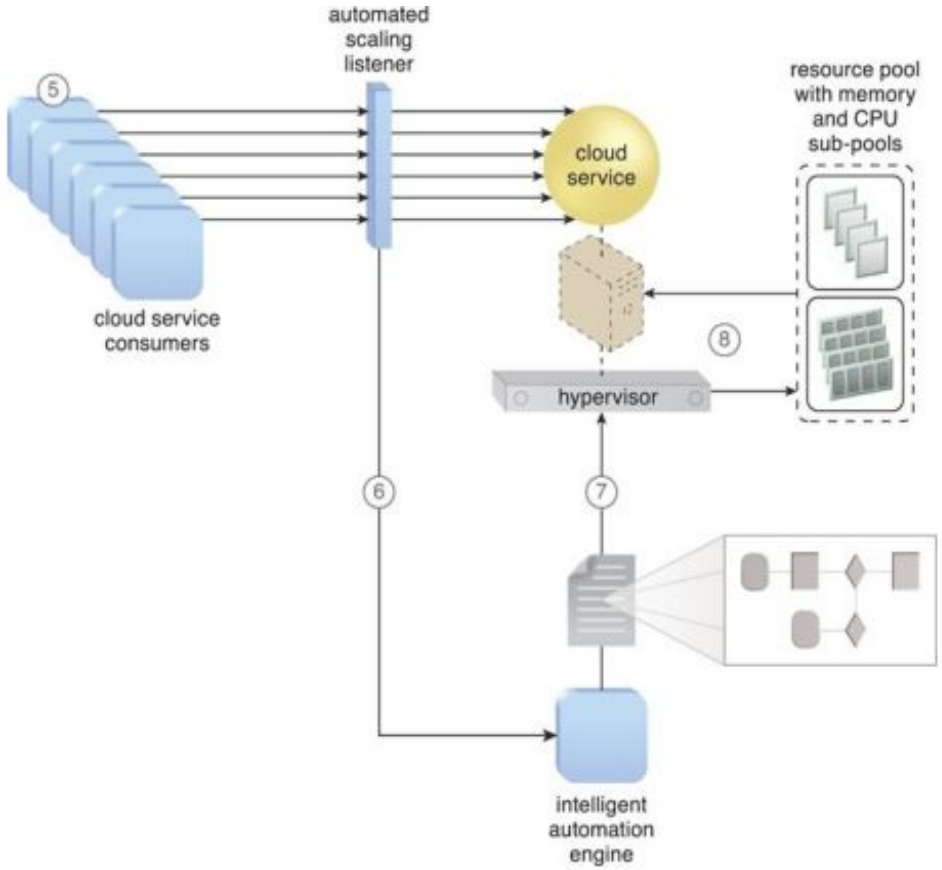


Figure 11.9. Cloud service consumer requests increase (5), causing the automated scaling listener to signal the intelligent automation engine to execute the script (6). The script runs the workflow logic that signals the hypervisor to allocate more IT resources from the resource pools (7). The hypervisor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

Lecture's Agenda

- Workload Distribution Architecture
- Resource Pooling Architecture
- Dynamic Scalability Architecture
- Elastic Resource Capacity Architecture
- **Service Load Balancing Architecture**
- Cloud Bursting Architecture
- Elastic Disk Provisioning Architecture

Service Load Balancing Architecture

- The service load balancing architecture can be considered a specialized variation of the workload distribution architecture that is geared specifically for scaling cloud service implementations.

- Redundant deployments of cloud services are created, with a load balancing system added to dynamically distribute workloads.

Service Load Balancing Architecture (Cont.)

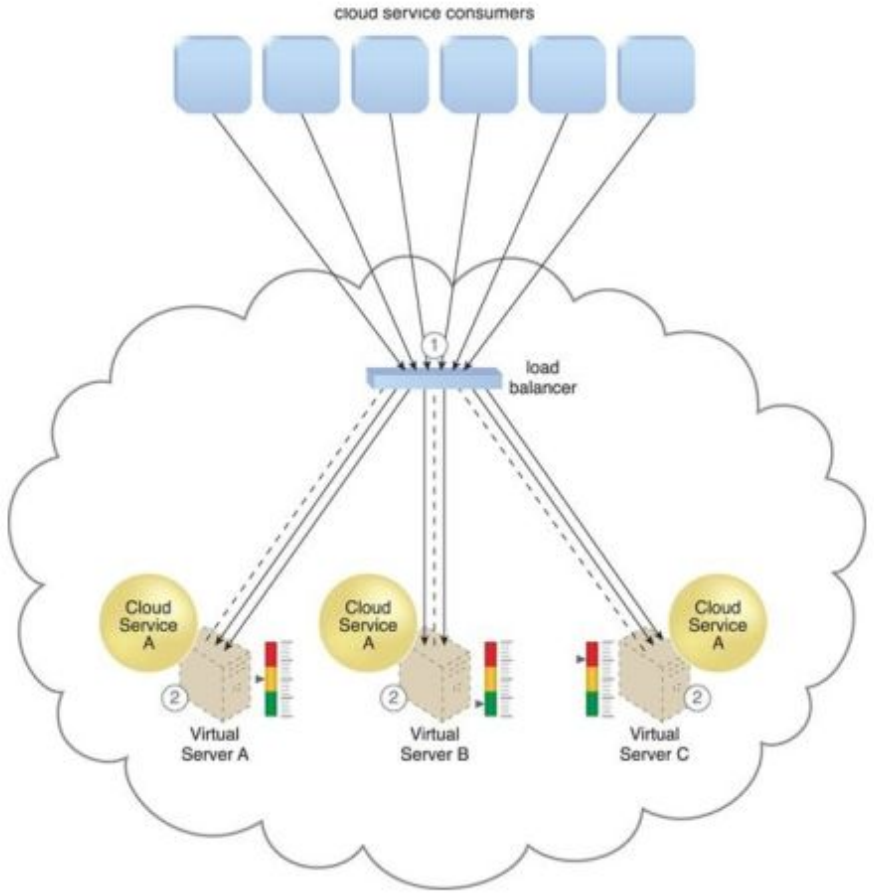


Figure 11.10. The load balancer intercepts messages sent by cloud service consumers (1) and forwards them to the virtual servers so that the workload processing is horizontally scaled (2).

Service Load Balancing Architecture (Cont.)

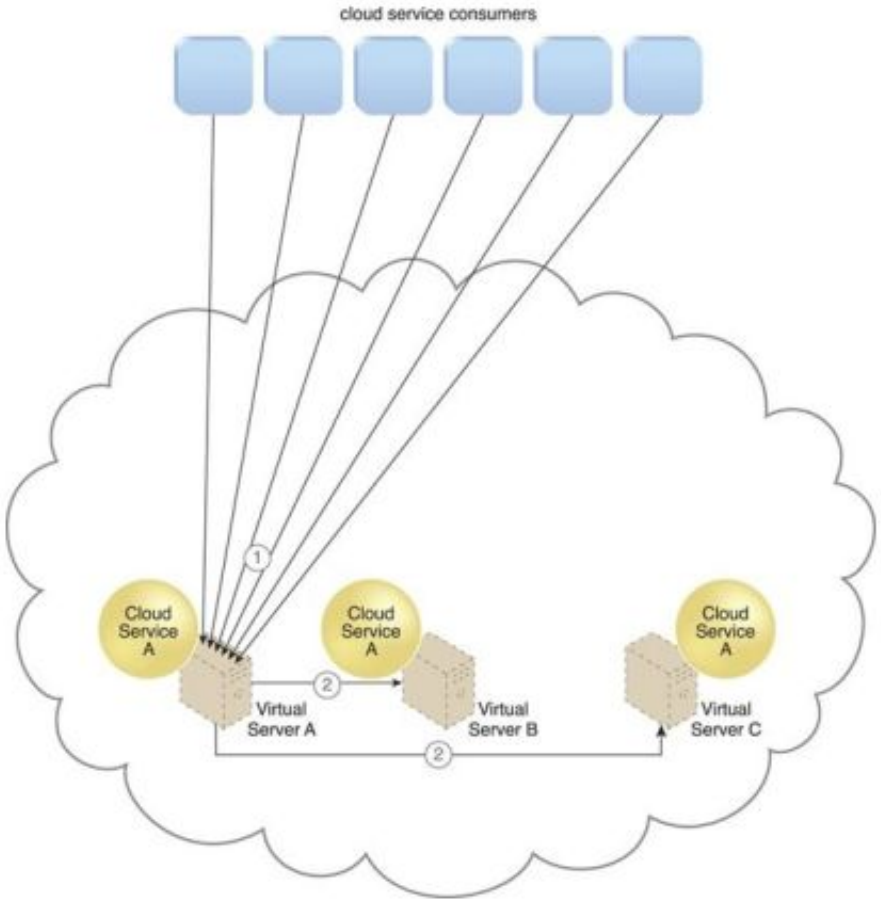


Figure 11.11. Cloud service consumer requests are sent to Cloud Service A on Virtual Server A (1). The cloud service implementation includes built-in load balancing logic that is capable of distributing requests to the neighboring Cloud Service A implementations on Virtual Servers B and C (2).

Lecture's Agenda

- Workload Distribution Architecture
- Resource Pooling Architecture
- Dynamic Scalability Architecture
- Elastic Resource Capacity Architecture
- Service Load Balancing Architecture
- **Cloud Bursting Architecture**
- Elastic Disk Provisioning Architecture

Cloud Bursting Architecture

- The cloud bursting architecture establishes a form of dynamic scaling that scales or "bursts out" on-premise IT resources into a cloud whenever predefined capacity thresholds have been reached.

- The corresponding cloud-based IT resources are redundantly pre-deployed but remain inactive until cloud bursting occurs.

- After they are no longer required, the cloud-based IT resources are released and the architecture "bursts in" back to the on-premise environment.

Cloud Bursting Architecture (Cont.)

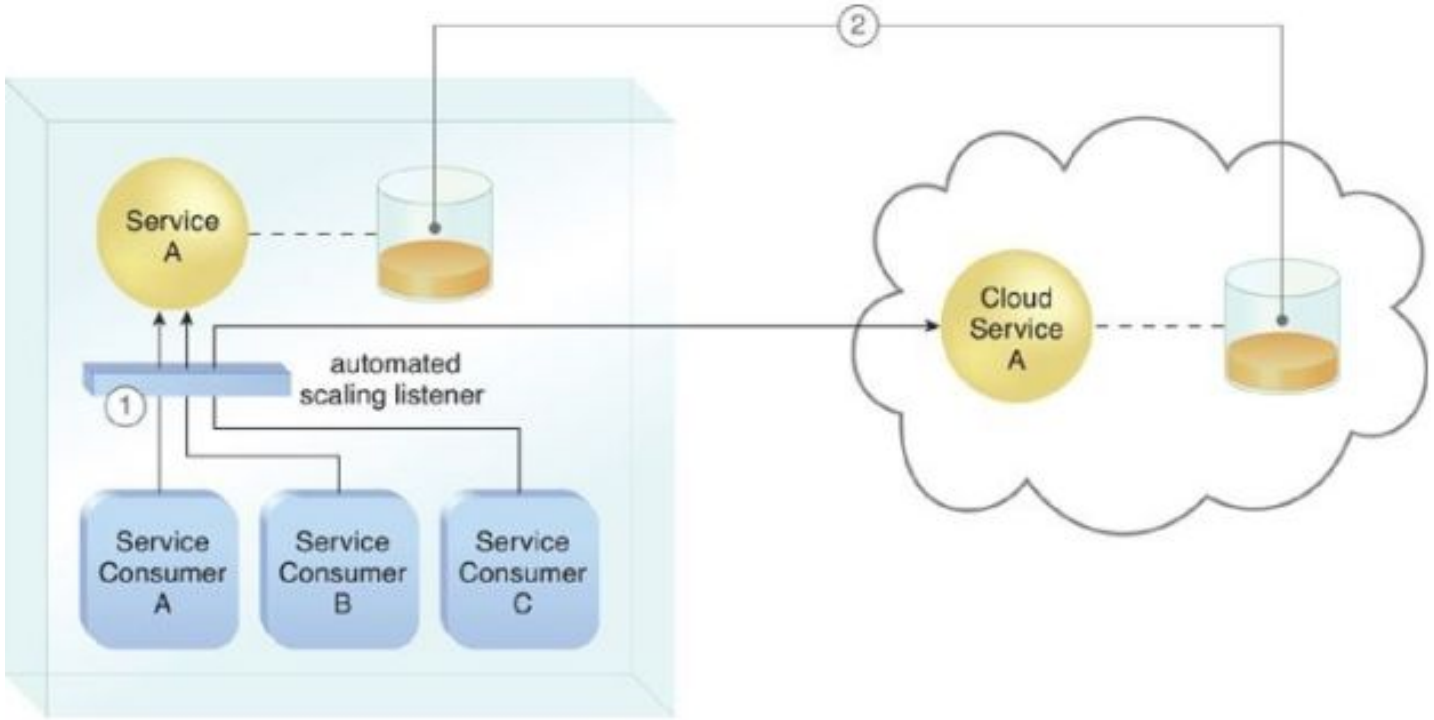


Figure 11.12. An automated scaling listener monitors the usage of on-premise Service A, and redirects Service Consumer C's request to Service A's redundant implementation in the cloud (Cloud Service A) once Service A's usage threshold has been exceeded (1). A resource replication system is used to keep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

Lecture's Agenda

- Workload Distribution Architecture
- Resource Pooling Architecture
- Dynamic Scalability Architecture
- Elastic Resource Capacity Architecture
- Service Load Balancing Architecture
- Cloud Bursting Architecture
- **Elastic Disk Provisioning Architecture**

Elastic Disk Provisioning Architecture

- The elastic disk provisioning architecture establishes a dynamic storage provisioning system that ensures that the cloud consumer is granularly billed for the exact amount of storage that it actually uses.

- This system uses thin-provisioning technology for the dynamic allocation of storage space, and is further supported by runtime usage monitoring to collect accurate usage data for billing purposes.

Elastic Disk Provisioning Architecture (Cont.)

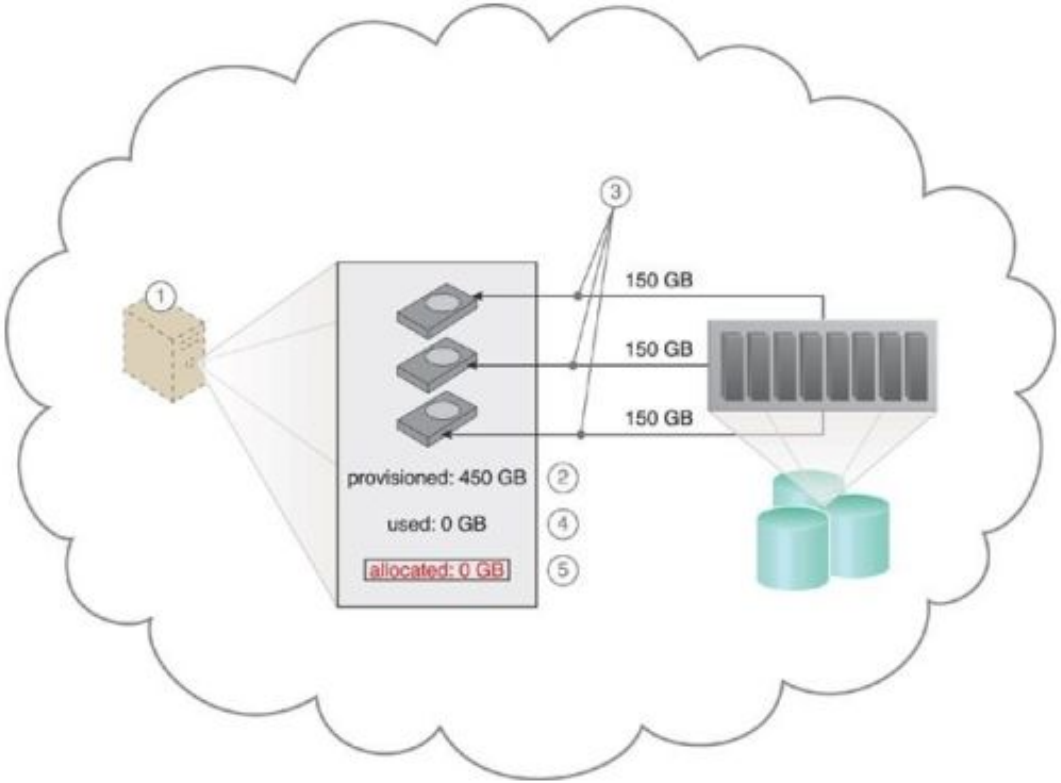


Figure 11.14. The cloud consumer requests a virtual server with three hard disks, each with a capacity of 150 GB (1). The virtual server is provisioned by this architecture with a total of 450 GB of disk space (2). The 450 GB are set as the maximum disk usage that is allowed for this virtual server, 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).

Elastic Disk Provisioning Architecture (Cont.)

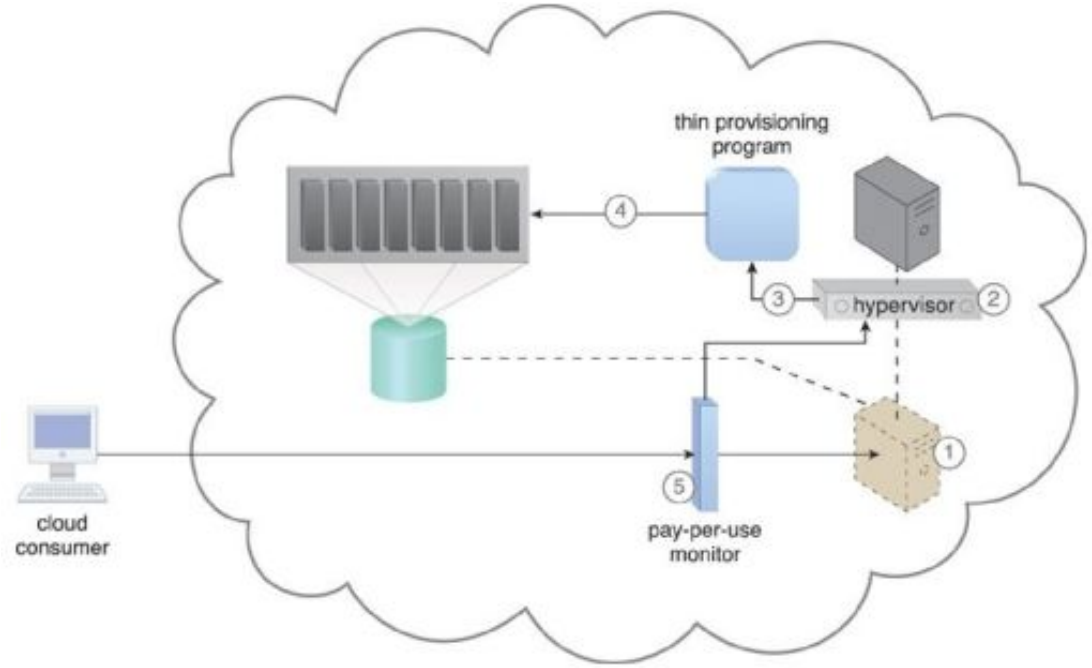


Figure 11.15. A request is received from a cloud consumer, and the provisioning of a new virtual server instance begins (1). As part of the provisioning process, the hard disks are chosen as dynamic or thin-provisioned disks (2). The hypervisor calls a dynamic disk allocation component to create thin disks for the virtual server (3). Virtual server disks are created via the thin-provisioning program and saved in a folder of near-zero size. The size of this folder and its files grow as operating applications are installed and additional files are copied onto the virtual server (4). The pay-per-use monitor tracks the actual dynamically allocated storage for billing purposes (5).

Additional Resources

- **Cloud Computing – Concepts, Technology, and Architecture by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini**

  - **Chapter 11: Fundamental Cloud Architectures**

Questions?