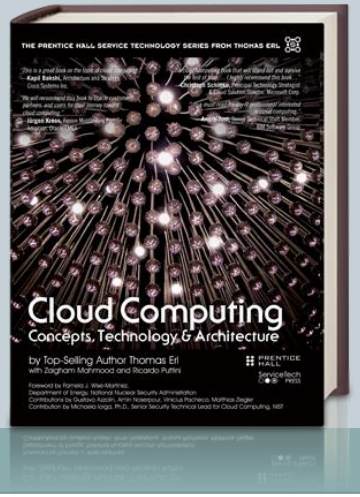
Cloud Computing

Concept, Technology & Architecture

Chapter 11
Cloud Computing Architecture

課程名稱：雲端管理系統

授課教師：高勝助



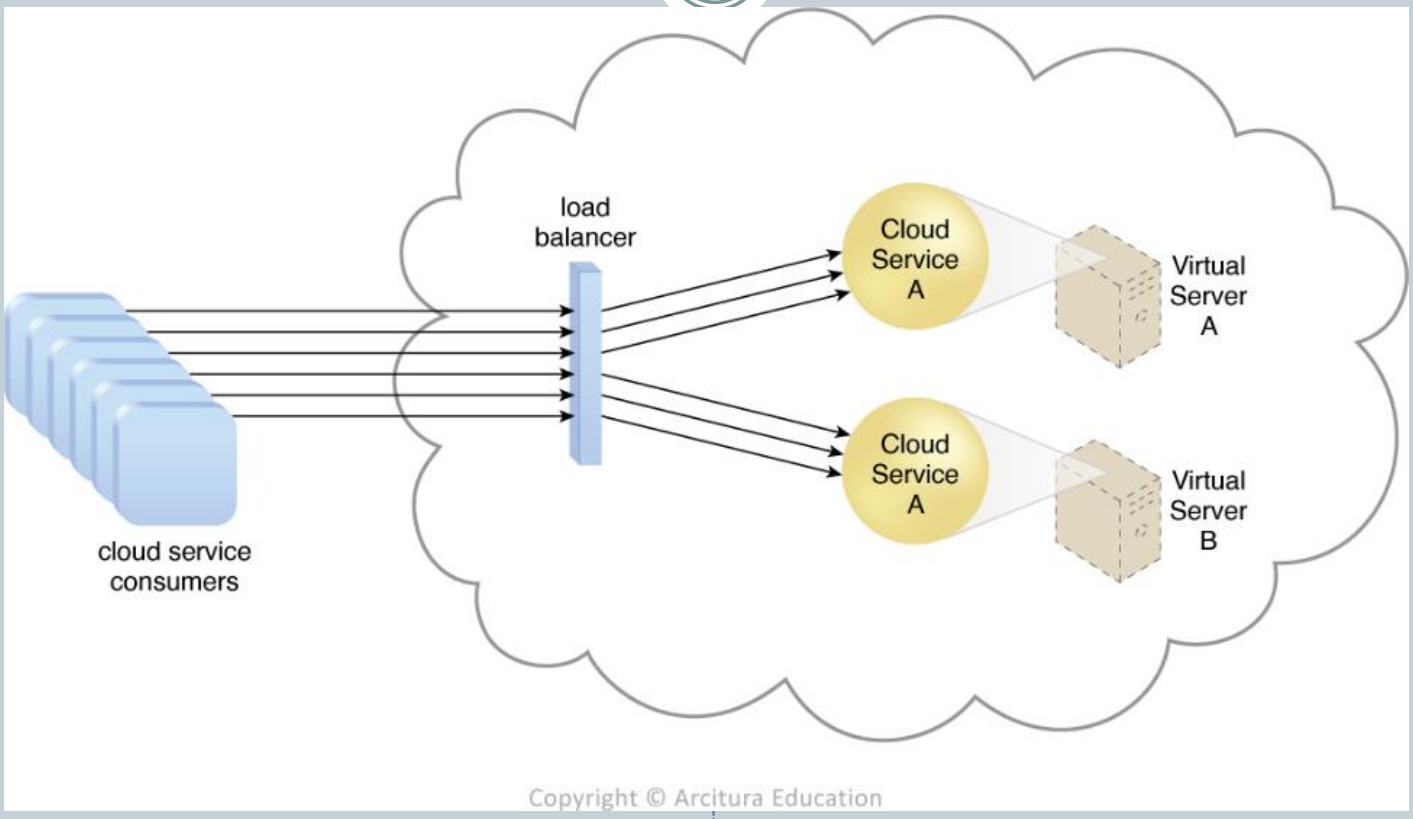
Contents

- Fundamental cloud architectural models establish baseline functions and capabilities. There are totally 29 architectures covered in this book.
  - 11.1 Workload Distribution Architecture
  - 11.2 Resource Pooling Architecture
  - 11.3 Dynamic Scalability Architecture
  - 11.4 Elastic Resource Capacity Architecture
  - 11.5 Service Load Balancing Architecture
  - 11.6 Cloud Bursting Architecture
  - 11.7 Elastic Disk Provisioning Architecture
  - 11.8 Redundant Storage Architecture
  - 11.9 Case Study Example

11.1 Workload Distribution Architecture (1/2)

- IT resources are horizontally scaled via the addition of one or more identical IT resources, and a load balancer that provides runtime logic capable of evenly distributing the workload among the available IT resources.
- The resulting workload distribution architecture reduces both IT resource over-utilization and under-utilization to an extent dependent upon the sophistication of the load balancing algorithms and runtime logic.

Figure 11.1



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

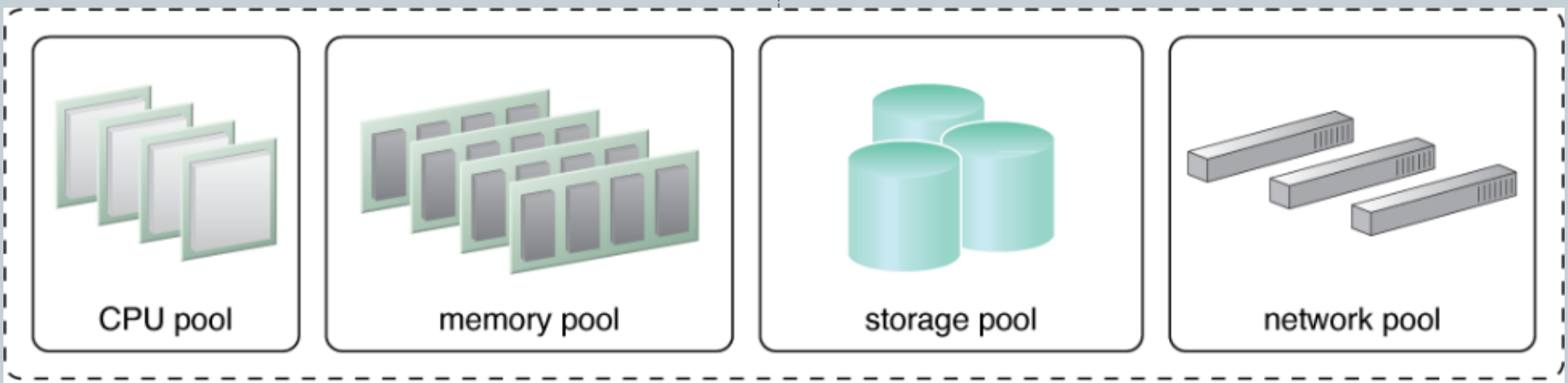
11.1 Workload Distribution Architecture (2/2)

- Workload distribution commonly carried out in support of distributed virtual servers, cloud storage devices, and cloud services.
- The following mechanisms can also be part of workload distribution architecture:
  - Audit Monitor
  - Cloud Usage Monitor
  - Hypervisor
  - Logical Network Perimeter
  - Resource Cluster
  - Resource Replication

11.2 Resource Pooling Architecture (1/3)

- A resource pooling architecture is based on the use of one or more resource pools, in which identical IT resources are grouped and maintained be a system that automatically ensures that they remain synchronized.
- Common examples of resource pools:
  - physical server pool
  - virtual server pool
  - storage pool
  - network pool
  - CPU pool
  - memory pool

Figure 11.2

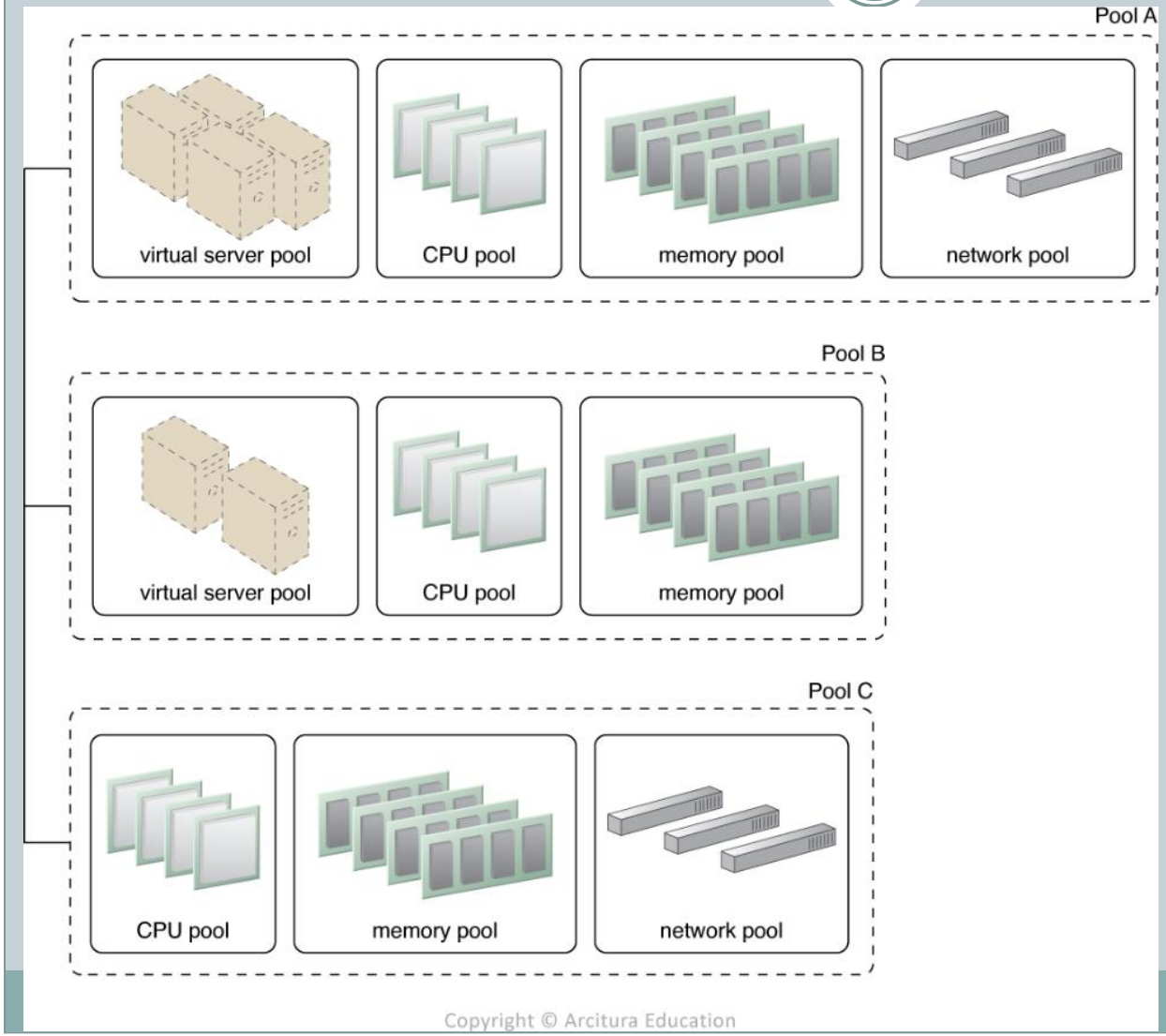


- Figure 11.2 - A sample resource pool that is comprised of four sub-pools of CPUs, memory, cloud storage devices, and virtual network devices.

11.2 Resource Pooling Architecture (2/3)

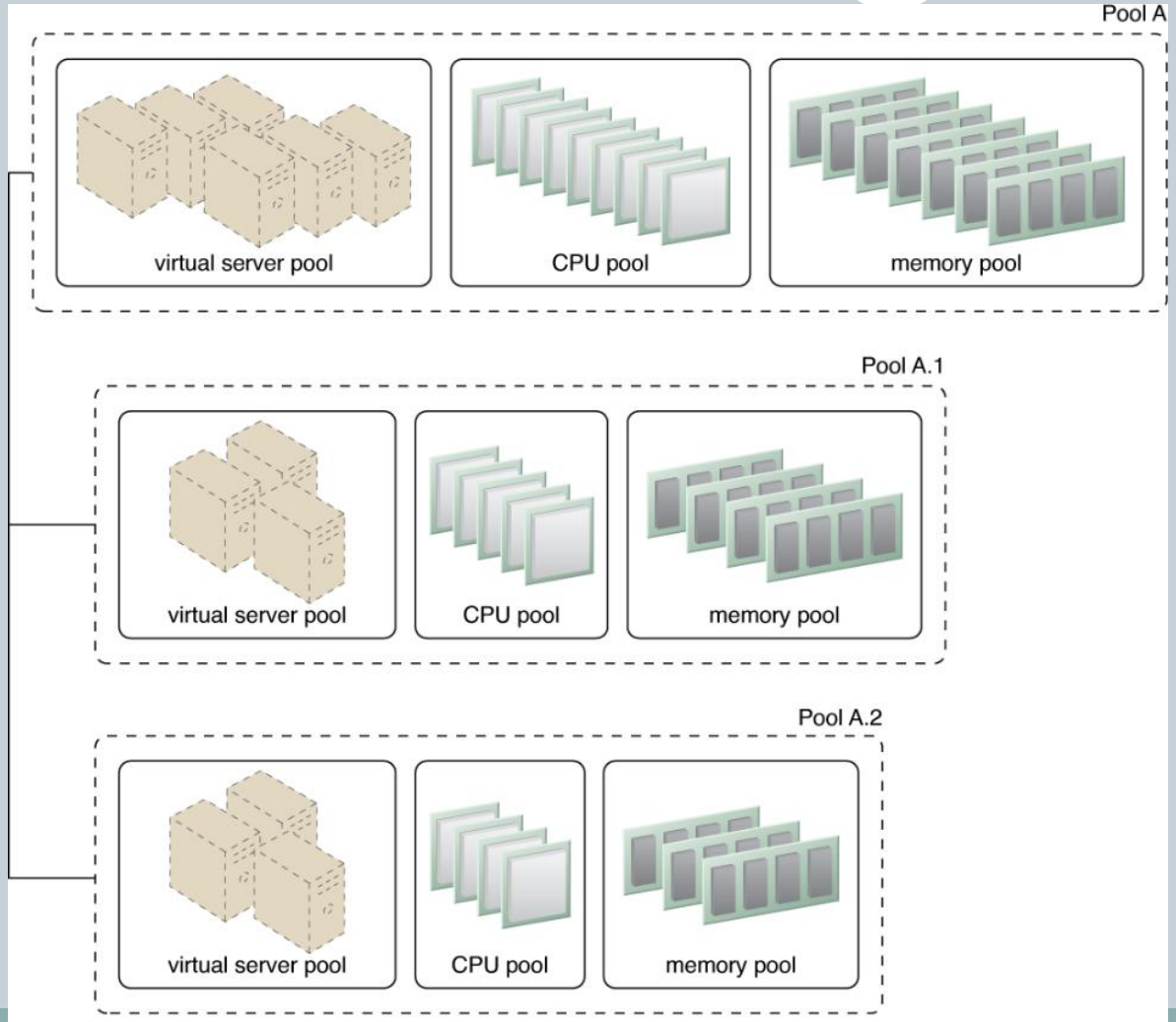
- A hierarchical structure can be established to form parent, sibling, and nested pools in order to facilitate the organization of diverse resource pooling requirement.
- Sibling pools are isolated from one another so that each cloud consumer is only provided access to its respective pool.
- Nested pools can be used to assign resource pools to different departments or groups in the same cloud consumer organization. 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.

Figure 11.3



- 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.

Figure 11.4



- 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.

11.2 Resource Pooling Architecture (3/3)

- In addition to commonly pooled mechanisms of cloud storage devices and virtual servers, the following mechanisms can be part of this resource pooling architecture:
  - Audit monitor
  - Cloud usage monitor
  - Hypervisor
  - Logical network perimeter
  - Pay-per-use monitor
  - Remote administration system
  - Resource management system
  - Resource replication

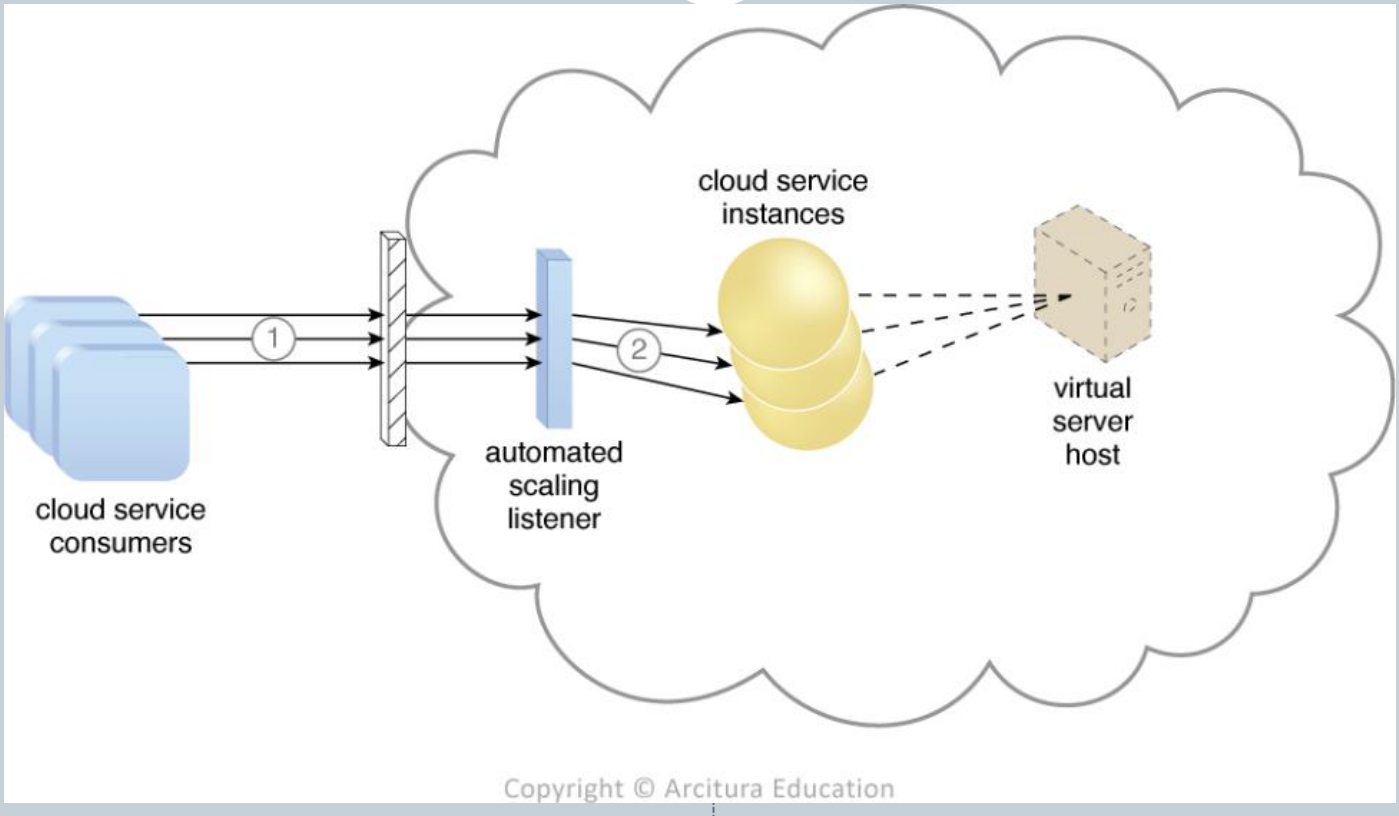
11.3 Dynamic Scalability Architecture (1/2)

- 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 depicted by usage demand fluctuations, science unnecessary IT resources are efficiently reclaimed without requiring manual interaction.

11.3 Dynamic Scalability Architecture (2/2)

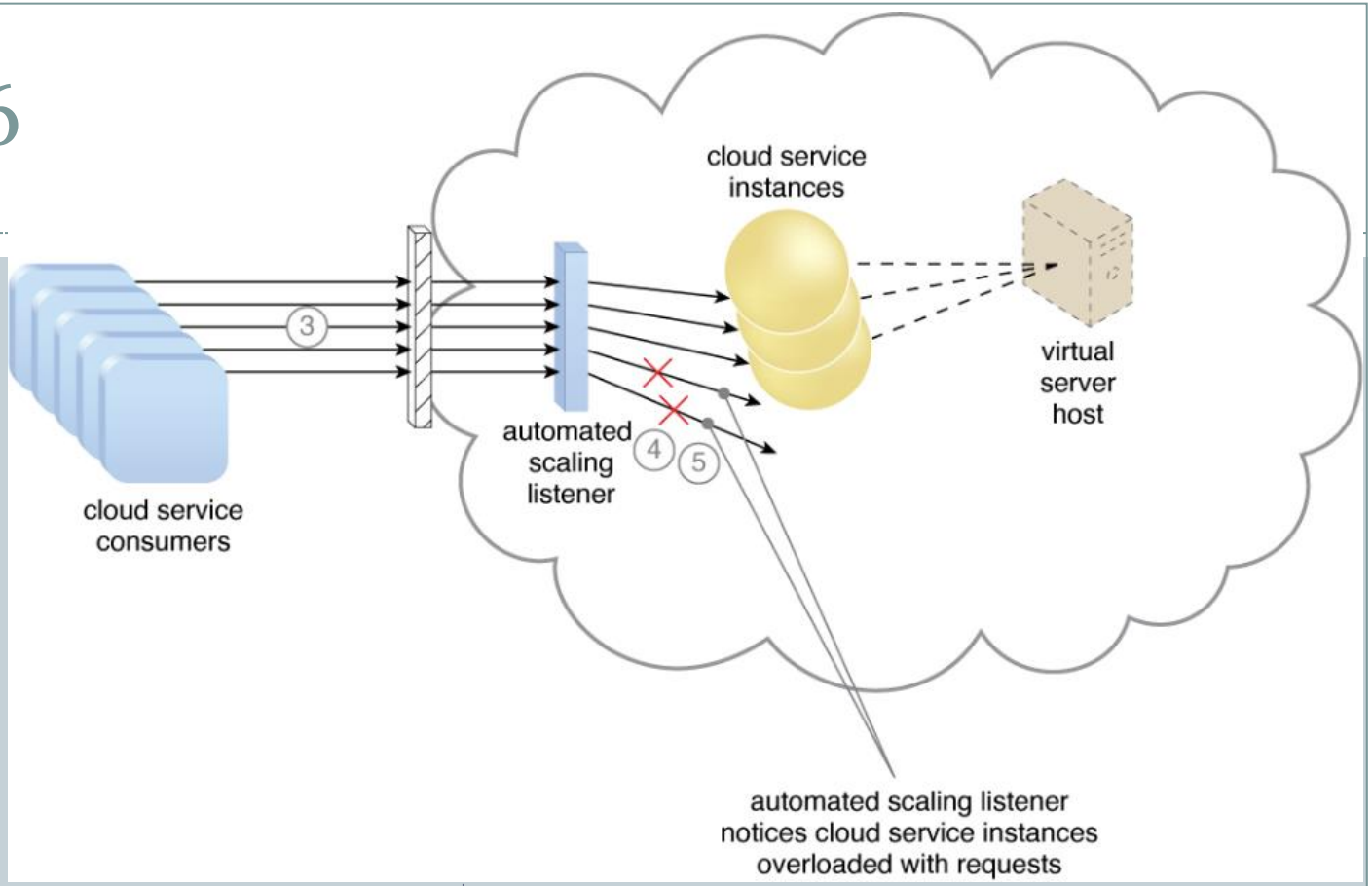
- The automated scaling listener is configured with workload thresholds that dictate when new IT resources need to be added to the workload processing.
- The following types of dynamic scaling are commonly used:
  - Dynamic Horizontal Scaling
  - Dynamic Vertical Scaling
  - Dynamic Relocation (resources are relocated with more capacity)

Figure 11.5



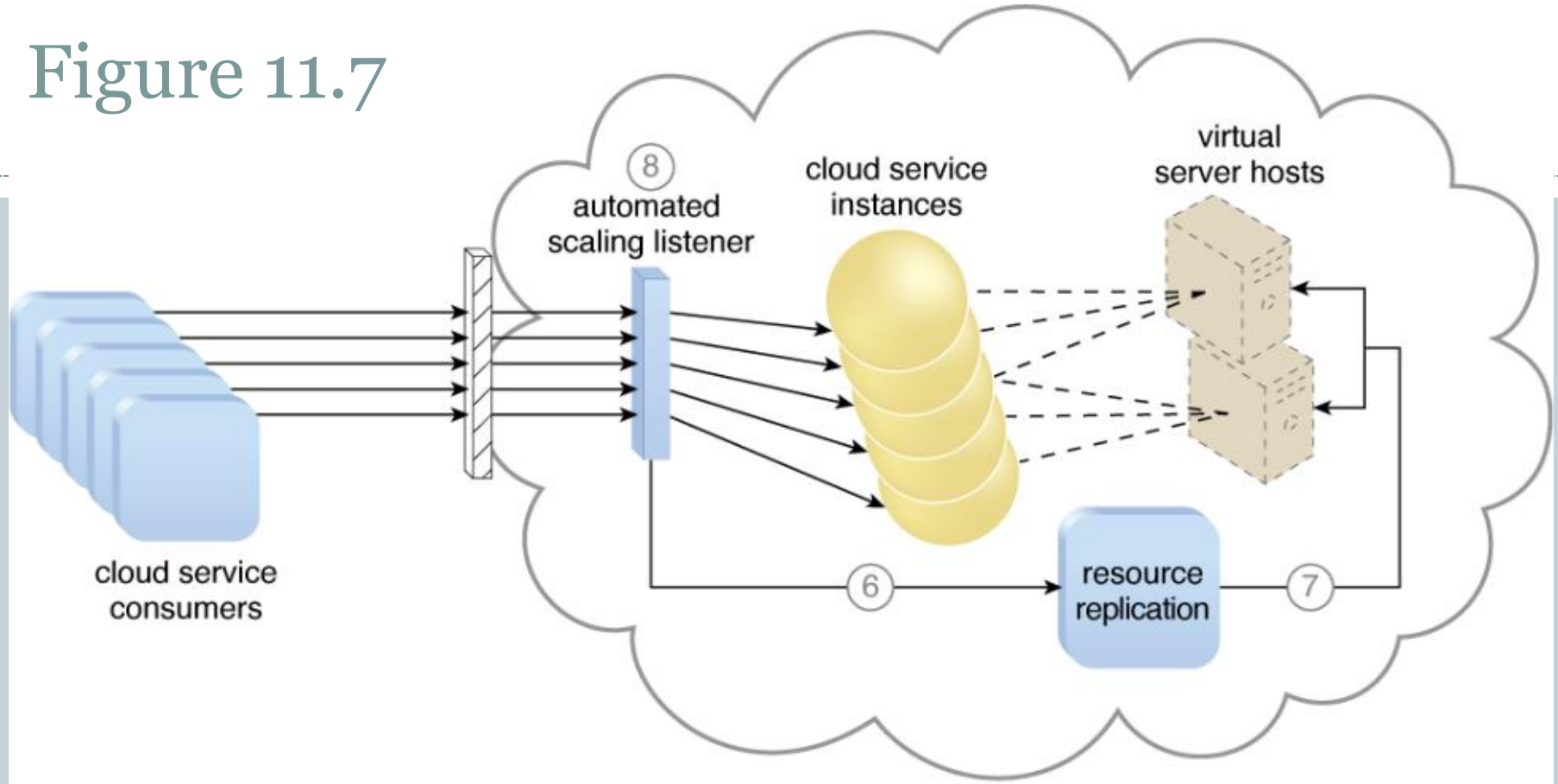
- Cloud service consumers are sending requests to a cloud service (1).
- The automated scaling listener monitors the cloud service to determine if predefined capacity thresholds are being exceeded (2).

Figure 11.6



- The number of service requests coming from cloud service consumers further 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).

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).

11.4 Elastic Resource Capacity Architecture (1/2)

- The elastic resource 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.
- The virtual server, its hosted applications, and IT resources are vertically scaled in response to dynamic processing power request.

11.4 Elastic Resource Capacity Architecture (2/2)

- Elastic resource capacity architecture can be designed so that the intelligent automation engine script sends its scaling request via the VIM instead of to the hypervisor directly.
- Virtual servers that participate in elastic resource allocation systems may require rebooting the system.
- Cloud usage monitor, pay-per-use monitor, and resource replication may be included in this architecture.

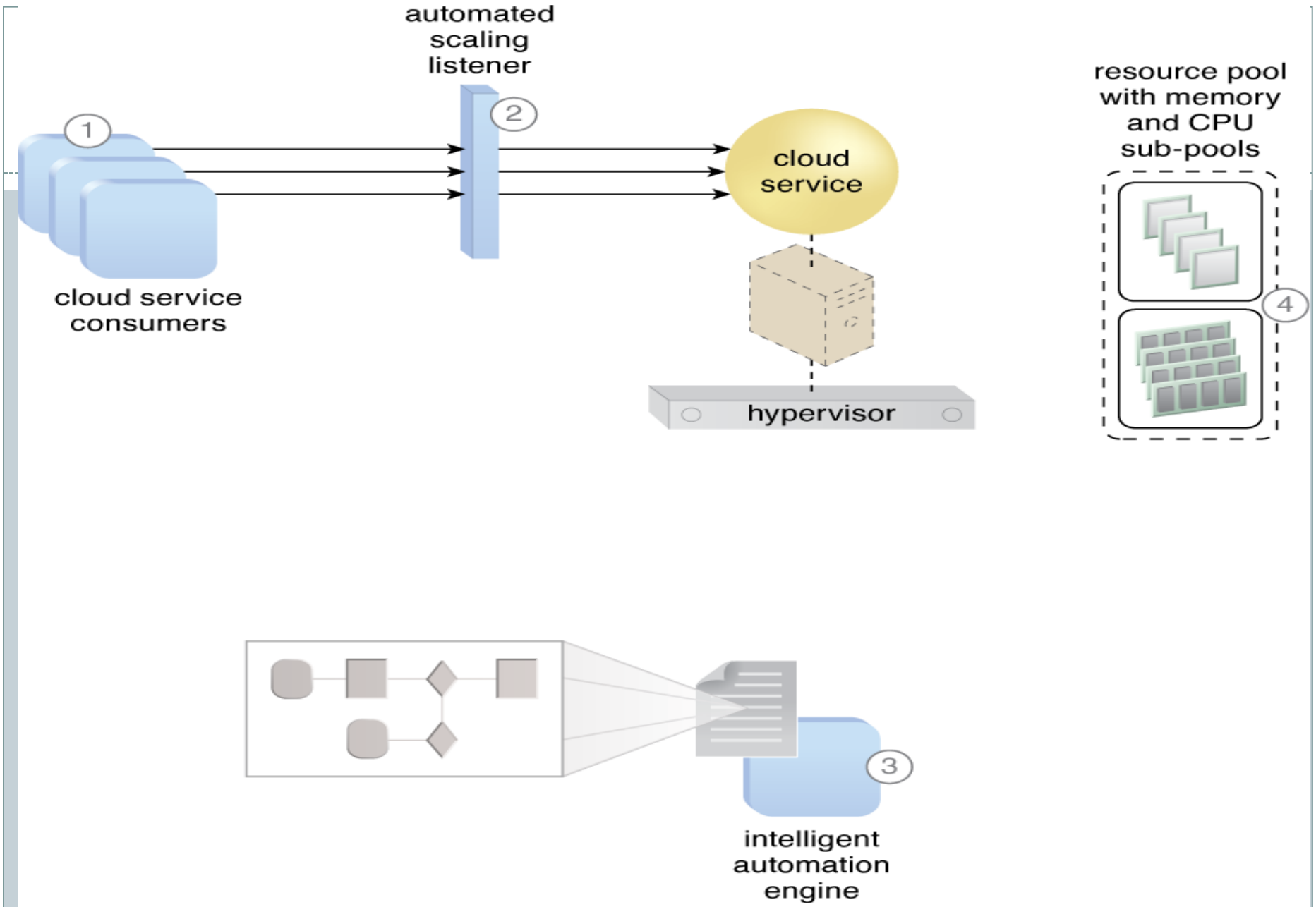


Figure 11.8

- Cloud service consumers are actively sending requests to a cloud service (1), which are monitored by the 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).

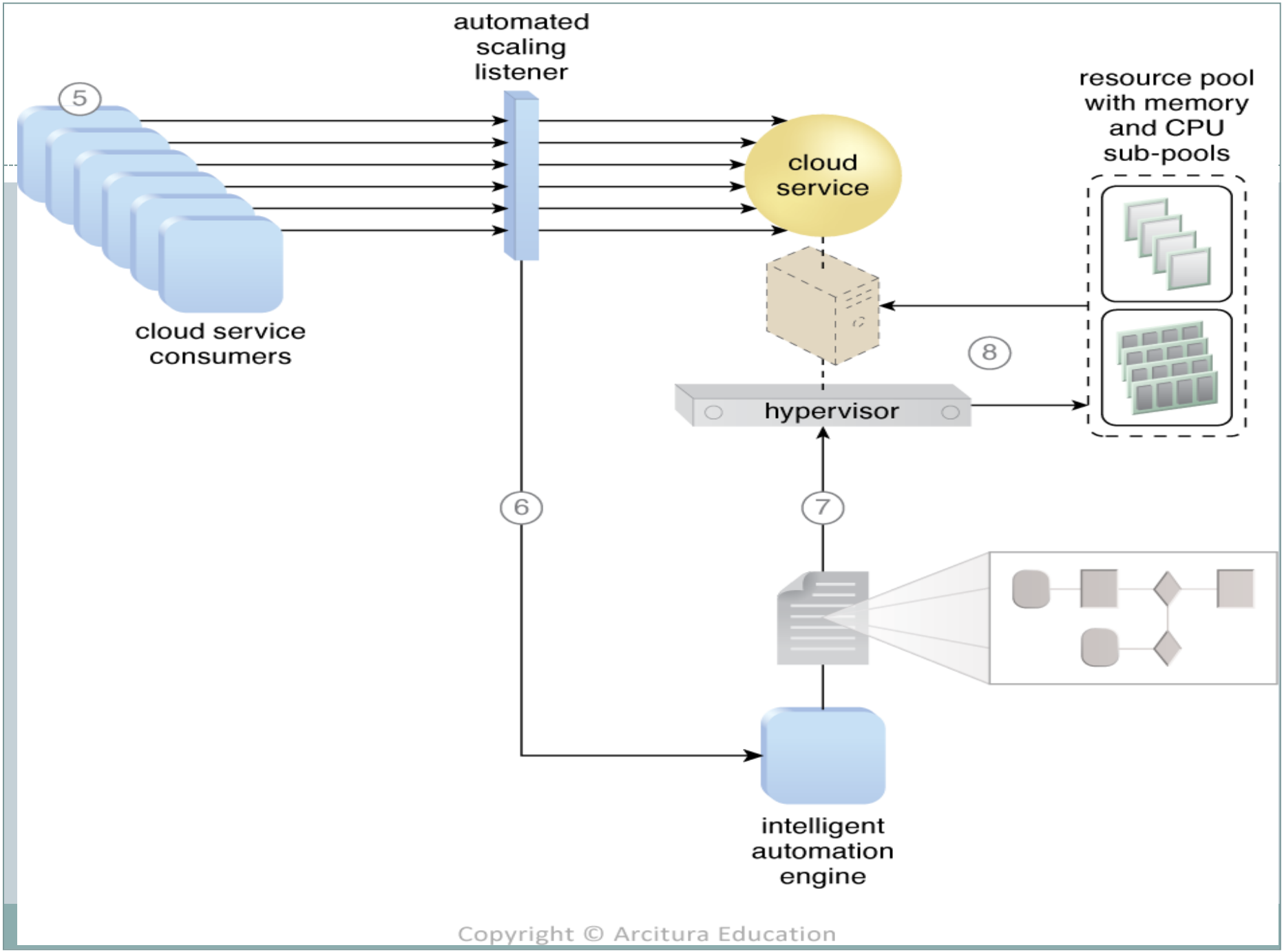


Figure 11.9

- 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).

11.5 Service Load Balancing Architecture (1/2)

- 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.
- The duplicate cloud service implementations are organized into a resource pool, while the load balancer is positioned as either an external or built-in component to allow the host servers to balance the workloads.

11.5 Service Load Balancing Architecture (2/2)

- The load balancer can be positioned either independent of the cloud services and their host servers, or built-in as part of the application or server's environment.
- The service load balancing architecture can involve the following mechanisms in addition to load balancing:
  - Cloud usage monitor
  - Resource monitor
  - Resource replication

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).

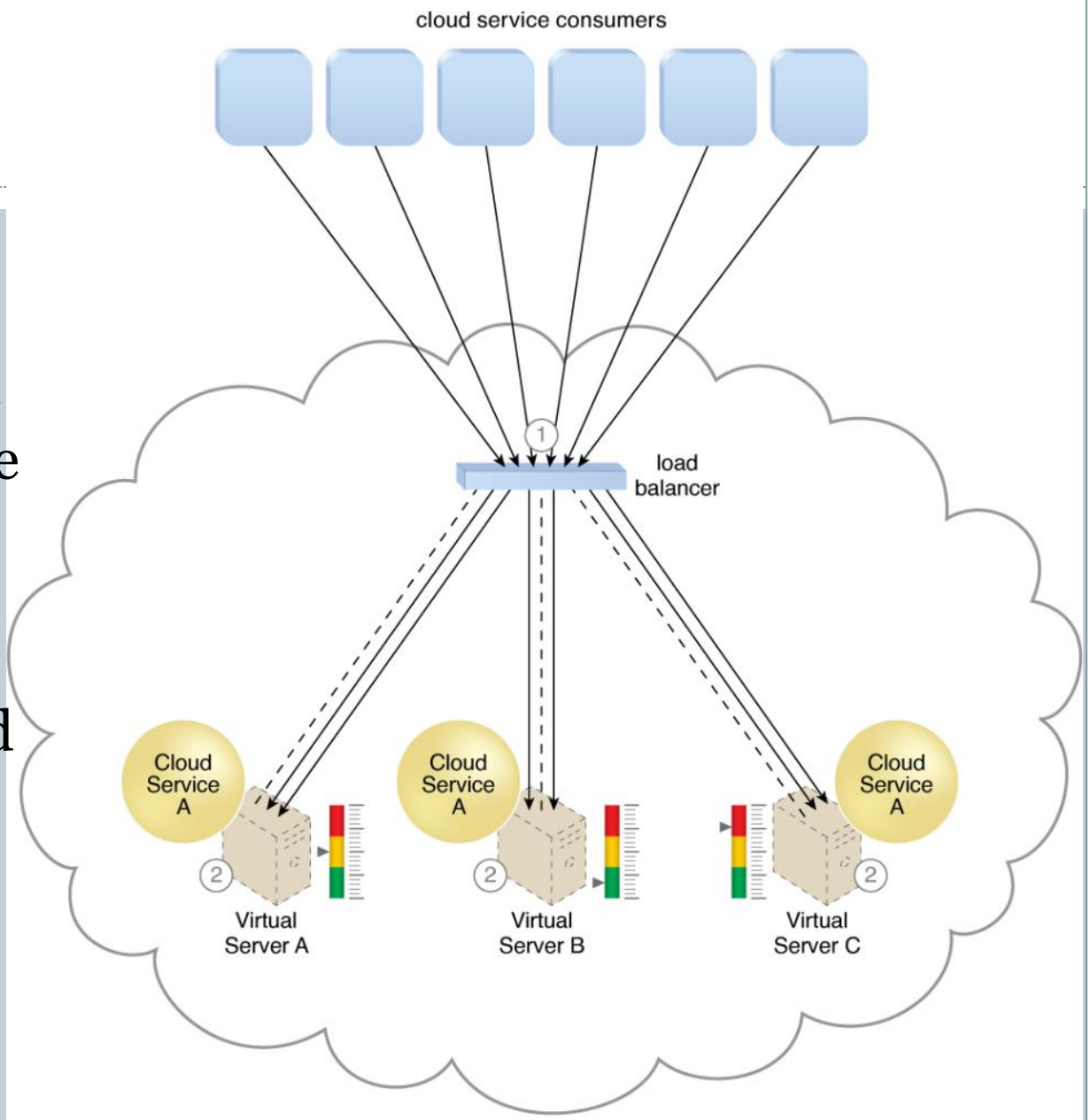
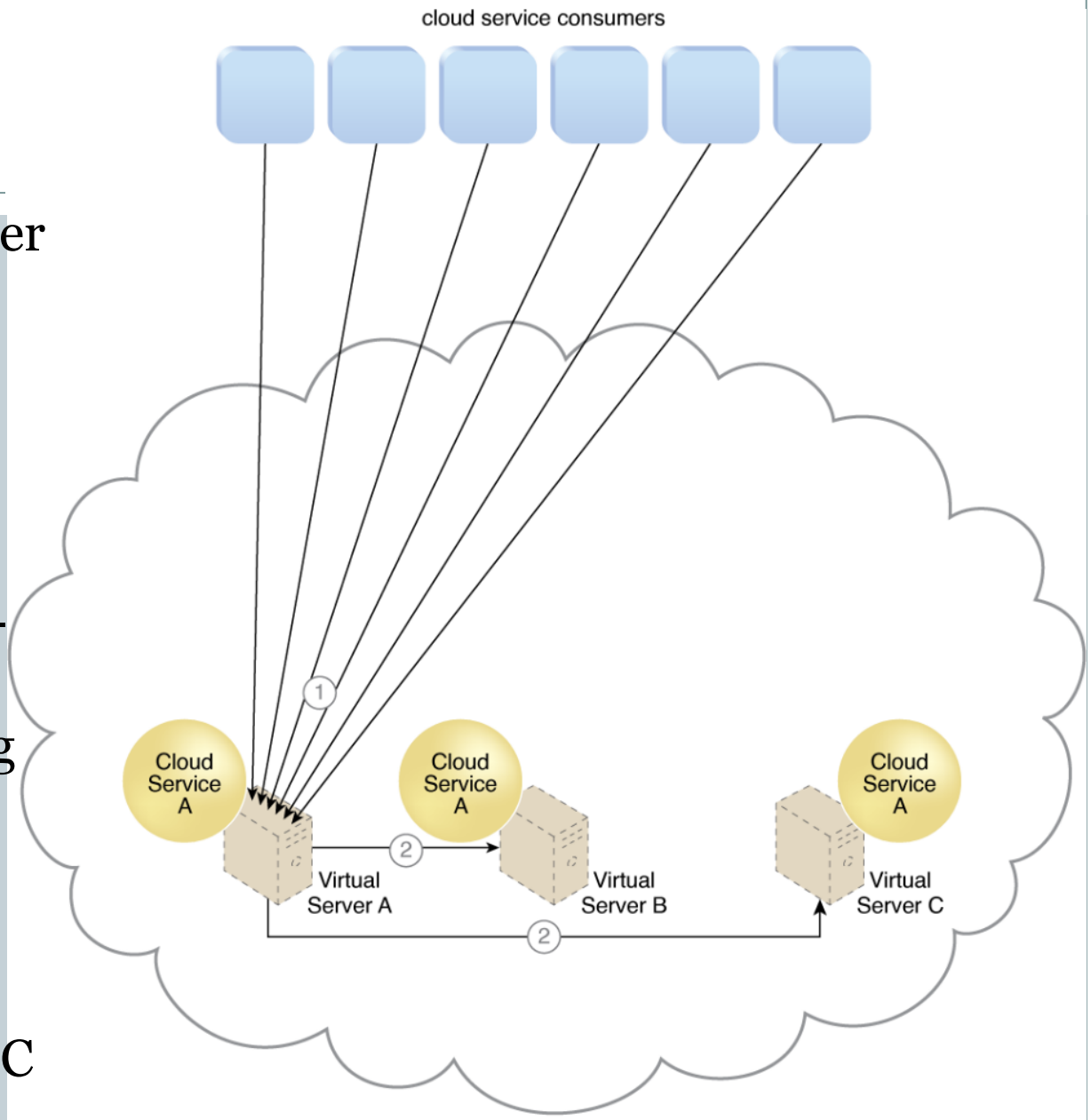


Figure 11.11

- 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).



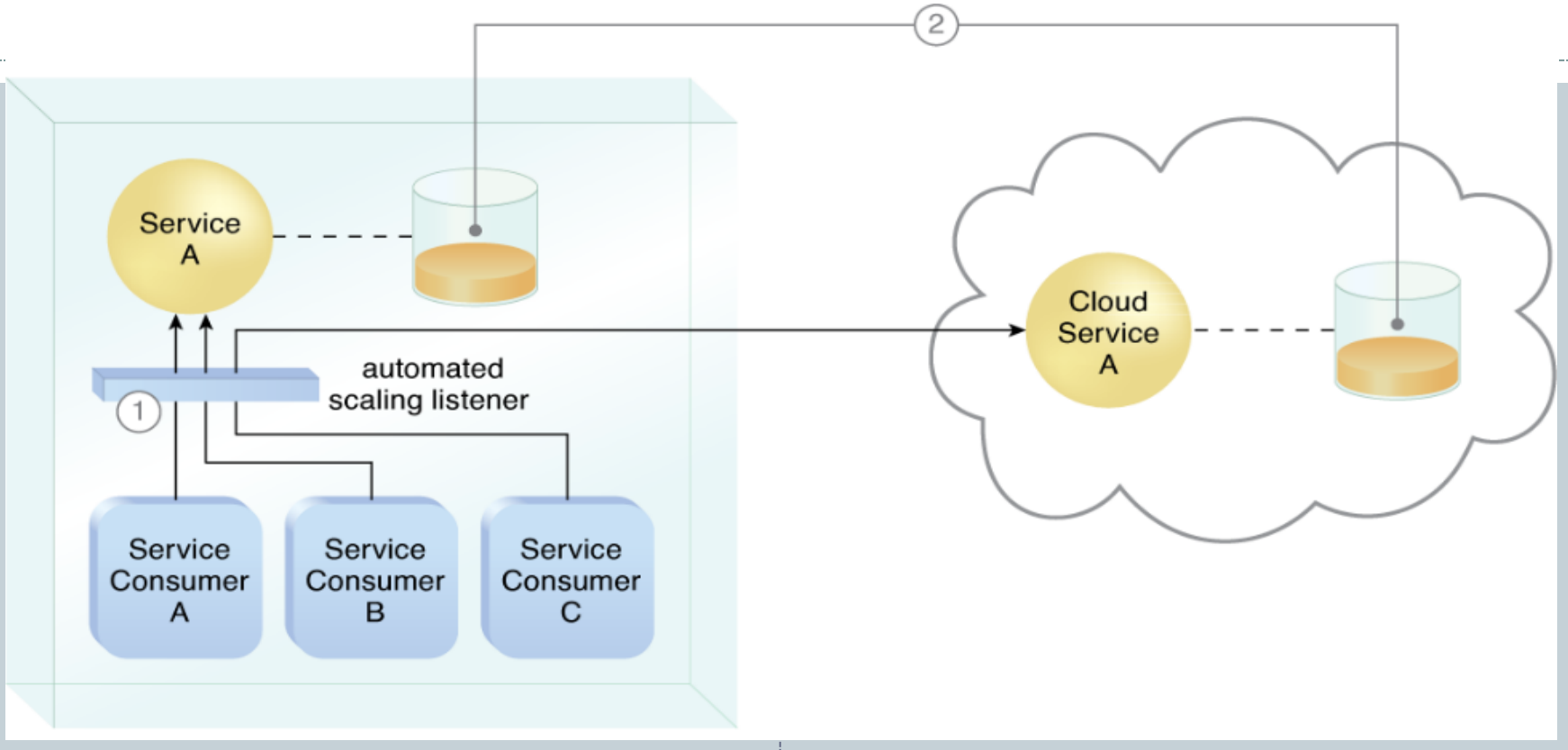
11.6 Cloud Bursting Architecture (1/2)

- 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, while burst-in when they are no longer required.
- The foundation of this architectural model is based on the automated scaling listener and resource replication mechanisms.

11.6 Cloud Bursting Architecture (2/2)

- Cloud bursting is a flexible scaling architecture that provides cloud consumers with the option of using cloud-based IT resources only to meet higher usage demands.
- The foundation of this architectural model is based on the automated scaling listener, to determine when to redirect requests, and resource replication, to maintain synchronicity between on-premise and cloud-based IT resources in relation to state information.

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).

11.7 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.
- Oppositely, cloud consumers are commonly charged for cloud-based storage space based on disk capacity allocation.

Figure 11.13

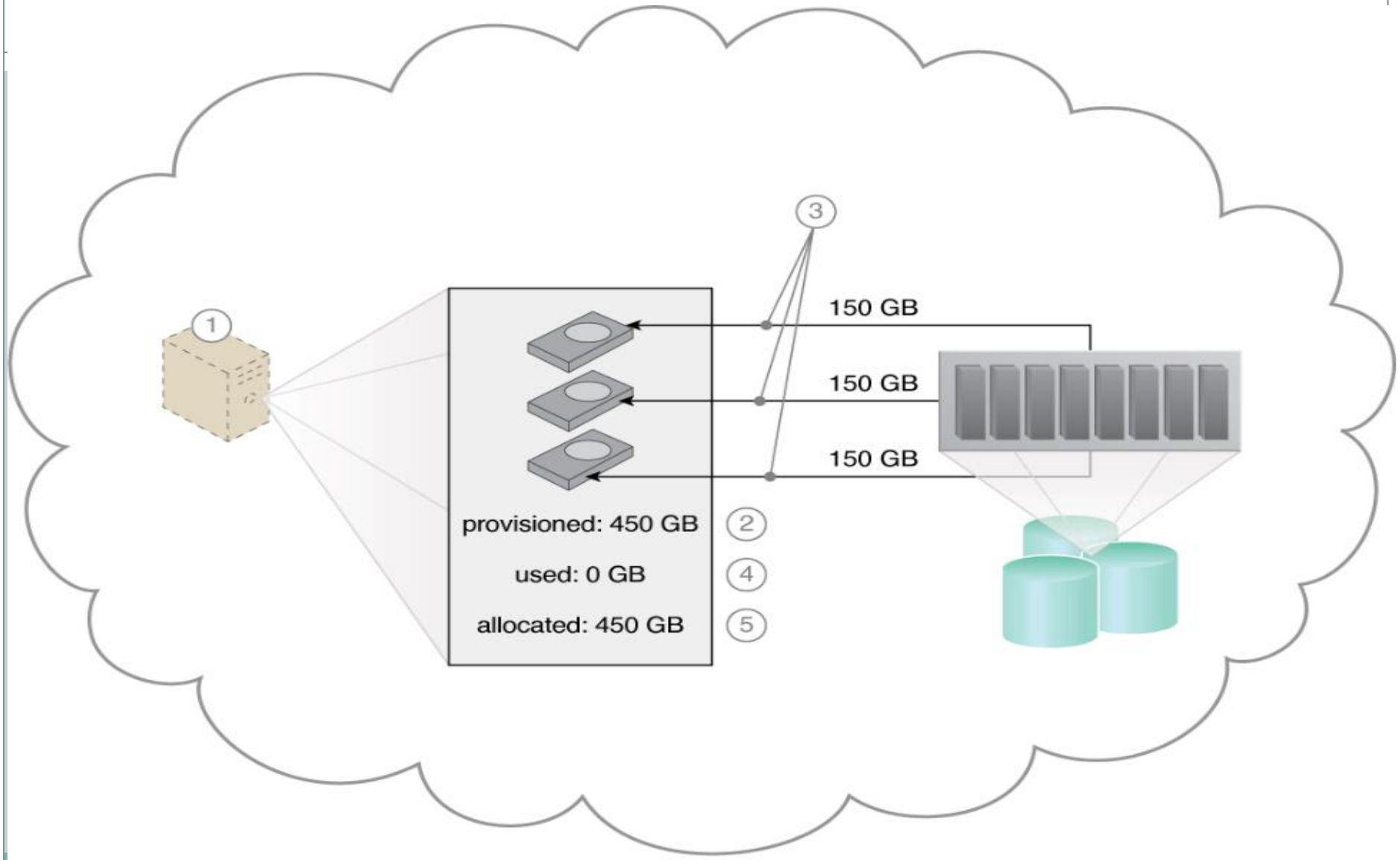


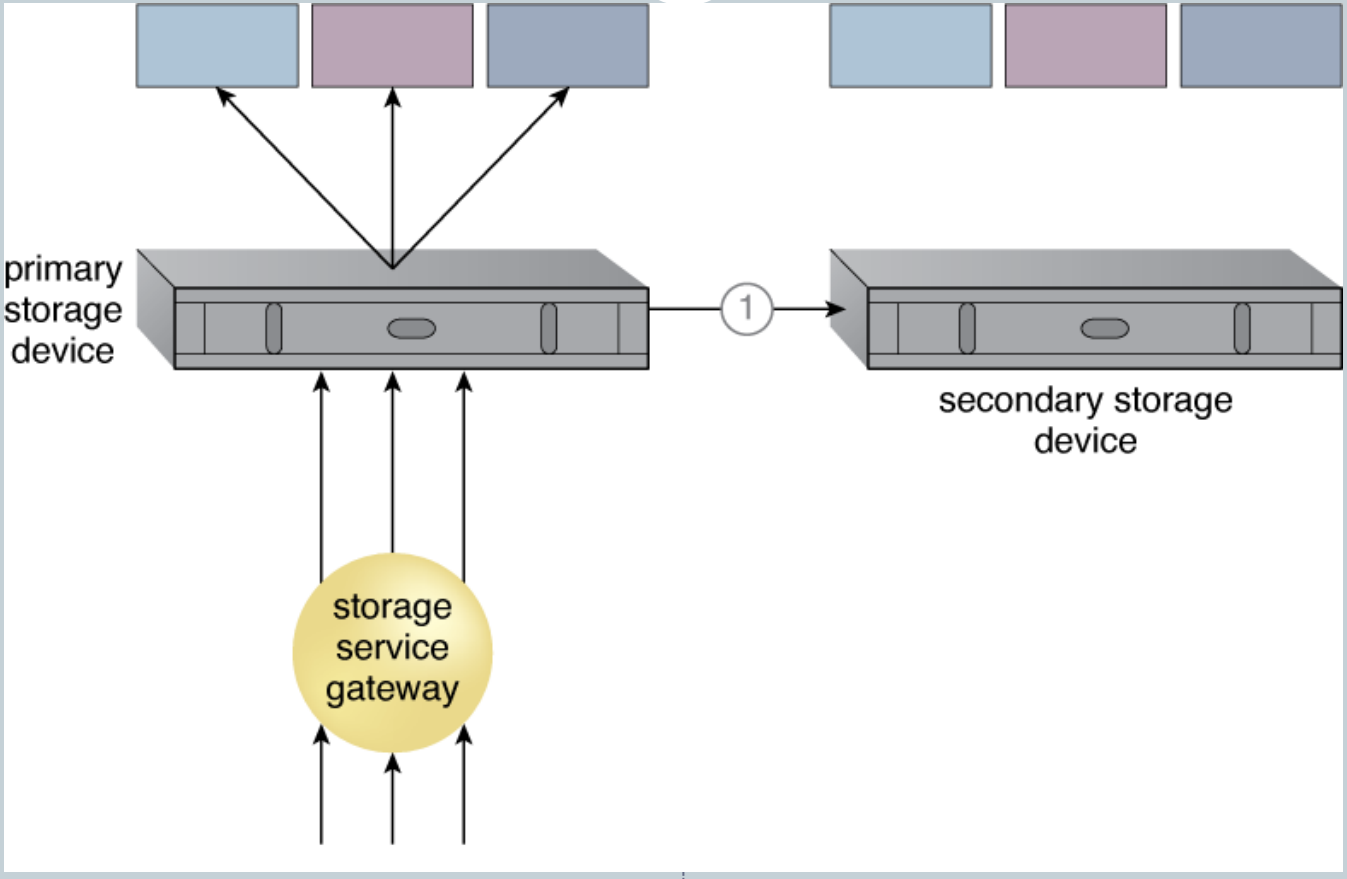
Figure 11.13, 11.14

- 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).

11.8 Redundant Storage Architecture (1/2)

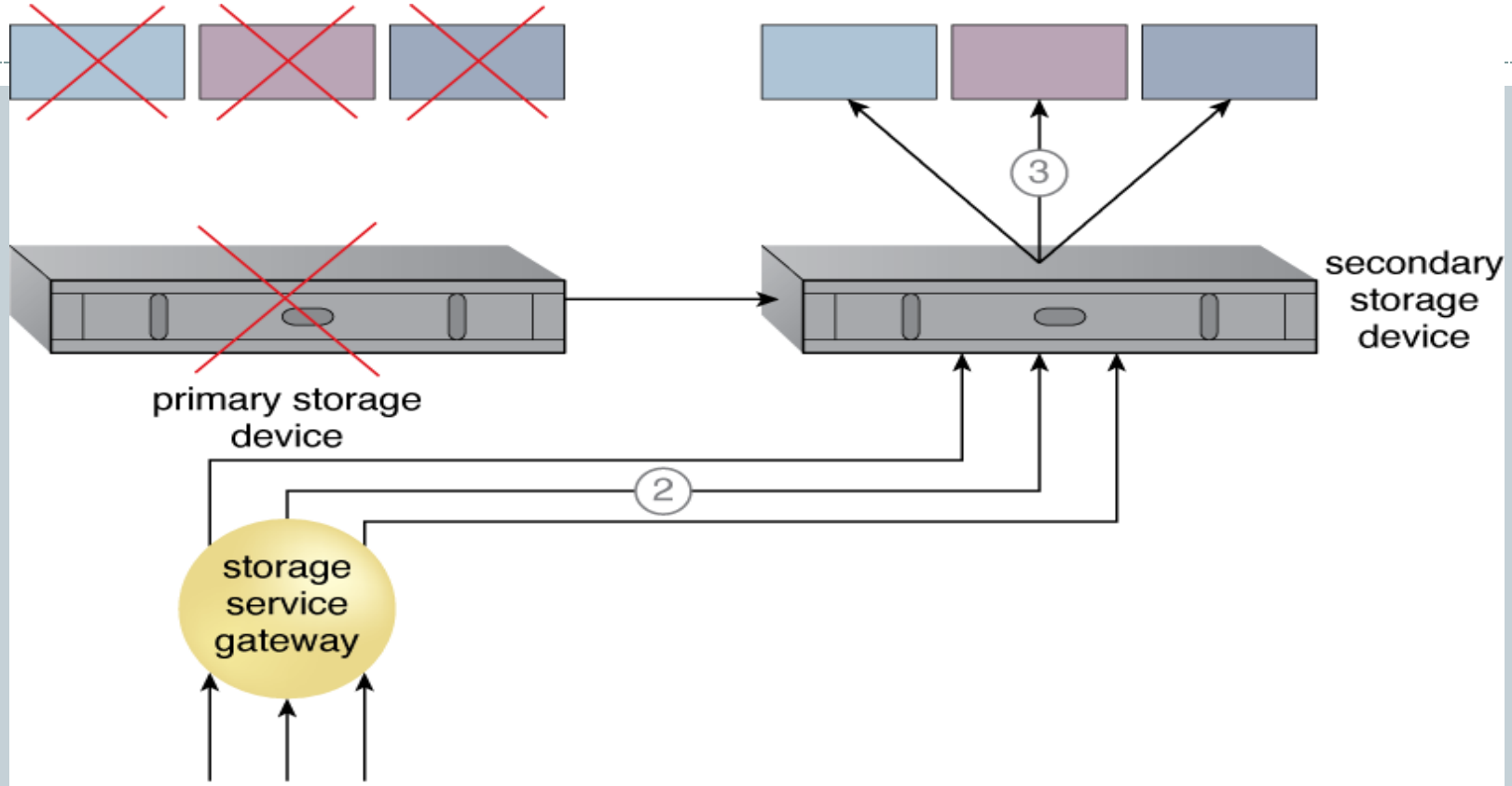
- The redundant storage architecture introduces a secondary duplicate cloud storage device as part of a failover system that synchronizes its data with the data in the primary cloud storage device. A storage requests to the secondary device whenever the primary device fails.
- The storage service gateway is a component that acts as the external interface to cloud storage services, and is capable of automatically redirecting cloud consumer requests whenever necessary.

Figure 11.16



- The primary cloud storage device is routinely replicated to the secondary cloud storage device .

Figure 11.17

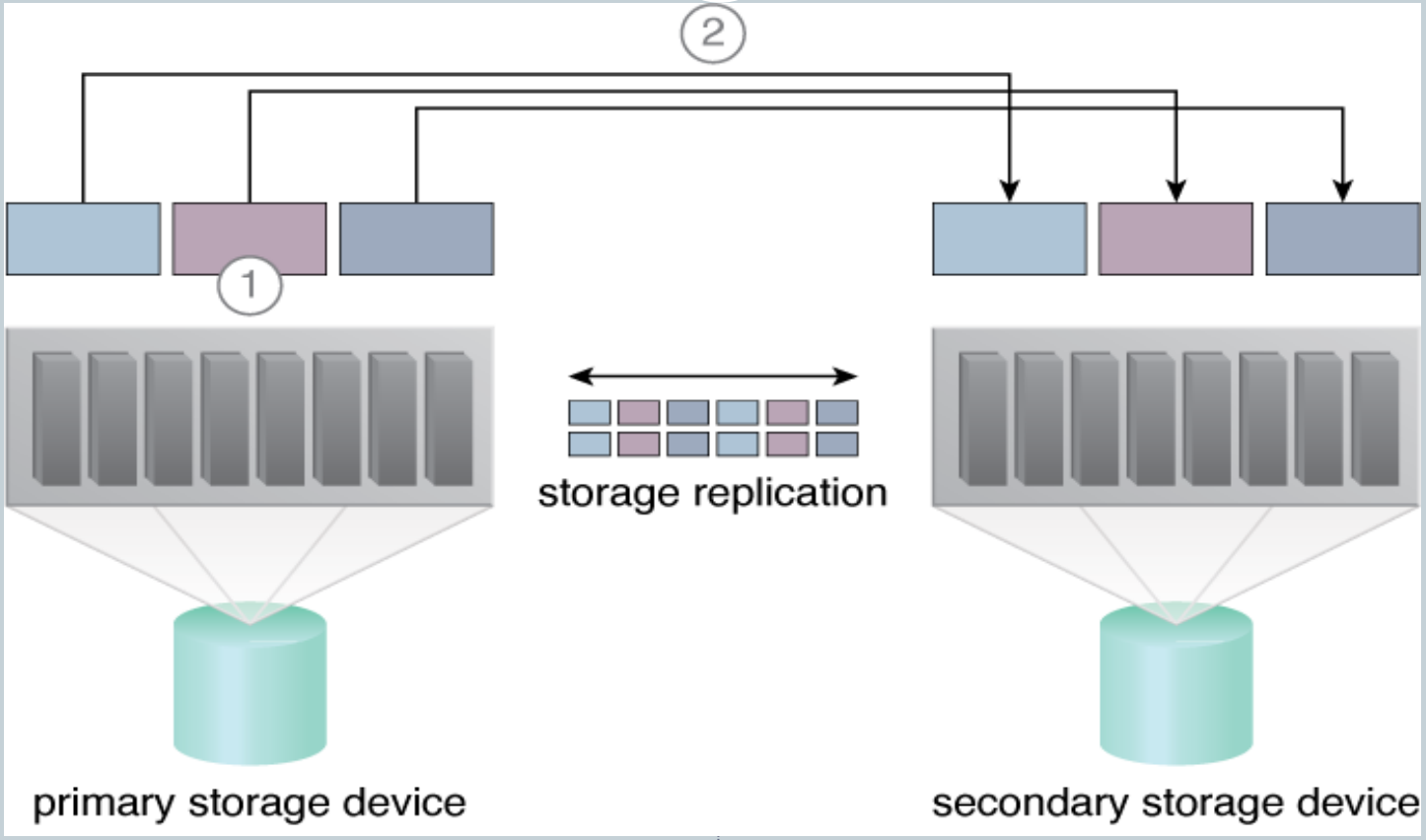


- The primary storage becomes unavailable and the storage service gateway forwards the cloud consumer requests to the secondary storage device (2).
- The secondary storage device forwards the requests to the LUNs, allowing cloud consumers to continue to access their data (3).

11.8 Redundant Storage Architecture (2/2)

- This architecture primarily relies on a storage replication system that keeps the primary cloud storage device synchronized with secondary devices.
- Cloud providers may locate secondary cloud storage devices in a different geographical region than the primary cloud storage device, usually for economic reasons.
- Some cloud providers use storage devices with dual array and storage devices in a different physical location for cloud balancing and disaster recovery purposes.

Figure 11.18



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