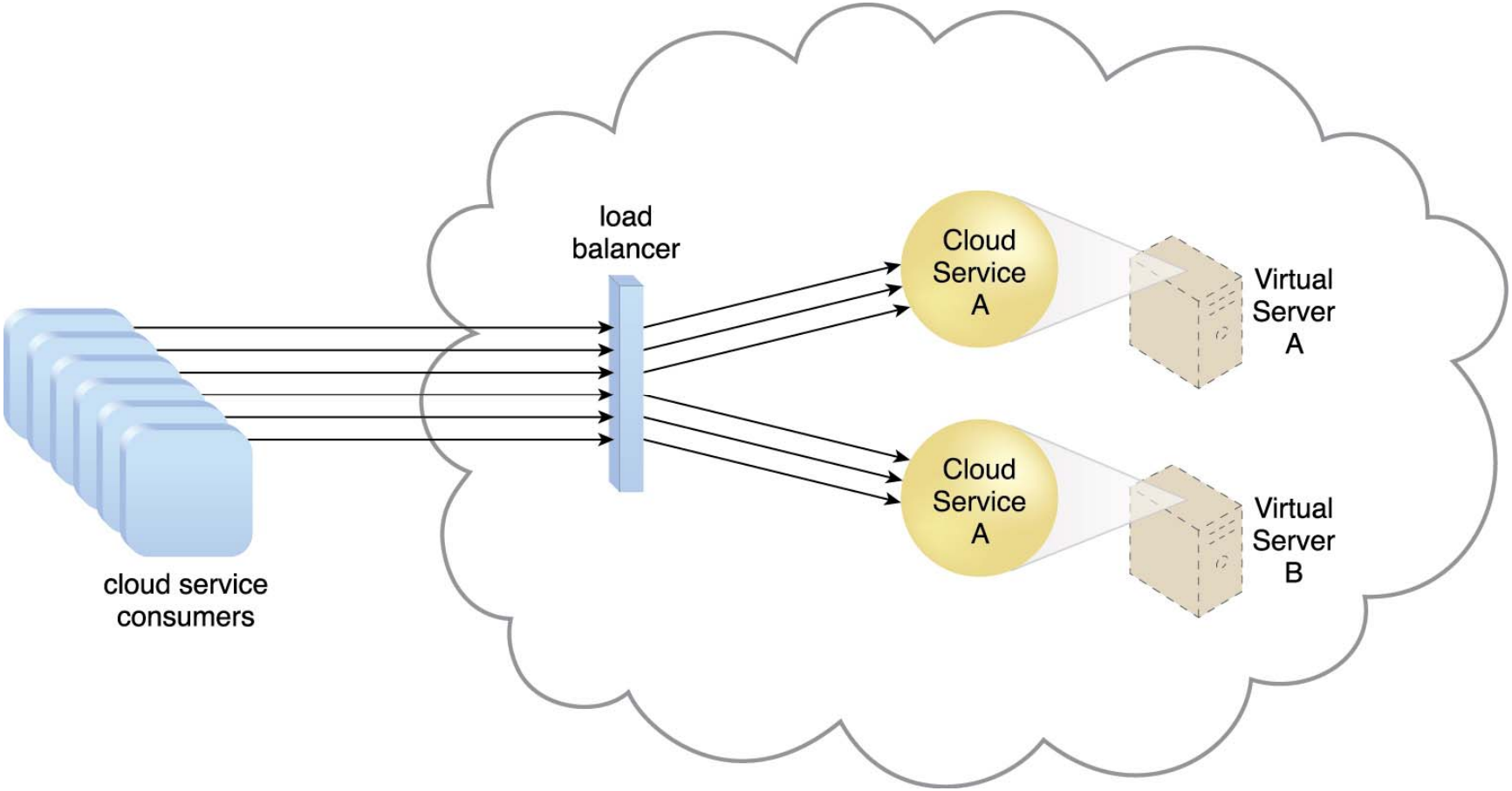
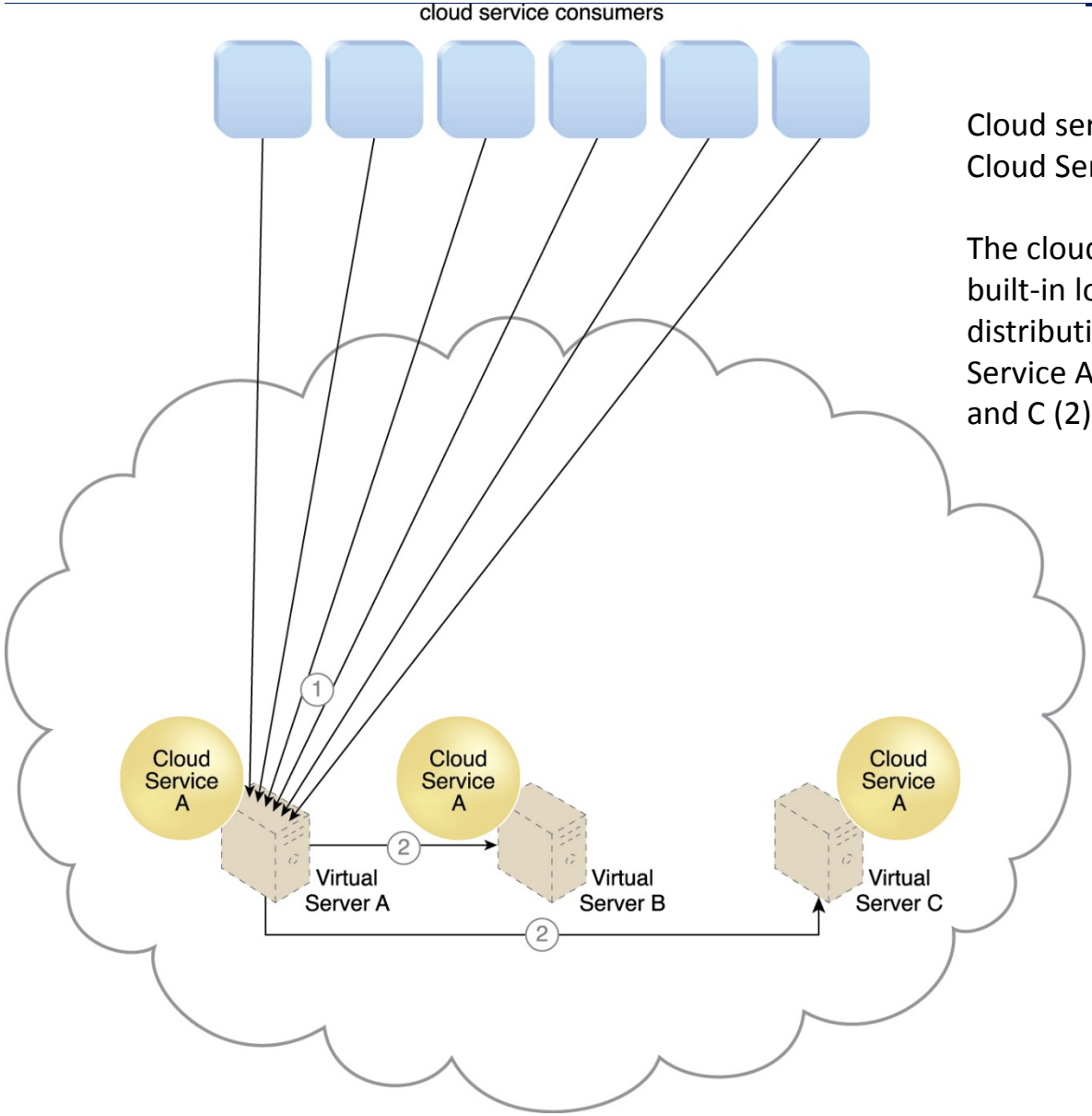
Cloud Computing Architecture

1. Workload distribution architecture



- Scale up/down the IT resources
- Distribute workload among IT resource evenly

A variant

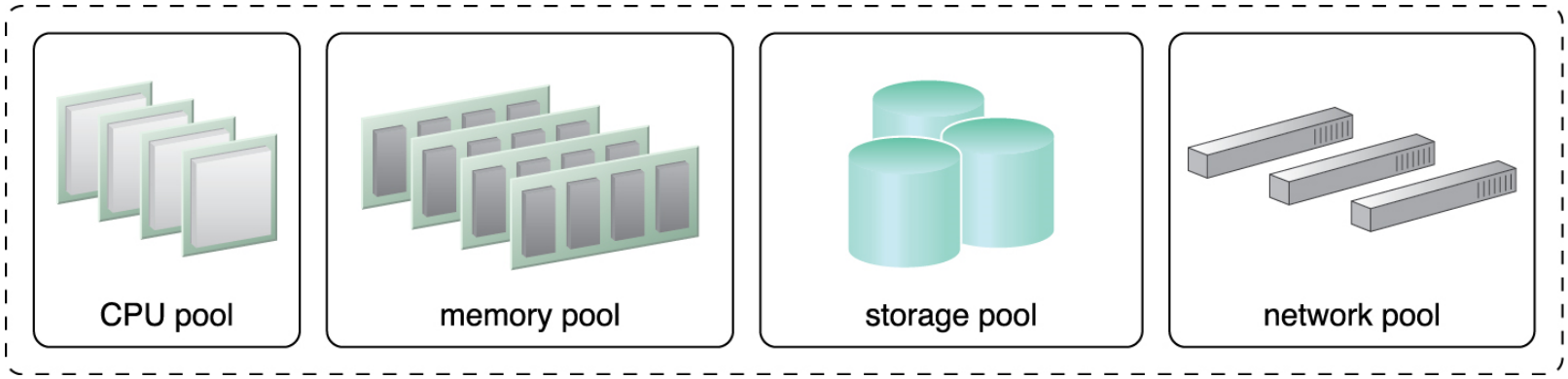


Cloud service consumer requests are sent to Cloud Service A on Virtual Server A (1).

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).

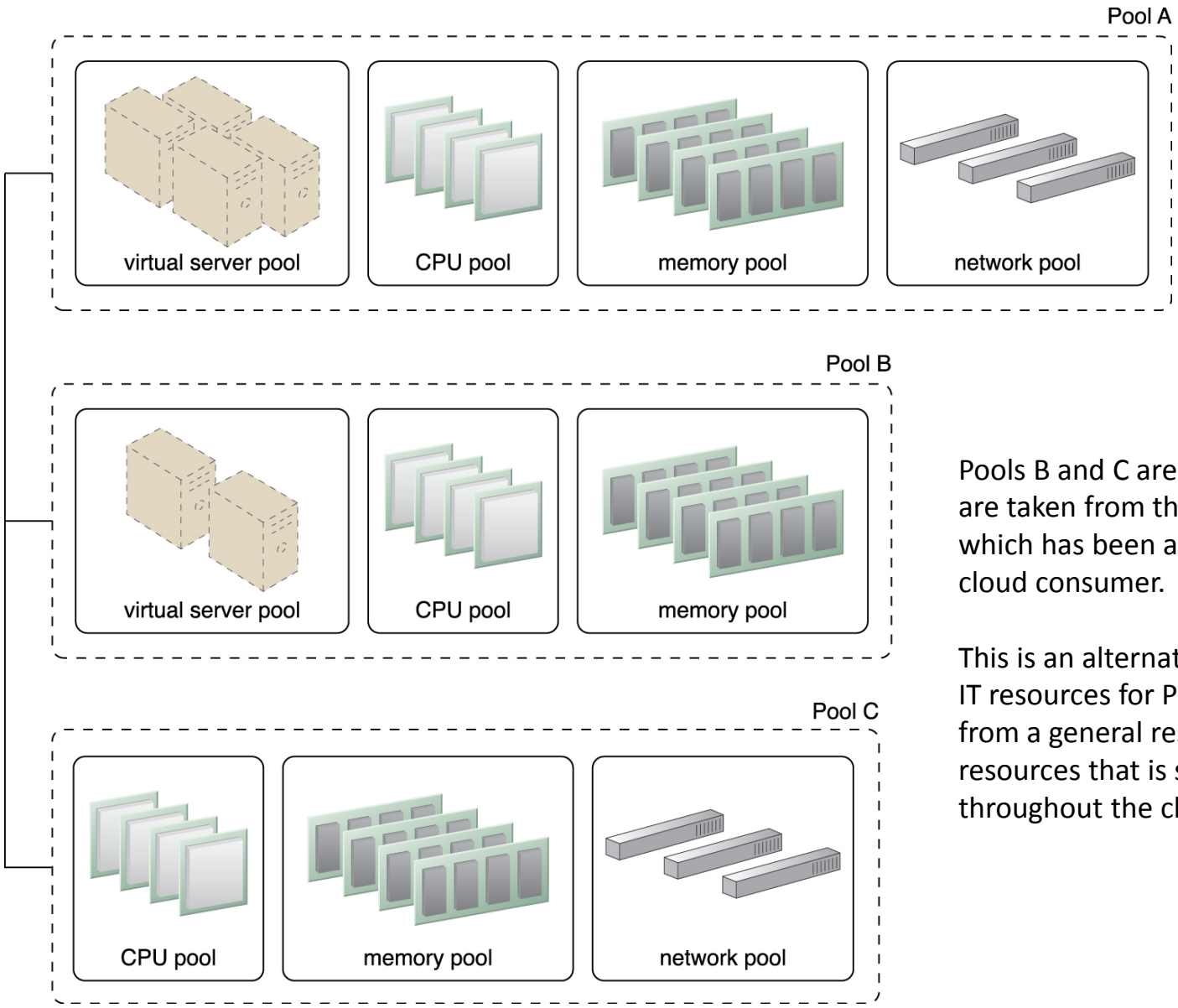
2. Resource pooling architecture

- A resource pooling architecture is based on the use of one or more resource pools



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

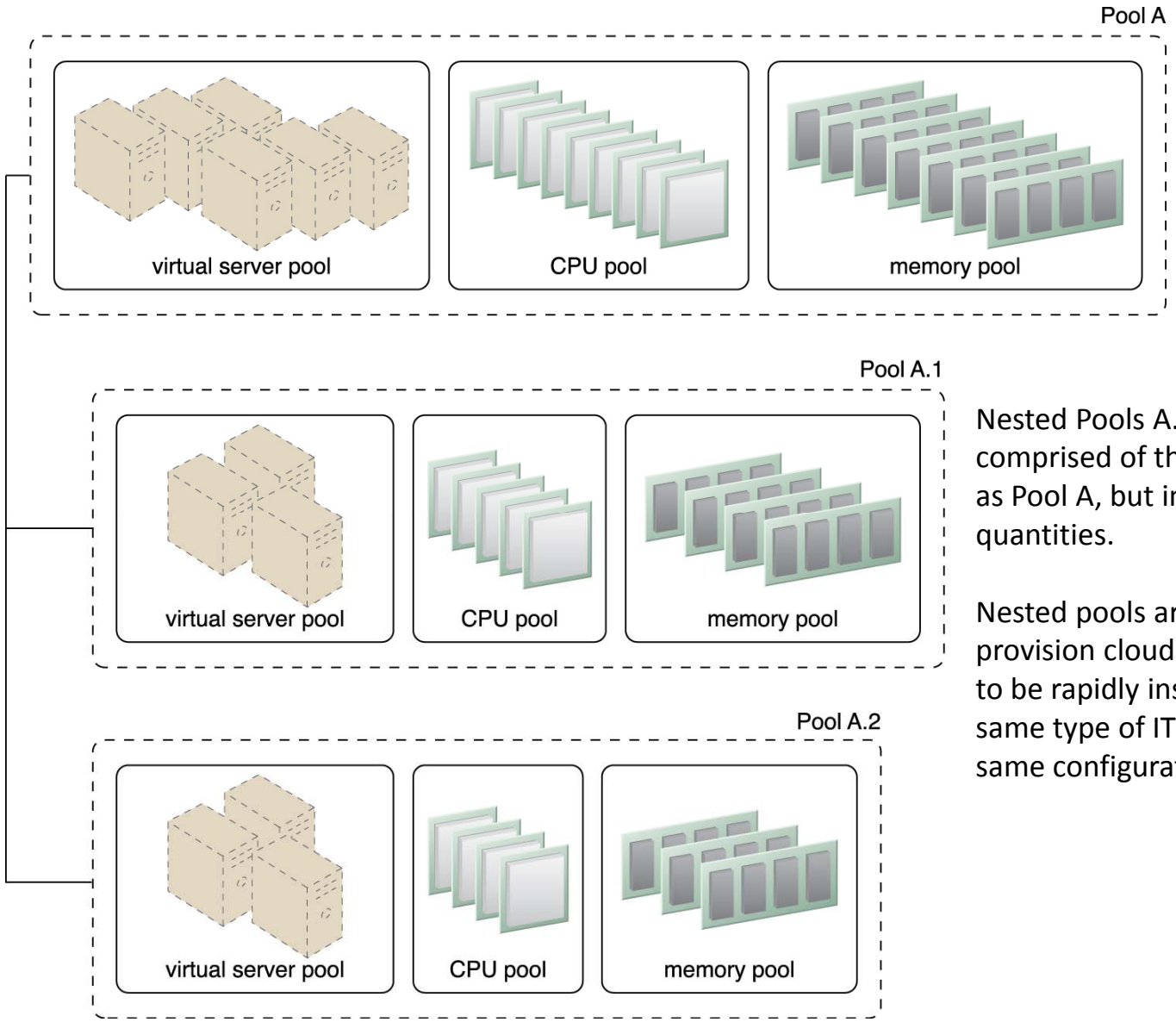
Sibling resource pools



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.

Nested pool model



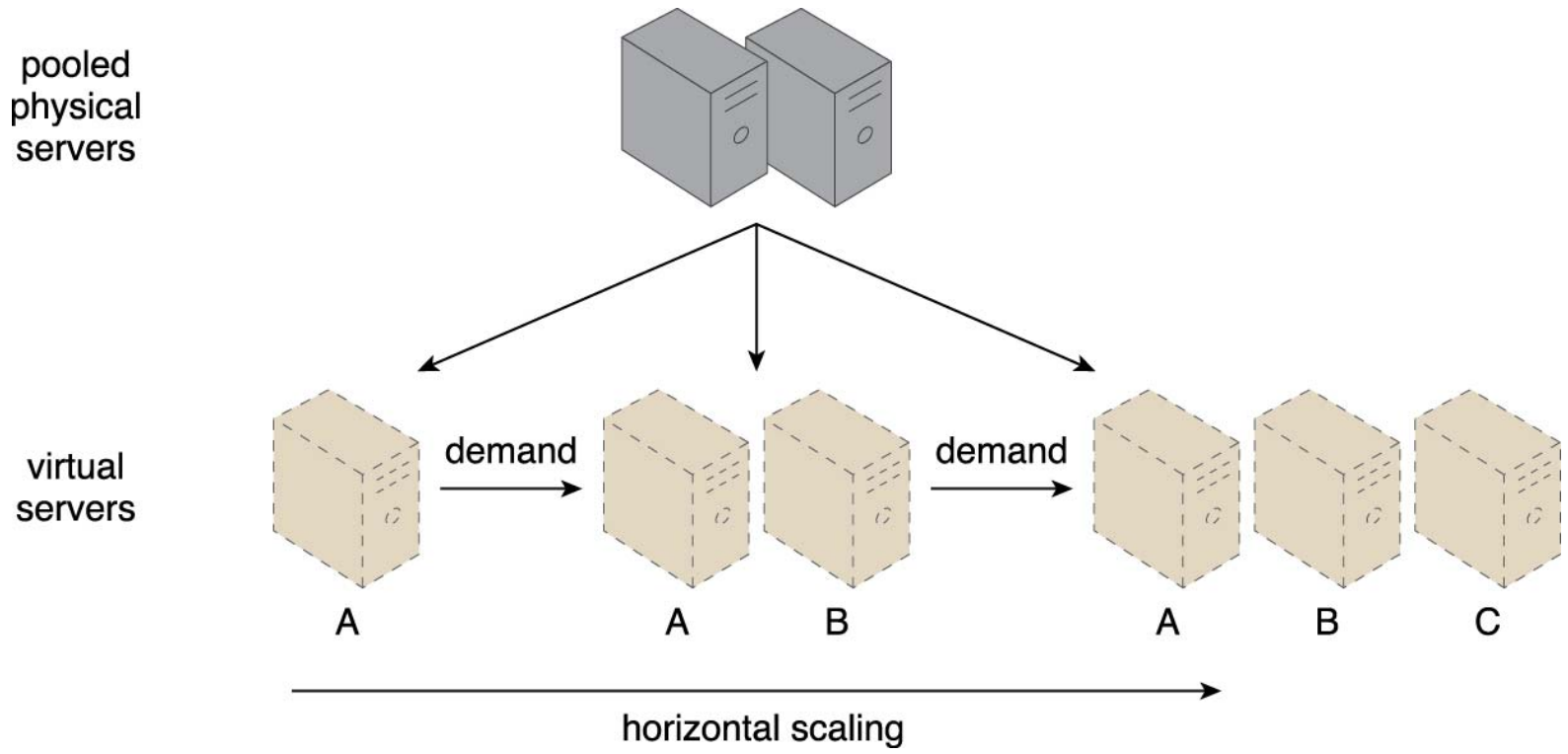
Nested Pools A.1 and Pool A.2 are comprised of the same IT resources as Pool A, but in different quantities.

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.

3. Dynamic scalability architecture

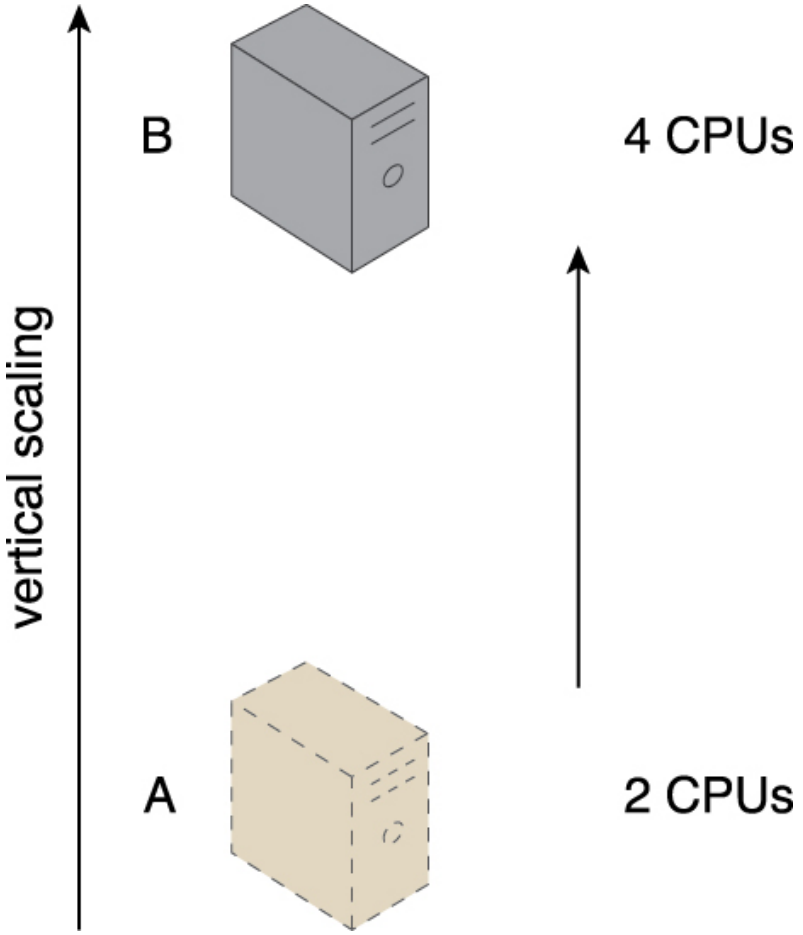
- An architecture model based on a system of predefined scaling conditions that trigger the dynamic allocation of IT resources from the pools
  - Dynamic horizontal scaling
    - IT resource instances are scaled out and in to handle fluctuating workloads
  - Dynamic vertical scaling
    - IT resource instances are scaled up and down when there is a need to adjust the processing capacity of a single IT resource
  - Dynamic relocation
    - IT resource is related to a host with more capacity

Horizontal scaling



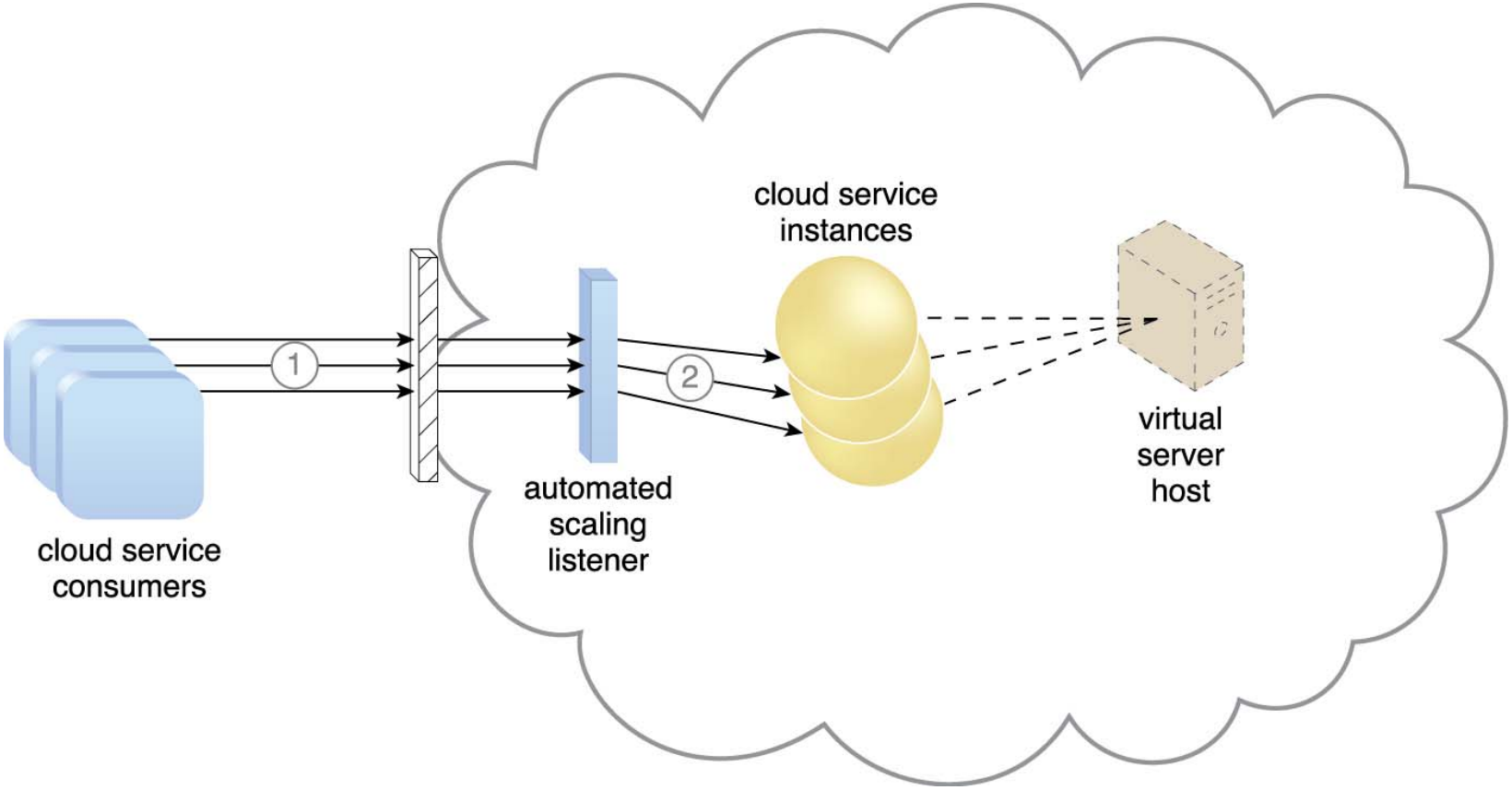
An IT resource (Virtual Server A) is scaled out by adding more of the same IT resources (Virtual Servers B and C).

Vertical scaling



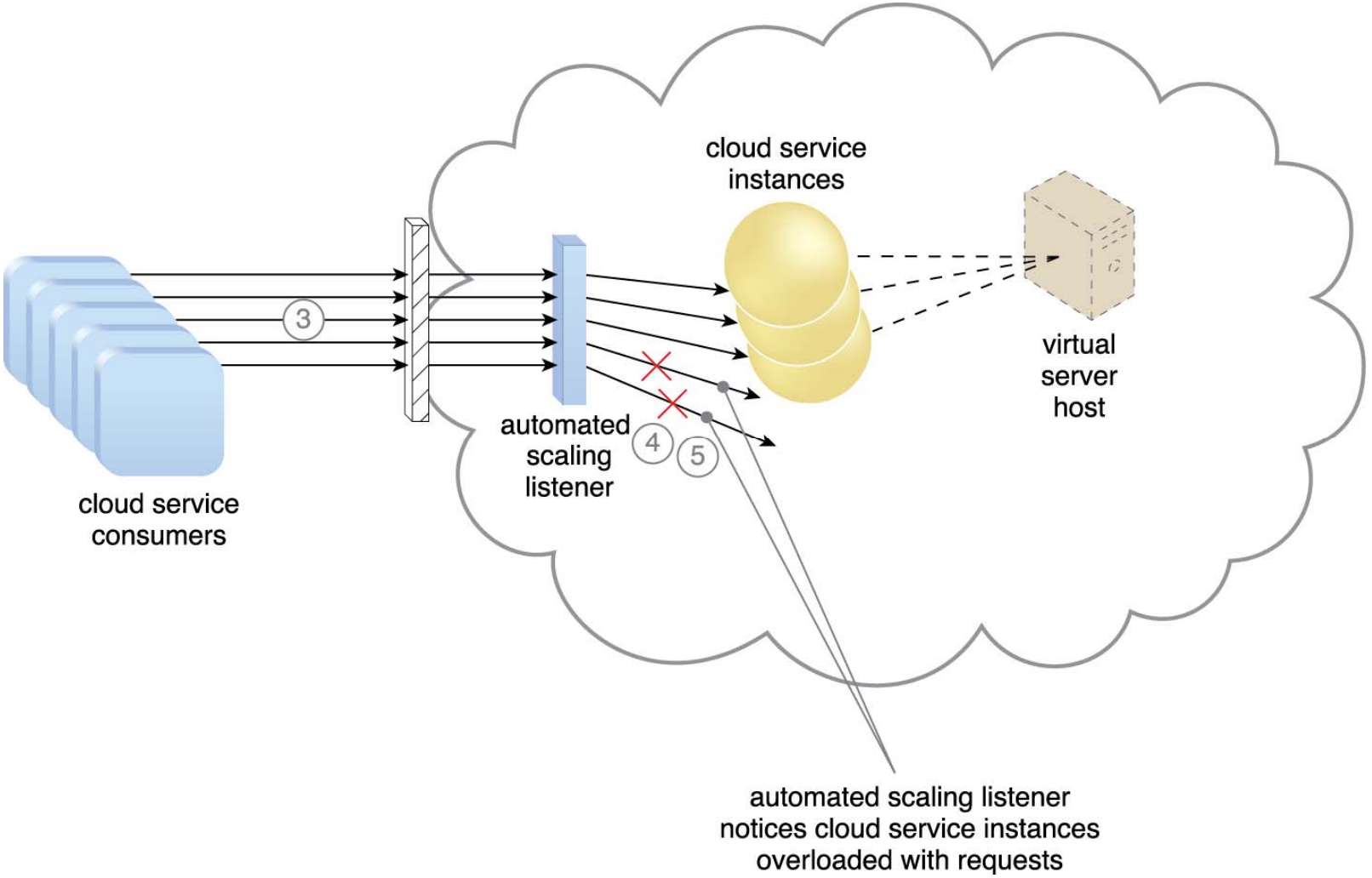
An IT resource (a virtual server with two CPUs) is scaled up by replacing it with a more powerful IT resource with increased capacity for data storage (a physical server with four CPUs).

Example



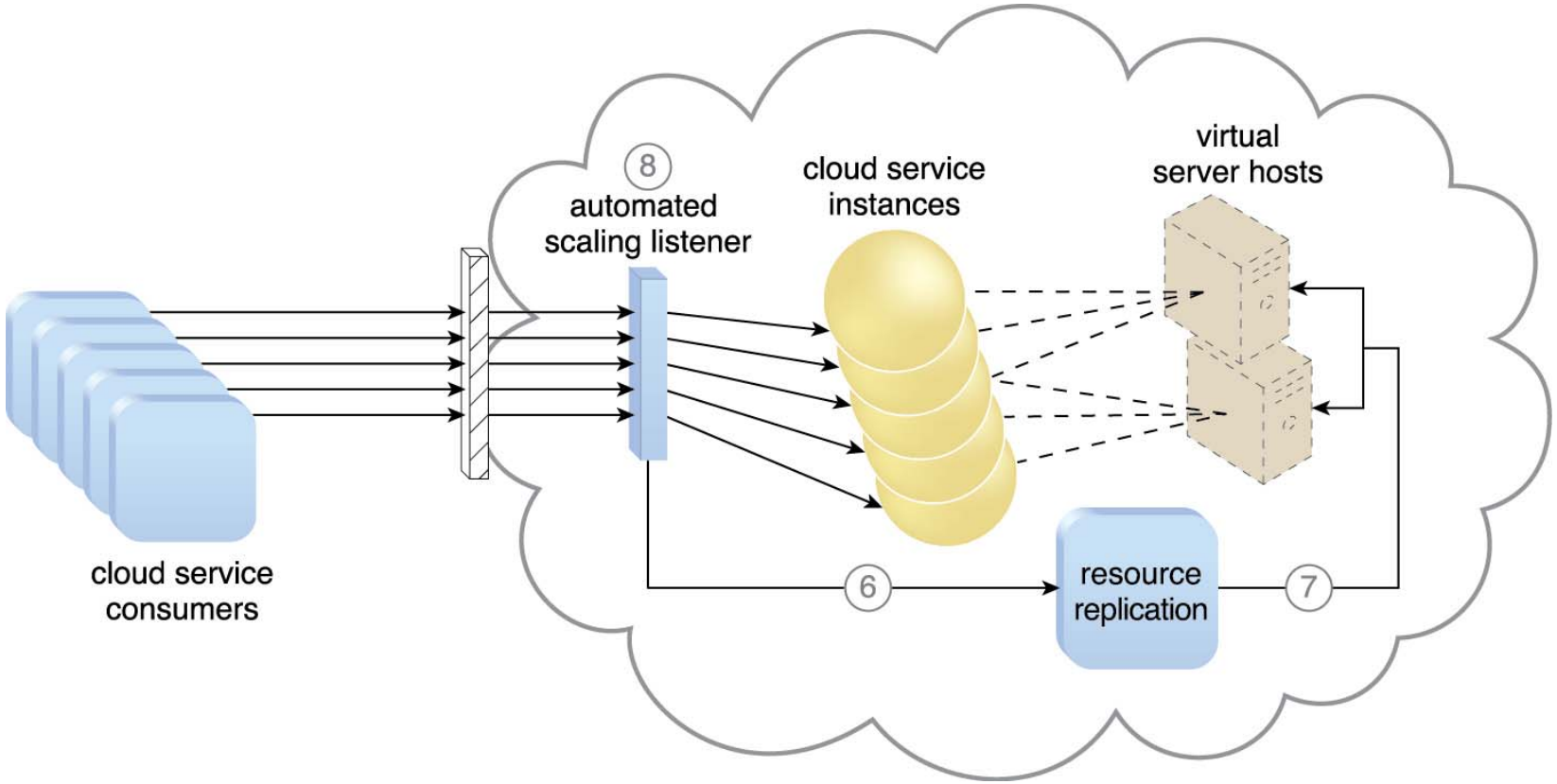
Cloud service consumers are sending requests to a cloud service (1). The automated scaling listener monitors the cloud service to determine if predefined capacity thresholds are being exceeded (2).

Example



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).

Example

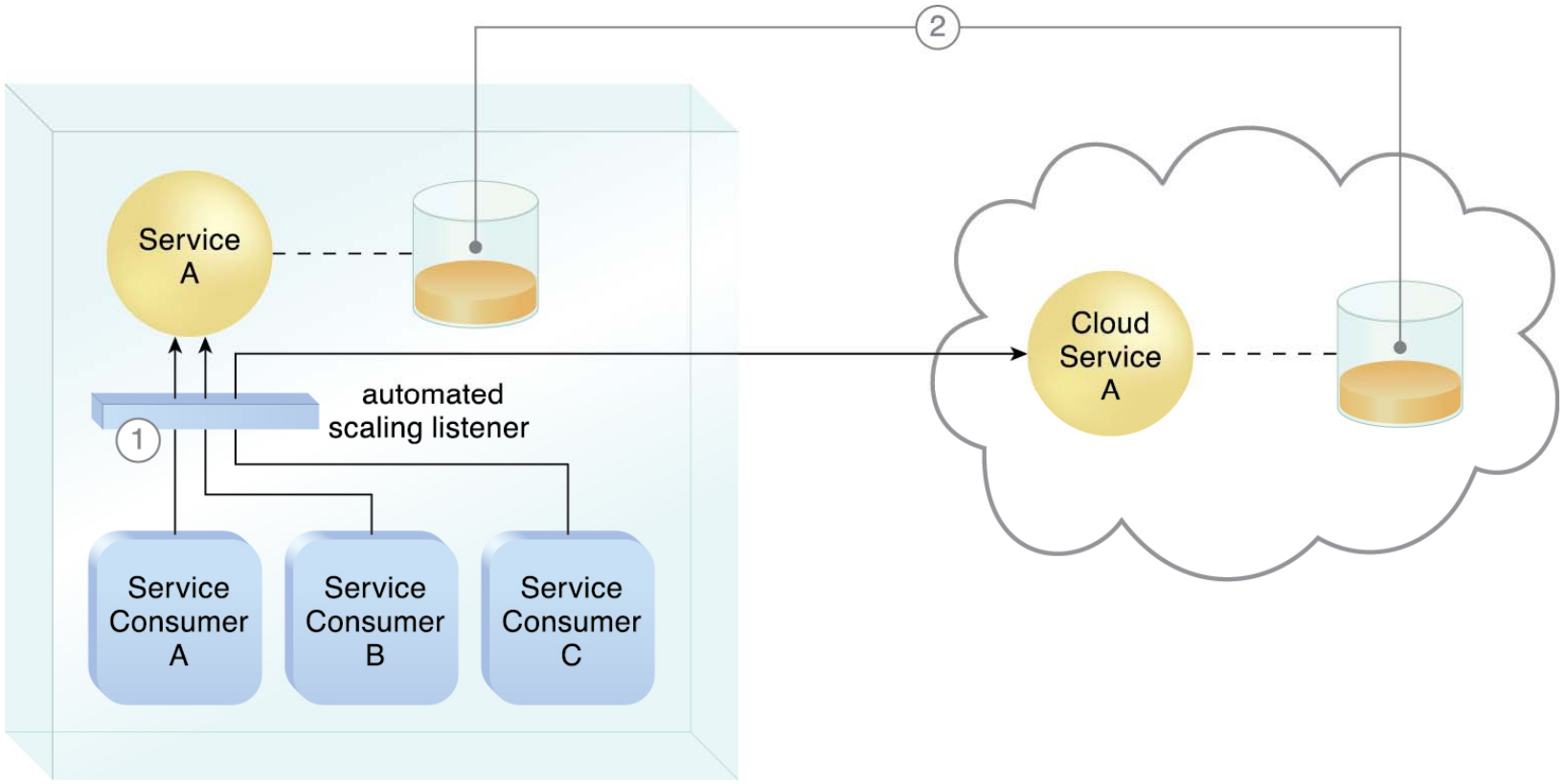


The automated scaling listener sends a signal to the resource replication mechanism (6), which creates more instances of the cloud service (7). Now that the increased workload has been accommodated, the automated scaling listener resumes monitoring and detracting and adding IT resources, as required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

4. Cloud bursting architecture

- A form of dynamic scaling that scales or “bursts out” on-premise IT resources into a cloud whenever predefined capacity thresholds have been reached
- A flexible scaling architecture that provides cloud consumers with the option of using cloud-based IT resources only to meet higher usage demands

Example

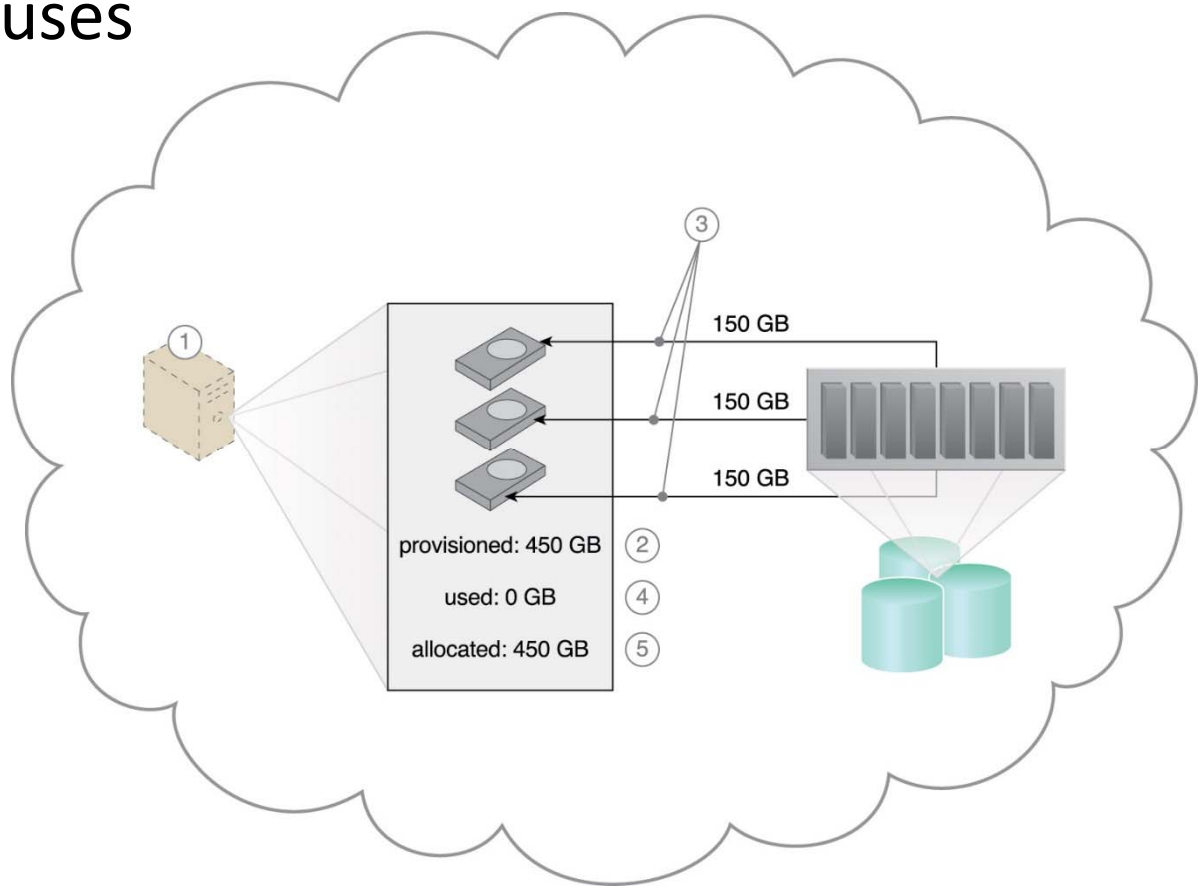


An automated scaling listener monitors the usage of on-premise Service A, and redirects Service Consumer C's request to Service A's redundant implementation in the cloud (Cloud Service A) once Service A's usage threshold has been exceeded (1).

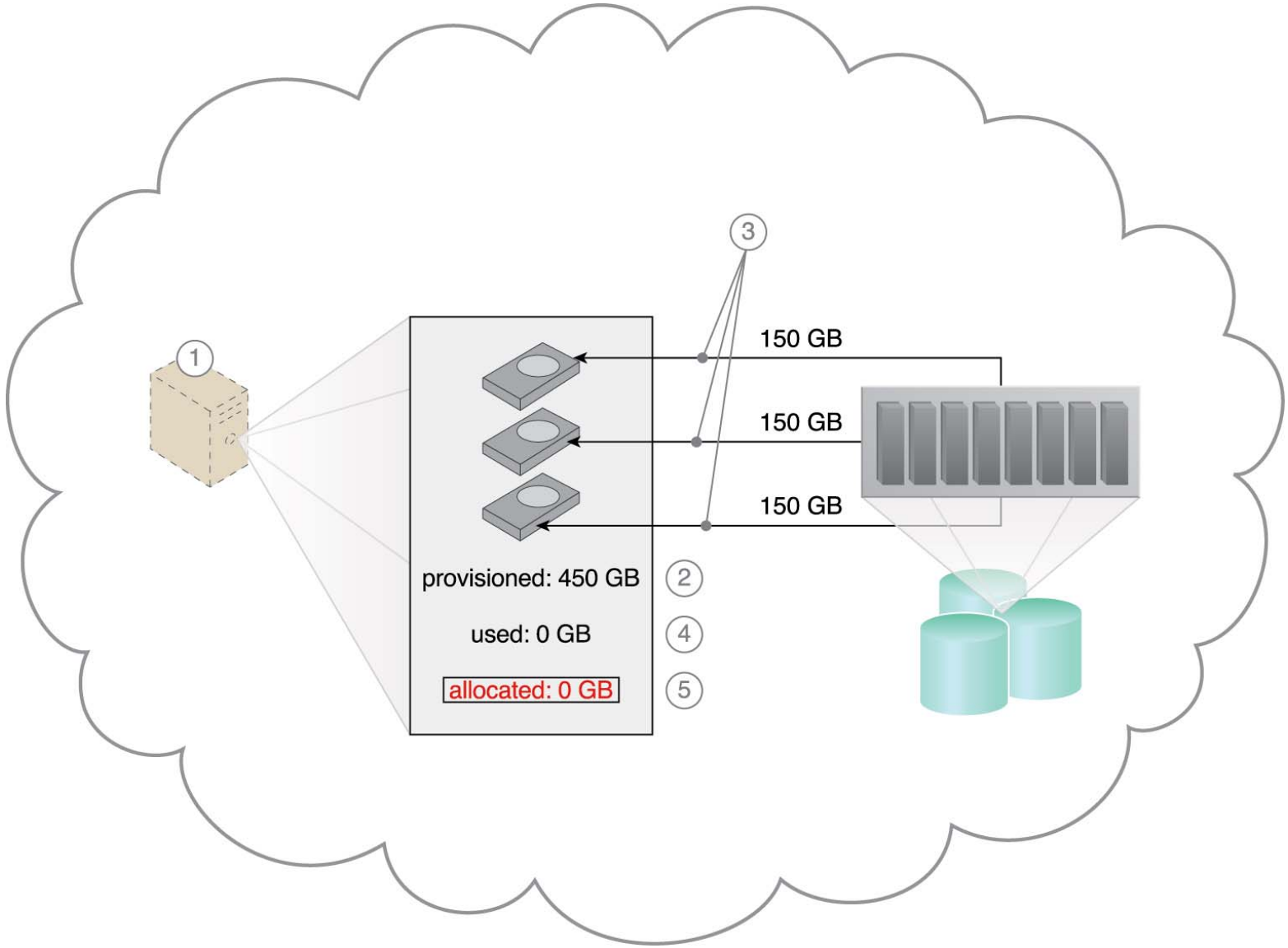
A resource replication system is used to keep state management databases synchronized (2).

5. Elastic disk provisioning architecture

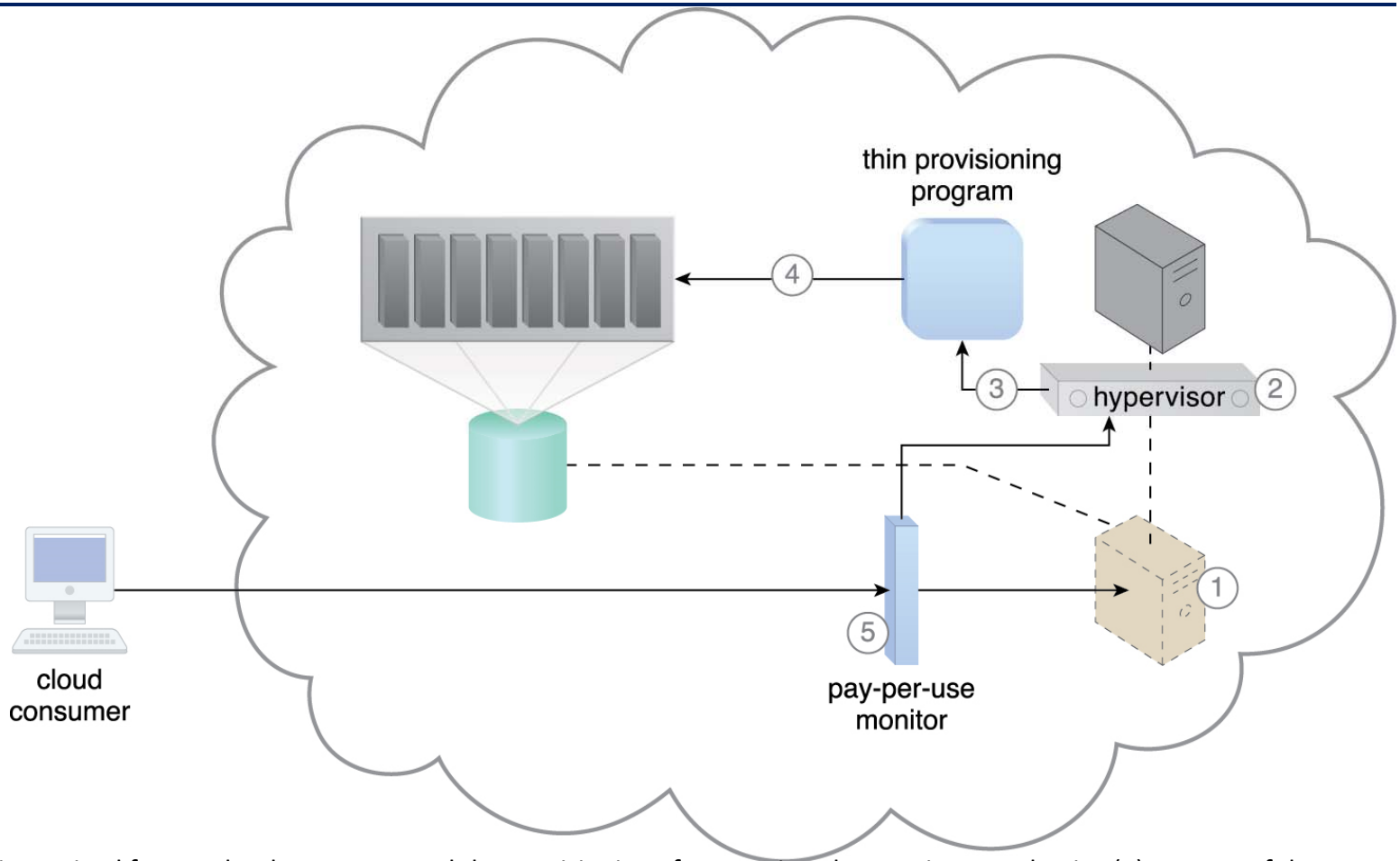
- Establish a dynamic storage provisioning system that ensures that the cloud consumer is granularly billed for the exact amount of storage that it actually uses



Example


1
2
3
4
5
150 GB
150 GB
150 GB
provisioned: 450 GB
used: 0 GB
allocated: 0 GB

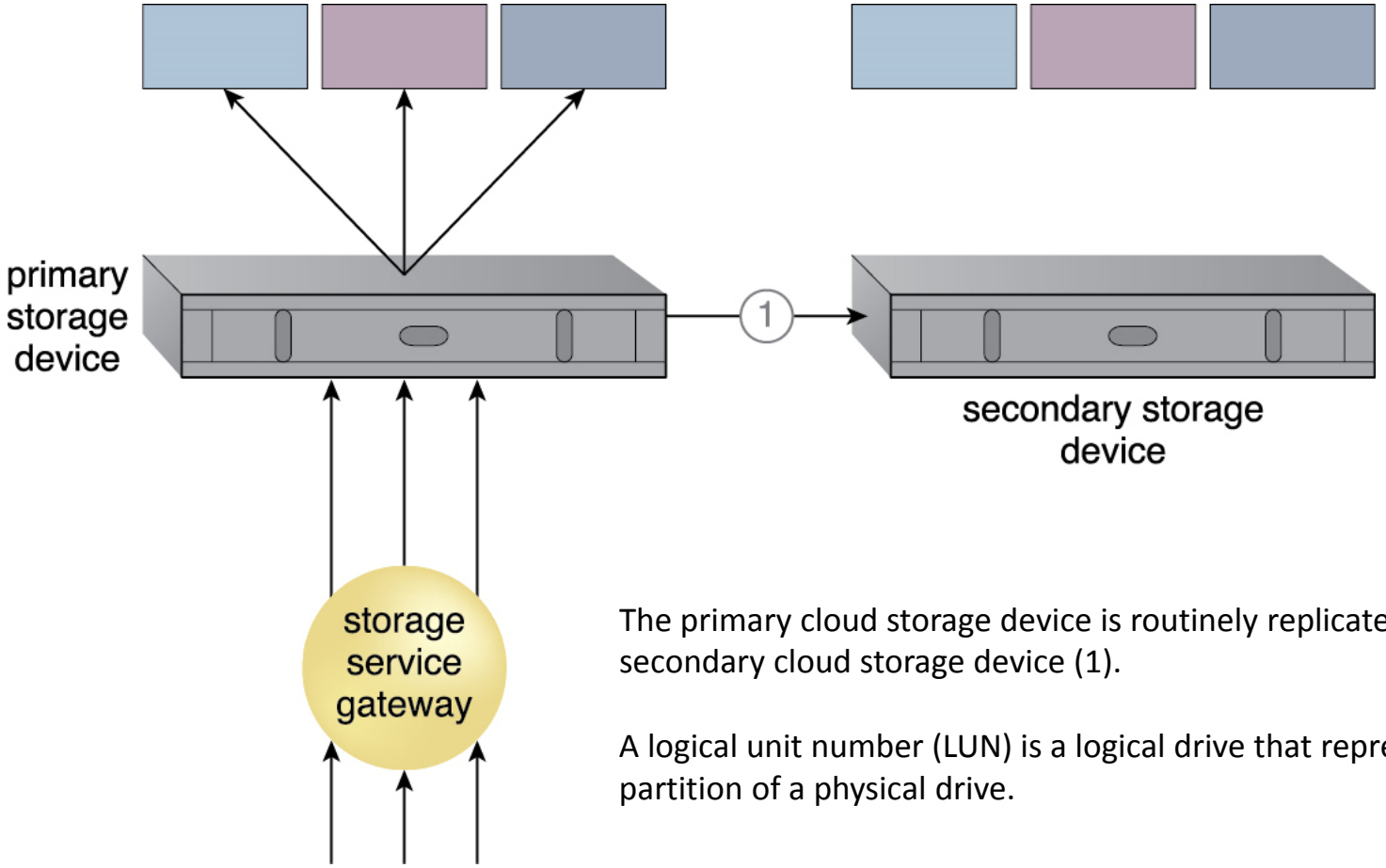
Example



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).

6. Redundant storage architecture

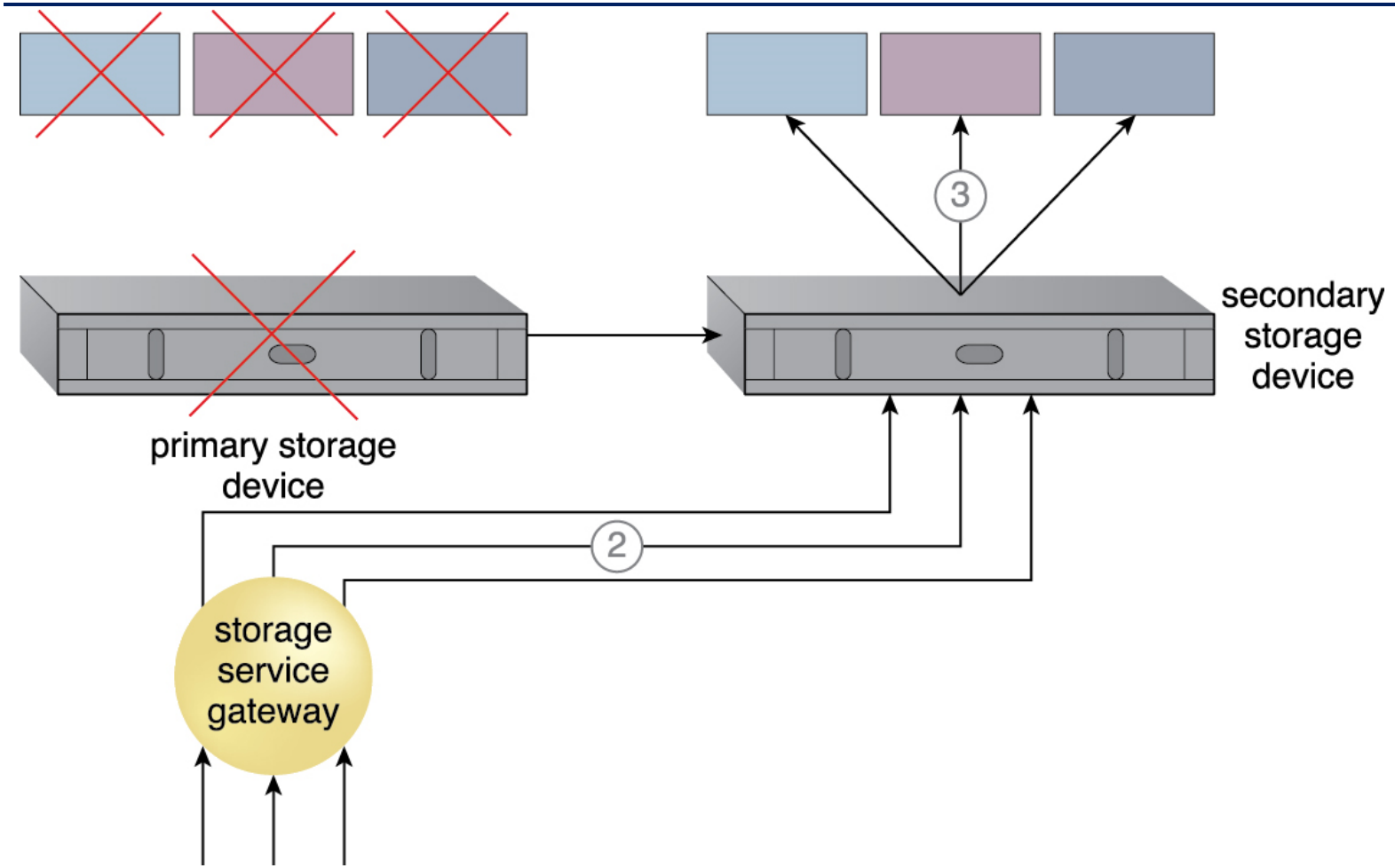
- Introduce a secondary duplicate cloud storage device as part of a failover system that synchronizes its data with the primary cloud storage device



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

A logical unit number (LUN) is a logical drive that represents a partition of a physical drive.

Example



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).