CS 4037
Introduction to Cloud Computing
Lecture 8

Danyal Farhat

FAST School of Computing

NUCES Lahore

Specialized Cloud Mechanisms

Lecture's Agenda

- **Load Balancer**
- SLA Monitor
- Pay-Per-Use Monitor
- Failover System
- Hypervisor

Load Balancer

- A load balancer is programmed or configured with a set of performance and QoS rules and parameters with the general objectives of optimizing IT resource usage, avoiding overloads, and maximizing throughput.
- The load balancer is located on the communication path between the IT resources generating the workload and the IT resources performing the workload processing.
- This mechanism can be designed as a transparent agent that remains hidden from the cloud service consumers, or as a proxy component that abstracts the IT resources performing their workload.

Load Balancer (Cont.)

- The load balancer mechanisms can exist as a:
  - Multi-layer network switch
  - Dedicated hardware appliance
  - Dedicated software-based system (common in server operating systems)
  - Service agent (usually controlled by cloud management software)

Load Balancer (Cont.)

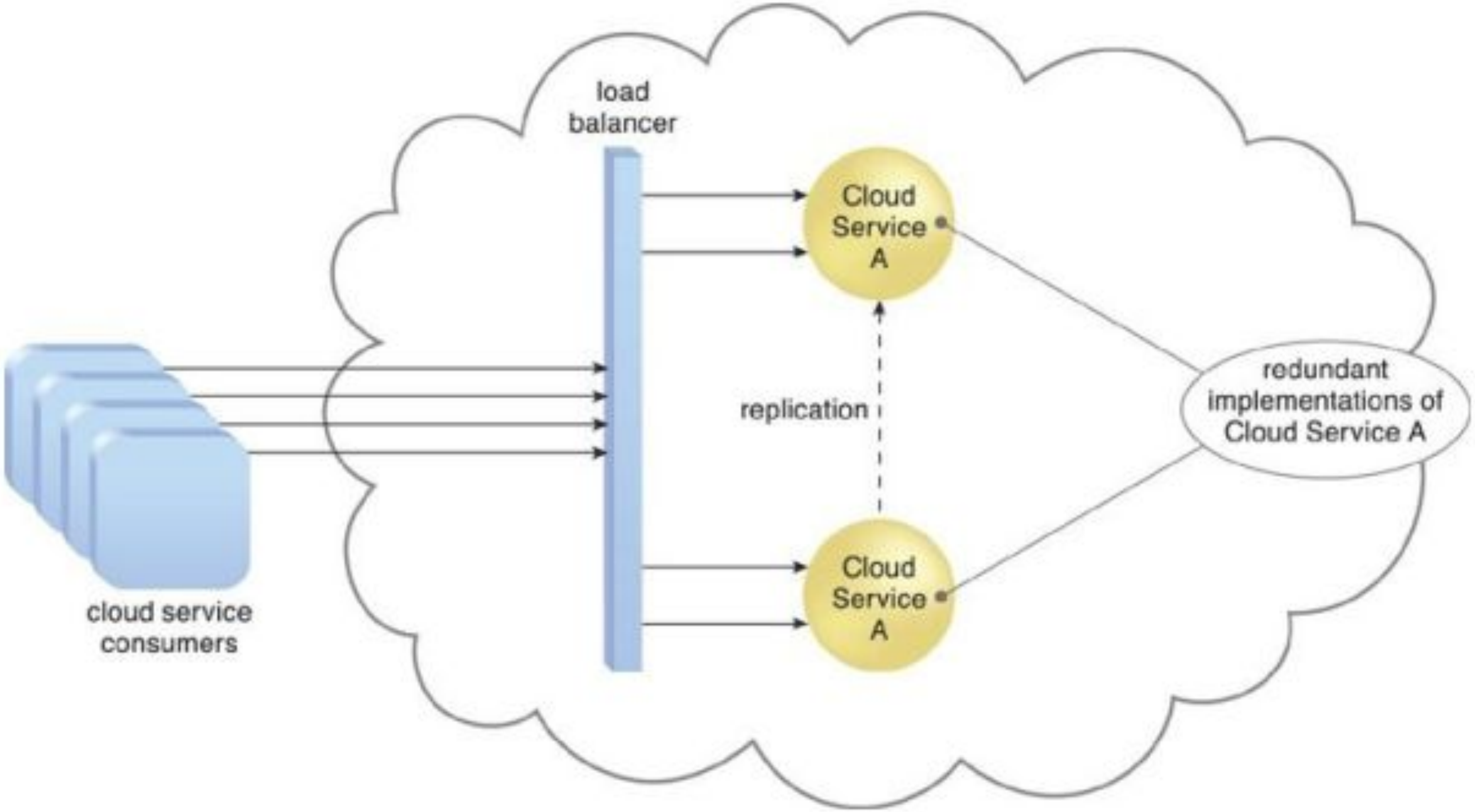


Figure 8.5. A load balancer implemented as a service agent transparently distributes incoming workload request messages across two redundant cloud service implementations, which in turn maximizes performance for the cloud service consumers.

Load Balancer (Cont.)

- Beyond simple division of labor algorithms, load balancers can perform a range of specialized runtime workload distribution functions that include:

Asymmetric Distribution:

- Larger workloads are issued to IT resources with higher processing capacities

Load Balancer (Cont.)

Workload Prioritization:

- Workloads are scheduled, queued, discarded, and distributed workloads according to their priority levels

Content-Aware Distribution:

- Requests are distributed to different IT resources as dictated by the request content

Lecture’s Agenda

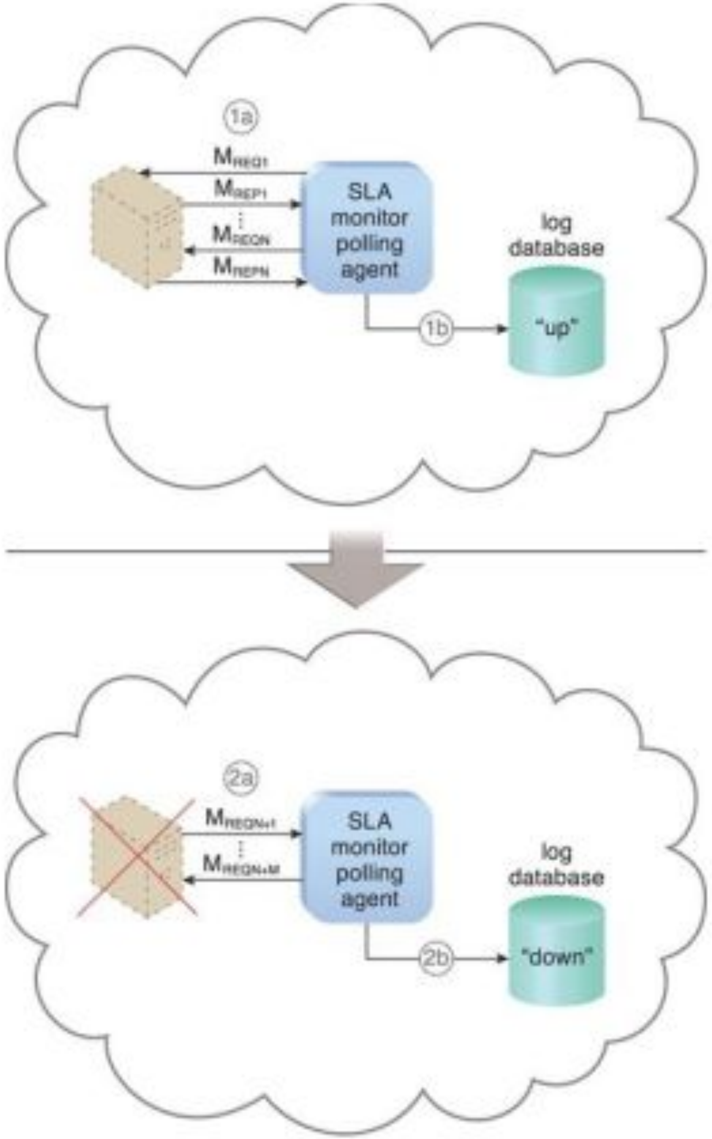
- Load Balancer
- **SLA Monitor**
- Pay-Per-Use Monitor
- Failover System
- Hypervisor

SLA Monitor

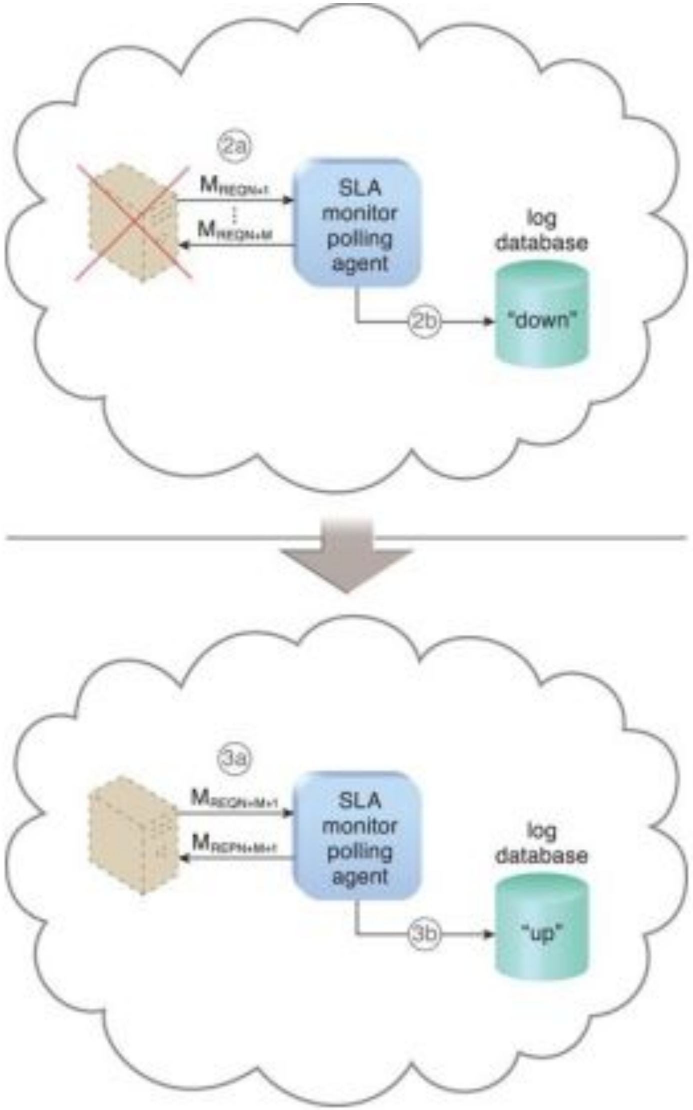
- **The SLA monitor mechanism is used to specifically observe the runtime performance of cloud services to ensure that they are fulfilling the contractual QoS requirements that are published in SLAs.**

- **The data collected by the SLA monitor is processed by an SLA management system to be aggregated into SLA reporting metrics.**

SLA Monitor (Cont.)



SLA Monitor (Cont.)



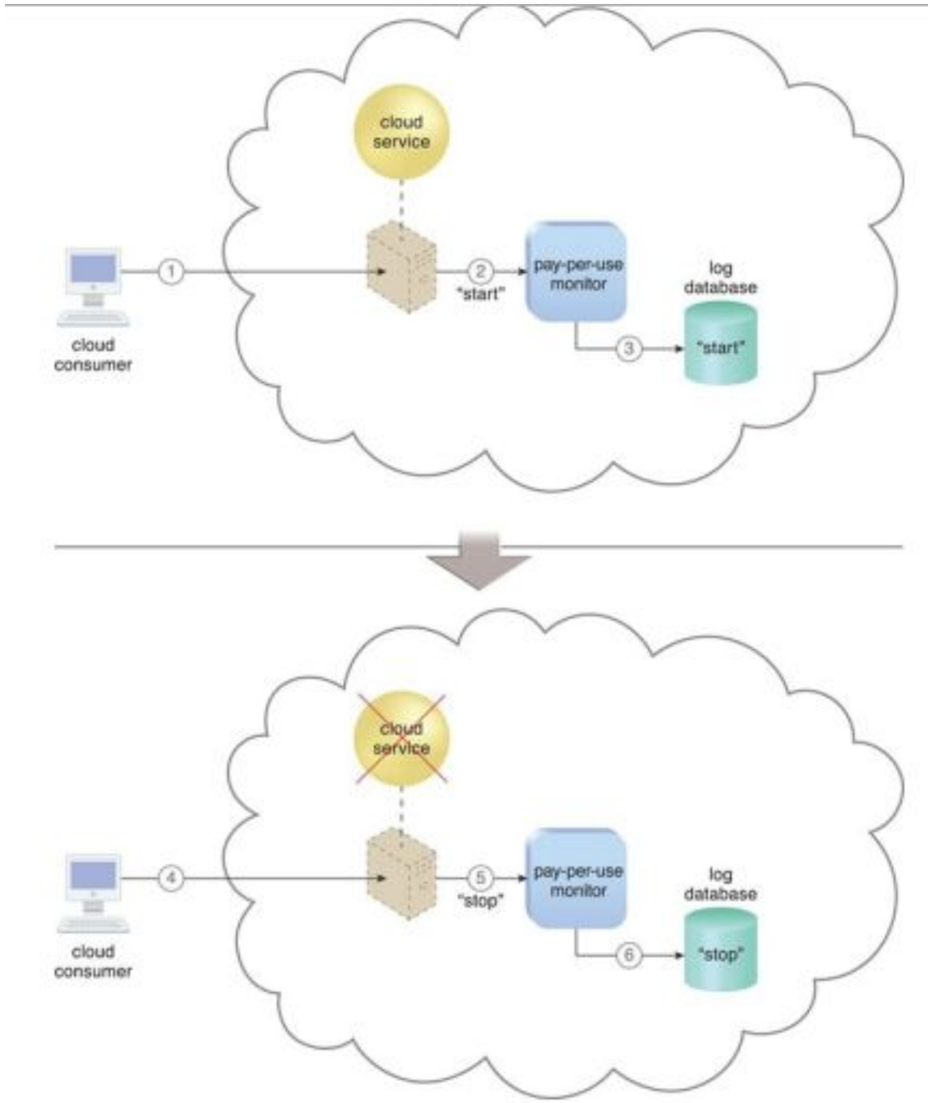
Lecture’s Agenda

- Load Balancer
- SLA Monitor
- **Pay-Per-Use Monitor**
- Failover System
- Hypervisor

Pay-Per-Use Monitor

- The pay-per-use monitor mechanism measures cloud-based IT resource usage in accordance with predefined pricing parameters and generates usage logs for fee calculations and billing purposes.
- The data collected by the pay-per-use monitor is processed by a billing management system that calculates the payment fees.
- Some typical monitoring variables are:
  - Request/response message quantity
  - Transmitted data volume
  - Bandwidth consumption
- The pay-per-use monitor can work as monitoring agent, resource agent and/or polling agent.

Pay-Per-Use Monitor As Resource Agent



Pay-Per-Use Monitor As Resource Agent (Cont.)

- A cloud consumer requests the creation of a new instance of a cloud service (1).
- The IT resource is instantiated and the pay-per-use monitor receives a “start” event notification from the resource software (2).
- The pay-per-use monitor stores the value timestamp in the log database (3).
- The cloud consumer later requests that the cloud service instance be stopped (4).
- The pay-per-use monitor receives a “stop” event notification from the resource software (5) and stores the value timestamp in the log database (6).

Pay-Per-Use Monitor As Monitoring Agent

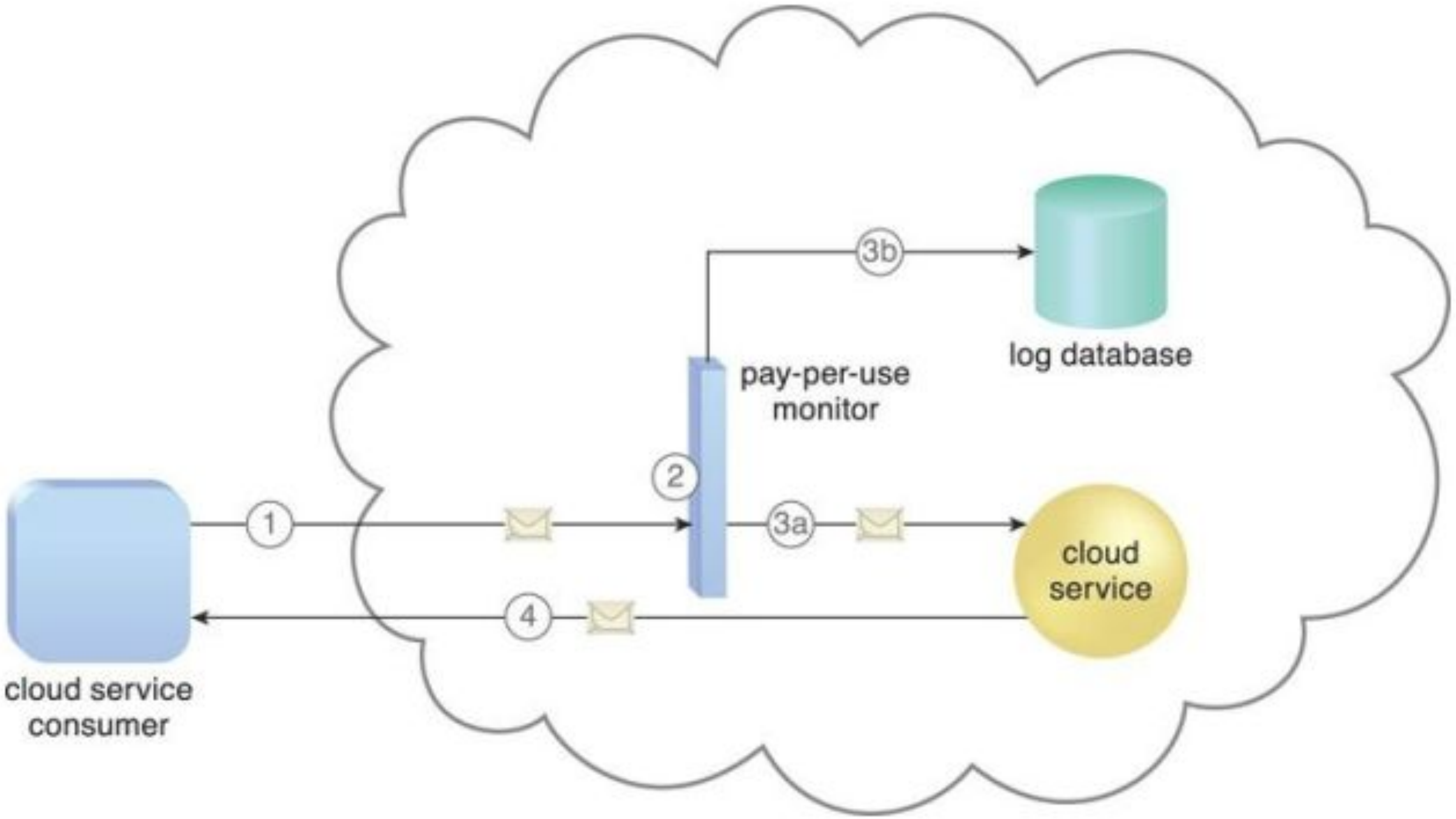


Figure 8.13. A cloud service consumer sends a request message to the cloud service (1). The pay-per-use monitor intercepts the message (2), forwards it to the cloud service (3a), and stores the usage information in accordance with its monitoring metrics (3b). The cloud service forwards the response messages back to the cloud service consumer to provide the requested service (4).

Lecture’s Agenda

- Load Balancer
- SLA Monitor
- Pay-Per-Use Monitor
- **Failover System**
- Hypervisor

Failover System

- The failover system mechanism is used to increase the reliability and availability of IT resources by using established clustering technology to provide redundant implementations.
- A failover system is configured to automatically switch over to a redundant or standby IT resource instance whenever the currently active IT resource becomes unavailable.
- Failover systems are commonly used for mission-critical programs and reusable services that can introduce a single point of failure for multiple applications.

Active-Active Failover System

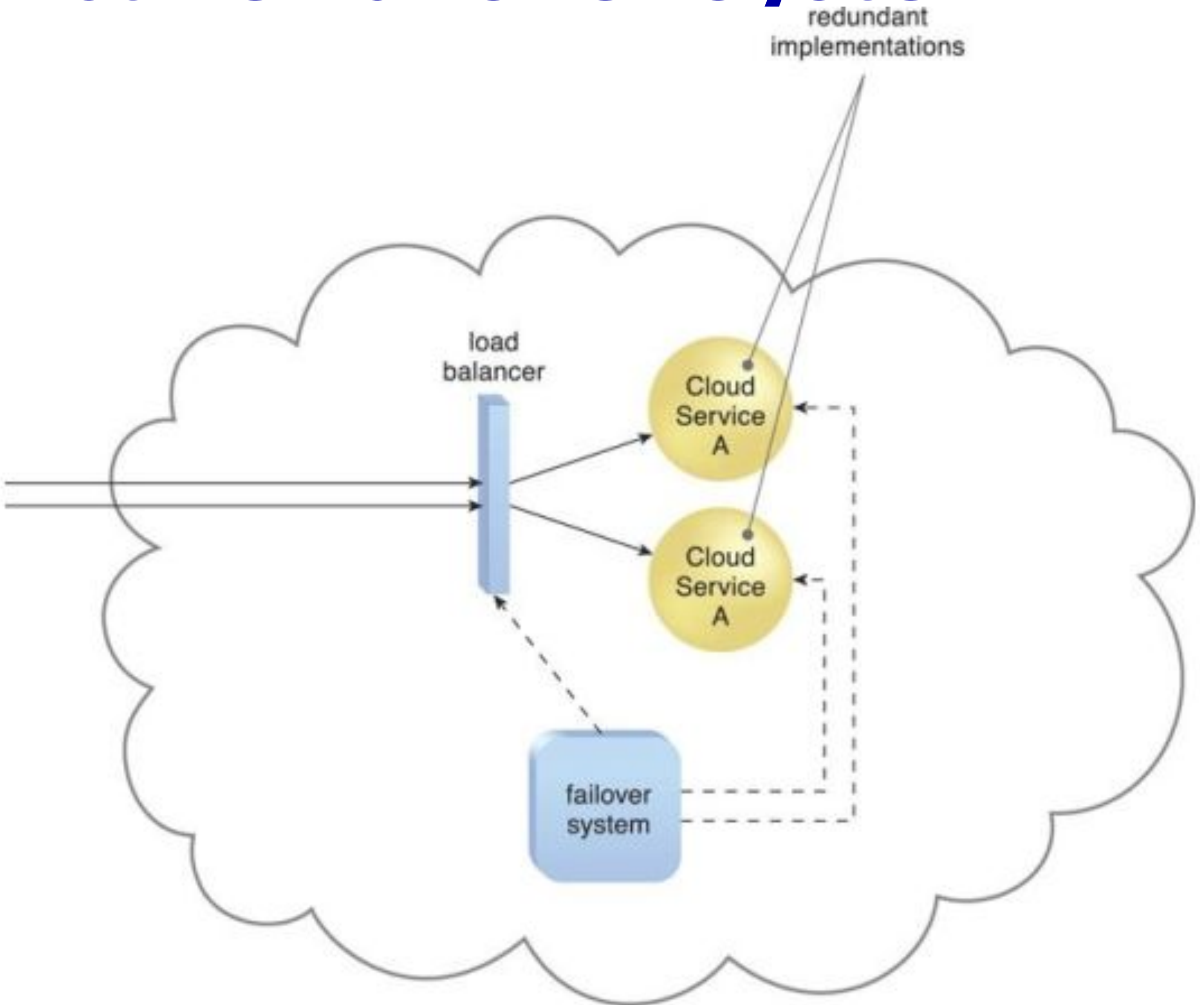


Figure 8.17. The failover system monitors the operational status of Cloud Service A.

Active-Active Failover System (Cont.)

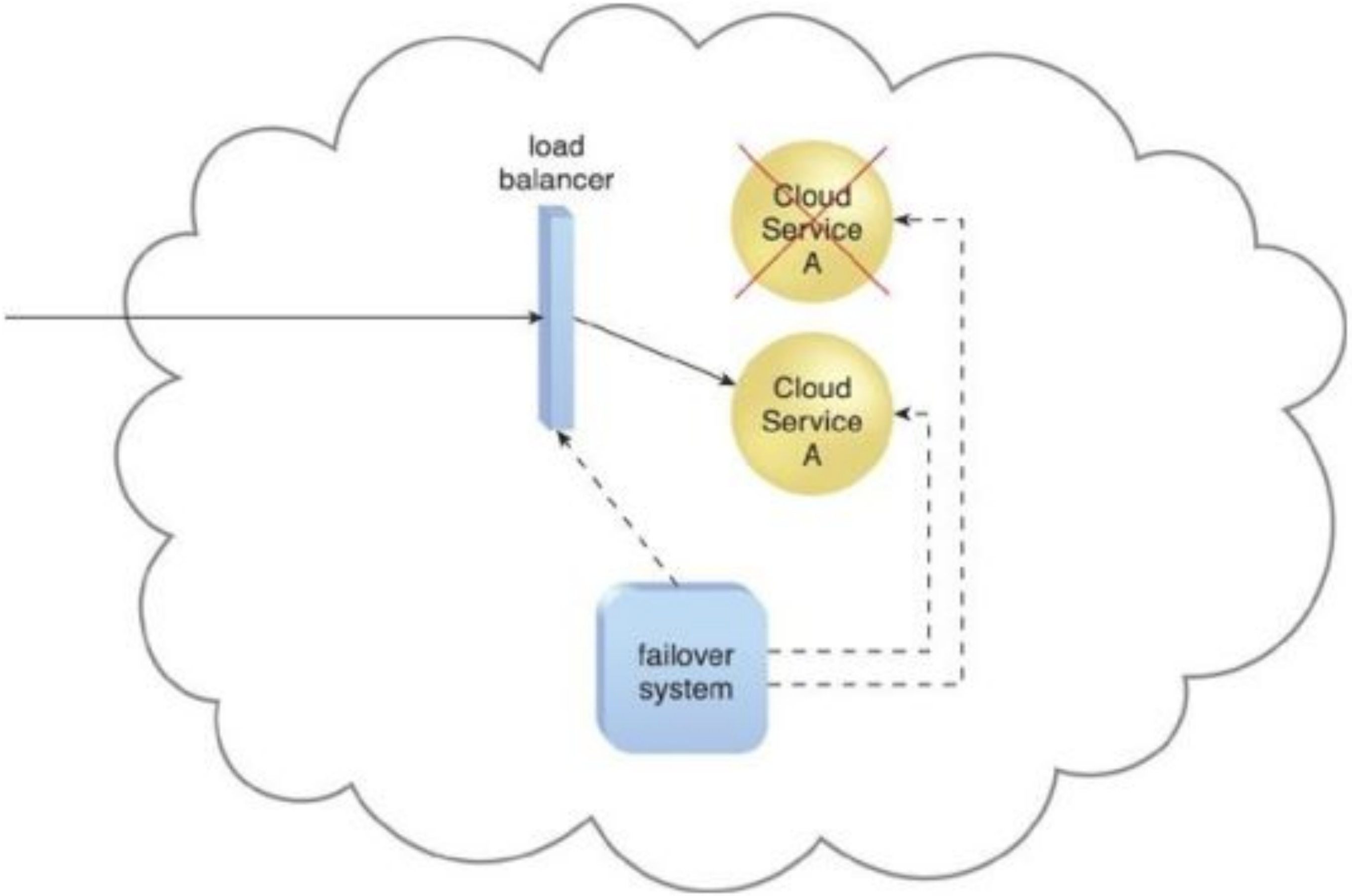


Figure 8.18. When a failure is detected in one Cloud Service A implementation, the failover system commands the load balancer to switch over the workload to the redundant Cloud Service A implementation.

Active-Active Failover System (Cont.)

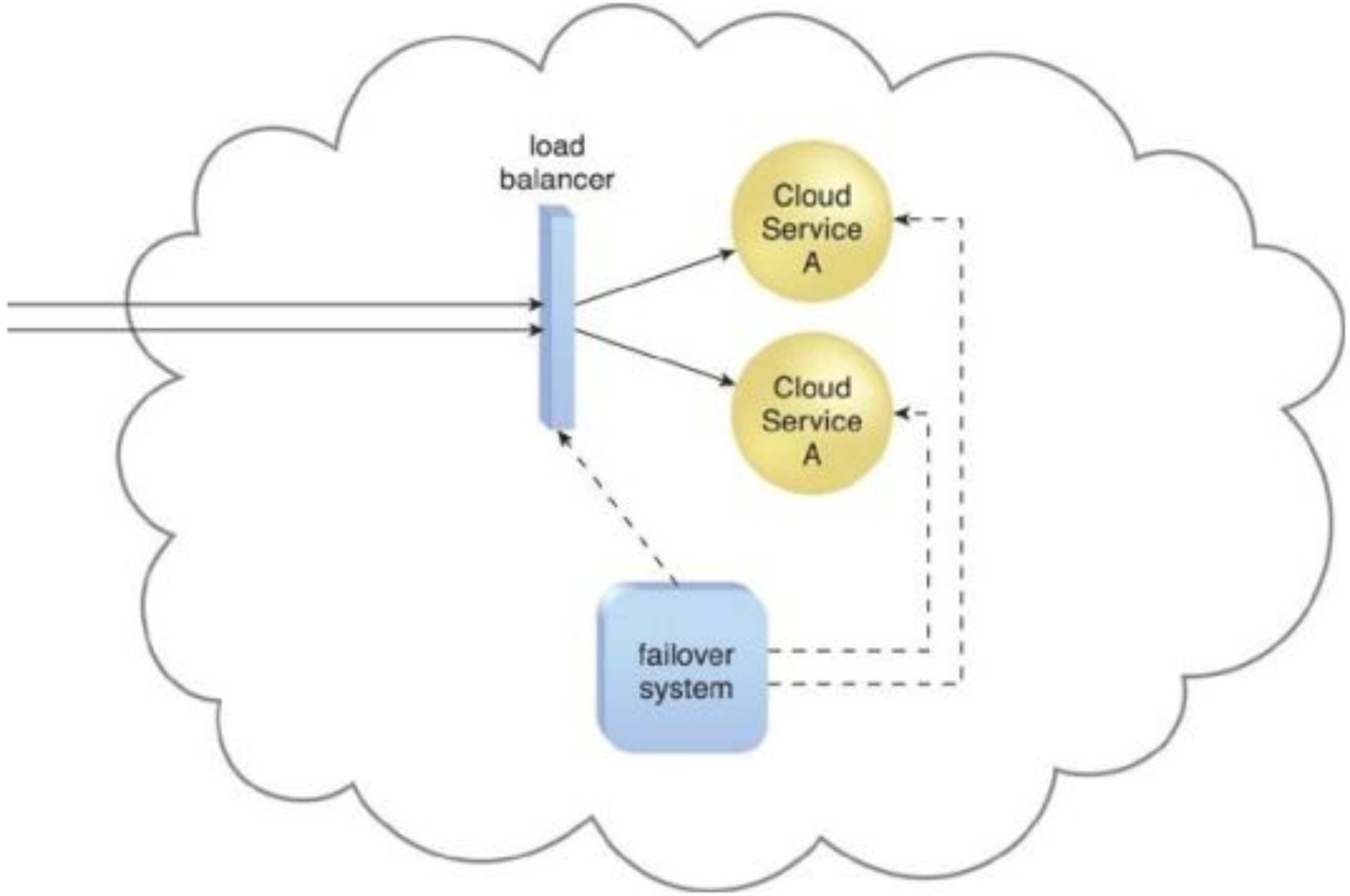


Figure 8.19. The failed Cloud Service A implementation is recovered or replicated into an operational cloud service. The failover system now commands the load balancer to distribute the workload again.

Active-Passive Failover System

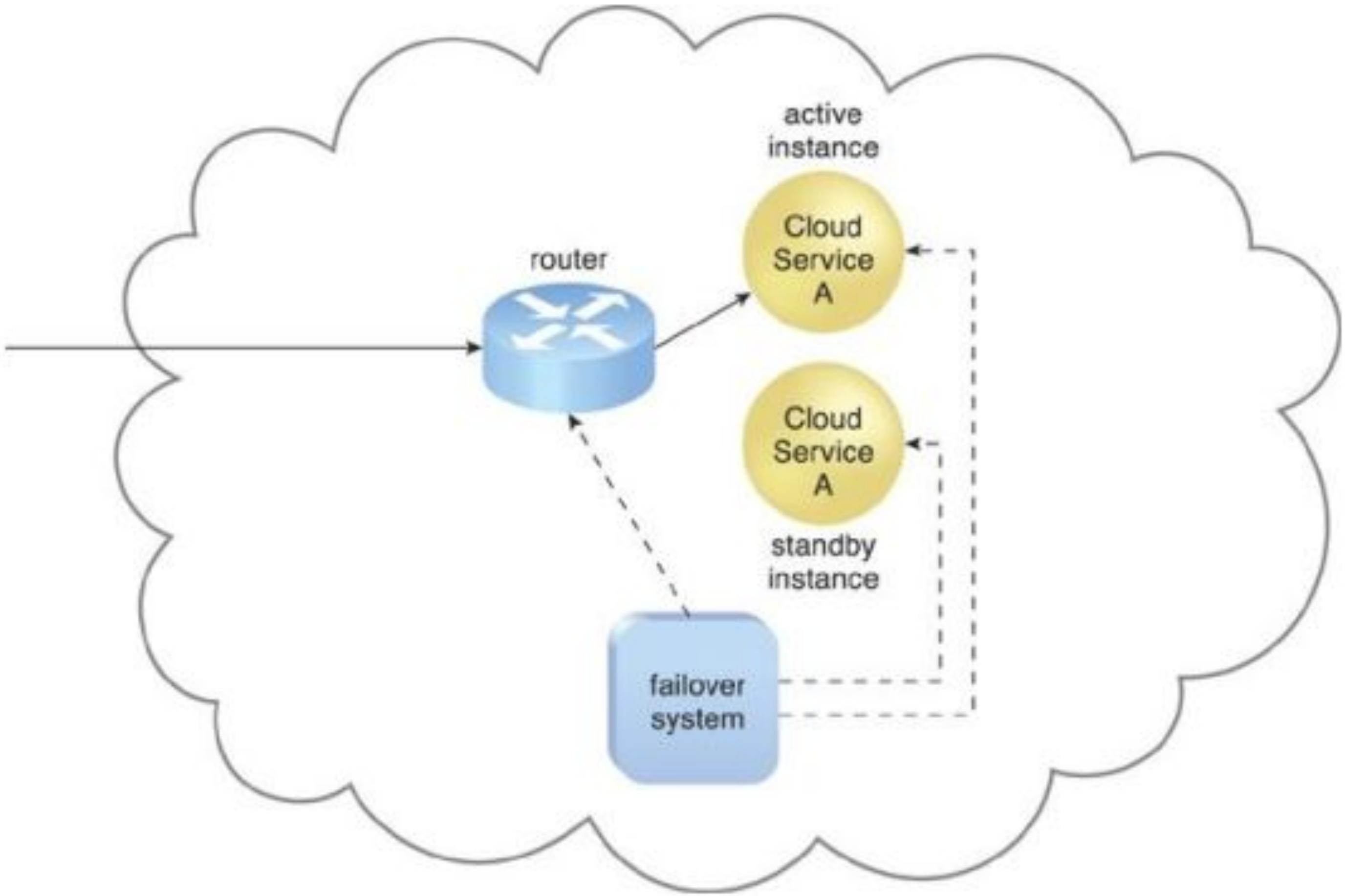


Figure 8.20. The failover system monitors the operational status of Cloud Service A. The Cloud Service A implementation acting as the active instance is receiving cloud service consumer requests.

Active-Passive Failover System (Cont.)

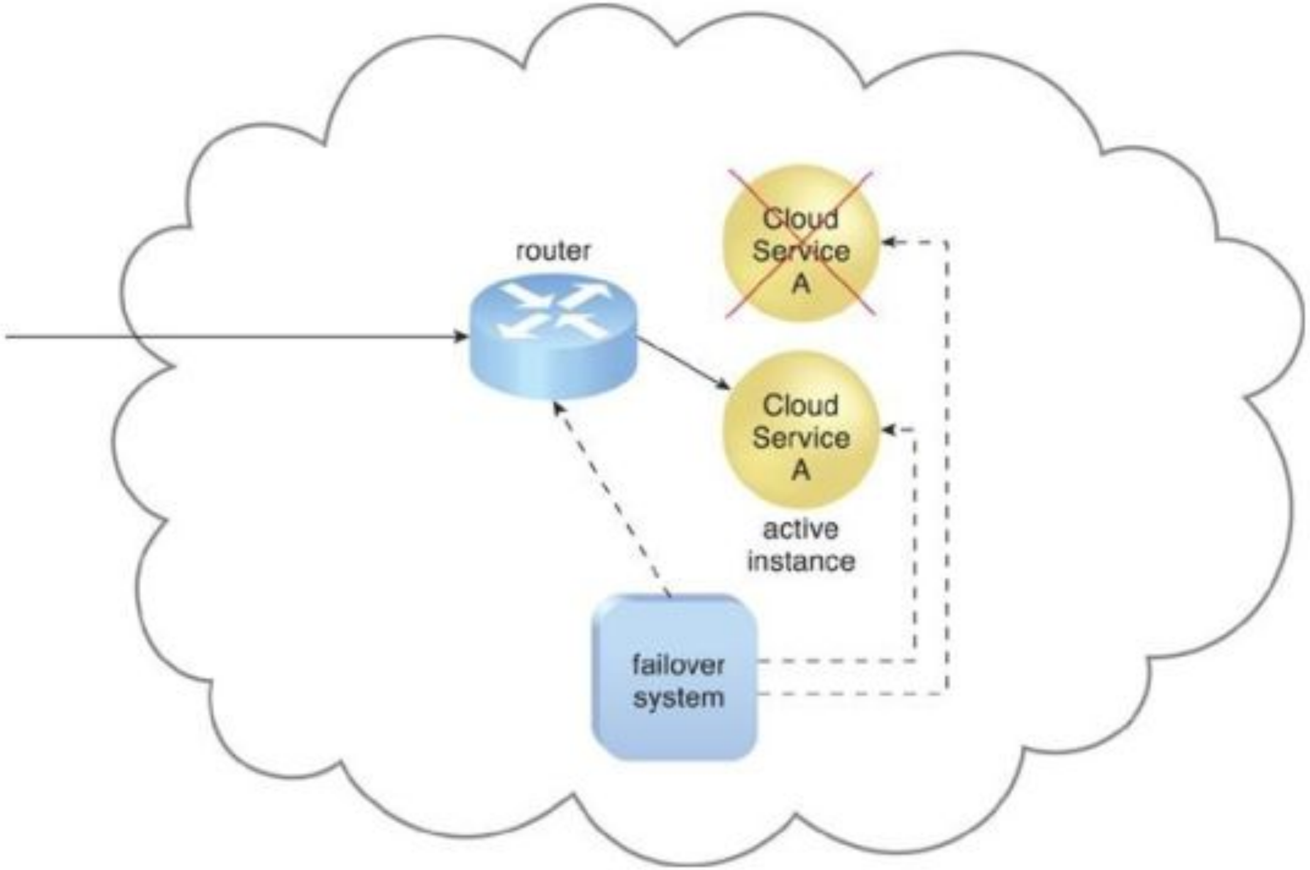


Figure 8.21. The Cloud Service A implementation acting as the active instance encounters a failure that is detected by the failover system, which subsequently activates the inactive Cloud Service A implementation and redirects the workload toward it. The newly invoked Cloud Service A implementation now assumes the role of active instance.

Active-Passive Failover System (Cont.)

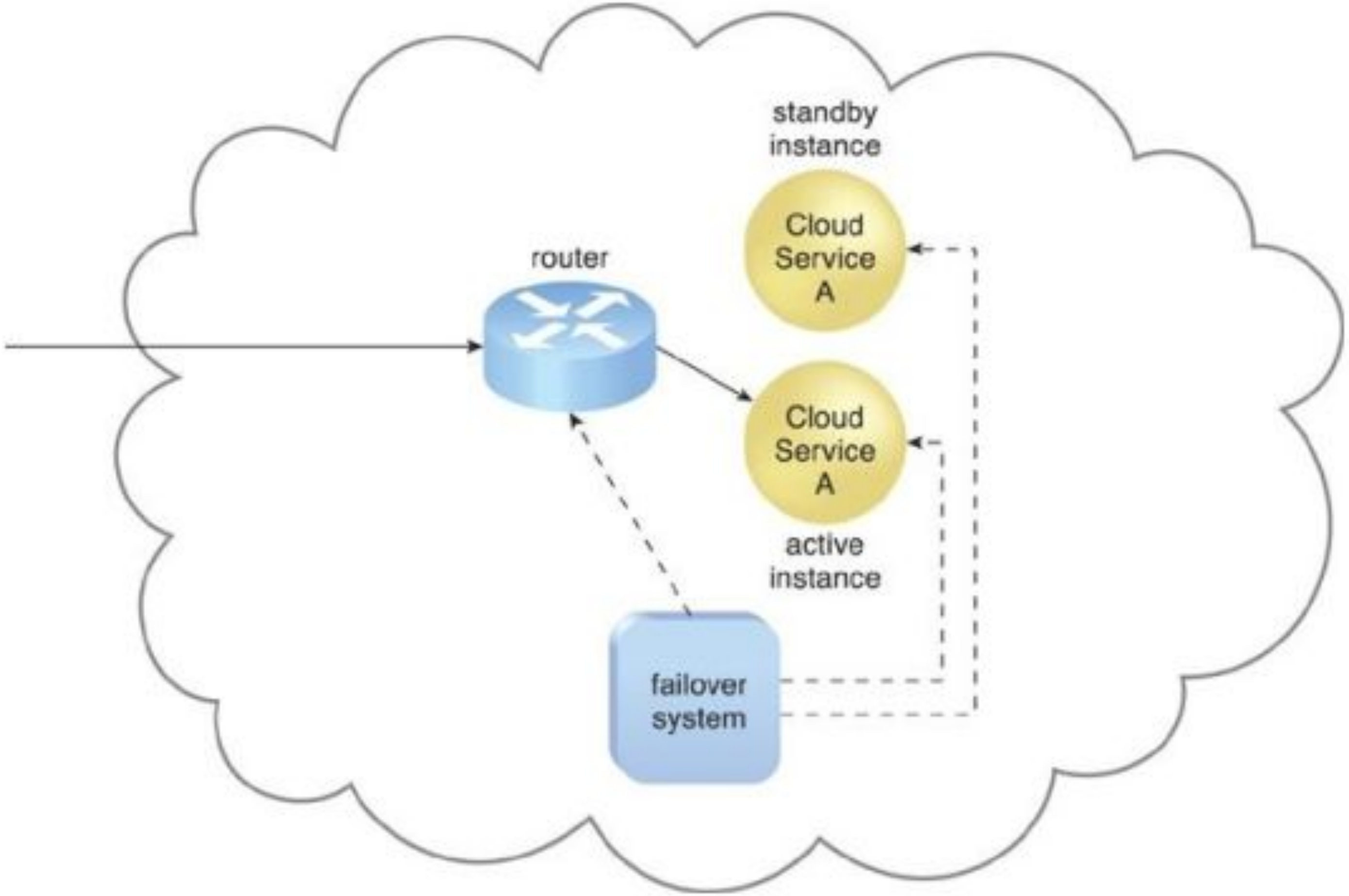


Figure 8.22. The failed Cloud Service A implementation is recovered or replicated an operational cloud service, and is now positioned as the standby instance, while the previously invoked Cloud Service A continues to serve as the active instance.

Lecture’s Agenda

- Load Balancer
- SLA Monitor
- Pay-Per-Use Monitor
- Failover System
- **Hypervisor**

Hypervisor

- The hypervisor mechanism is a fundamental part of virtualization infrastructure that is primarily used to generate virtual server instances of a physical server.
- A hypervisor is generally limited to one physical server and can therefore only create virtual images of that server.
- A hypervisor has limited virtual server management features, such as increasing the virtual server’s capacity or shutting it down.
- The VIM provides a range of features for administering multiple hypervisors across physical servers.
  - In some cloud environments, hypervisor and VIM are same software.

Hypervisor (Cont.)

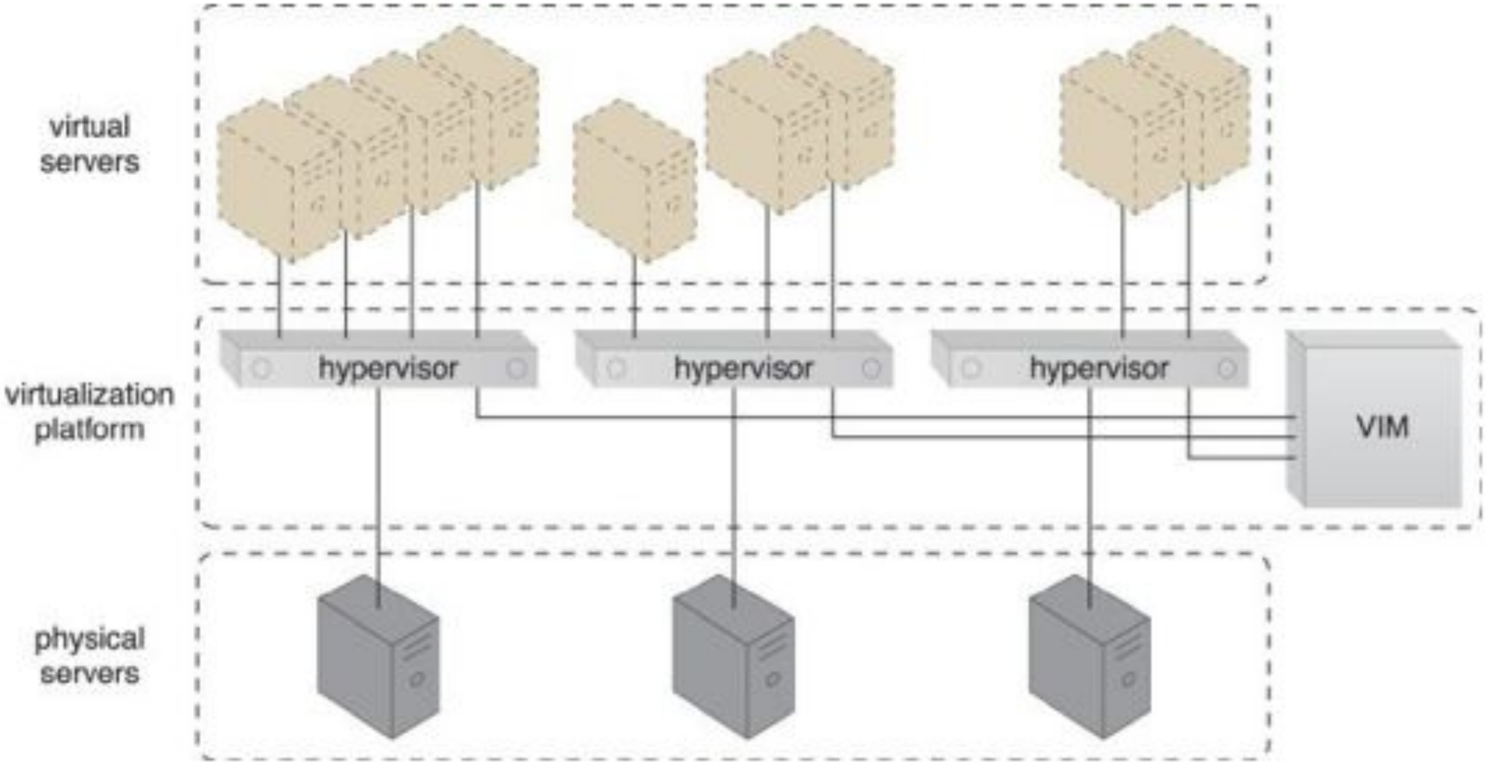


Figure 8.27. Virtual servers are created via individual hypervisor on individual physical servers. All three hypervisors are jointly controlled by the same VIM.

Hypervisor (Cont.)

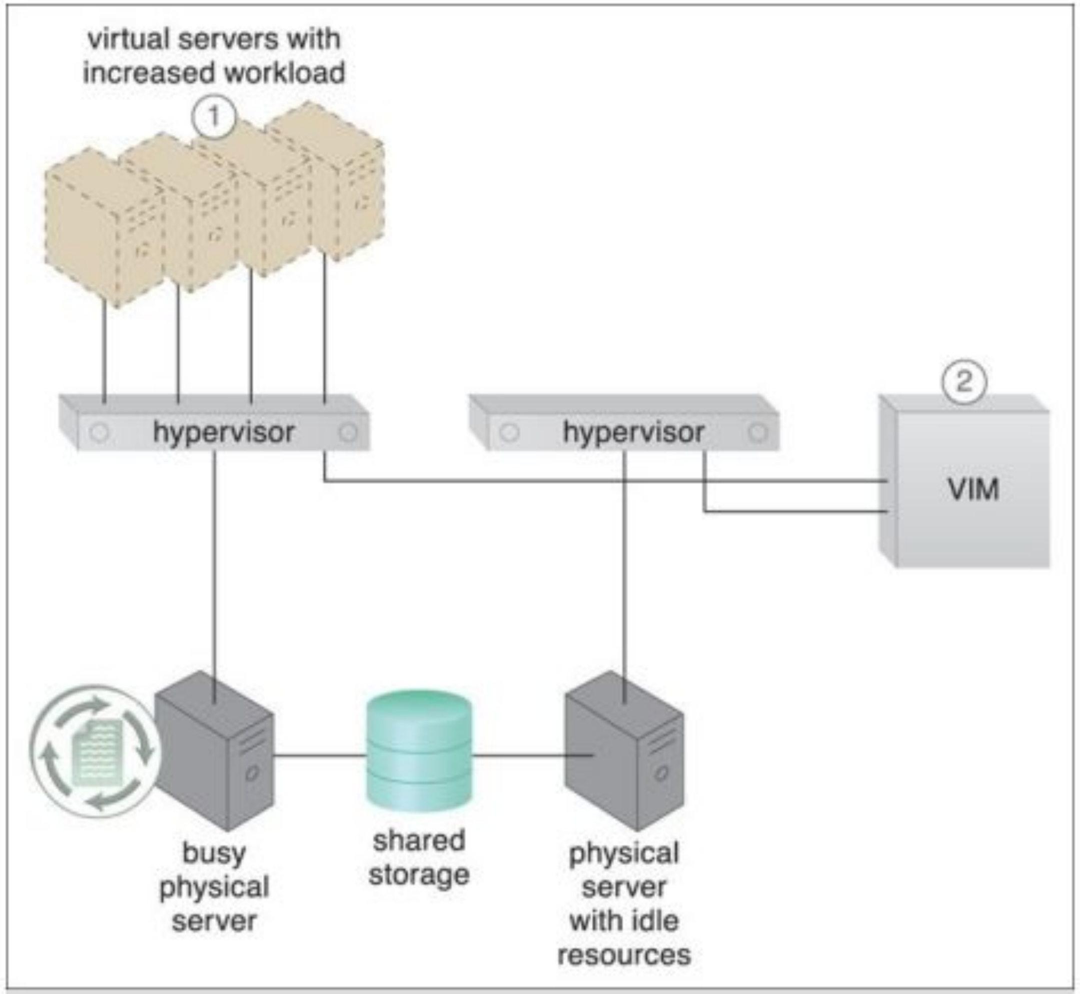

virtual servers with increased workload
1
hypervisor
hypervisor
2
VIM
busy physical server
shared storage
physical server with idle resources

Figure 8.28. A virtual server capable of auto-scaling experiences an increase in its workload (1). The VIM decides that the virtual server cannot scale up because its underlying physical server host is being used by other virtual servers (2).

Hypervisor (Cont.)

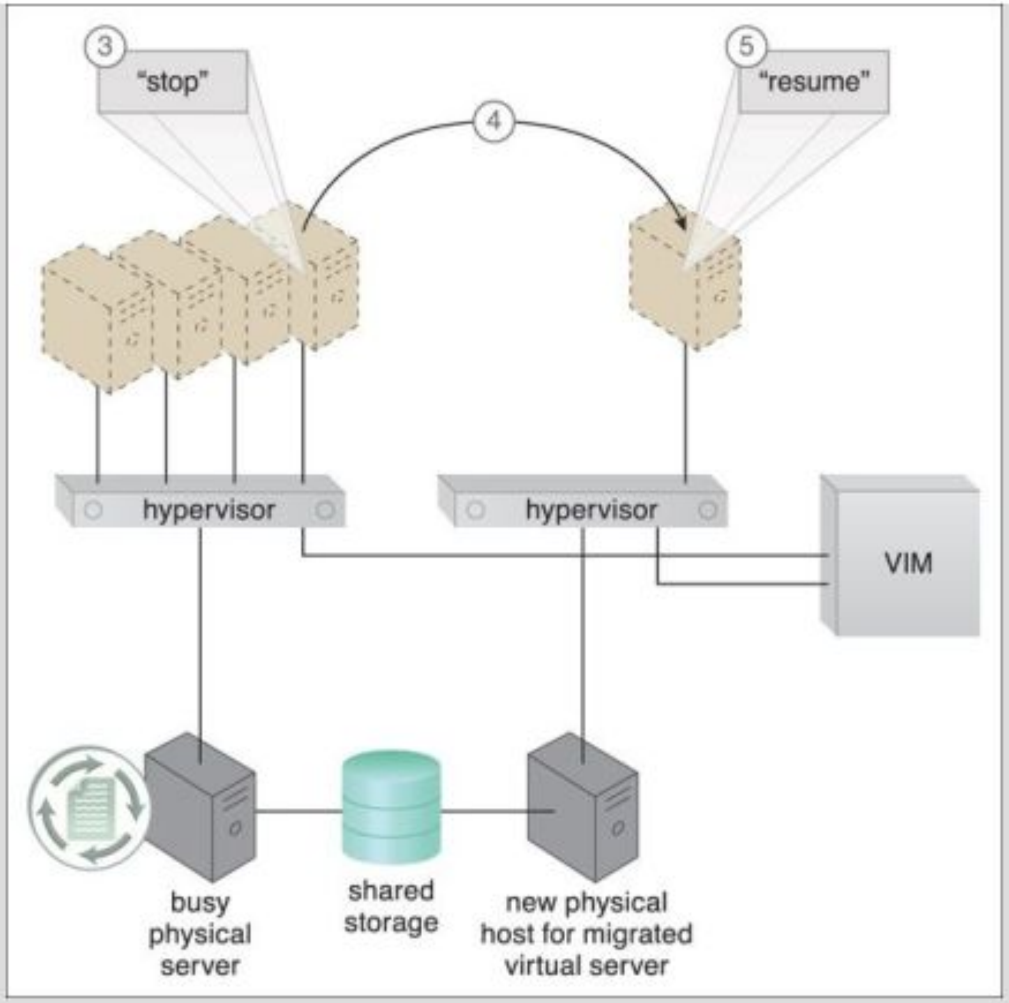


Figure 8.29. The VIM commands the hypervisor on the busy physical server to suspend execution of the virtual server (3). The VIM then commands the instantiation of the virtual server on the idle physical server. State information (such as dirty memory pages and processor registers) is synchronized via a

Additional Resources

- **Cloud Computing – Concepts, Technology, and Architecture by Thomas Erl, Zaigham Mahmood, and Ricardo Puttini**
  - **Chapter 8: Specialized Cloud Mechanisms**

Questions?