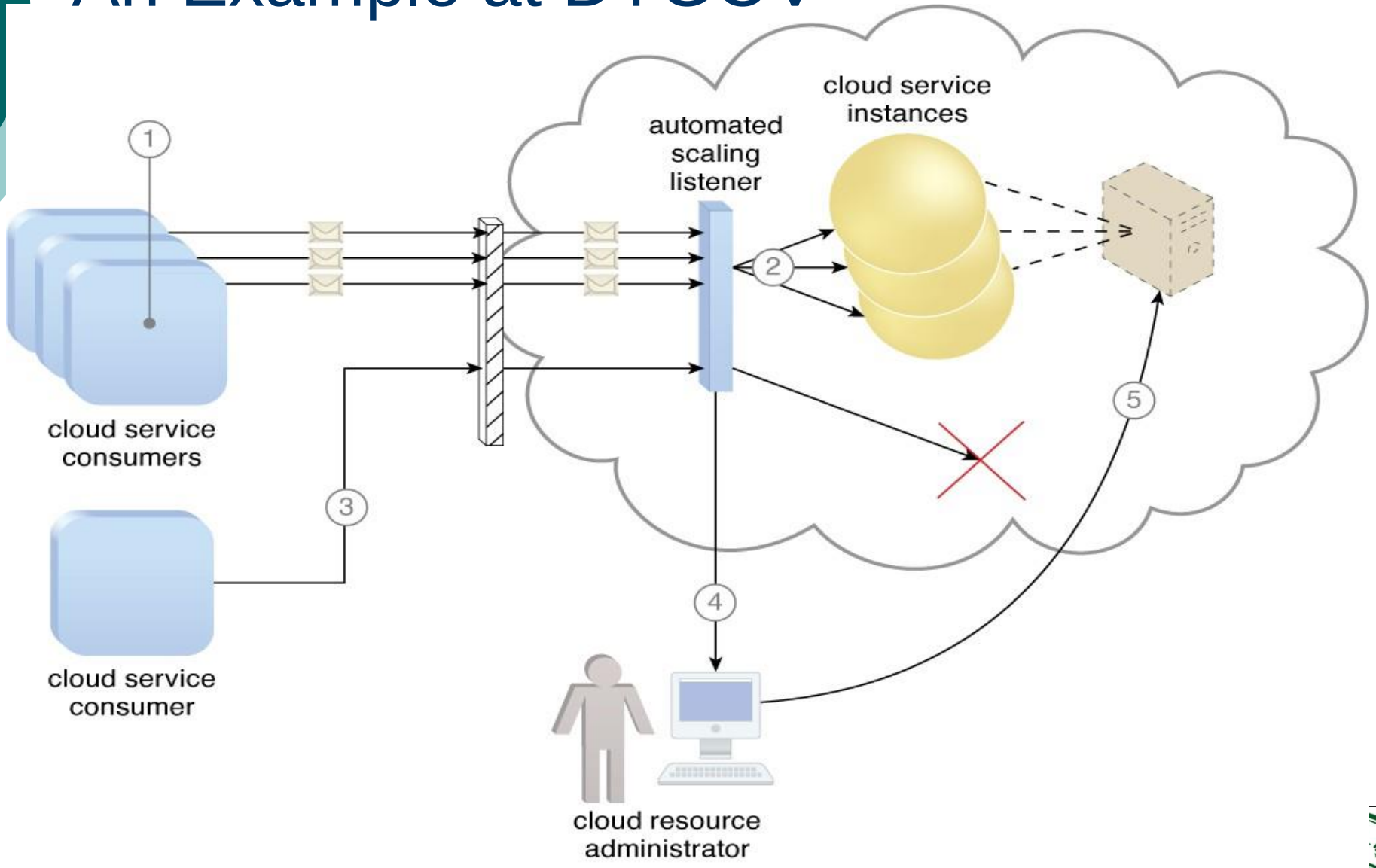
第8章 特殊云机制

§ 8.1 自动伸缩监听器
§ 8.2 负载均衡器
§ 8.3 SLA监控器
§ 8.4 按使用付费监控器
§ 8.5 审计监控器
§ 8.6 故障转移系统
§ 8.7 虚拟机监控器
§ 8.8 资源集群
§ 8.9 多设备代理
§ 8.10 状态管理数据库

§8.1自动伸缩监听器

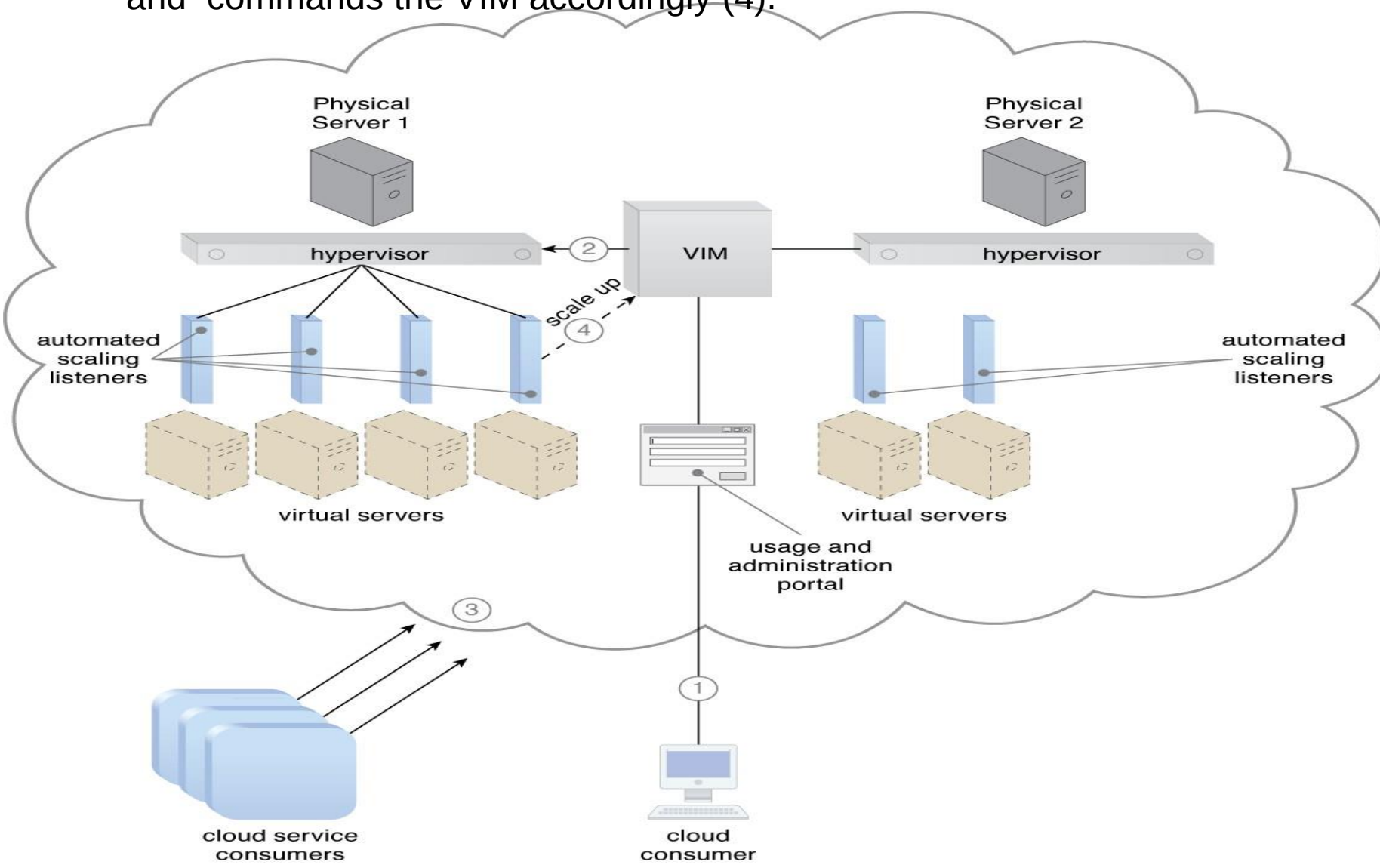
- 自动伸缩监听器(automated scaling listener mechanism)
  - 一个服务代理，监控和追踪云服务用户和云服务之间的通讯，用以动态自动伸缩。
- 通常部署在靠近防火墙的位置，来自动追踪负载状态信息。
- 对应负载波动的条件，可以提供不同类型的响应：
  - 自动伸缩IT资源(auto-scaling)：根据事先定义的参数
  - 自动通知云用户(auto-notification)：负载过高或过低时

An Example at DTGOV

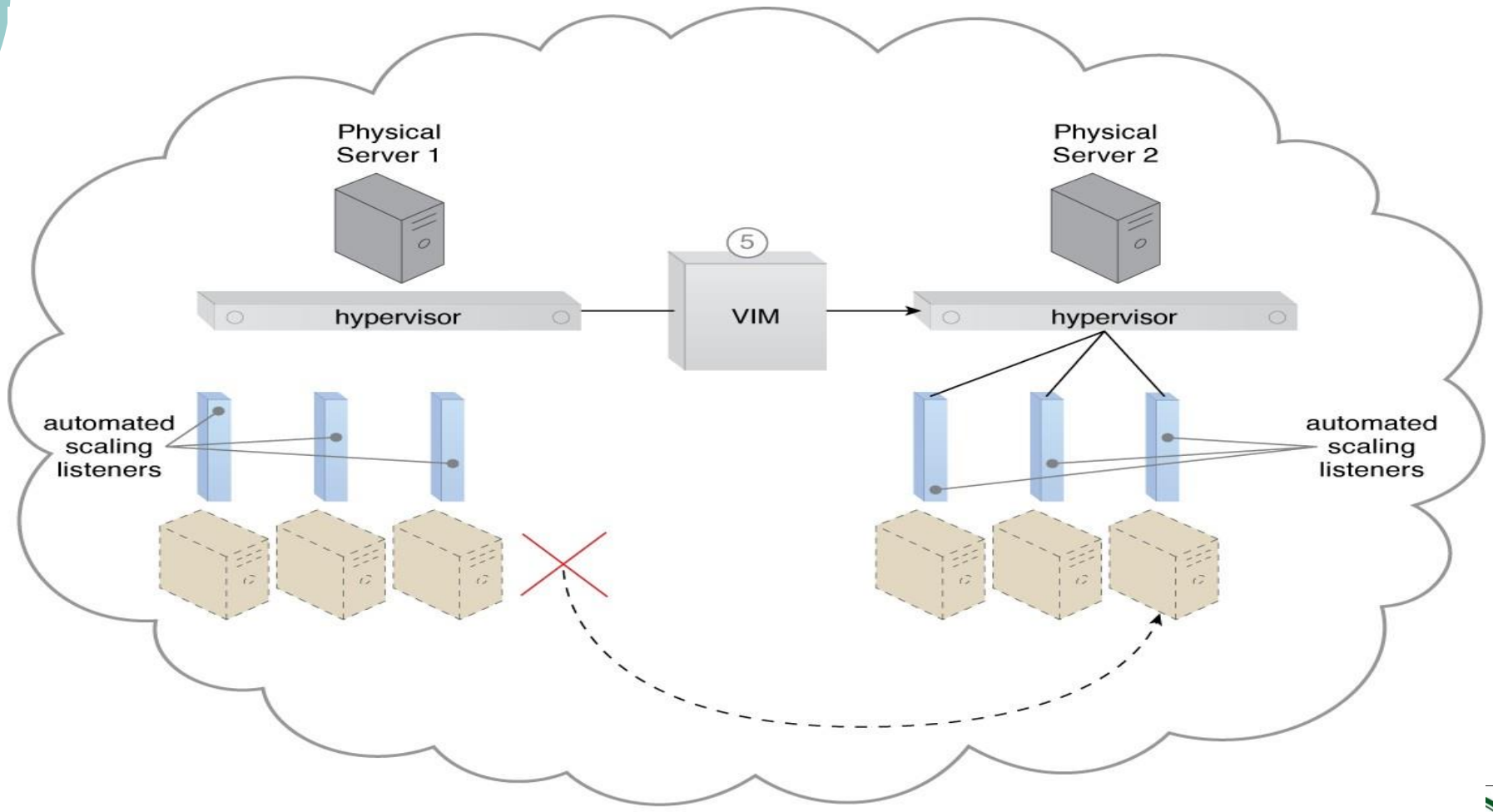


The fourth cloud service consumer is rejected due to workload constraints; Administrator can change the constraint upon notification.

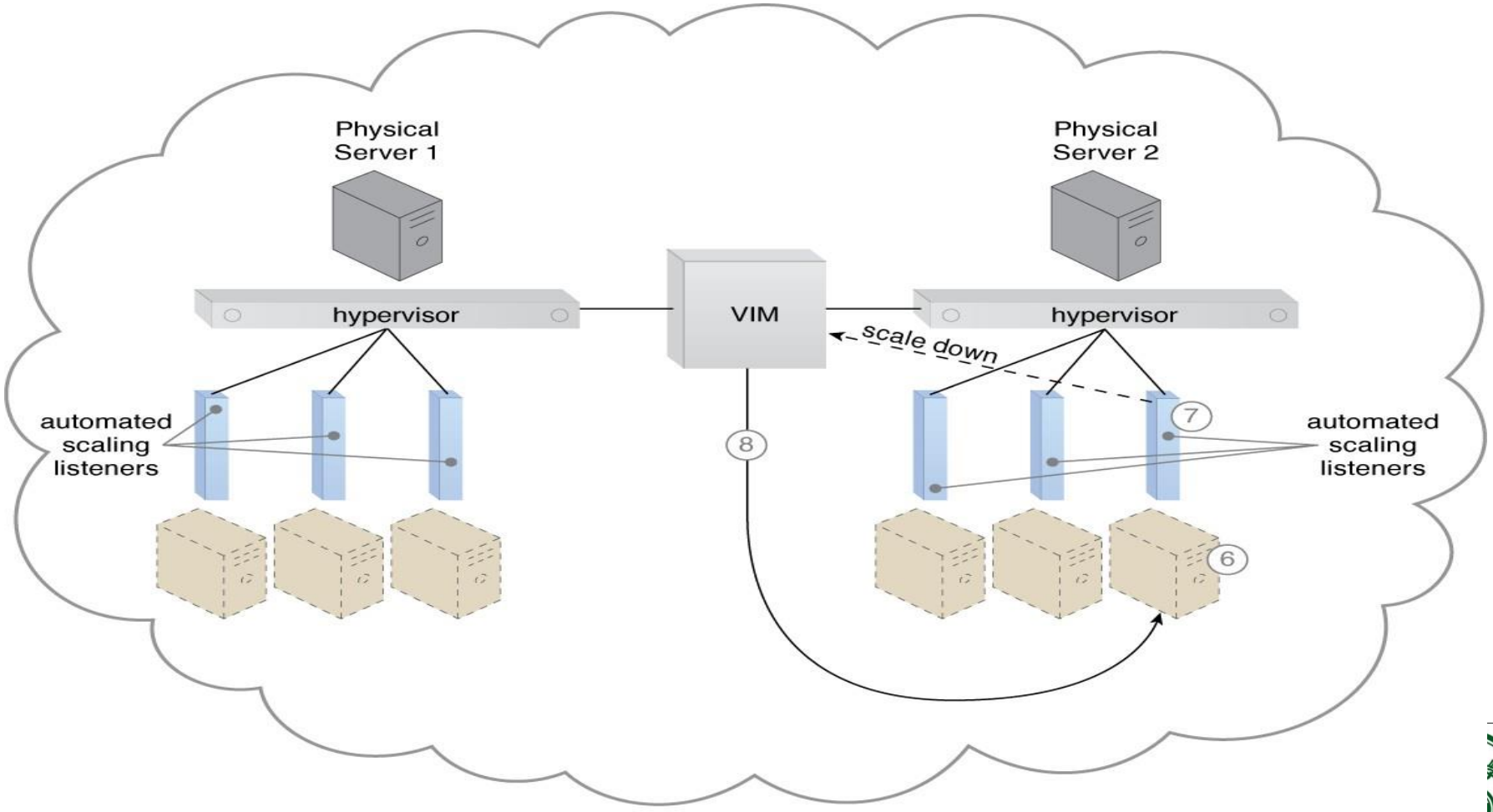
A virtual server usage increased over 80% of the CPU capacity for 60 consecutive seconds (3).
The automated scaling listener detects the need to scale up and commands the VIM accordingly (4).



The VIM determines that scaling up on Physical Server 1 is not possible, and proceeds to live migrate it to Physical Server 2 (5).



The virtual server's CPU/RAM usage remains below 15% capacity for 60 consecutive seconds (6).
The automated scaling listener detects the need to scale down and commands the VIM (7), which scales down the virtual server (8) while it remains active on Physical Server 2.



§8.2负载均衡器

- 负载均衡器(load balancer)机制是一个运行时代理
- 主要用于把负载在两个或更多的IT 资源上做负载
- 负载均衡器可以执行不同的运行时负载分配功能：
  - 非对称分配(Asymmetric distribution)
    - Direct Server Return
  - 负载优先级(Workload prioritization)
  - 上下文感知的分配(Content-aware distribution)
    - Relationship among requests/jobs

负载均衡器的实体和部署

- 负载均衡器通常位于产生负载的IT资源和执行负载处理的IT资源之间的通讯路径上。
- 负载均衡器可以被设计成一个透明的代理或是一个代理的组件

负载均衡器机制的载体

- 多层网络交换机(multi-layer network switch)
- 专门的硬件设备(dedicated hardware appliance)
- 专门的基于软件的系统(dedicated software-based system (in server OS))
- 服务代理(service agent (controlled by cloud management software))

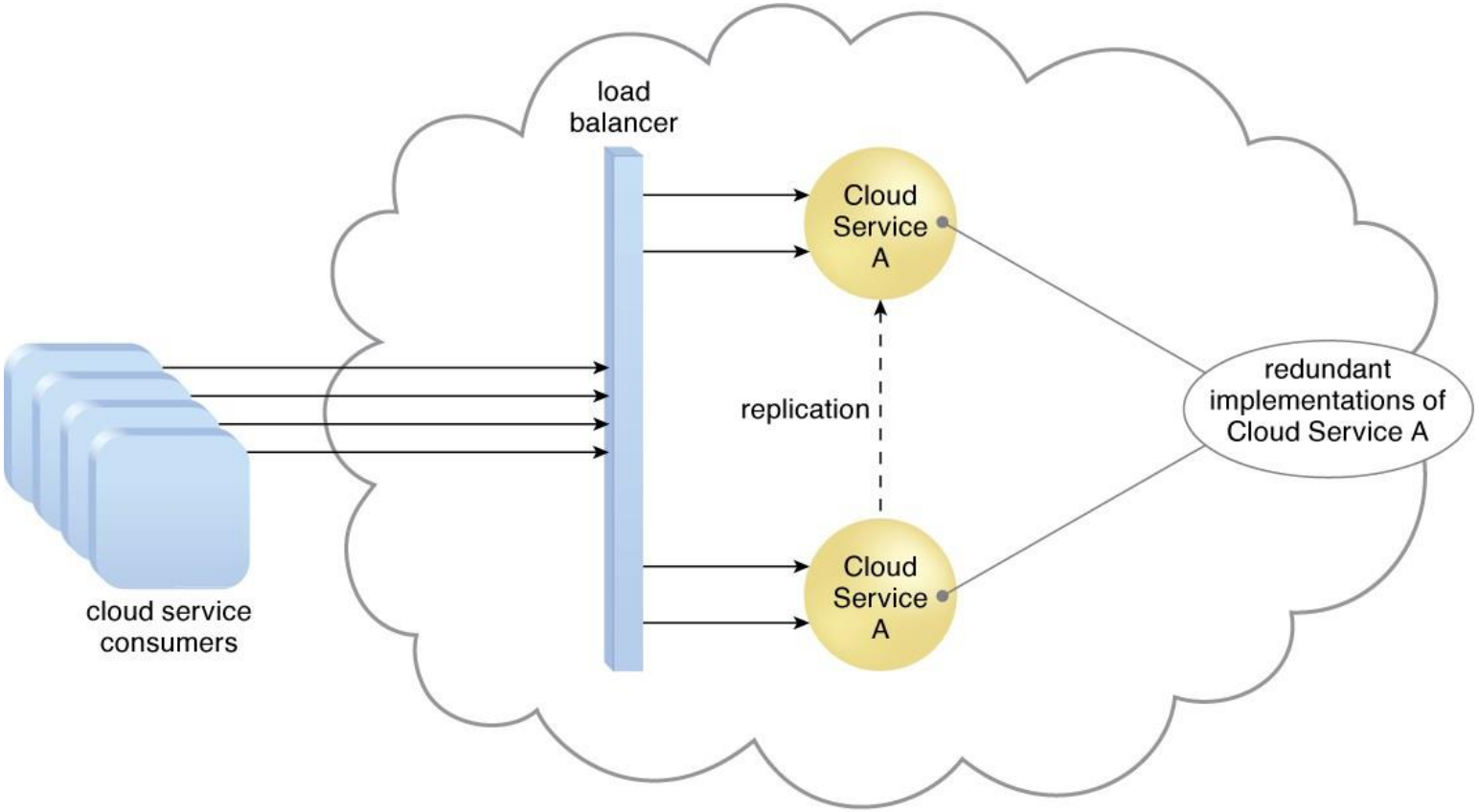
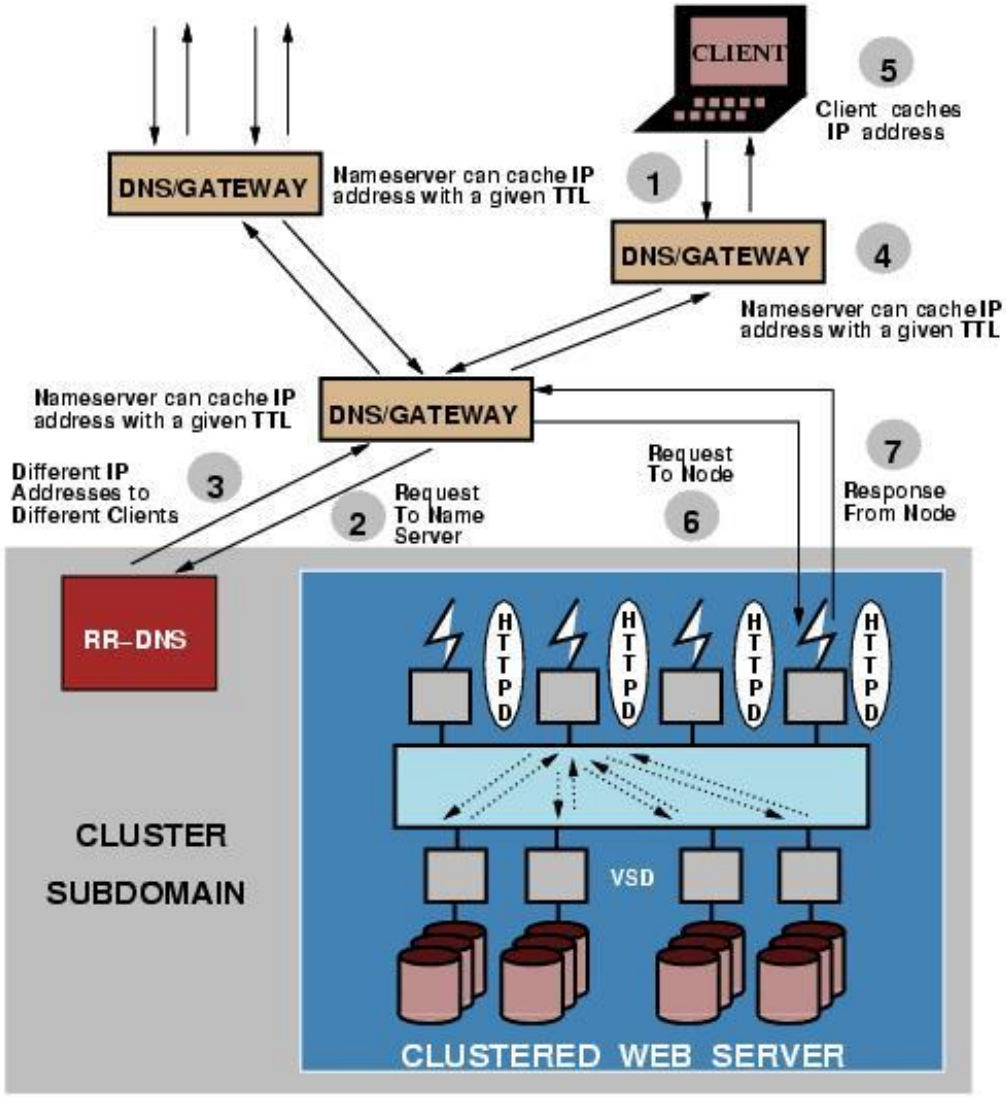


Figure 8.5 - A load balancer implemented as a service agent transparently distributes incoming workload request messages across two redundant cloud service implementations.

调度技术

- 调度方法
  - 基于DNS
  - 基于虚拟IP
  - 基于链路聚合：用于整合链路提高网络传输能力
  - 基于应用：用于分配到分布式调度器
- 调度策略
  - 轮转、负载水平，…
  - 同一用户的多个请求调度到同一服务器
  - 同一租户的请求调度到尽量少的一组服务器
  - 尽量实现不同类型负载的互补
  - 。。。

RR-DNS

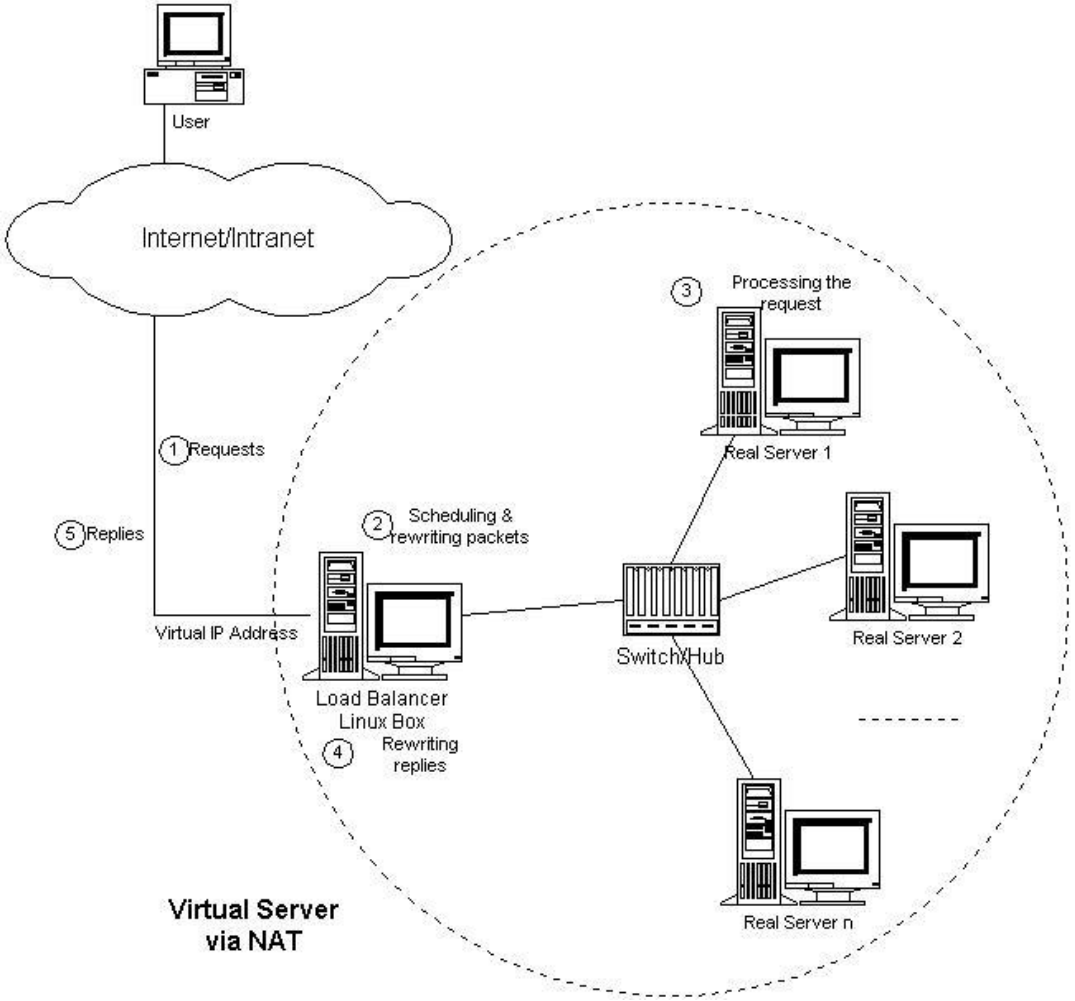


一组WEB服务器，通过分布式文件系统 AFS(Andrew File System)来共享所有的HTML文档。

这组服务器拥有相同的域名（如www.ncsa.uiuc.edu）。

当用户访问时，RR-DNS服务器会把域名轮流解析到不同服务器的IP地址，从而将访问负载均衡分布。

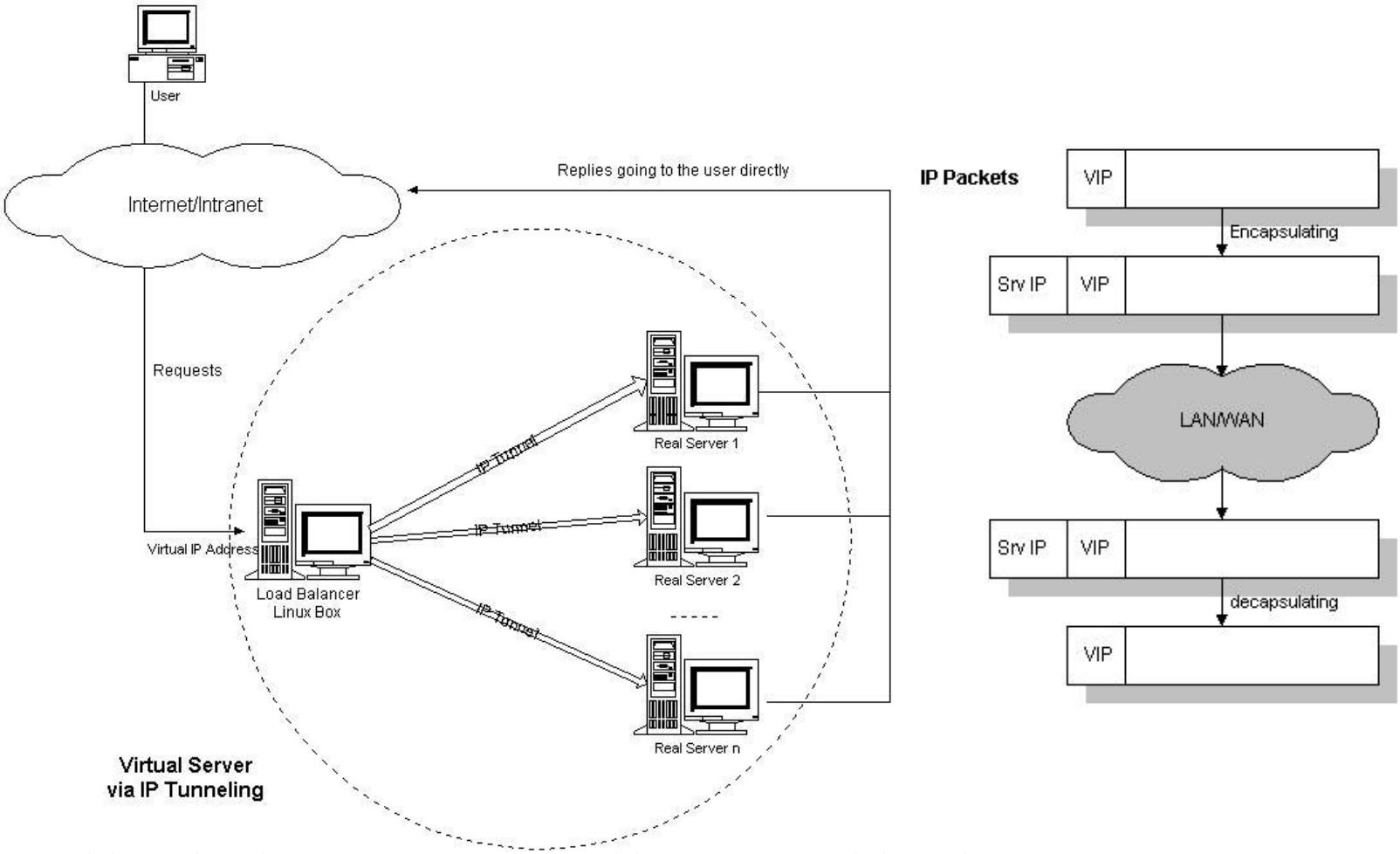
Virtual IP -- NAT



- 客户通过Virtual IP访问网络服务时，调度器根据连接调度算法从一组真实服务器中选出一台服务器，将报文的目标地址Virtual IP改写成选定服务器的地址，发送给选出的服务器。
- 调度器在连接Hash表中记录这个连接，用于处理同一个连接的多个报文。
- 来自真实服务器的响应报文经过调度器时，调度器将报文的源地址和源端口改为Virtual IP，发给用户。

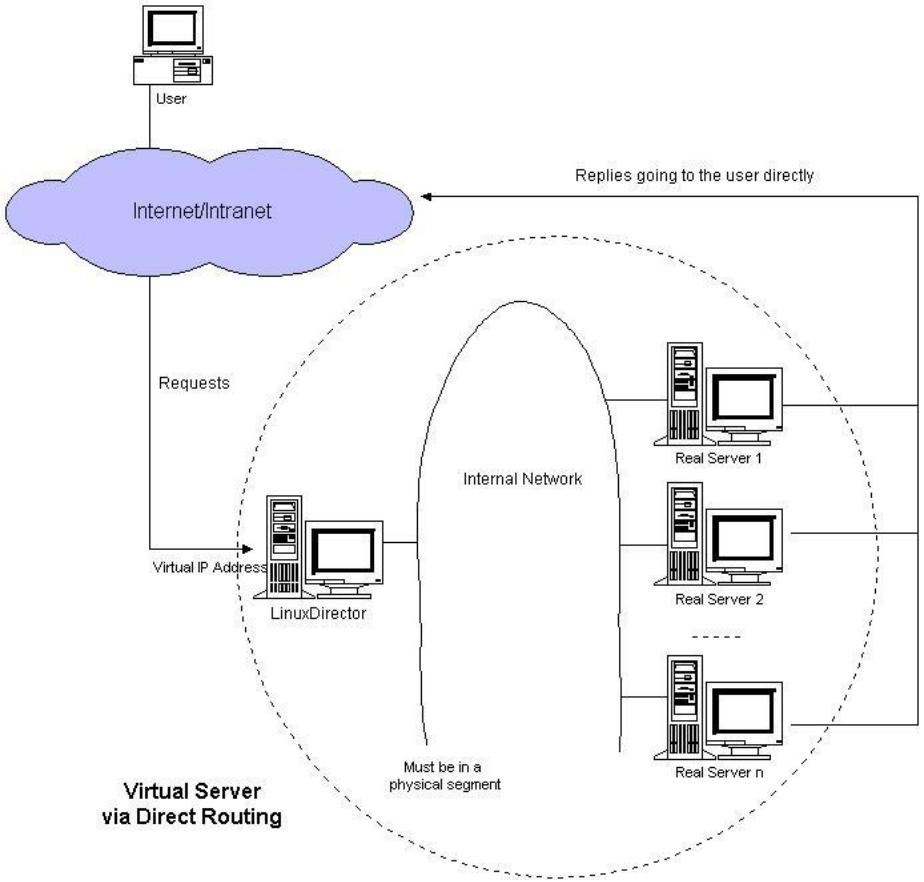
服务端口也要做相应修改。

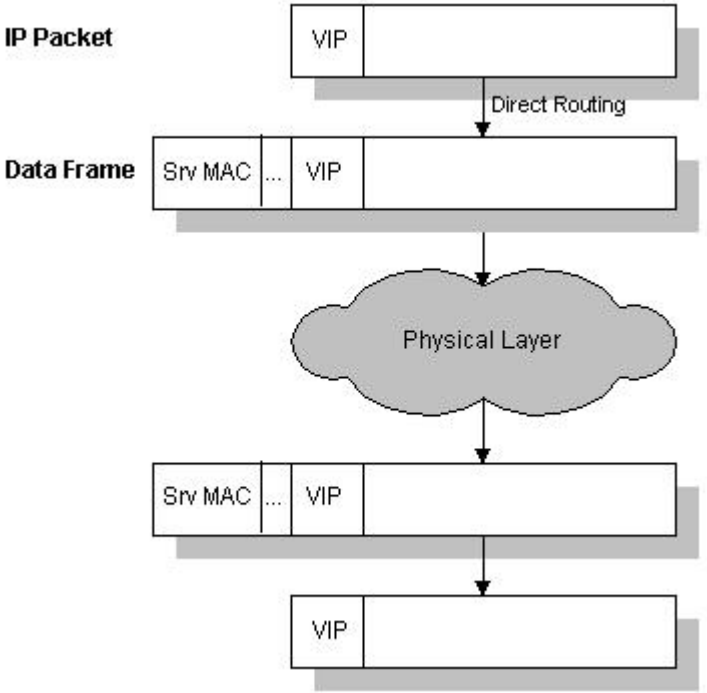
Virtual IP – IP tunneling



- 连接调度和管理与VS/NAT中的一样，只是报文转发方法不同。
- 调度器根据各个服务器的负载情况，动态地选择一台服务器，将请求报文封装在另一个IP报文中，转发给选出的服务器。
- 服务器收到报文后，先将报文解封获得原来目标地址为VIP的报文，处理这个请求，然后根据路由表将响应报文直接返回给客户。

Virtual IP – direct routing





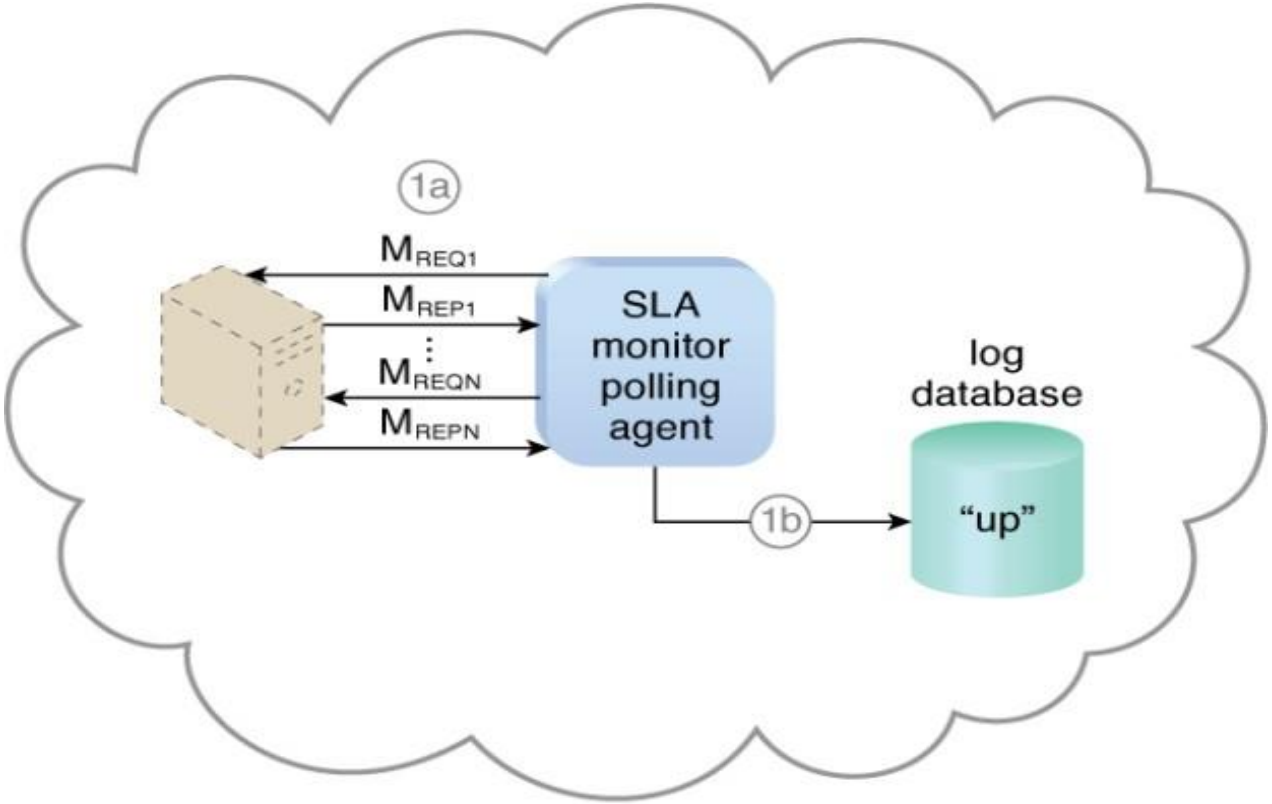
- 连接调度和管理与VS/NAT和VS/TUN一样。
- 调度器不修改也不封装IP报文，而是将数据帧的MAC地址改为选出服务器的MAC地址，再在局域网上发送。
- 以服务器收到这个数据帧，获得该IP报文，处理这个报文，然后根据路由表将响应报文直接返回给客户。

§8.3 SLA监控器

- SLA监控器(SLA monitor)
- 用来专门观察云服务的运行时性能，确保它们履行了SLA中报告的约定的QoS需求。
- SLA监控器收集的数据由SLA管理系统处理并集成到SLA报告的标准中

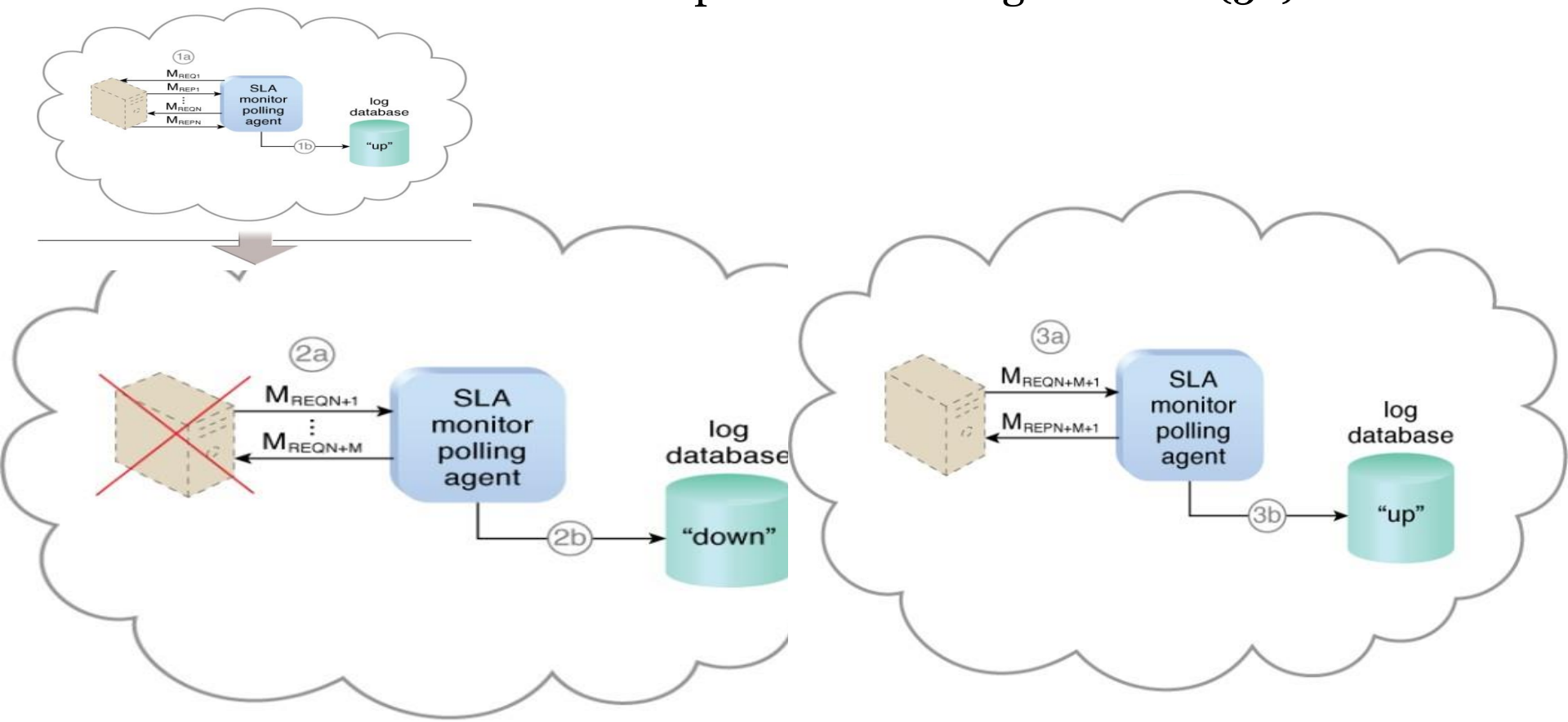
An Illustration

- The SLA monitor polls the cloud service: MREQ1 to MREQN.
- Cloud service response: MREP1 to MREPN, to report "up" (1a).
- The SLA monitor stores the "up" time (time period of all polling cycles 1 to N) in the log database (1b).



An Illustration

- Polling response (MREQN+1 MREQN+M)are not received (2a).
- The SLA monitor stores the "down" time in the log database (2b).
- Polling response message (MREPN+M+1) is received (3a).
- The SLA monitor stores the "up" time in the log database (3b).



§8.4按使用付费监控器

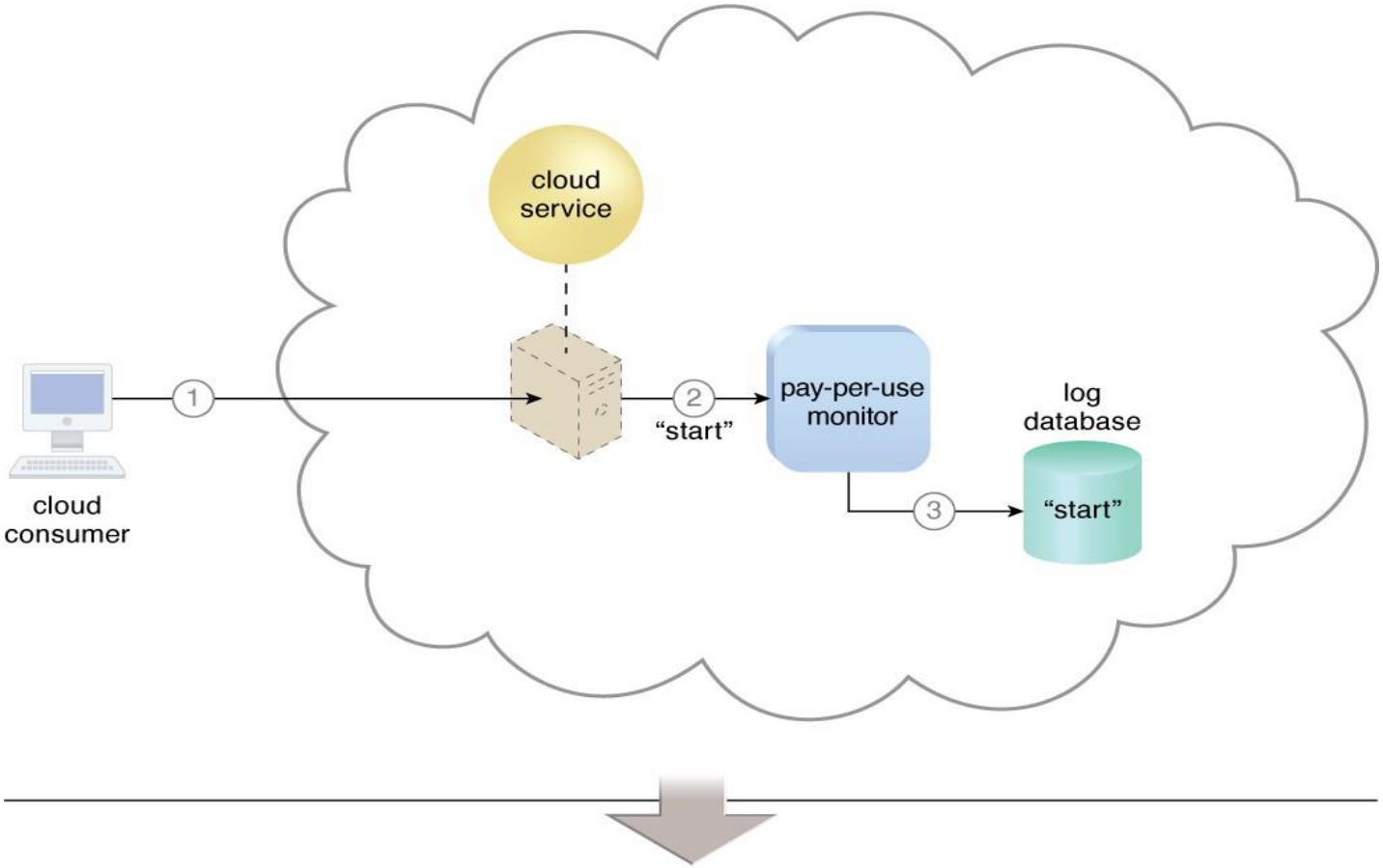
- 按使用付费监控器(pay-per-use monitor)
- 按照预先定义好的定价参数测量云资源使用，并生成使用日志用于计算费用。
- 使用数据由计费管理系统(billing management system)进行处理。

按使用付费监控器

- 一些典型的监控变量包括：
  - 请求/响应消息数量
  - 传送的数据量
  - 带宽消耗
- 实现方式
  - 资源代理
  - 监控代理

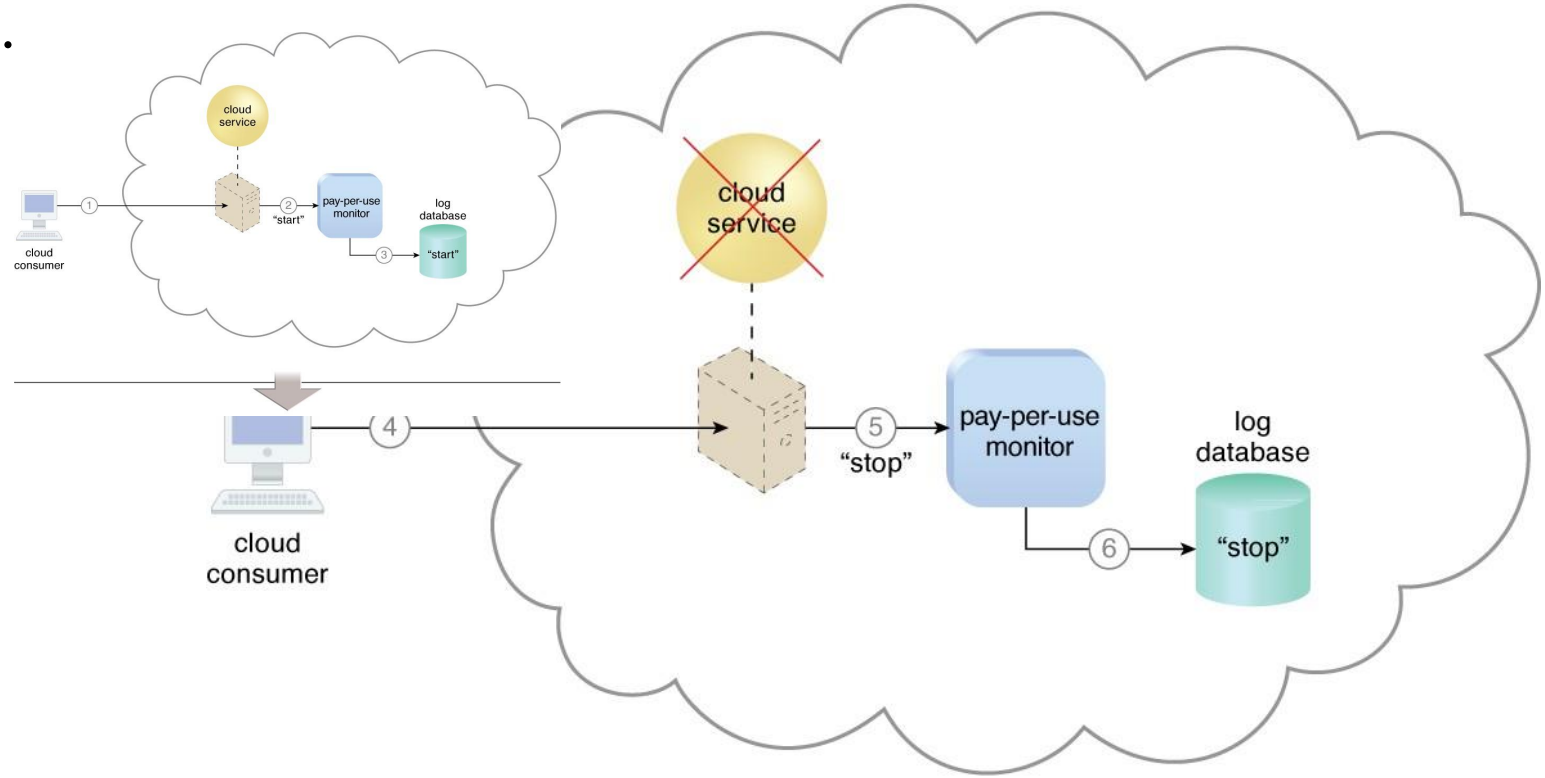
使用付费监控器作为资源代理

- A cloud consumer requests a new instance of a cloud service (1).
- 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).



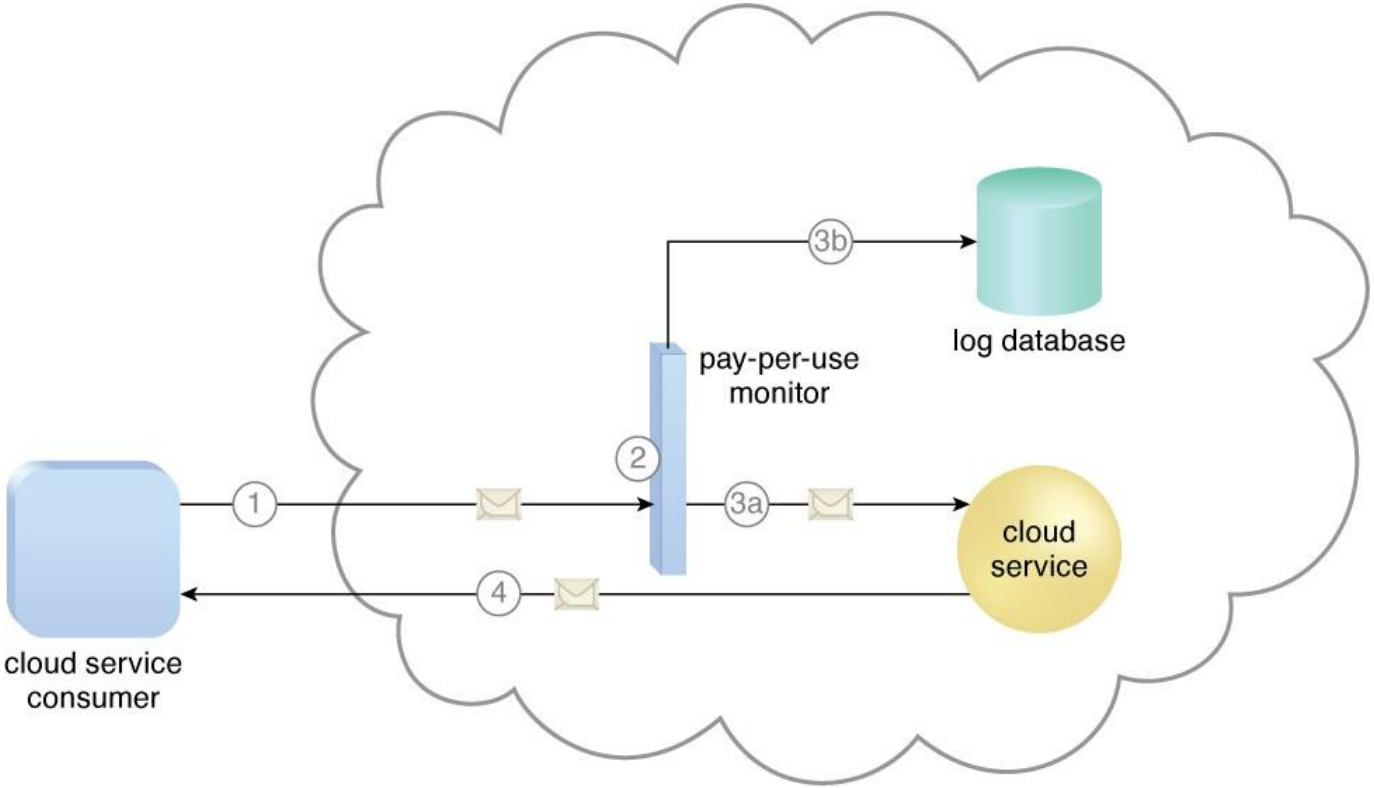
使用付费监控器作为资源代理

- 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).



使用付费监控器作为监控代理

- 云服务用户向云服务发送请求消息（1）
- **按使用付费监控器**截获该消息（2），将他转发给云服务（3a），按照监控指标把使用信息存储起来（3b）
- 云服务讲响应消息转发回云服务用户，提供所请求的服务

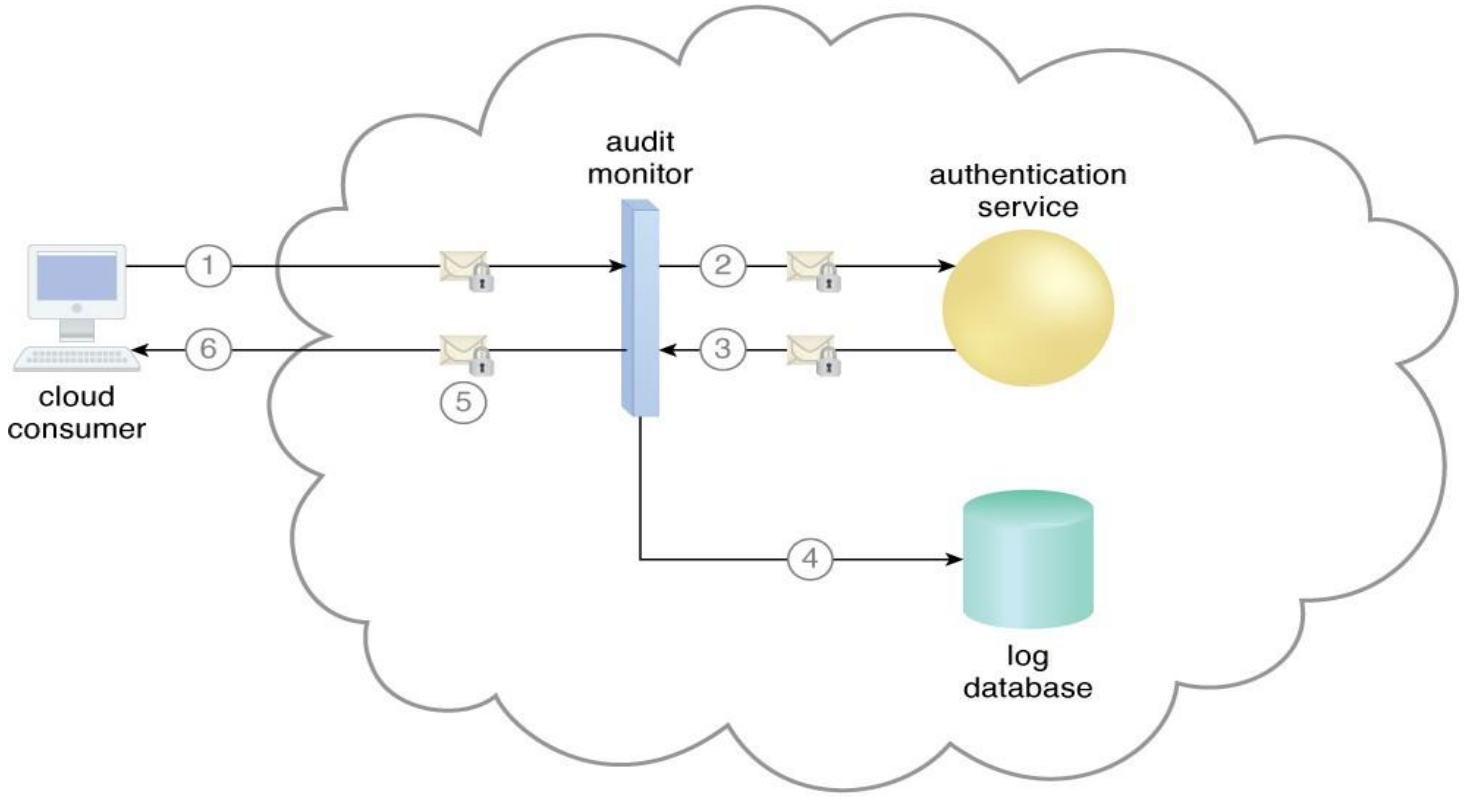


§8.5审计监控器

- 审计监控器(audit monitor)
- 用来收集网络和IT 资源的审计记录数据
- 用以满足管理需要或者合同义务。

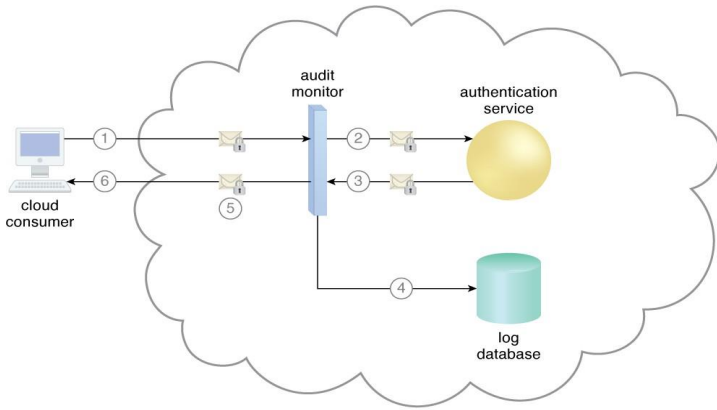
一个实现为监控代理的审计监控器

它截获“登录”请求，在日志数据库中存储请求者的安全证书，以及成功和失败的登录尝试，以供今后审计报告用



一个实现为监控代理的审计监控器

1. A cloud service consumer requests access to a cloud service by sending a login request message with security credentials (1).
2. The audit monitor intercepts the message (2) and forwards the message to the authentication service (3).
3. The authentication service processes the security credentials. A response message is generated for the cloud service consumer, in addition to the results from the login attempt (4).
4. The audit monitor intercepts the response message and stores the entire collected login event details in the log database, as per the organization's audit policy requirements (5).
5. Access has been granted, and a response is sent back to the cloud service consumer (6).



§8.6故障转移系统

- 故障转移系统(failover system)
- 通过使用现有的集群技术提供冗余的实现来增加IT资源的可靠性和可用性。
- 只要当前活跃的IT资源变得不可用时，便会自动切换到冗余的或待机IT资源实例上。

故障转移系统的类型

- 主动-主动
  - 多个实例都处于活动状态，同时提供服务
  - IT 资源的冗余实现和负载均衡器是必须要的
- 主动-被动
  - 一个处于活动状态，一个待机或闲置
  - 当IT资源变得不可用的时候，就会激活待机实例来接管工作
  - 相应的工作负载就会被重定向到接管操作的这个实例上

主动-主动

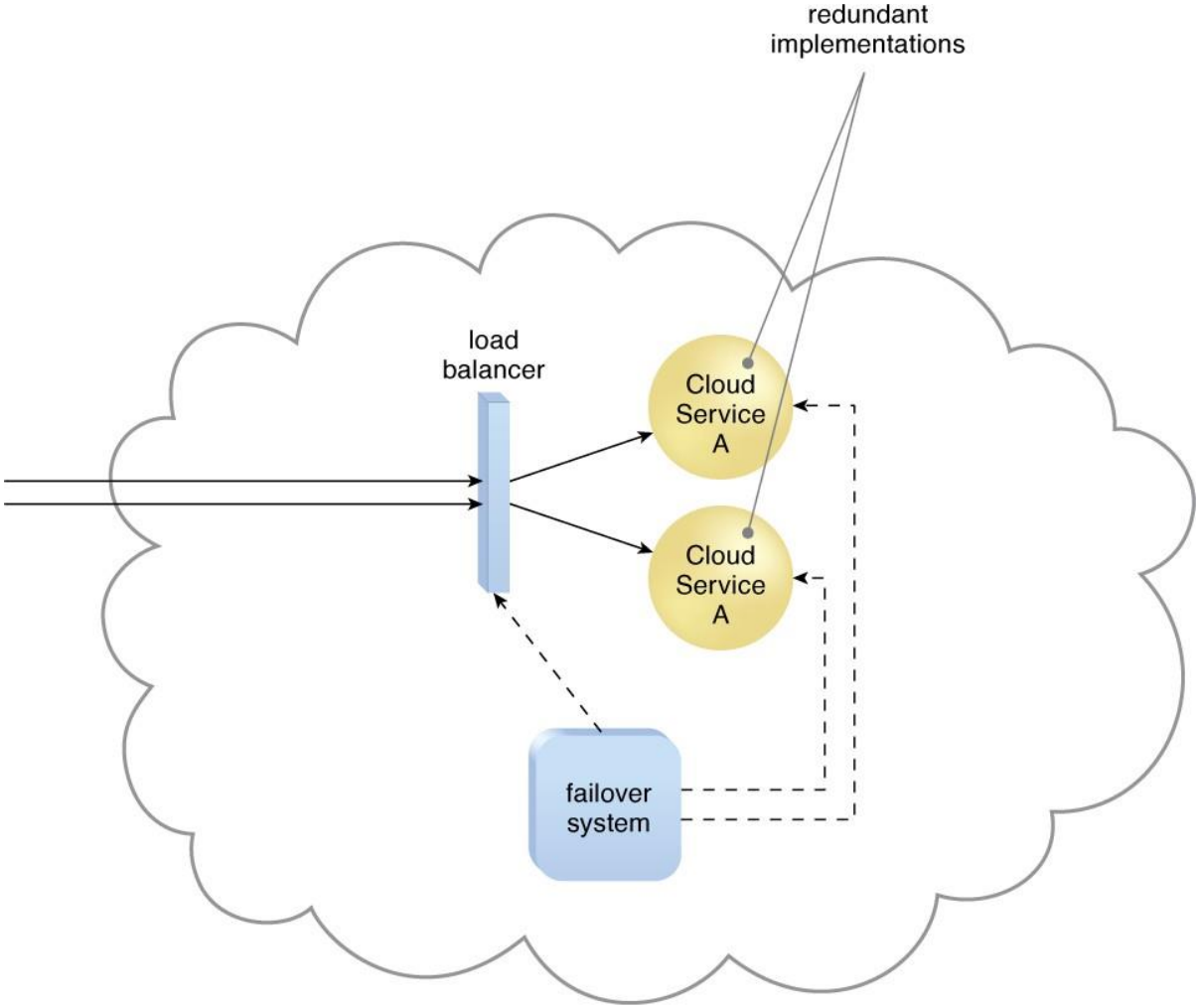
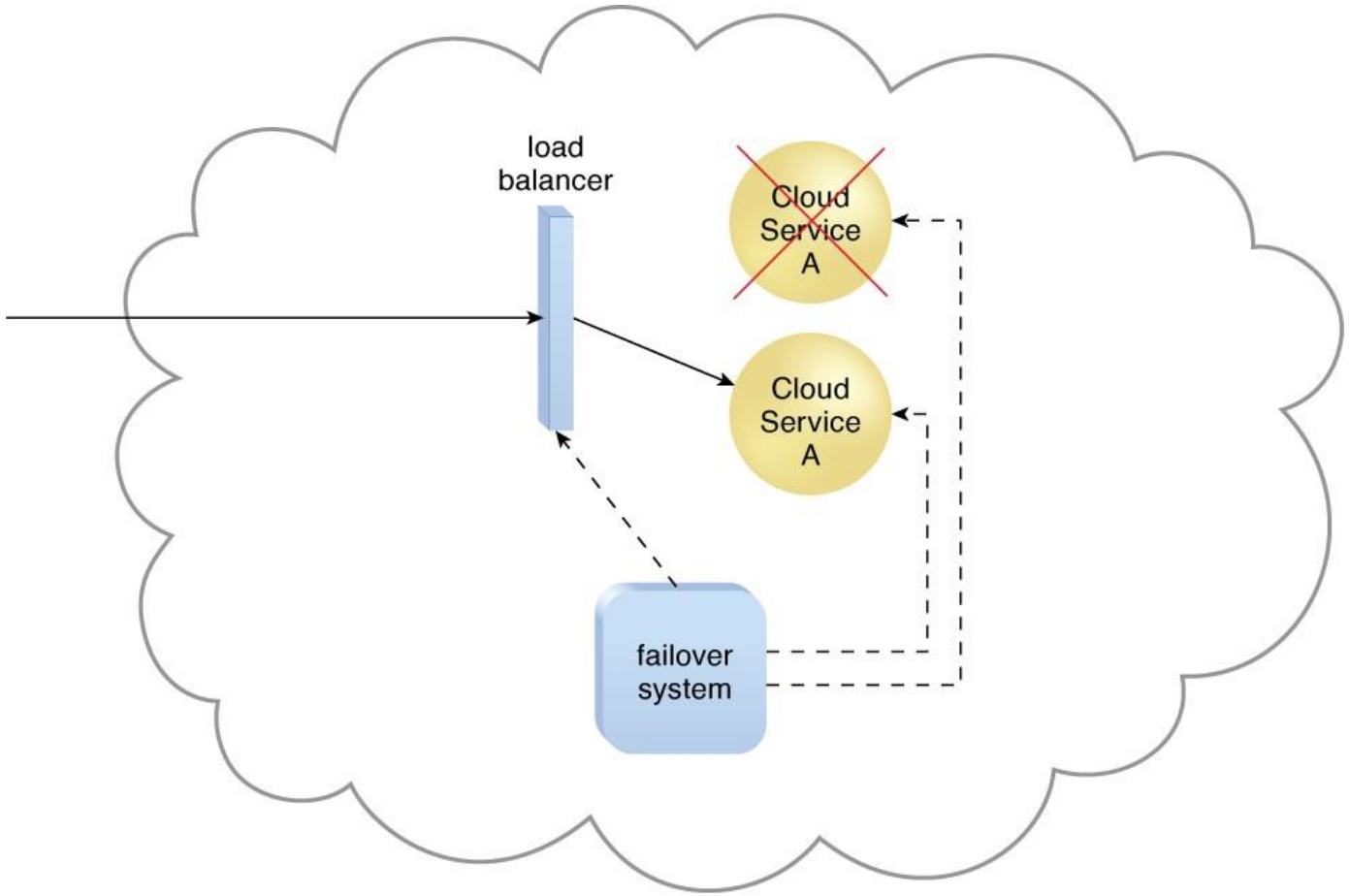


Figure 8.17 - The failover system monitors the operational status of Cloud Service A.



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

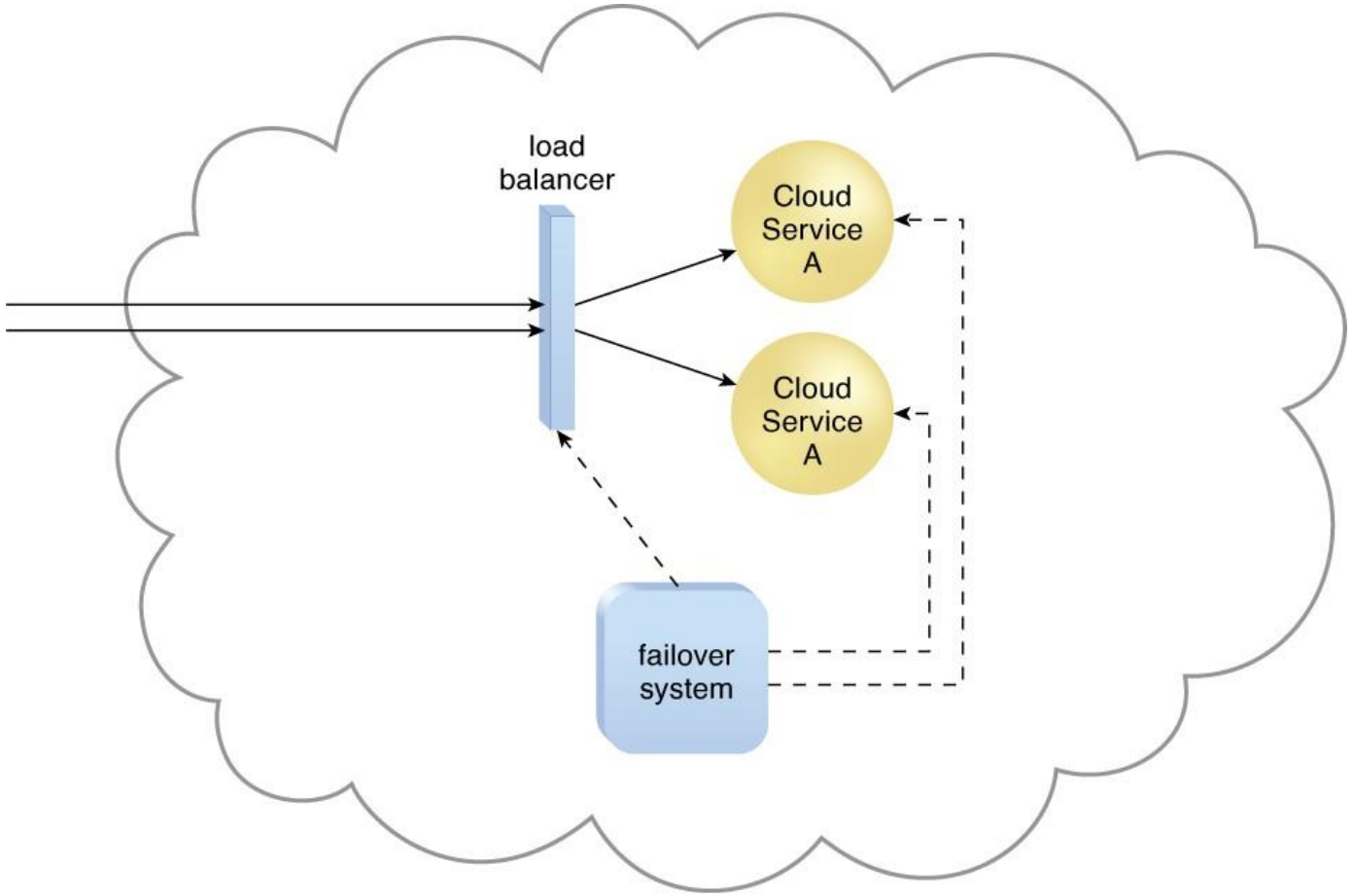
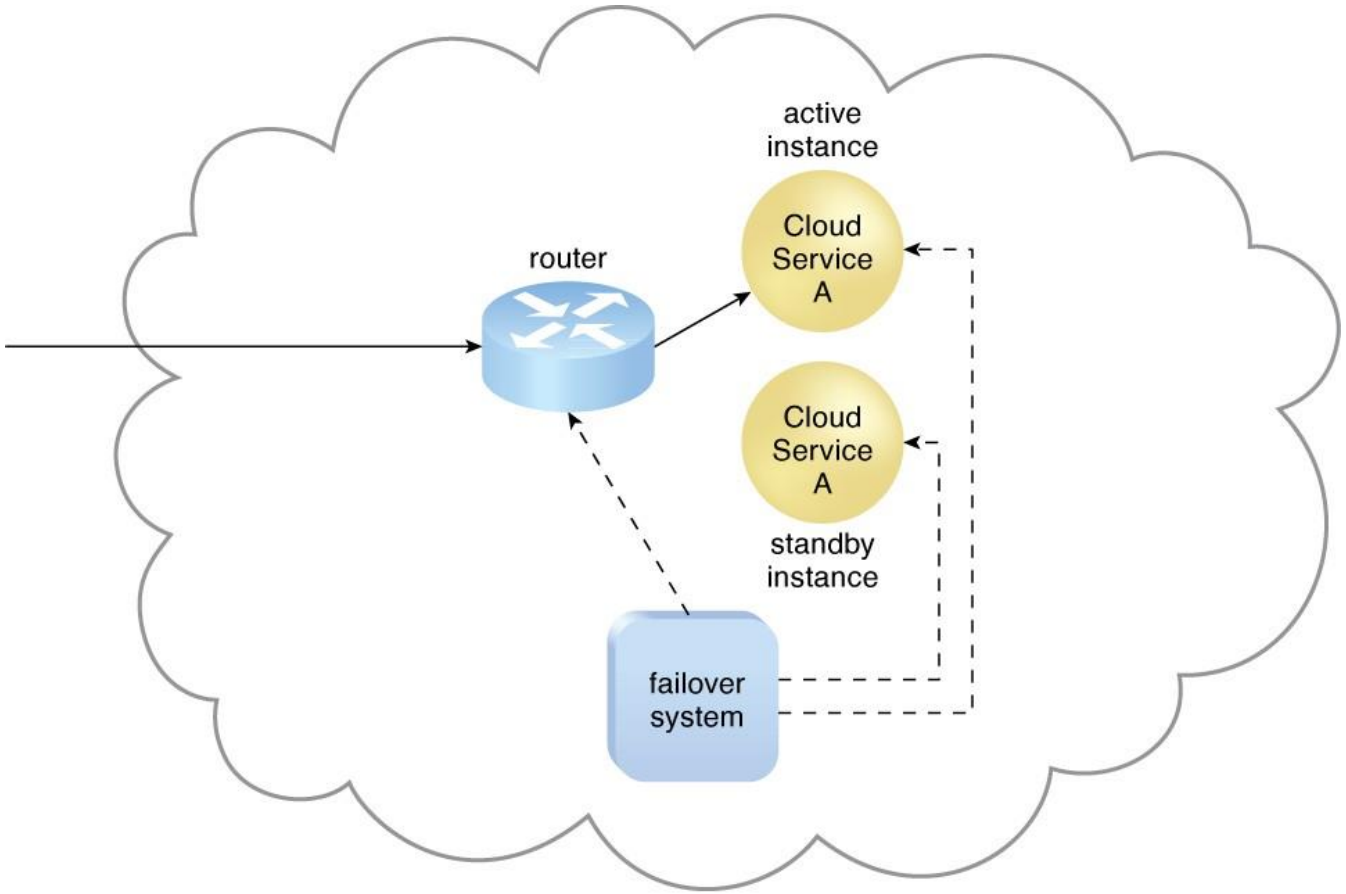


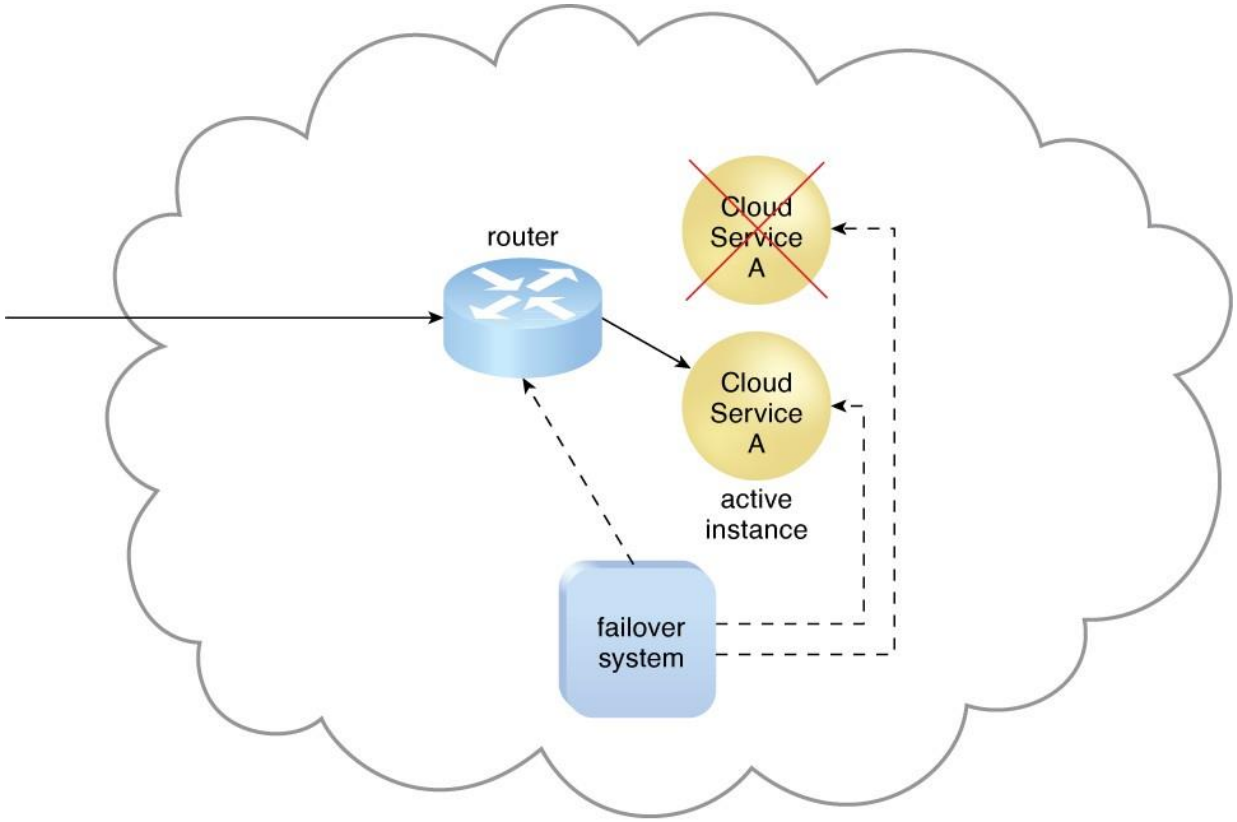
Figure 8.19 - The failed Cloud Service A implementation is recovered or replicated into another operational resource. The failover system now commands the load balancer to distribute the workload again.

主动-被动

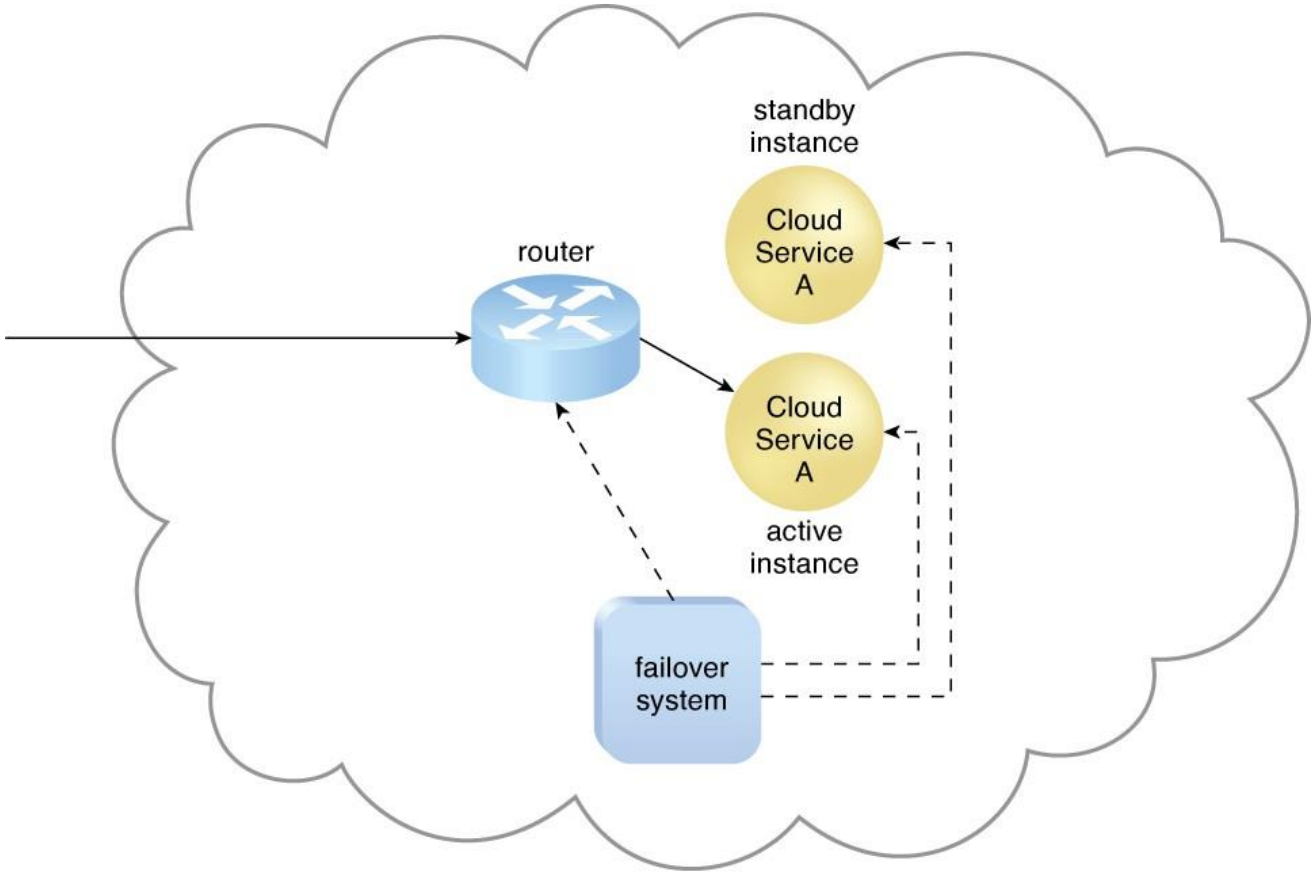


- *Figure 8.20 - The failover system monitors the operational status of Cloud Service A. The Cloud Service A implementation acting as the active instance is receiving cloud service consumer requests.*

Figure 8.21



- *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.*

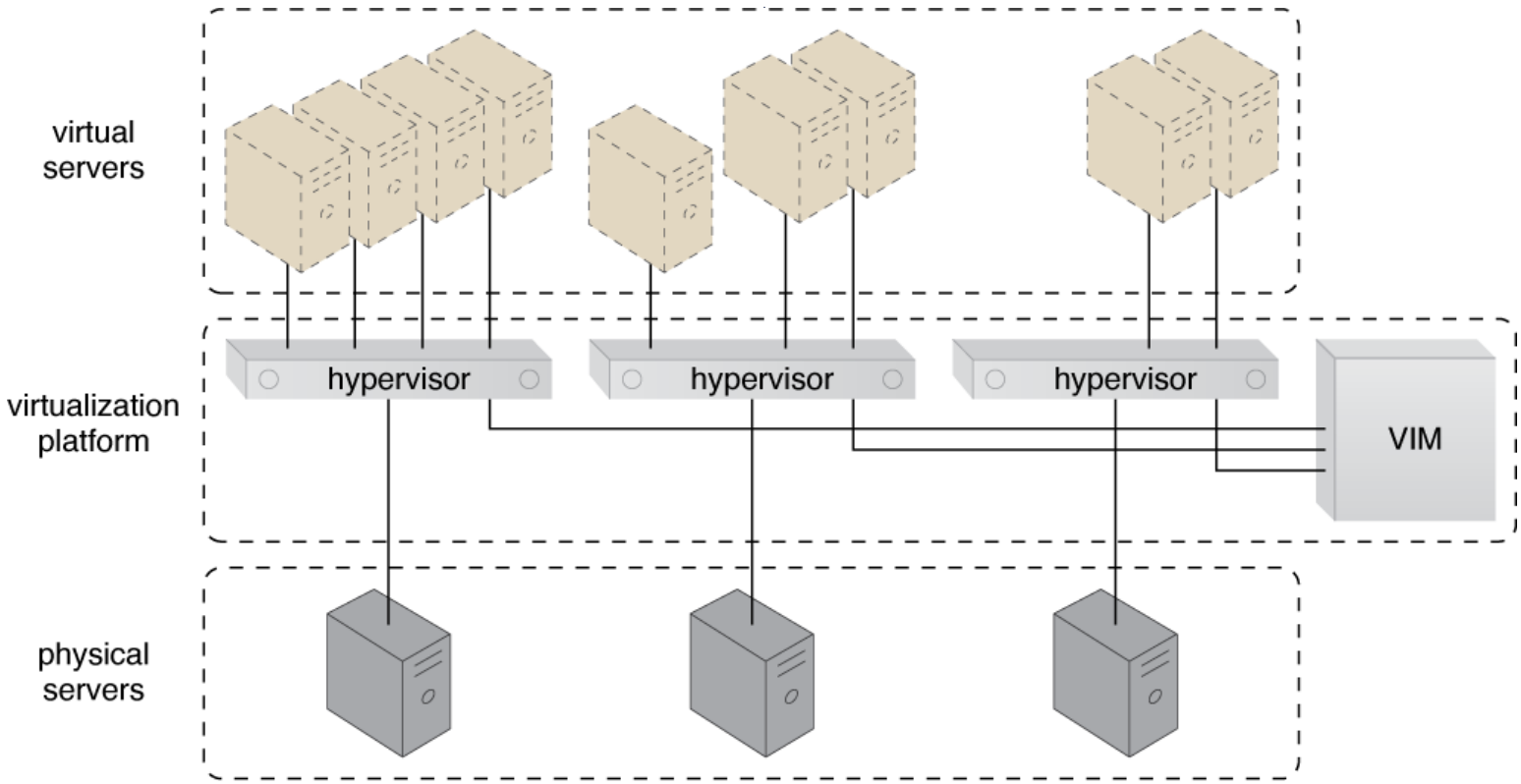


- *Figure 8.22 - The failed Cloud Service A implementation is recovered or replicated into another operational resource, and is now positioned as the standby instance while the previously invoked Cloud Service A continues to serve as the active instance.*

§8.7虚拟机监控器

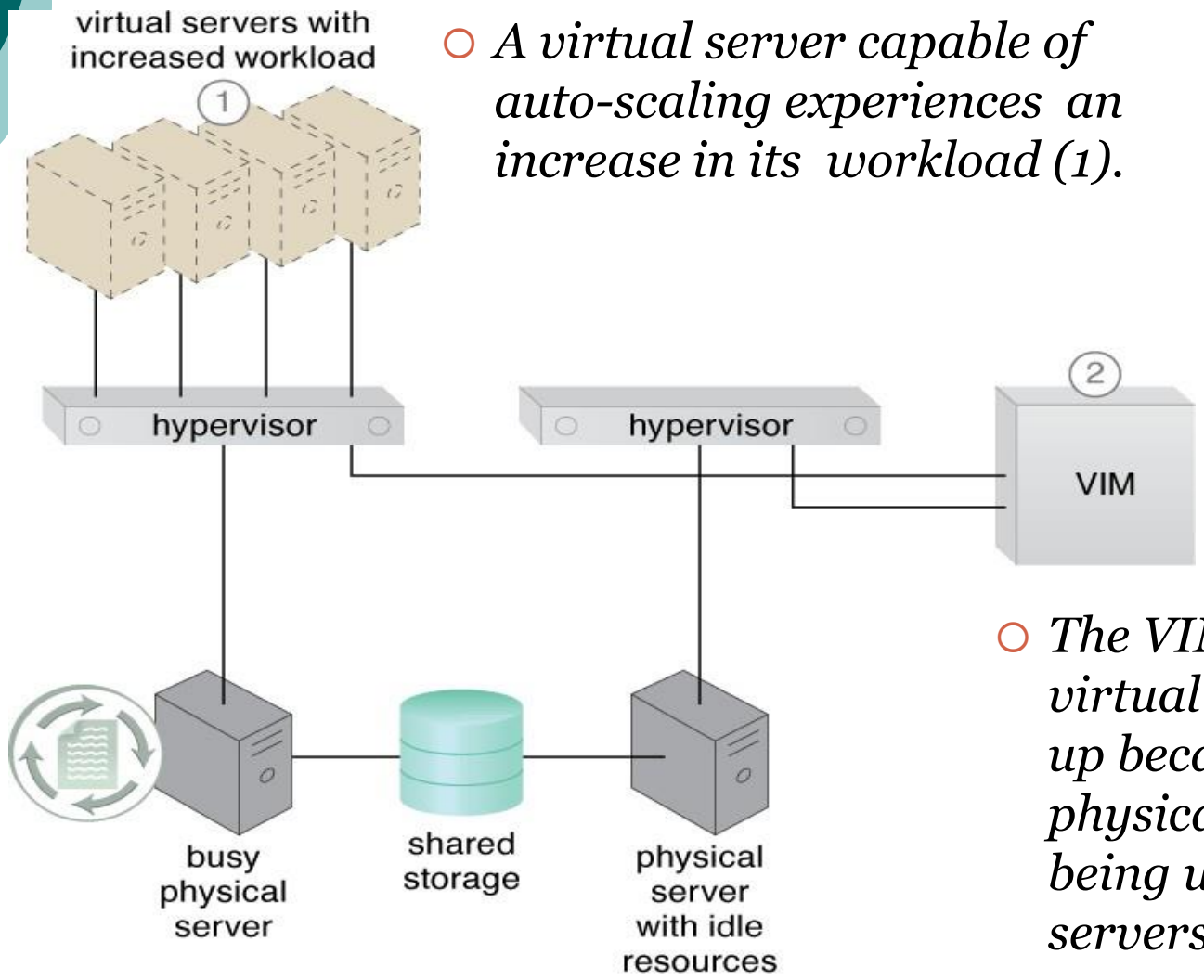
- 虚拟机监控器(hypervisor)
- 是虚拟化基础设施的最基本部分，主要用来在物理服务器上生个虚拟服务器实例。
- 虚拟机监控器通常受限于一台物理服务器
- VIM提供了一组特性来管理跨物理服务器的多虚拟机监控器

虚拟机监控器



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

An Example of VM Imigration



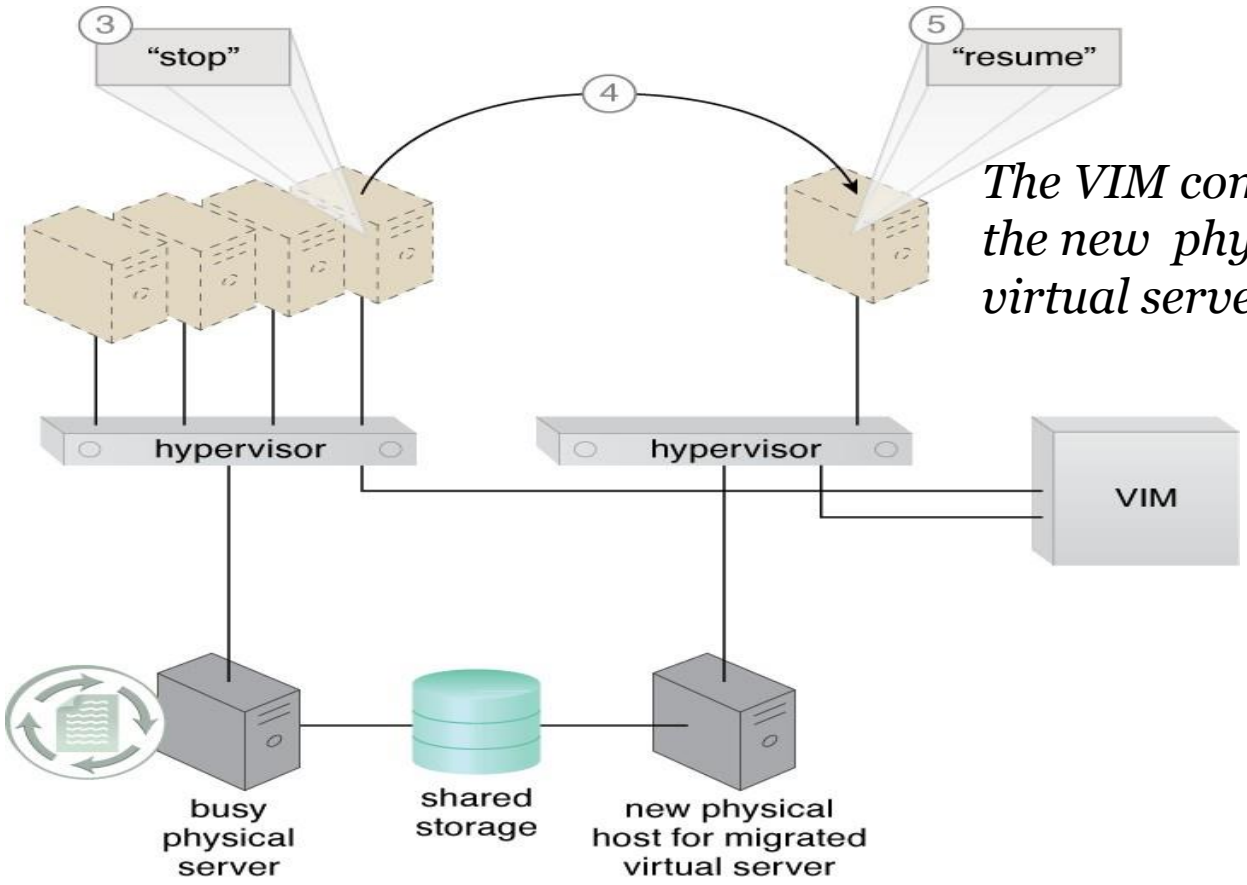
- *A virtual server capable of auto-scaling experiences an increase in its workload (1).*

- *The VIM decides that the virtual server cannot scale up because its underlying physical server host is being used by other virtual servers (2).*

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 (4).



The VIM commands the hypervisor at the new physical server to resume the virtual server processing (5).

§8.8资源集群

- 资源集群(resource cluster)
- 把多个IT资源实例分为一组，使得他们能像一个IT资源那样进行操作。

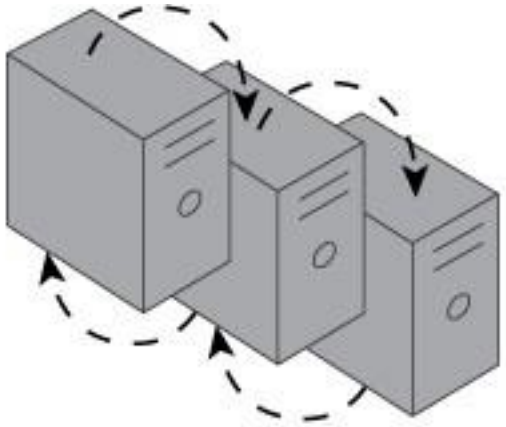


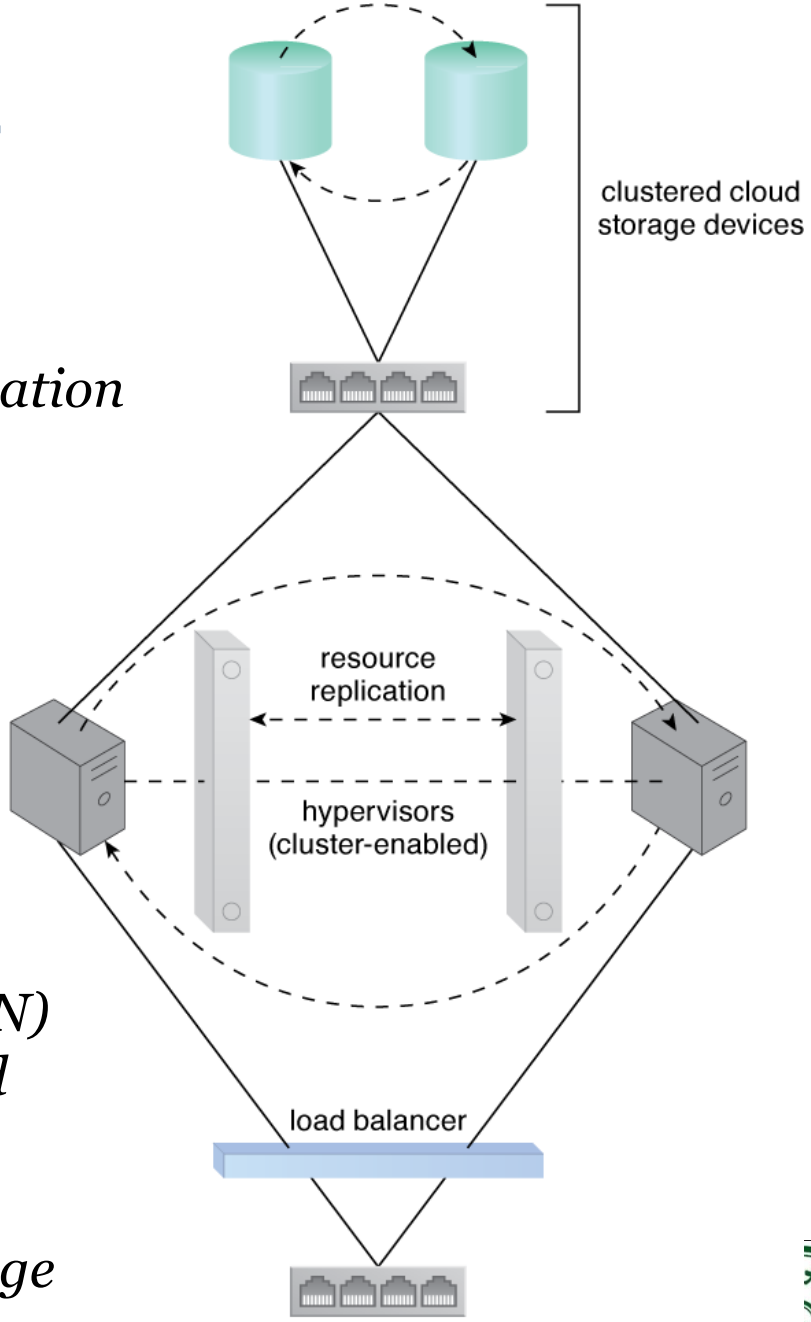
Figure 8.30 - The curved dashed lines are used to indicate that IT resources are clustered.

资源集群

- 通过高速专用网络链接或者集群结点实现工作负载、任务调度、数据共享和系统同步等通讯要求。
- 常见的资源集群类型包括：
  - 服务器集群——提高性能和可用性
  - 数据库集群——提高数据可用性，维持数据的一致性
  - 大数据集集群——数据的分区和分布
  - 负载均衡的集群——保持集中管理的特性下实现了在集群结点中的分布式工作负载。
  - 高可用集群——在多节点失效的情况下保持系统的可用性，需要冗余实现和故障转移机制。

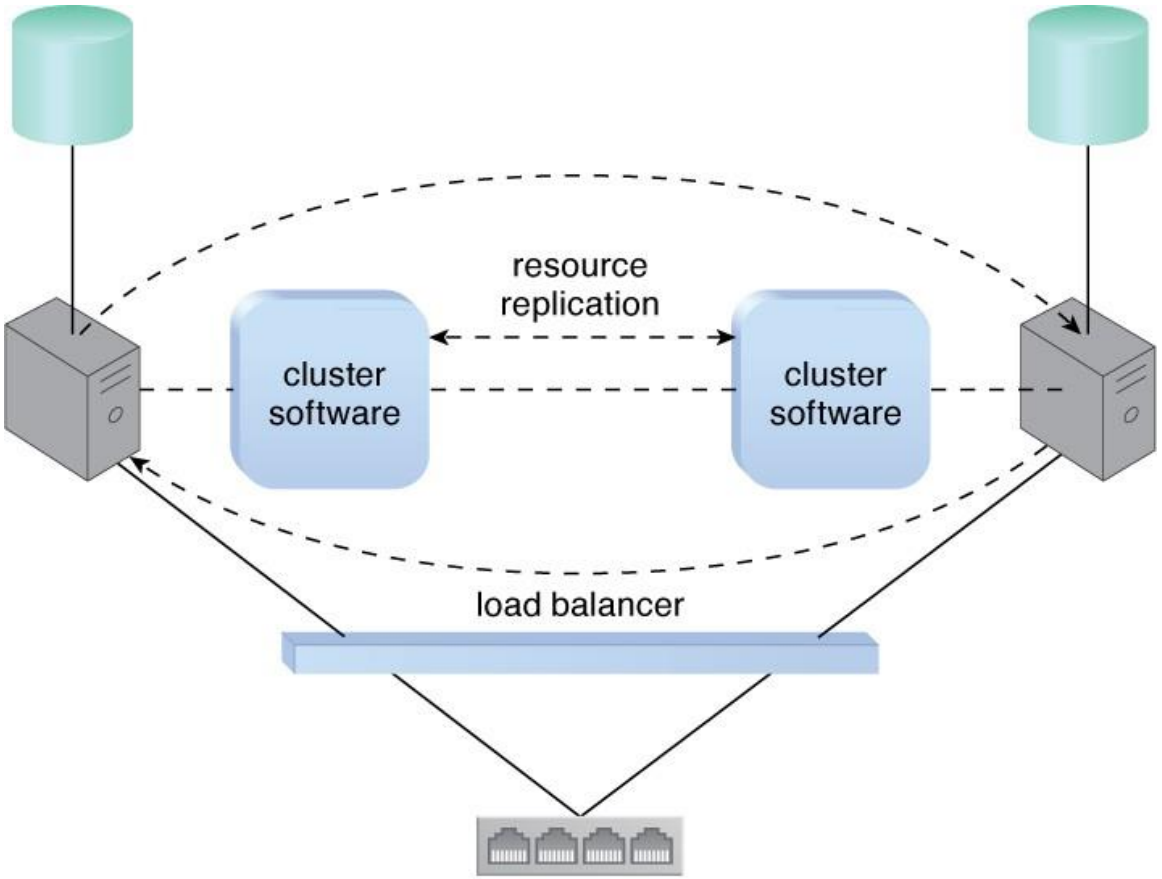
A Example of Cluster

- *Load balancing and resource replication via a cluster-enabled hypervisor.*



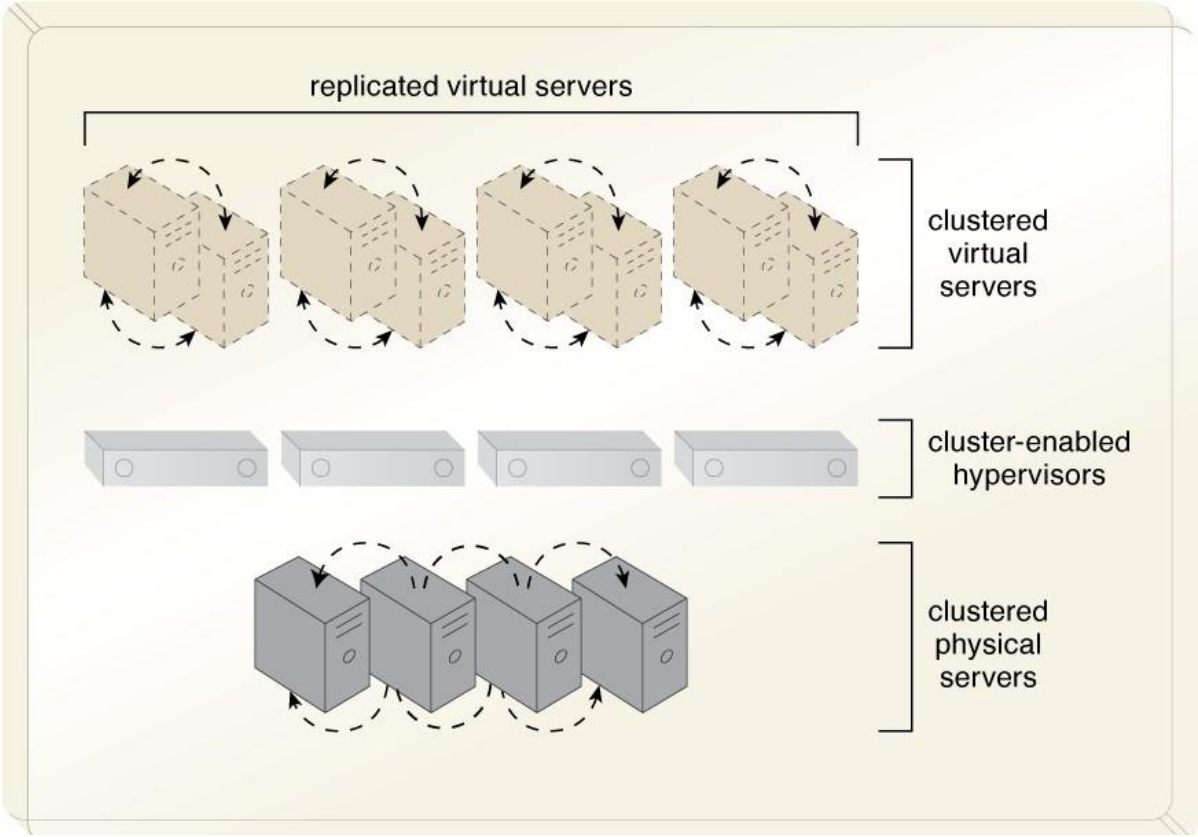
- *A dedicated storage area network (SAN) is used to connect storage devices and clustered servers*
- *Storage replication process is independently carried out at the storage cluster.*

A Server Cluster with Distributed Storage



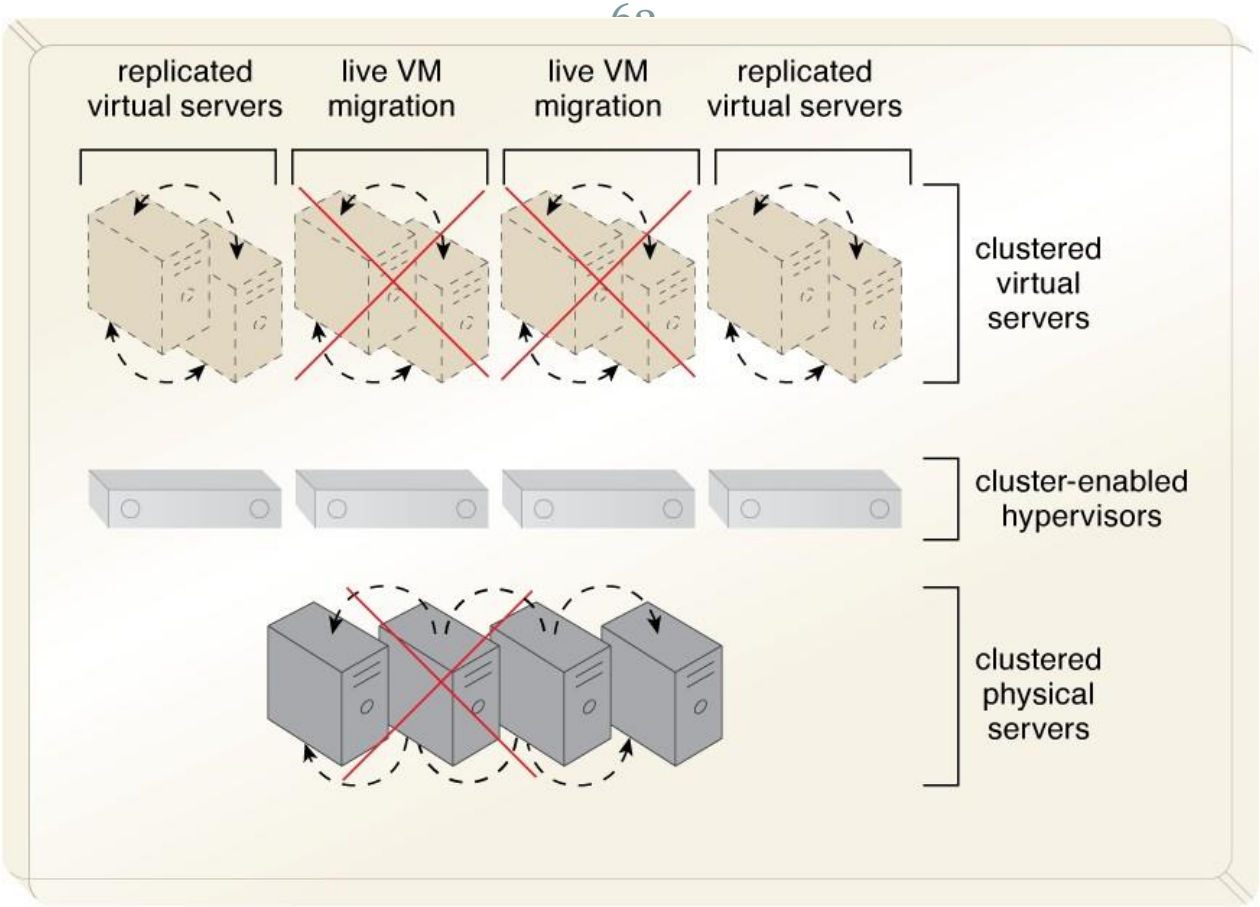
- *A loosely coupled server cluster that incorporates a load balancer. There is no shared storage.*
- *Resource replication is used to replicate cloud storage devices through the network by the cluster software.*

HA Cluster (DTGOV's Example)



- *An HA virtualization cluster of physical servers is deployed using a cluster-enabled hypervisor, which guarantees that the physical servers are constantly in sync.*
- *Every virtual server that is instantiated in the cluster is automatically replicated in at least two physical servers.*

HA Cluster (DTGOV's Example)



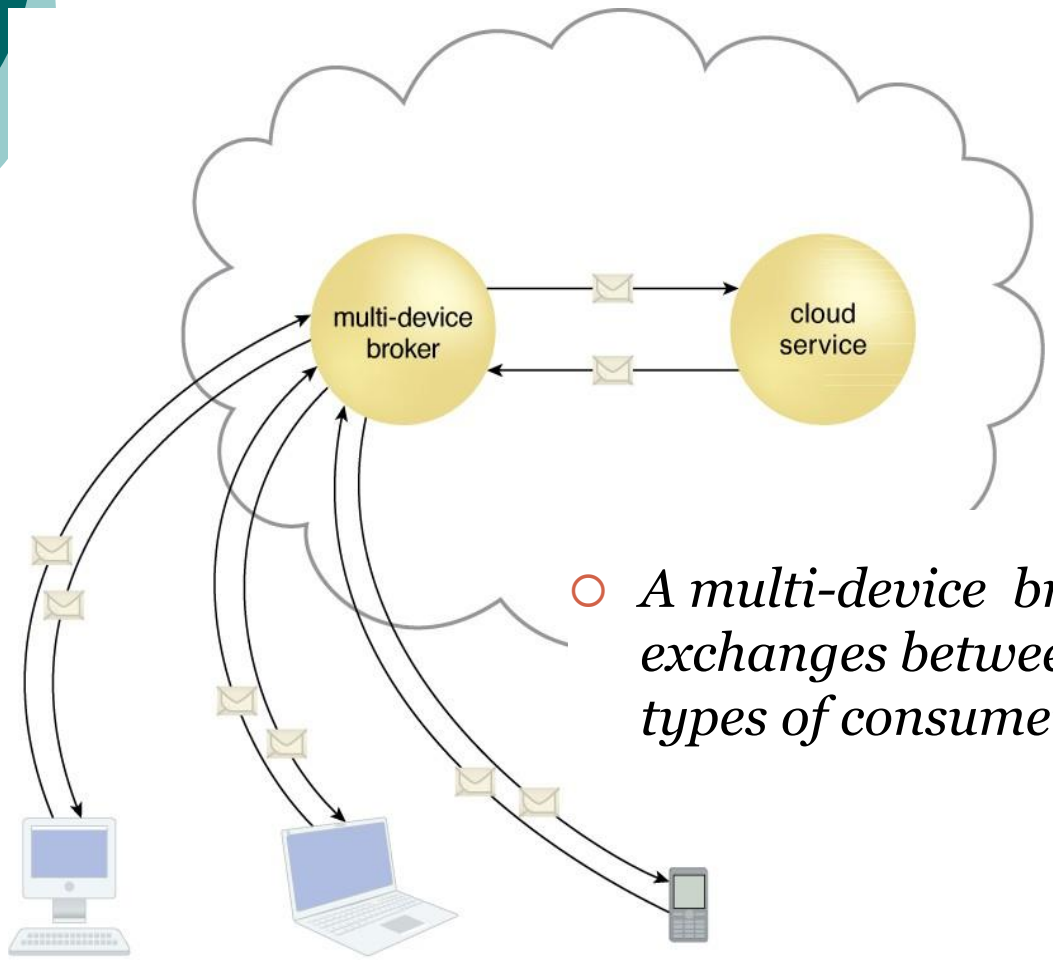
HA virtualization cluster

All of the virtual servers that are hosted on a physical server experiencing failure are automatically migrated to other physical servers by the hypervisor.

§8.9多设备代理

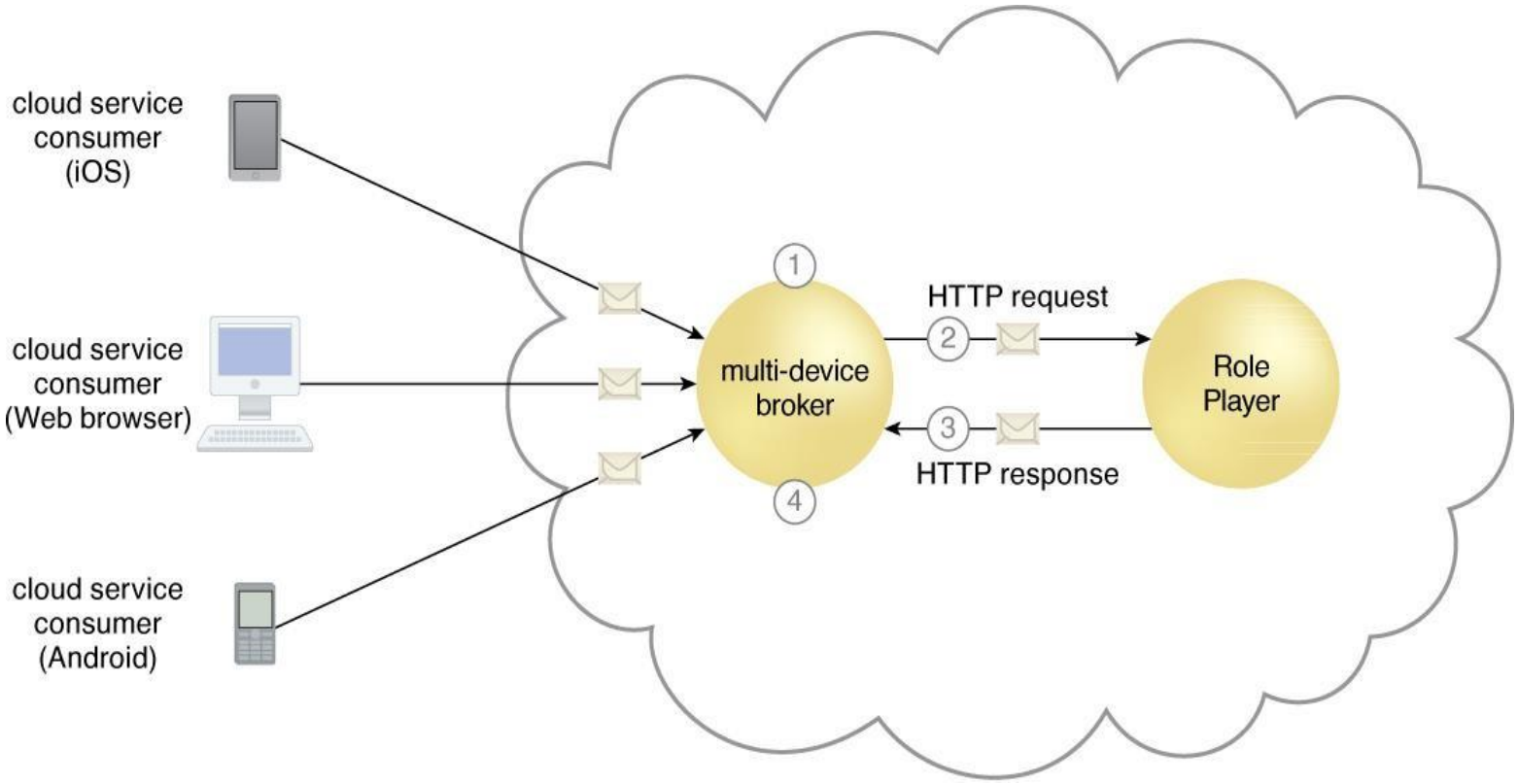
- 多设备代理(multiple-device broker)
- 用于运行时的数据转换。
- 克服云服务和多样性的云服务用户之间的不兼容性
- 使得云服务能够被更广泛的云服务用户程序和设备所使用
- 需要创建映射逻辑(mapping logic)来改变运行时交换的信息。
- 多设备代理通常是作为网关存在的：
  - XML 网关
  - 云存储网关
  - 移动设备网关

A Multi-device Broker



- *A multi-device broker to transform data exchanges between a cloud service and different types of consumer devices.*

Innovartus's Example



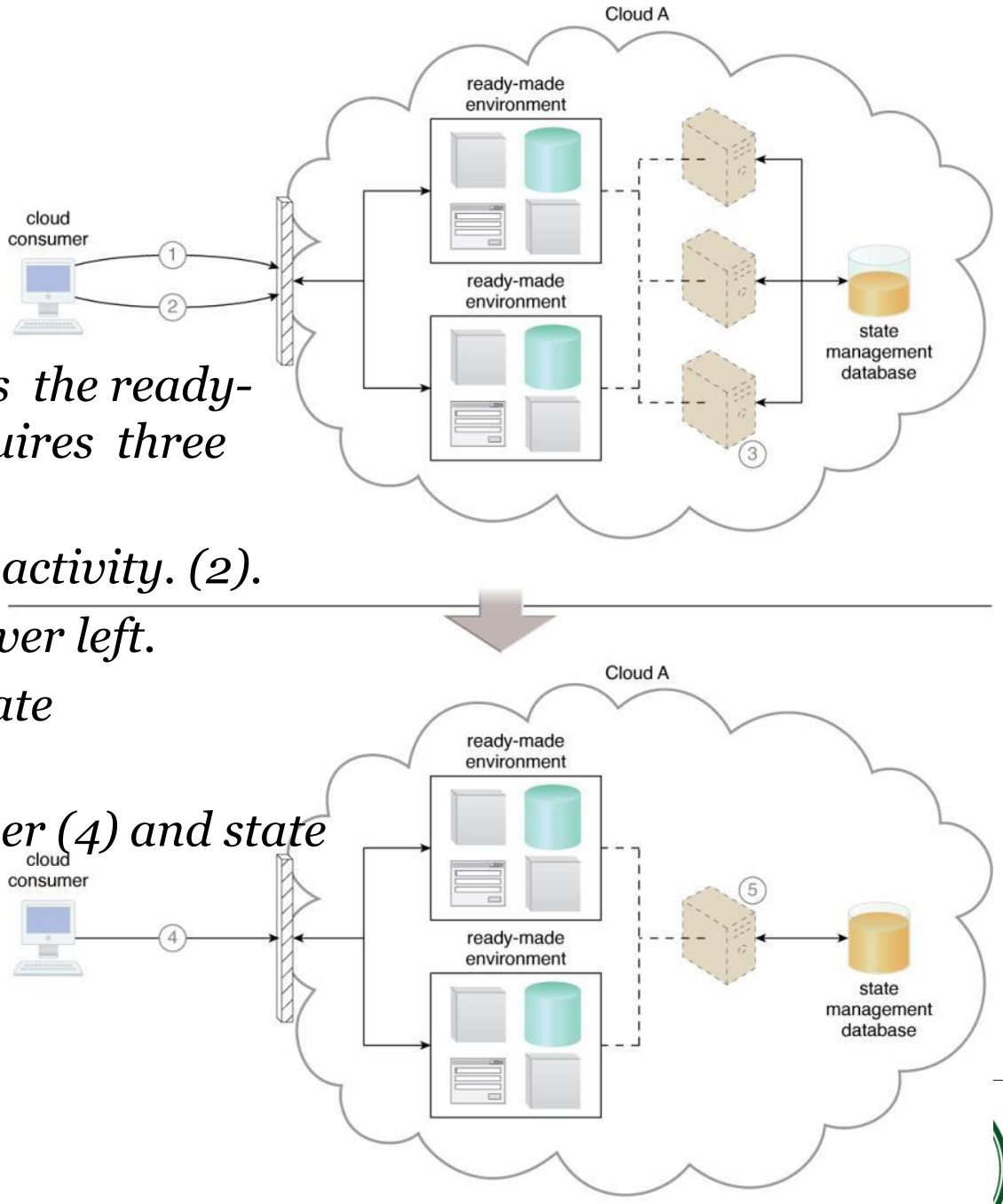
- *A message format transformers*
- *Intercepts incoming messages and detects the platform (Web browser, iOS, Android) of the source device*

§8.10状态管理数据库

- 状态管理数据库(state management database)
- 是一种存储设备，用来暂时地持久化软件程序的状态数据,软件程序可以把状态数据卸载到数据库中
- 替代状态数据缓存在内存中的一种方法，用以降低程序占用的运行时的内存量
- 状态管理数据库使得软件程序和周边的基础设施都具有更大的可扩展性

An Example

- *The cloud consumer accesses the ready-made environment and requires three virtual servers (1).*
- *The cloud consumer pauses activity. (2).*
- *Scaling in by one virtual server left.*
- *State data is saved in the state management database (3).*
- *Later login by cloud consumer (4) and state data is saved (5).*



本章小结

- 云计算平台中用到的一些非核心但重要、相对独立的辅助性机制
- 多为资源管理与监控性方面的功能
  - 自动伸缩、LB、SLA、Pay-per-Use、Audit、transforming、failover…

课后题

1、讨论分析可以用于自动伸缩的判定条件和机制。

2、分析讨论按使用付费监控器的两种实现方式的优缺点。