Lecture 6

Cloud Computing I

(Concepts, Technology & Architecture)

Dr. Lydia Wahid

Agenda

- What is cloud computing?
- Cloud Delivery Models
- Cloud Deployment Models
- Cloud-Enabling Technology
- Cloud Architecture

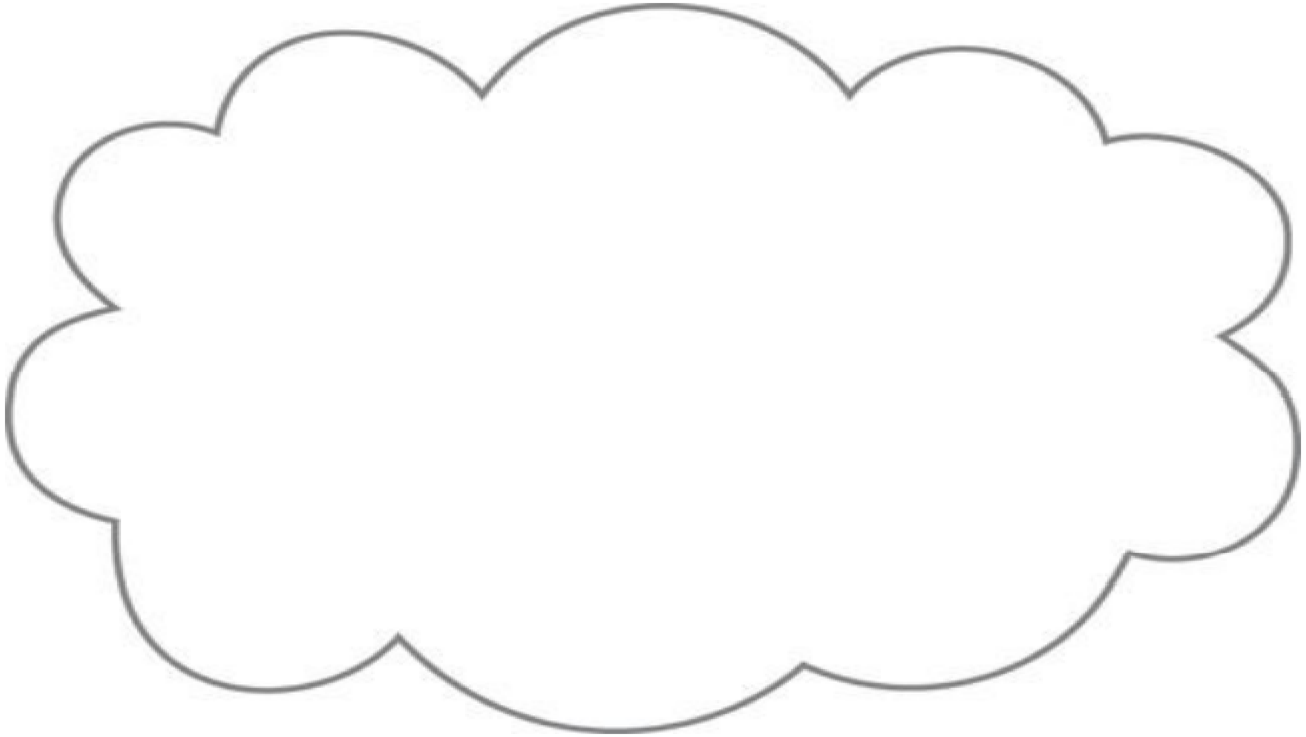
What is cloud computing?

Definitions

- **Cloud Computing:**
  - "…a style of computing in which scalable and elastic IT-enabled capabilities are delivered as a service to external customers using Internet technologies." *Gartner*
  - "Cloud computing is a specialized form of distributed computing that introduces utilization models for remotely provisioning scalable and measured resources." *National Institute of Standards and Technology (NIST)*

Definitions

- The symbol used to denote the boundary of a cloud environment:



Basic Concepts and Terminology

➢Cloud:

- A cloud refers to a distinct IT environment that is designed for the purpose of remotely provisioning scalable and measured IT resources.

➢IT Resource:

- An IT resource is a physical or virtual IT-related artifact that can be either software-based, such as a virtual server or a custom software program, or hardware-based, such as a physical server or a network device.

➢Cloud Consumers and Cloud Providers:

- The party that provides cloud-based IT resources is the cloud provider. The party that uses cloud-based IT resources is the cloud consumer.

Basic Concepts and Terminology

➢ **On-Premise:**

- An IT resource that is hosted in a conventional IT enterprise within an organizational boundary (that does not specifically represent a cloud) is considered to be located on the premises of the IT enterprise, or on-premise.
- This term is used to qualify an IT resource as an alternative to "cloud-based." An IT resource that is on-premise cannot be cloud-based, and vice-versa.
- An on-premise IT resource can access and interact with a cloud-based IT resource.
- An on-premise IT resource can be moved to a cloud, thereby changing it to a cloud-based IT resource.

Cloud Vs. Internet

- As a specific environment used to remotely provision IT resources, a cloud has a finite boundary.

- There are many individual clouds that are accessible via the Internet.

- Whereas the Internet provides open access to many Web-based IT resources, a cloud is typically privately owned and offers access to IT resources that is metered.

Cloud Vs. Internet

- Much of the Internet is dedicated to the access of content-based IT resources published via the World Wide Web.

- IT resources provided by cloud environments, on the other hand, are dedicated to supplying back-end processing capabilities and user-based access to these capabilities.

- It is not necessary for clouds to be Web-based. A cloud can be based on the use of any protocols that allow for the remote access to its IT resources.

Benefits to cloud consumers

- On-demand access to pay-as-you-go computing resources on a short-term basis.
- The ability to release these computing resources when they are no longer needed.
- Having unlimited computing resources that are available on demand.
- The ability to add or remove IT resources at a fine-grained level.
- Abstraction of the infrastructure so applications are not locked into devices or locations and can be easily moved if needed.

Risks and Challenges

- Increased Security Vulnerabilities: responsibility over data security becomes shared with the cloud provider.

- Longer geographic distances between the cloud consumer and cloud provider can require additional network hops that introduce fluctuating latency and potential bandwidth constraints.

- Limited Portability Between Cloud Providers.

Cloud Delivery Models

Cloud Delivery Models

- A *cloud delivery model* represents a specific, pre-packaged combination of IT resources offered by a cloud provider.
- Three common cloud delivery models have become widely established and formalized:
  - Infrastructure-as-a-Service (IaaS)
  - Platform-as-a-Service (PaaS)
  - Software-as-a-Service (SaaS)

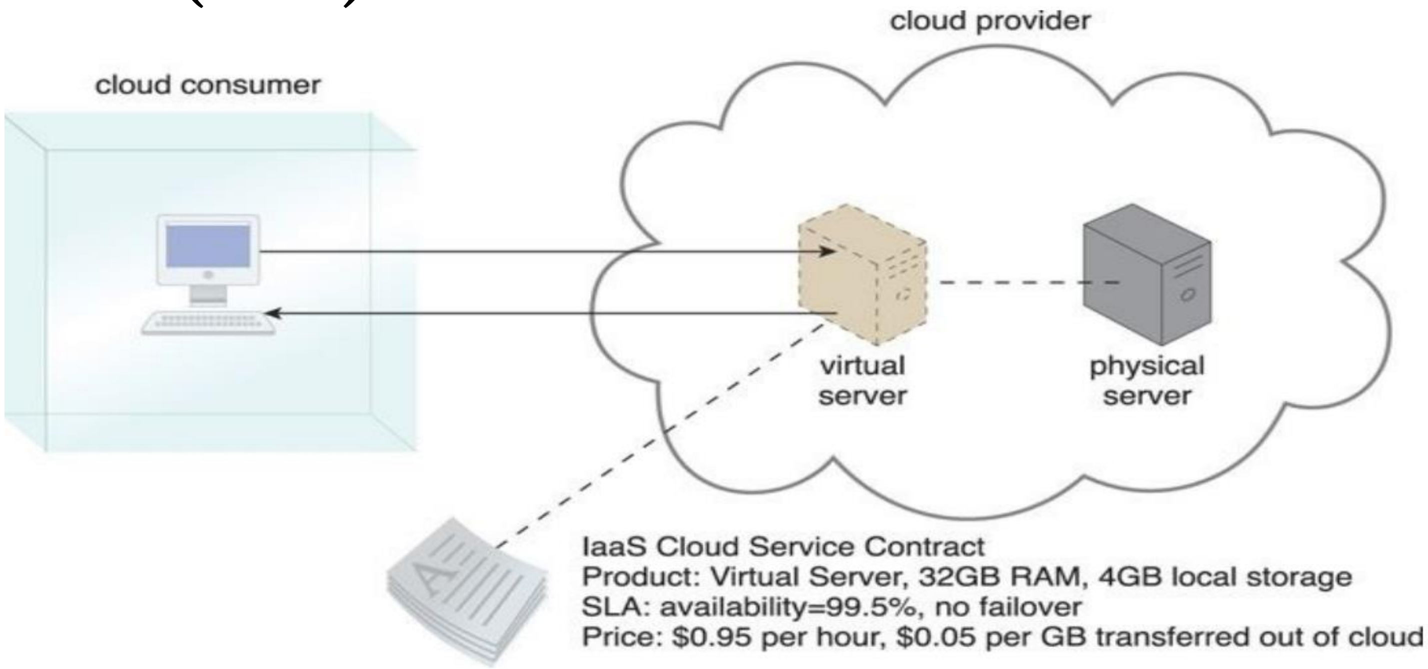
Cloud Delivery Models

- **Infrastructure-as-a-Service (IaaS):**
  - The IaaS delivery model represents a self-contained IT environment comprised of infrastructure-centric IT resources that can be accessed and managed via cloud service-based interfaces and tools.
  - This environment can include hardware, network, connectivity, operating systems, and so on.
  - The general purpose of an IaaS environment is to provide cloud consumers with a high level of control and responsibility over its configuration and utilization.
  - This model is therefore used by cloud consumers that require a high level of control over the cloud-based environment they intend to create.

Cloud Delivery Models

➢Infrastructure-as-a-Service (IaaS):

- A central and primary IT resource within a typical IaaS environment is the **virtual server**.
- Virtual servers are leased by specifying server hardware requirements, such as processor capacity, memory, and local storage space.



IaaS Cloud Service Contract
Product: Virtual Server, 32GB RAM, 4GB local storage
SLA: availability=99.5%, no failover
Price: $0.95 per hour, $0.05 per GB transferred out of cloud

Figure 4.11. A cloud consumer is using a virtual server within an IaaS environment. Cloud consumers are provided with a range of contractual guarantees by the cloud provider, pertaining to characteristics such as capacity, performance, and availability.

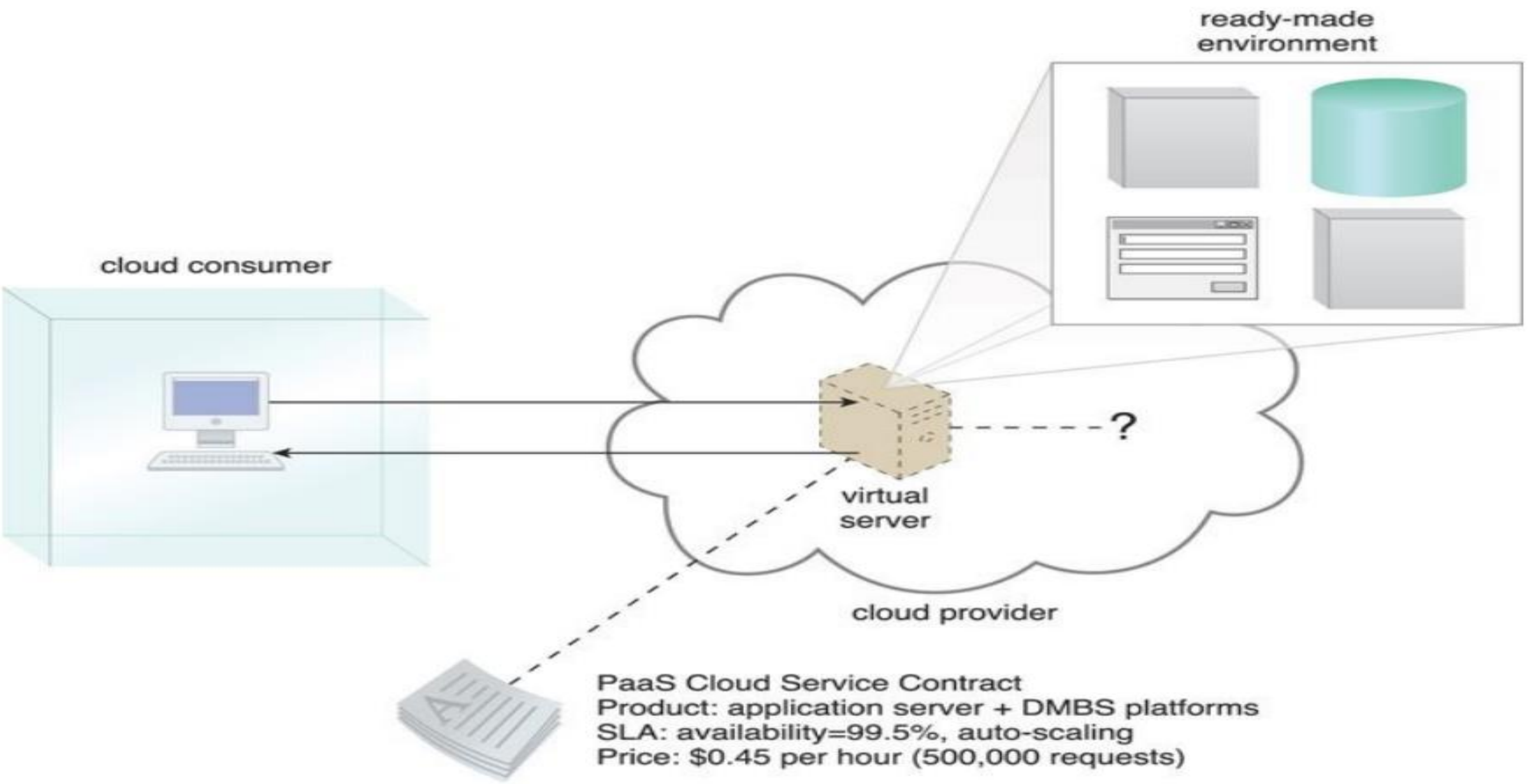
Cloud Delivery Models

➢ **Platform-as-a-Service (PaaS):**

- The PaaS delivery model represents a pre-defined "ready-to-use" environment typically comprised of already deployed and configured IT resources.
- By working within a ready-made platform, the cloud consumer is spared the administrative burden of setting up and maintaining the bare infrastructure IT resources provided via the IaaS model.

Cloud Delivery Models

- ➢ **Platform-as-a-Service (PaaS):**



PaaS Cloud Service Contract
Product: application server + DMBS platforms
SLA: availability=99.5%, auto-scaling
Price: $0.45 per hour (500,000 requests)

Figure 4.12. A cloud consumer is accessing a ready-made PaaS environment. The question mark indicates that the cloud consumer is intentionally shielded from the implementation details of the platform.

Cloud Delivery Models

- **Software-as-a-Service (SaaS):**
  - A software program positioned as a shared cloud service and made available as a "product" or generic utility represents the typical profile of a SaaS offering.
  - A cloud consumer is generally granted very limited administrative control over a SaaS implementation. It is most often provisioned by the cloud provider.

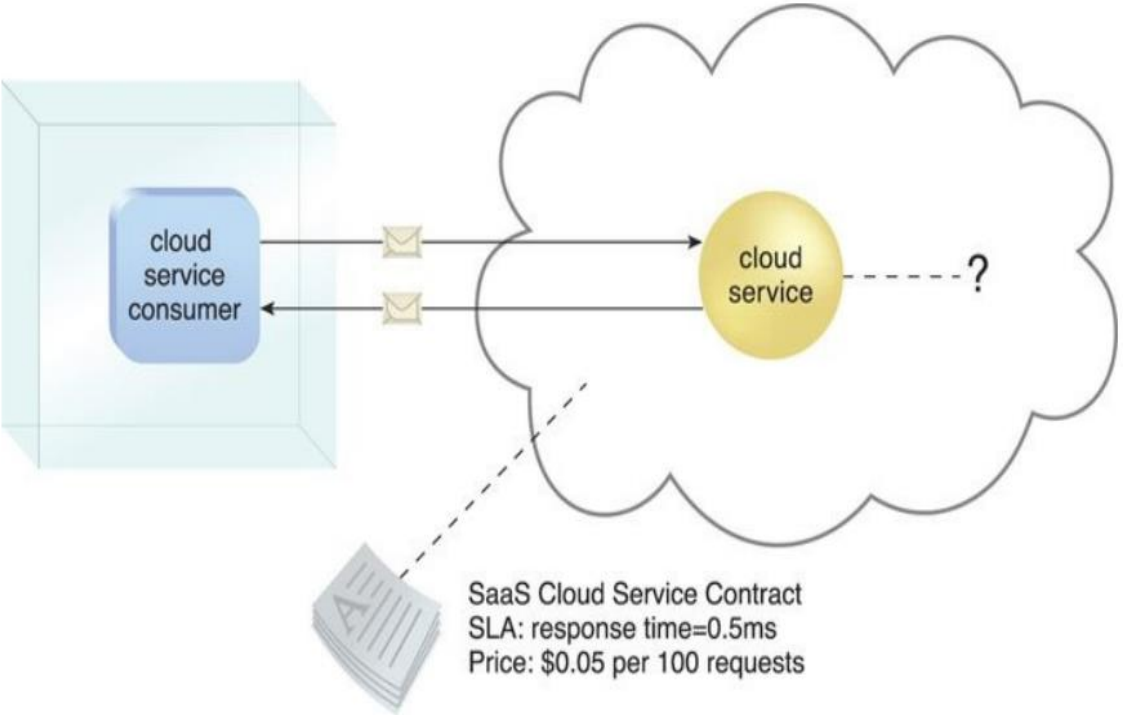
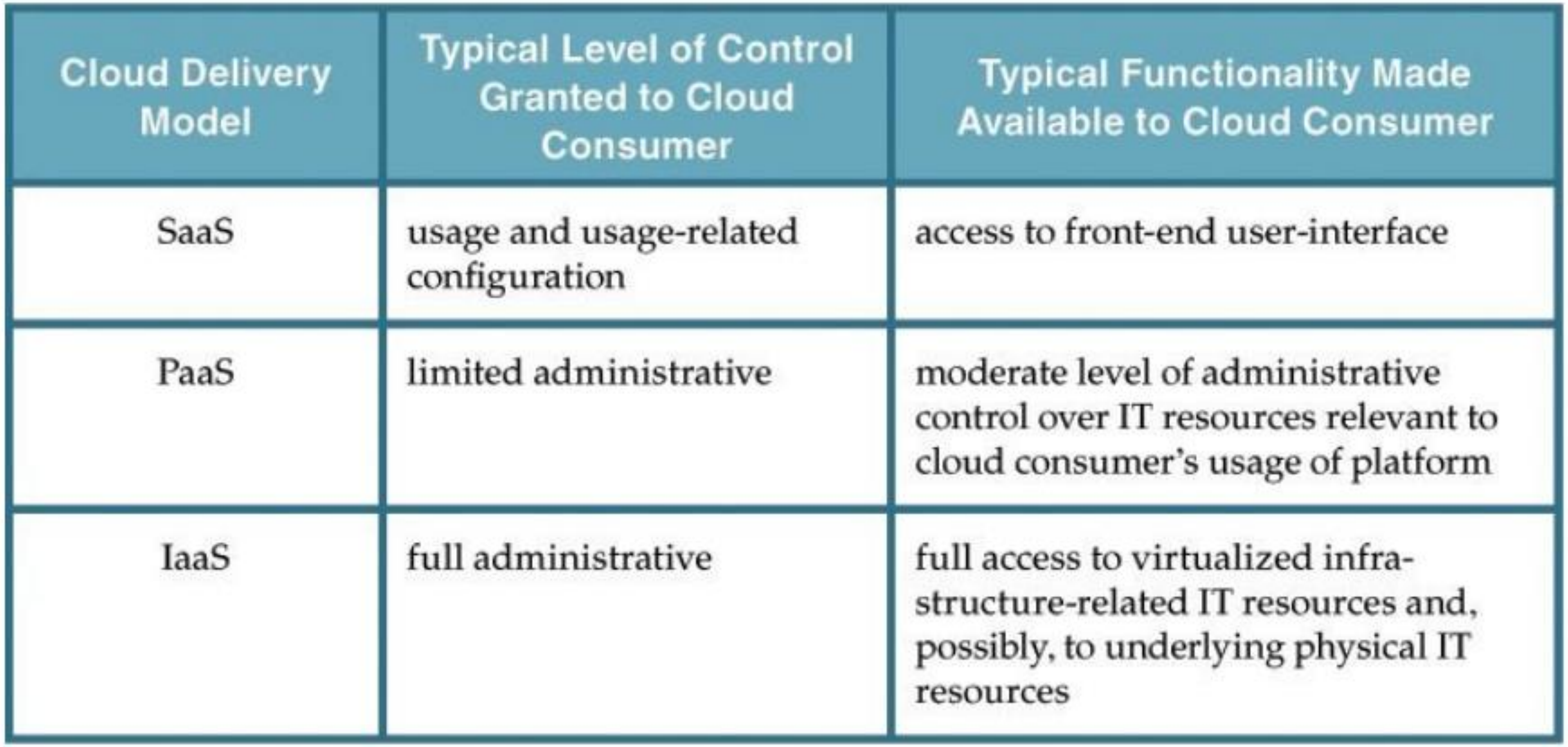


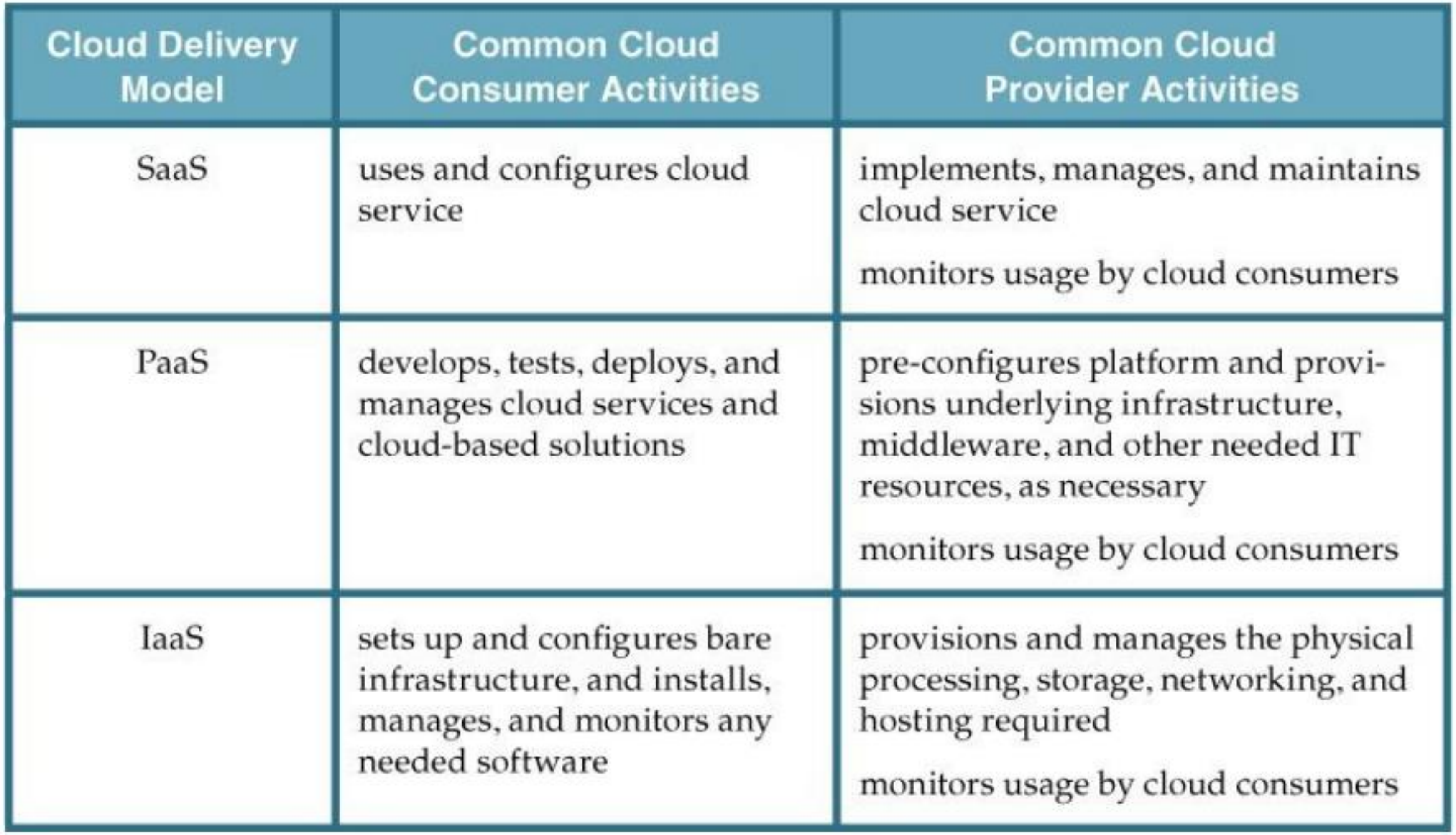
Figure 4.13. The cloud service consumer is given access the cloud service contract, but not to any underlying IT resources or implementation details.

Cloud Delivery Models



Cloud Delivery Model	Typical Level of Control Granted to Cloud Consumer	Typical Functionality Made Available to Cloud Consumer
SaaS	usage and usage-related configuration	access to front-end user-interface
PaaS	limited administrative	moderate level of administrative control over IT resources relevant to cloud consumer's usage of platform
IaaS	full administrative	full access to virtualized infra-structure-related IT resources and, possibly, to underlying physical IT resources

Cloud Delivery Models



Cloud Delivery Model	Common Cloud Consumer Activities	Common Cloud Provider Activities
SaaS	uses and configures cloud service	implements, manages, and maintains cloud service monitors usage by cloud consumers
PaaS	develops, tests, deploys, and manages cloud services and cloud-based solutions	pre-configures platform and provisions underlying infrastructure, middleware, and other needed IT resources, as necessary monitors usage by cloud consumers
IaaS	sets up and configures bare infrastructure, and installs, manages, and monitors any needed software	provisions and manages the physical processing, storage, networking, and hosting required monitors usage by cloud consumers

Cloud Deployment Models

Cloud Deployment Models

➢ A cloud deployment model represents a specific type of cloud environment, primarily distinguished by ownership, size, and access. There are four common cloud deployment models:

- Public cloud
- Private cloud
- Community cloud
- Hybrid cloud

Cloud Deployment Models

➢ **Public cloud:**

- A public cloud is a publicly accessible cloud environment owned by a third-party cloud provider.
- The IT resources on public clouds are usually provisioned via the previously described cloud delivery models and are generally offered to cloud consumers at a cost or are commercialized via other avenues (such as advertisement).
- The cloud provider is responsible for the creation and on-going maintenance of the public cloud and its IT resources.

Cloud Deployment Models

➢**Private cloud:**

- A private cloud is owned by a single organization. Private clouds enable an organization to use cloud computing technology as a means of centralizing access to IT resources by different parts, locations, or departments of the organization.
- With a private cloud, the same organization is technically both the cloud consumer and cloud provider.
- Even though the private cloud may physically reside on the organization's premises, IT resources it hosts are still considered "cloud-based" as long as they are made remotely accessible to cloud consumers.

Cloud Deployment Models

➢ **Community cloud:**

- A community cloud access is limited to a specific community of cloud consumers.
- The community cloud may be jointly owned by the community members or by a third-party cloud provider that provisions a public cloud with limited access.
- The member cloud consumers of the community typically share the responsibility for defining and evolving the community cloud.
- Membership in the community does not necessarily guarantee access to or control of all the cloud’s IT resources.
- Parties outside the community are generally not granted access unless allowed by the community.

Cloud Deployment Models

- **Hybrid cloud:**
  - A hybrid cloud is a cloud environment comprised of two or more different cloud deployment models.
  - For example, a cloud consumer may choose to deploy cloud services processing sensitive data to a private cloud and other, less sensitive cloud services to a public cloud.

Cloud Deployment Models

➢ **Other Cloud Deployment Models:**

- **Virtual Private Cloud** – Also known as a “dedicated cloud” or “hosted cloud,” this model results in a self-contained cloud environment hosted and managed by a public cloud provider, and made available to a cloud consumer.

- **Inter-Cloud** – This model is based on an architecture comprised of two or more inter-connected clouds.

Cloud-Enabling Technology

Cloud-Enabling Technology

- Modern-day clouds are supported by a set of primary technology components that collectively enable key features and characteristics associated with contemporary cloud computing such as:
  - **Internet Architecture**
  - **Data Center Technology**
  - **Virtualization Technology**
  - **Web Technology**
  - **Multitenant Technology**

Cloud-Enabling Technology

➢**Internet Architecture:**

- All clouds must be connected to a network.
- The Internet, allow for the remote provisioning of IT resource.

➢**Data Center Technology:**

- At the foundation of any cloud is the data center hardware on which workloads run—servers, storage, and networking.
- Grouping IT resources in close proximity with one another, rather than having them geographically dispersed, allows for power sharing, higher efficiency in shared IT resource usage, and improved accessibility.

Cloud-Enabling Technology

➢ Virtualization Technology:

- **Virtualization is the process of converting a physical IT resource into a virtual IT resource.**
- The first step in creating a new virtual server through **virtualization software** is the allocation of physical IT resources, followed by the **installation of an operating system**.
- Virtualization software (called **hypervisor**) runs on a physical server called a host or physical host. A hypervisor is generally limited to one physical server and can therefore only create virtual images of that server.
- Virtualized IT resource management is often supported by **virtualization infrastructure management (VIM)** tools that collectively manage virtual IT resources.

Cloud-Enabling Technology

➢ **Virtualization Technology:**

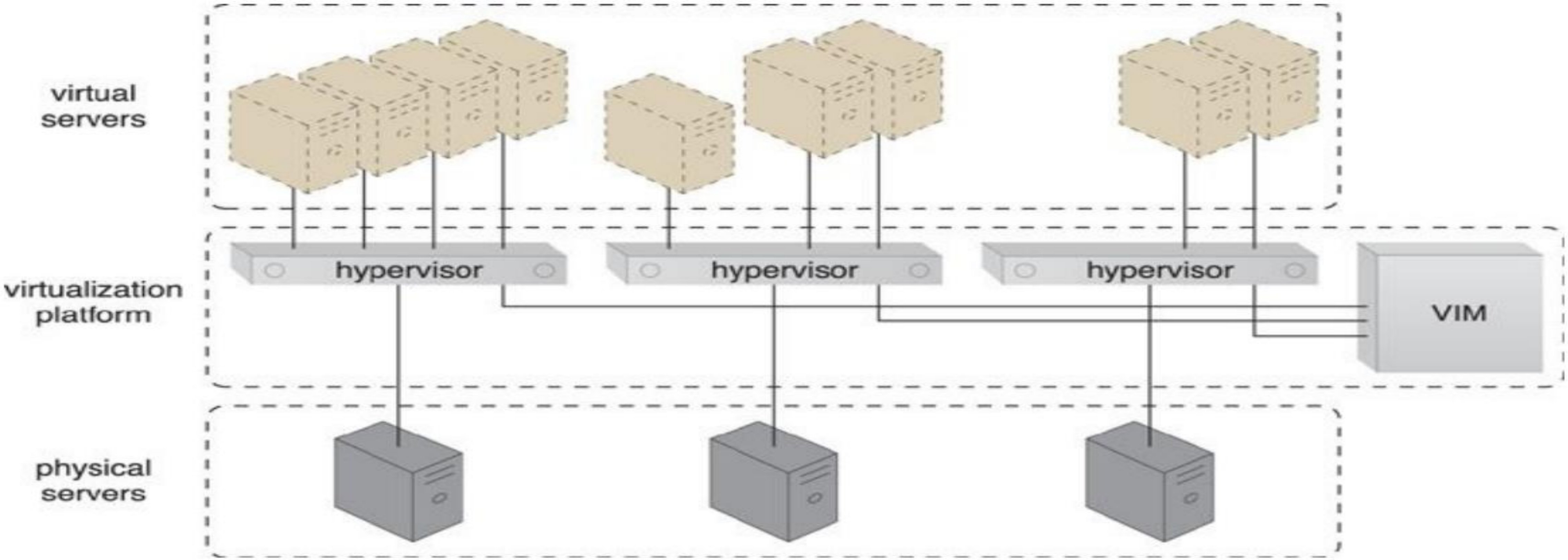


Figure 8.27. Virtual servers are created via individual hypervisor on individual physical servers. All three hypervisors are jointly controlled by the same VIM.

Cloud-Enabling Technology

- **Web Technology:**
  - Web technology is generally used as both the **implementation medium** and the **management interface** for cloud services.
  - The World Wide Web is a system of interlinked IT resources that are accessed through the Internet. The two basic components of the Web are **the Web browser client** and the **Web server**.
  - Three fundamental elements comprise the technology architecture of the Web:
    - **URL**: identifiers that point to Web-based resources
    - **HTTP**: communications protocol used to exchange content and data throughout the World Wide Web
    - **HTML, XML**: Markup languages provide a means of expressing Web data and metadata.

Cloud-Enabling Technology

➢ **Web Technology:**

- For example, a Web browser can request to execute an action like read, write, update, or delete on a Web resource on the Internet, and proceed to identify and locate the Web resource through its URL.
- The request is sent using HTTP to the resource host, which is also identified by a URL.
- The Web server locates the Web resource and performs the requested operation, which is followed by a response being sent back to the client.
- The response may be comprised of content that includes HTML and XML statements.

Cloud-Enabling Technology

➢ **Multitenant Technology:**

- The multitenant application design was created to enable multiple users (tenants) to access the same application logic simultaneously.
- Each tenant has its own view of the application that it uses.

Cloud Architecture

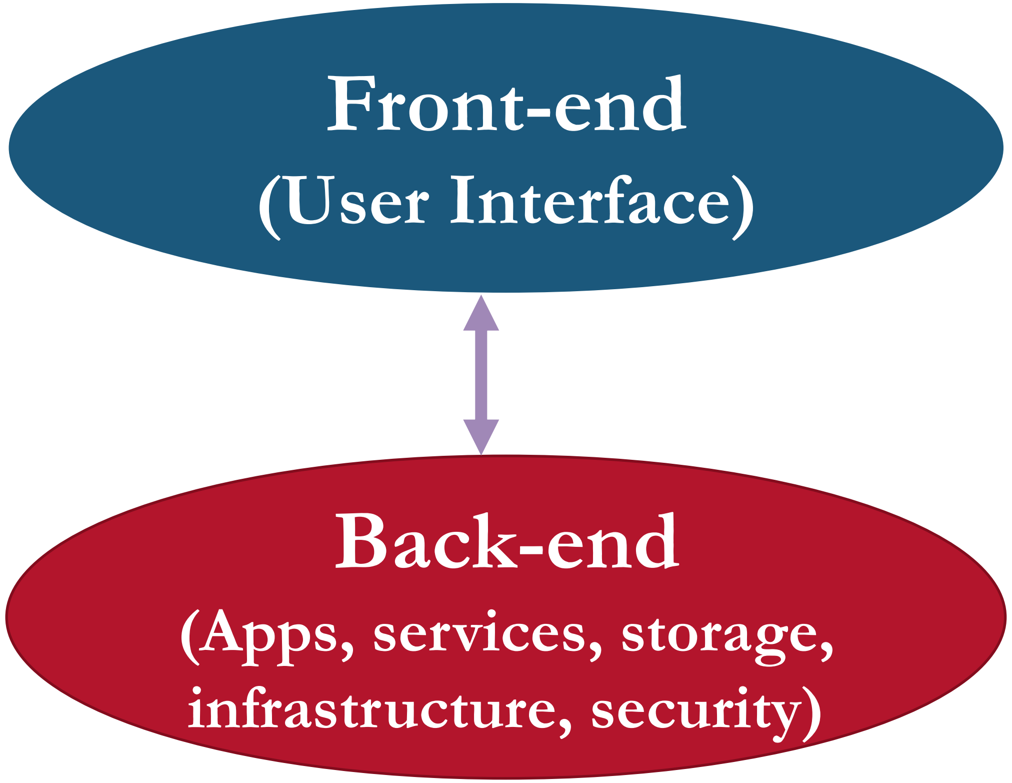
Cloud Architecture

➢ **Cloud Infrastructure Components:**

- There are two sides of the cloud environment. The **front end** is what's visible to the end user; in other words, it's the **user interface**. The **back-end** infrastructure is what runs the cloud.
- This back end is made up of data center hardware, virtualization, applications, and services.
- The front end communicates with the back end through **middleware**.
- **Middleware** is software that enables one or more kinds of communication or connectivity between applications or application components in a distributed network.

Cloud Architecture

- **Cloud Infrastructure Components:**



Front-end
(User Interface)

Back-end
(Apps, services, storage, infrastructure, security)

Cloud Architecture

- **Principles of cloud architecture:**
  - Before you can design your cloud, you must first assess your existing environment and business needs. Here are just some of the questions your team will need to explore:
    - What are your existing workloads and applications? Where do they currently run, and who uses them?
    - How is your overall cloud utilization? Is it lower than it should be because it was designed to accommodate peak loads? Do you need to scale up to support new workloads?
    - Are you running into any bottlenecks in compute performance, memory, or networking?

Cloud Architecture

- **Fundamental Cloud Architecture examples:**
  - Workload Distribution Architecture
  - Dynamic Scalability Architecture
  - Elastic Resource Capacity Architecture
  - Service Load Balancing Architecture
  - …..and many more

Cloud Architecture: Fundamental Cloud Architecture

➢ Workload Distribution Architecture:

- IT resources can be 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.

Cloud Architecture: Fundamental Cloud Architecture

➢Workload Distribution Architecture:

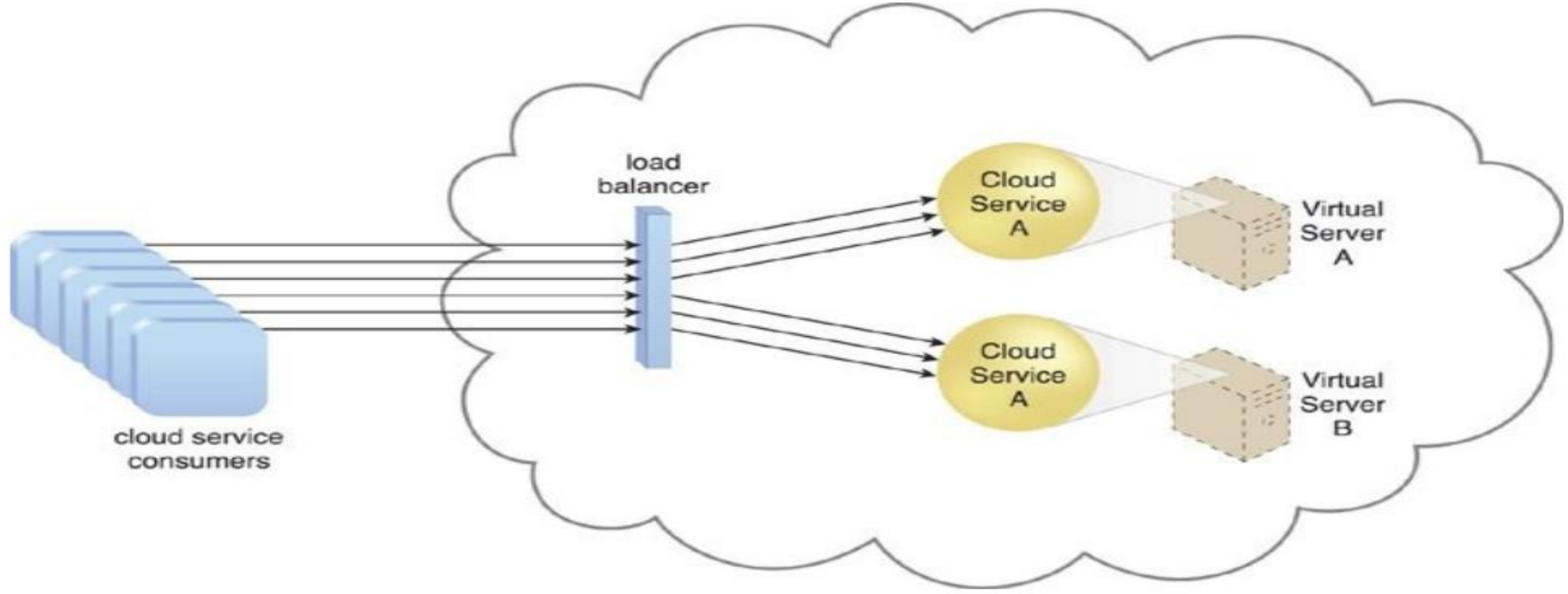


Figure 11.1. A redundant copy of Cloud Service A is implemented on Virtual Server B. The load balancer intercepts cloud service consumer requests and directs them to both Virtual Servers A and B to ensure even workload distribution.

Cloud Architecture: Fundamental Cloud Architecture

➢ Dynamic Scalability Architecture:

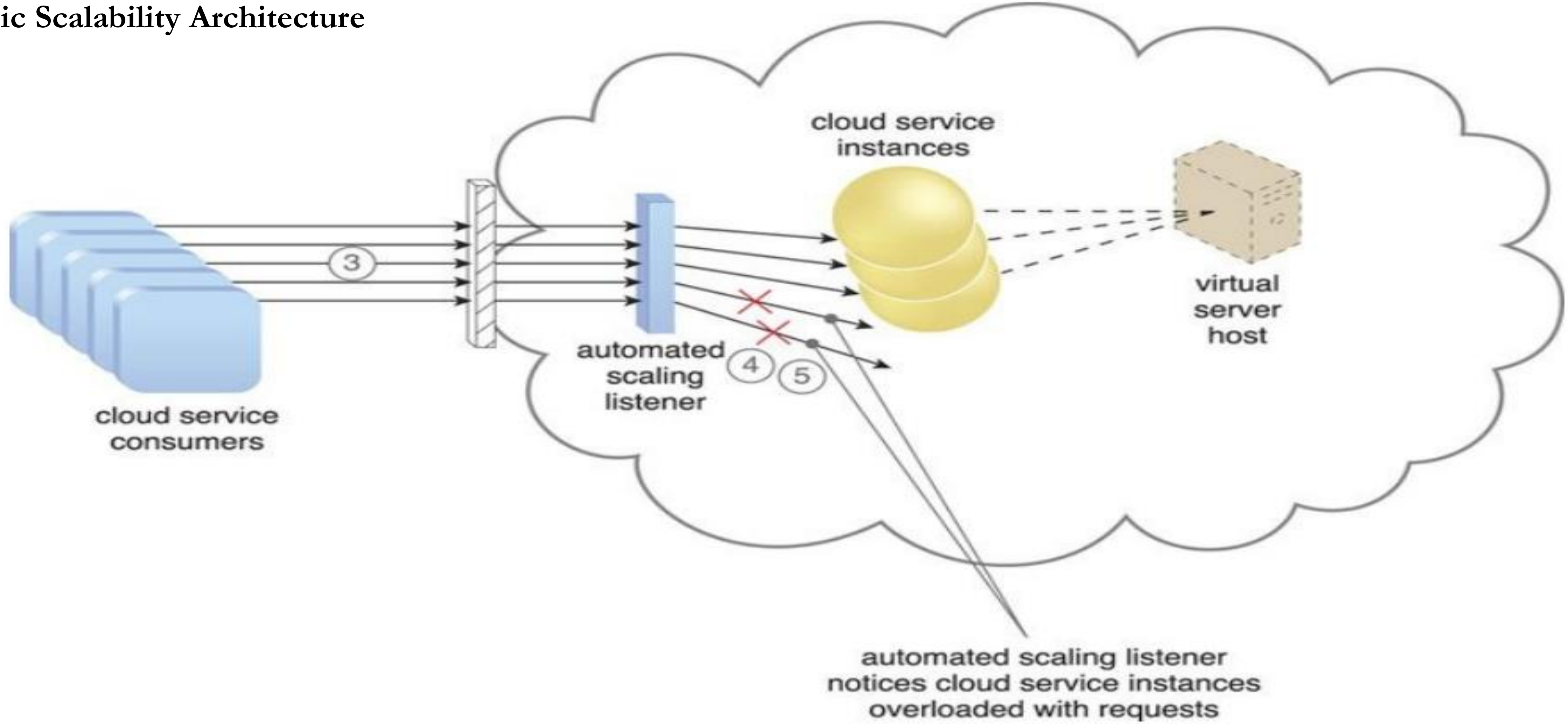
- Dynamic allocation enables **variable utilization** as dictated by usage demand fluctuations, since unnecessary IT resources are efficiently reclaimed without requiring manual interaction.
- The **automated scaling listener** is configured with **workload thresholds** that dictate when new IT resources need to be added to the workload processing. This mechanism can be provided with logic that determines how many additional IT resources can be dynamically provided, based on the terms of a given cloud consumer's provisioning contract.

Cloud Architecture: Fundamental Cloud Architecture

➢Dynamic Scalability Architecture:

- The following types of dynamic scaling are commonly used:
  - **Dynamic Horizontal Scaling:** IT resource instances are ***scaled out and in*** to handle fluctuating workloads. The automatic scaling listener monitors requests and signals resource replication to initiate IT resource duplication.
  - **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:** The IT resource is relocated to a host with more capacity.

Dynamic Scalability Architecture



cloud service instances
virtual server host
3
automated scaling listener
4
5
cloud service consumers
automated scaling listener notices cloud service instances overloaded with requests

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

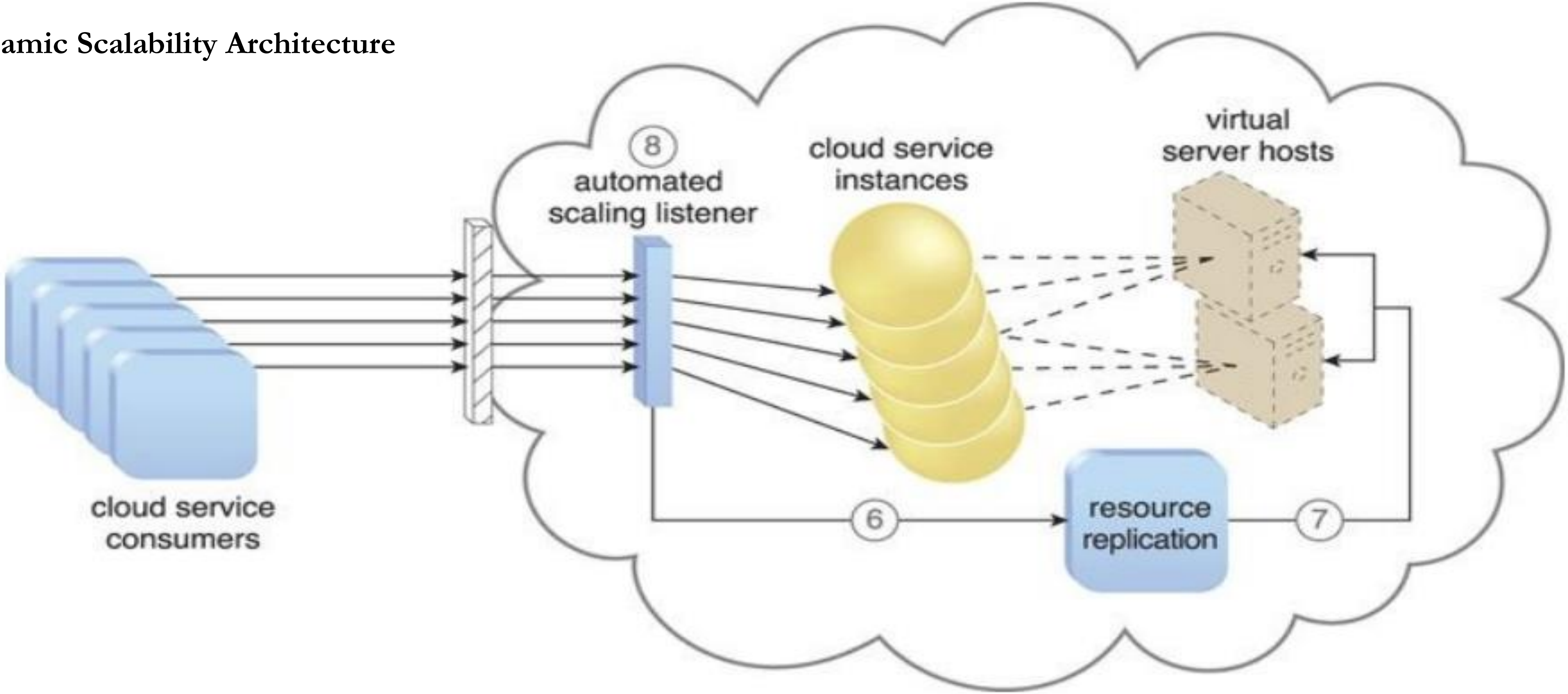


Figure 11.7. The automated scaling listener sends a signal to the resource replication mechanism (6), which creates more instances of the cloud service (7). Now that the increased workload has been accommodated, the automated scaling listener resumes monitoring and detracting and adding IT resources,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

Reference

- Cloud Computing: Concepts, Technology & Architecture by Thomas Erl et al. Prentice Hall, 2013.
  - The source of all images and Tables are this reference.

Thank You