CLOUD COMPUTING
(Undergraduate Course)

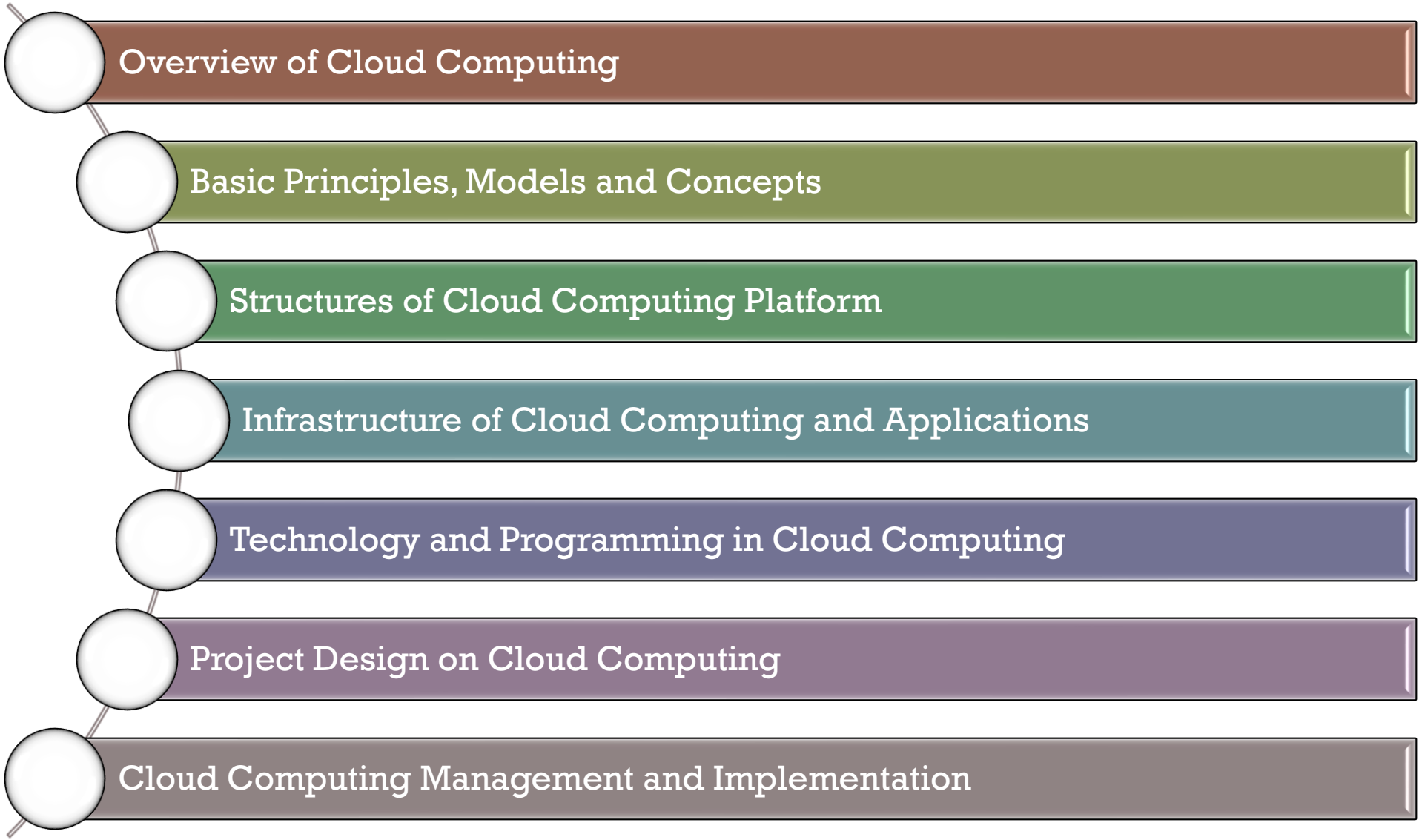
Chapter 4 – Cloud Infrastructure and Applications

Presenter: **Dr. Nguyen Dinh Long**

Email: dinhlonghcmut@gmail.com

Oct. 2022

Outline


Overview of Cloud Computing
Basic Principles, Models and Concepts
Structures of Cloud Computing Platform
Infrastructure of Cloud Computing and Applications
Technology and Programming in Cloud Computing
Project Design on Cloud Computing
Cloud Computing Management and Implementation

References

Main:

- Thomas Erl, Zaigham Mahmood, and Ricardo Puttini. 2013. *Cloud Computing Concepts, Technology & Architecture*. Prentice Hall.
- Michael J. Kavis. 2014. *Architecting the Cloud: Design Decisions for Cloud Computing Service Models*. Wiley
- Arshdeep Bahga, and Vijay Madisetti. 2013. *Cloud Computing: A Hands-On Approach*. CreateSpace Independent Publishing Platform

More:

- Rajkuma Buyya, Jame Broberg and Andrzej Goscinski. 2011. *Cloud Computing –Principles and paradigms*, Wiley
- Nick Antonopoulos, and Lee Gillam. 2010. *Cloud Computing - Principles, Systems and Applications*, Springer-Verlag London Limited.

- Slides here are modified from several sources in Universities and Internet.

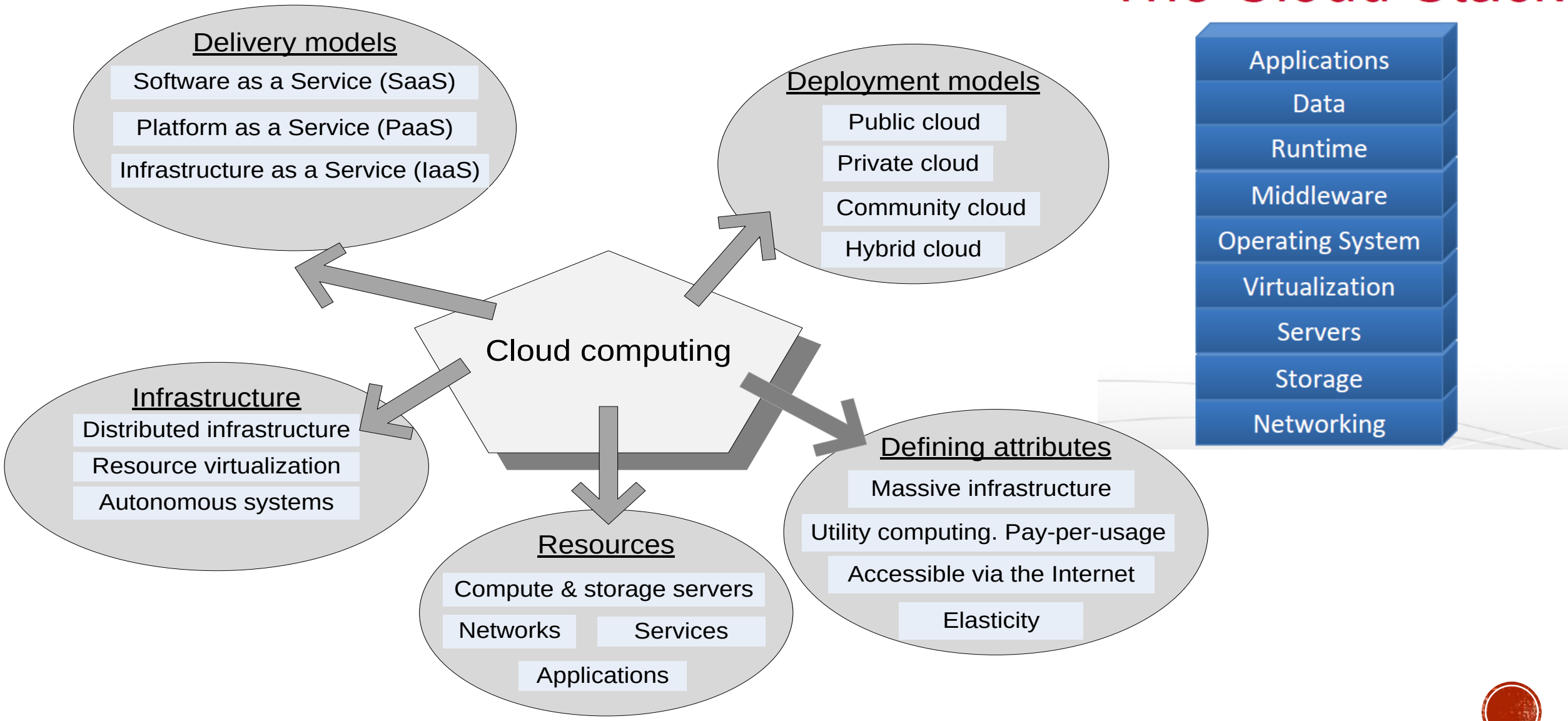
Content of Chapter 4

1. Cloud infrastructure

2. Models of AWS, AZURE, GOOGLE CLOUD

3. Applications

BASIC CONCEPTS AND PRINCIPLE COMPONENTS



What is Cloud computing?

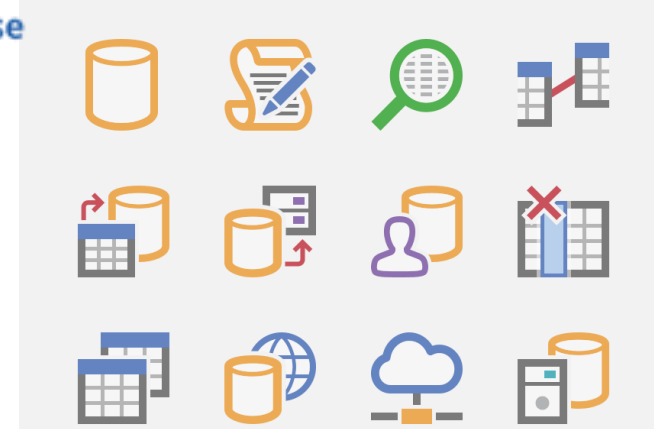
Linh hoạt lưu trữ

Đa ứng dụng, phương tiện

Đa tương tác
Low coding

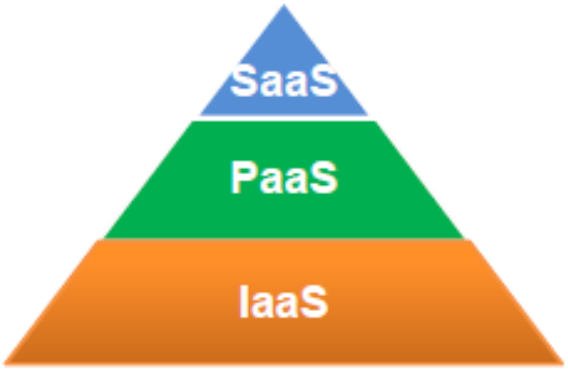
Máy chủ ảo
(serverless computing)

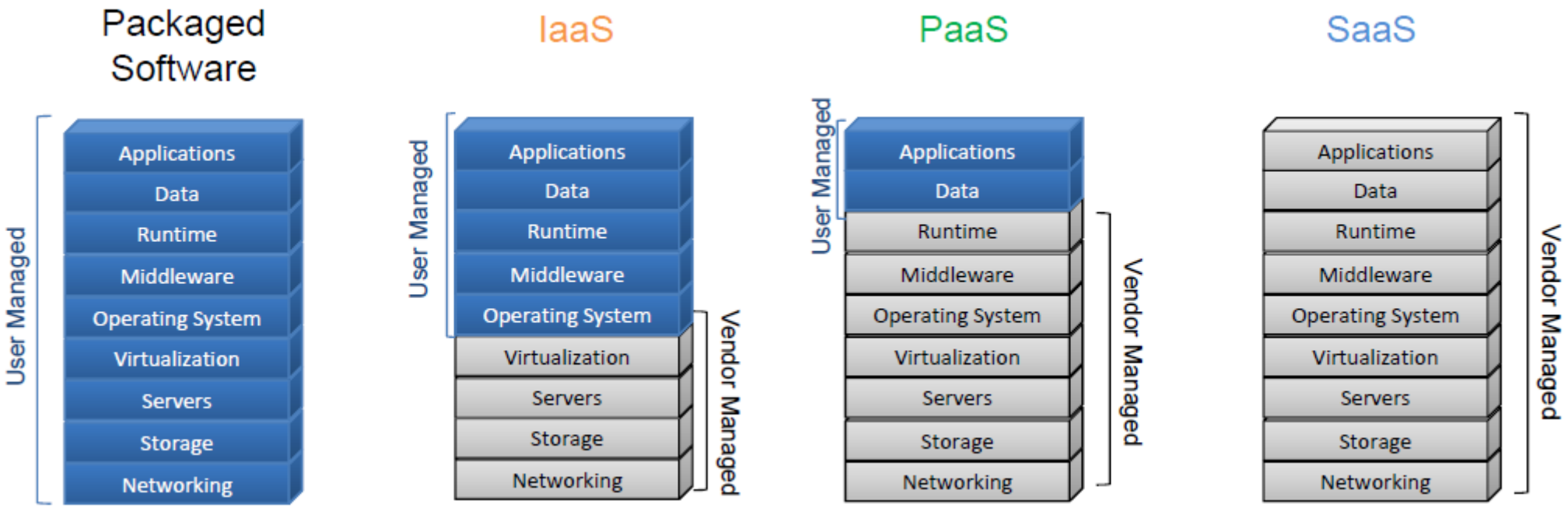
Cơ sở dữ liệu trực tuyến



Structure of Cloud computing

- Cloud Service Layers …





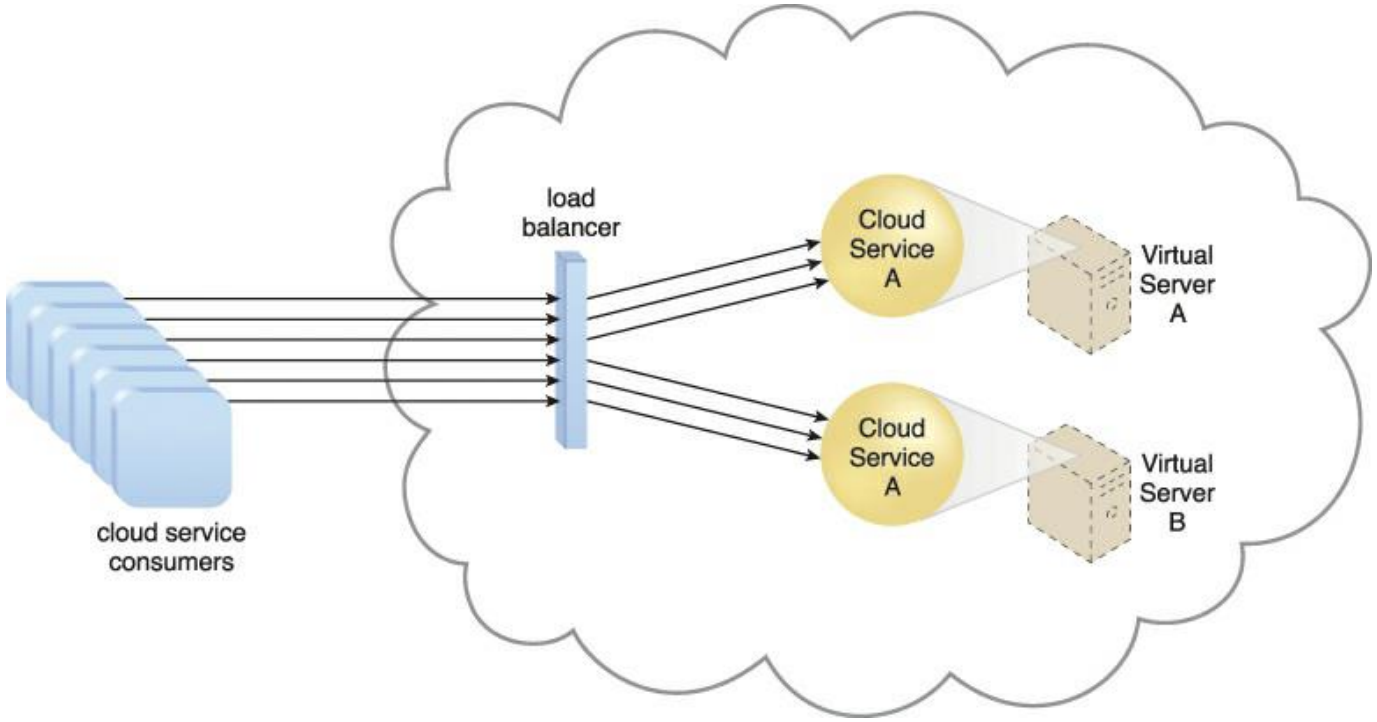
Cloud computing infrastructure

Workload Distribution Architecture:

The resulting workload distribution architecture reduces both IT resource overutilization and under-utilization to an extent dependent upon the sophistication of the load balancing algorithms and runtime logic.

Load balancing systems applied to specific IT resources usually produce specialized variations of this architecture that incorporate aspects of load balancing, such as:

- the service load balancing architecture explained later
- the load balanced virtual server architecture
- the load balanced virtual switches architecture



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 computing infrastructure

Workload Distribution Architecture:

The virtual server and cloud storage device mechanisms to which load balancing can be applied.

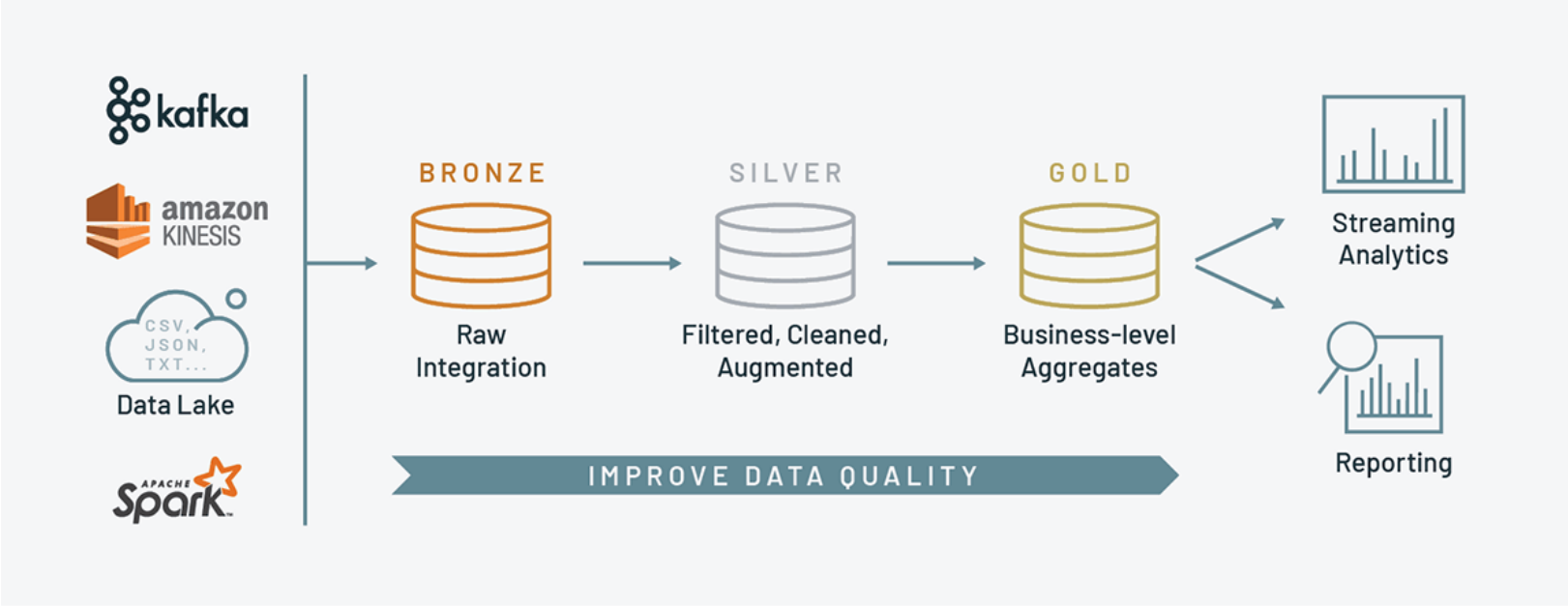
- Audit Monitor – When distributing runtime workloads, the type and geographical location of the IT resources that process the data can determine whether monitoring is necessary to fulfill legal and regulatory requirements.
- Cloud Usage Monitor – Various monitors can be involved to carry out runtime workload tracking and data processing.
- Hypervisor – Workloads between hypervisors and the virtual servers that they host may require distribution.
- Logical Network Perimeter – The logical network perimeter isolates cloud consumer network boundaries in relation to how and where workloads are distributed.
- Resource Cluster – Clustered IT resources in active/active mode are commonly used to support workload balancing between different cluster nodes.
- Resource Replication – This mechanism can generate new instances of virtualized IT resources in response to runtime workload distribution demands.

Cloud computing infrastructure

- **Workload Distribution Architecture:**

The virtual server and cloud storage device mechanisms to which load balancing can be applied.

- Audit Monitor – When distributing runtime workloads, the type and geographical location of the IT resources that process the data can determine whether monitoring is necessary to fulfill legal and regulatory requirements.

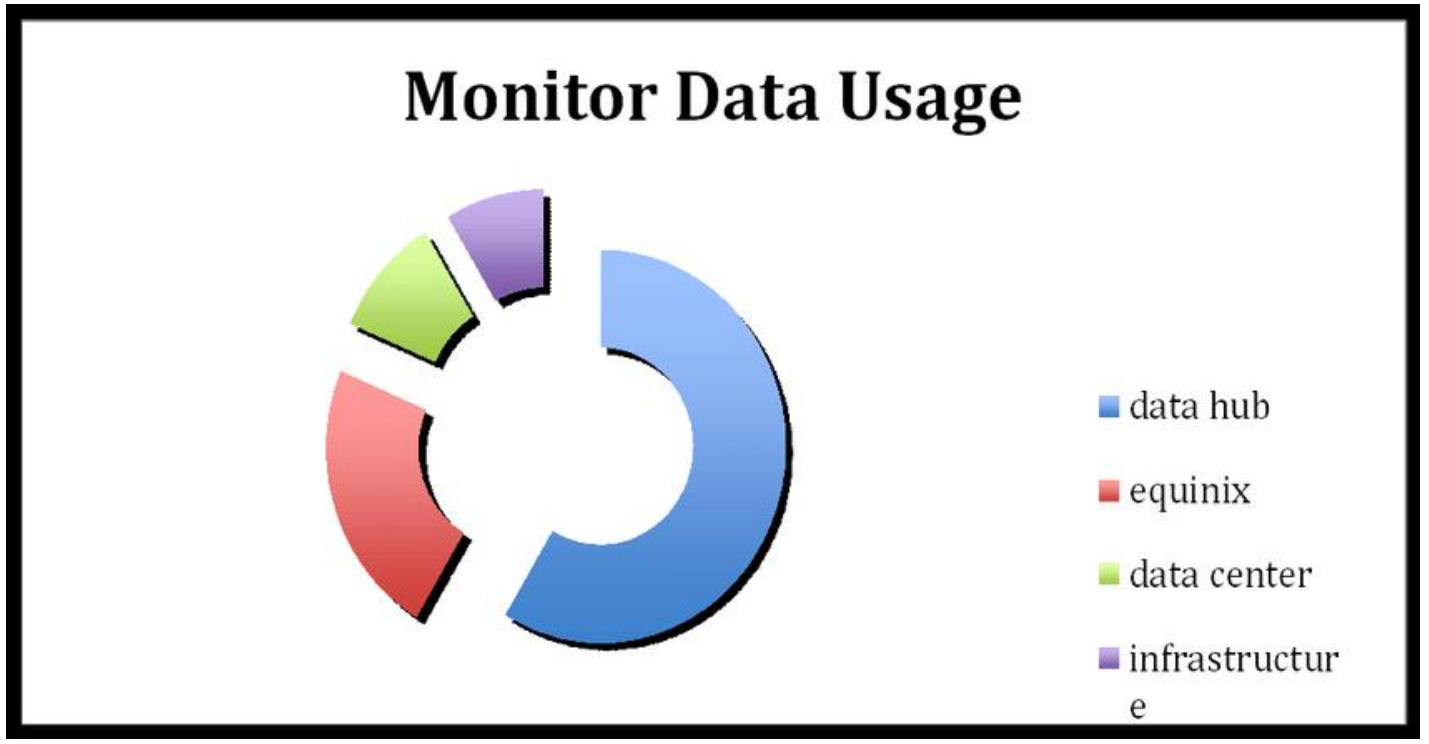


Cloud computing infrastructure

- **Workload Distribution Architecture:**

The virtual server and cloud storage device mechanisms to which load balancing can be applied.

- Cloud Usage Monitor – Various monitors can be involved to carry out runtime workload tracking and data processing.

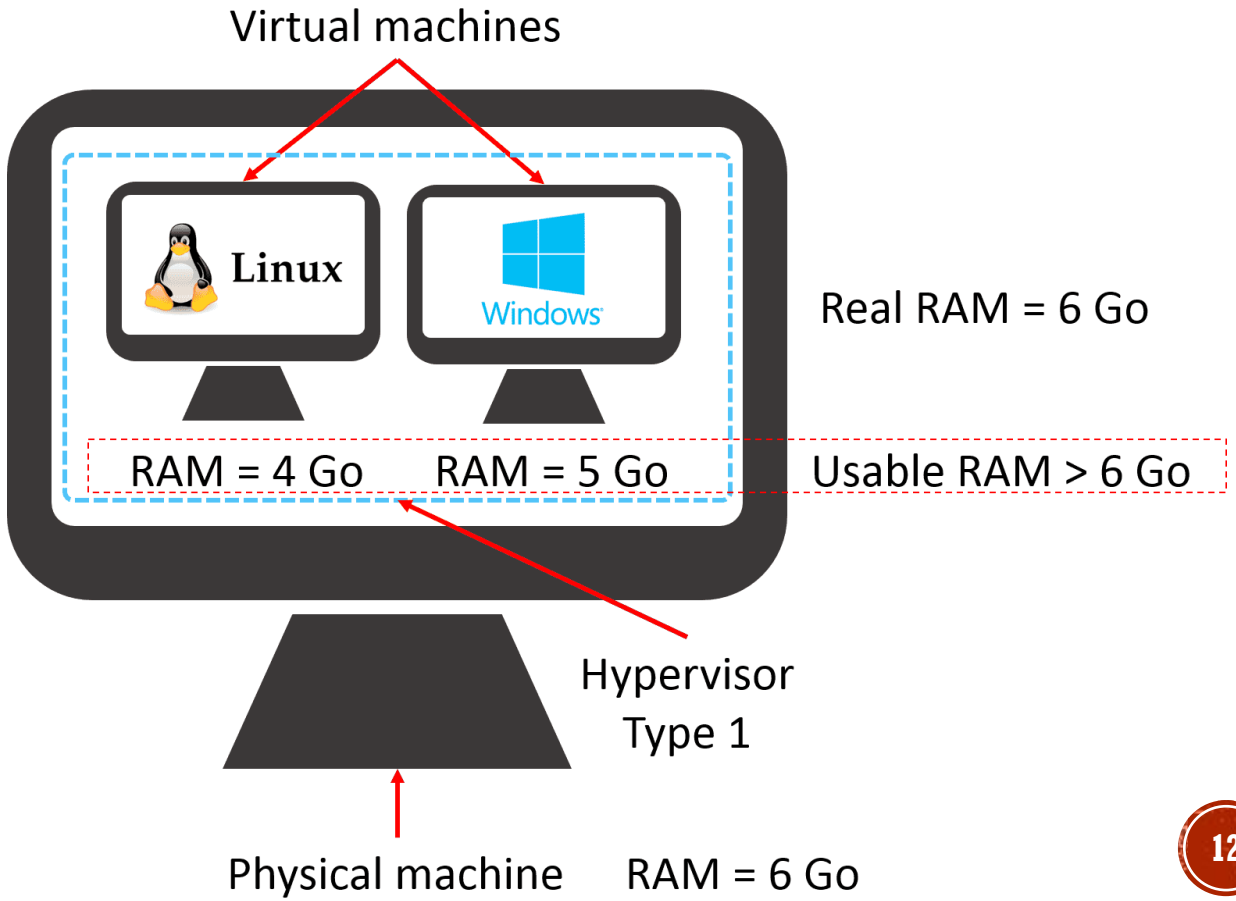


Cloud computing infrastructure

- **Workload Distribution Architecture:**

The virtual server and cloud storage device mechanisms to which load balancing can be applied.

- Hypervisor – Workloads between hypervisors and the virtual servers that they host may require distribution.

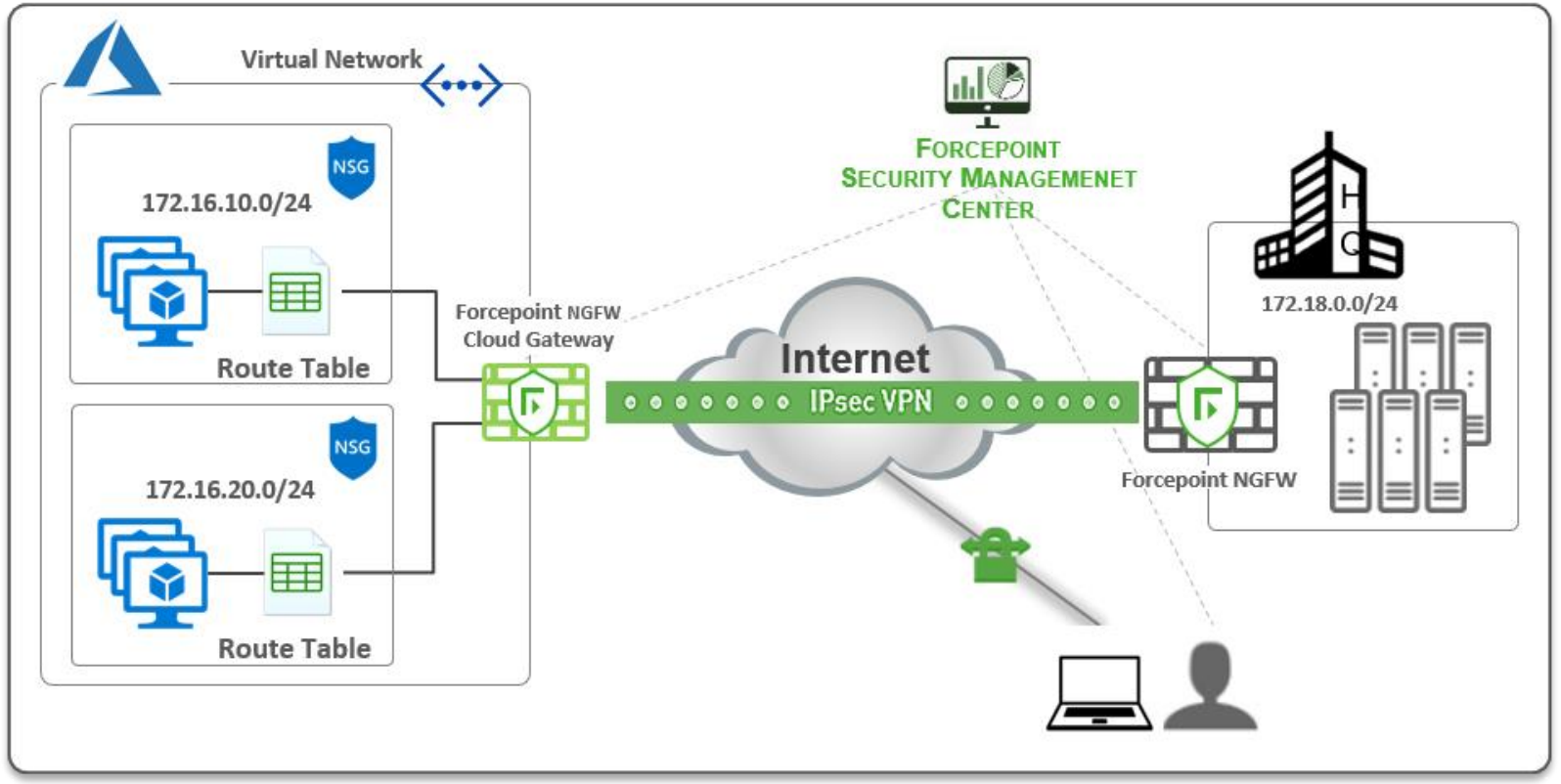


Cloud computing infrastructure

- **Workload Distribution Architecture:**

The virtual server and cloud storage device mechanisms to which load balancing can be applied.

- Logical Network Perimeter – The logical network perimeter isolates cloud consumer network boundaries in relation to how and where workloads are distributed.

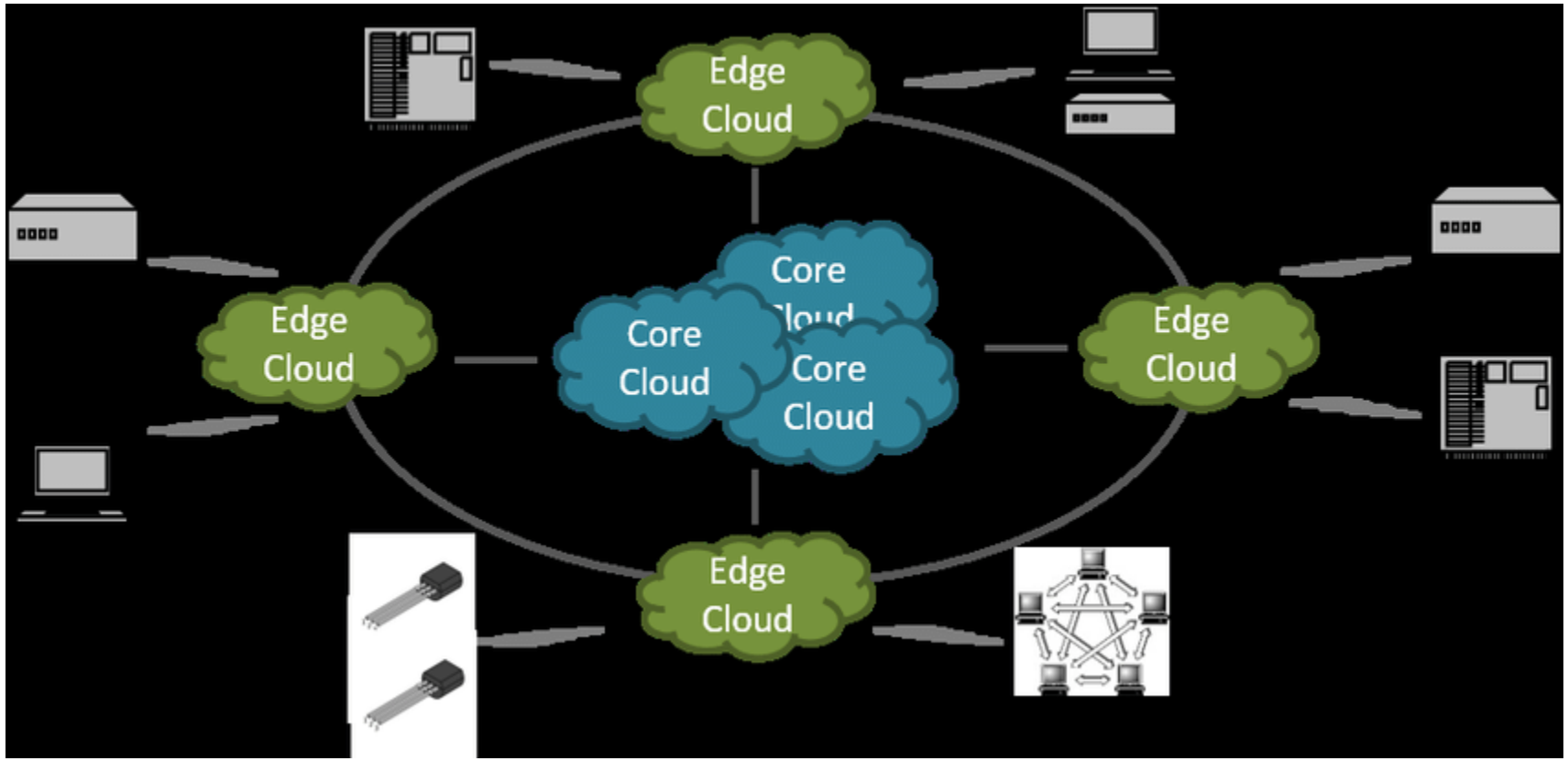


Cloud computing infrastructure

- **Workload Distribution Architecture:**

The virtual server and cloud storage device mechanisms to which load balancing can be applied.

- Resource Cluster – Clustered IT resources in active/active mode are commonly used to support workload balancing between different cluster nodes.

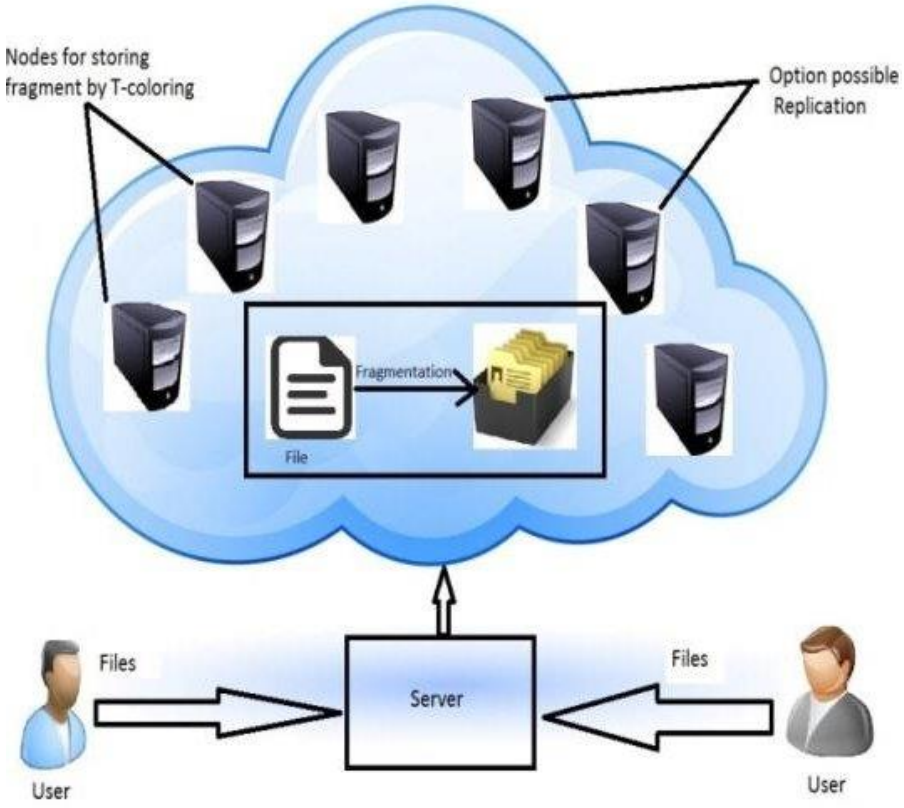


Cloud computing infrastructure

- **Workload Distribution Architecture:**

The virtual server and cloud storage device mechanisms to which load balancing can be applied.

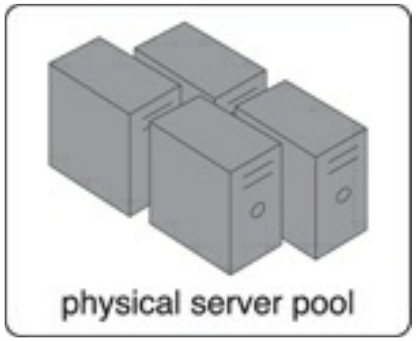
- Resource Replication – This mechanism can generate new instances of virtualized IT resources in response to runtime workload distribution demands.



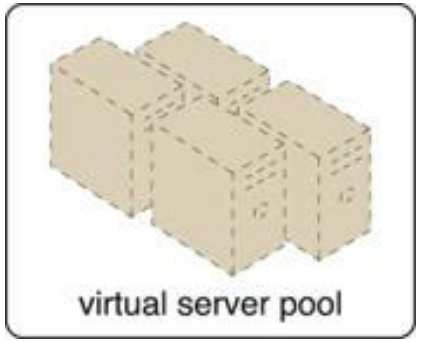
Cloud computing infrastructure

- **Resource Pooling Architecture:**

A resource pooling architecture is based on the use of one or more resource pools, in which identical IT resources are grouped and maintained by a system that automatically ensures that they remain synchronized.



Physical server pools are composed of networked servers that have been installed with operating systems and other necessary programs and/or applications and are ready for immediate use.

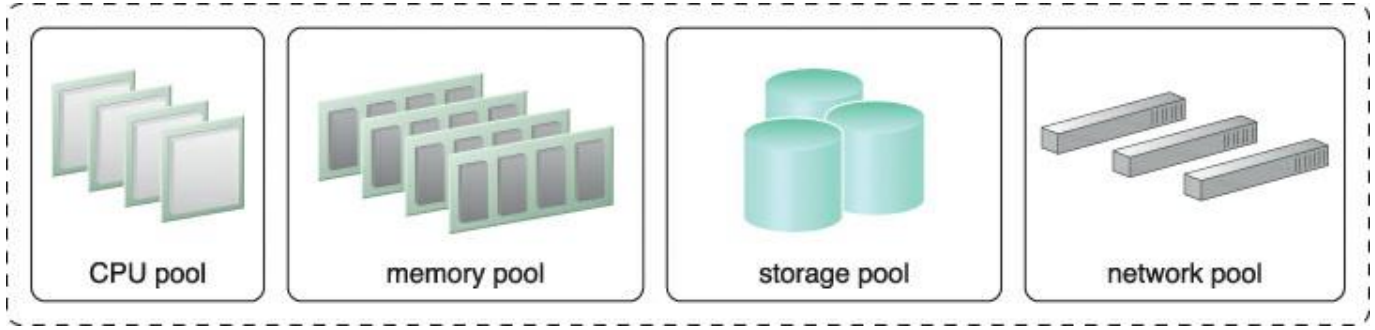


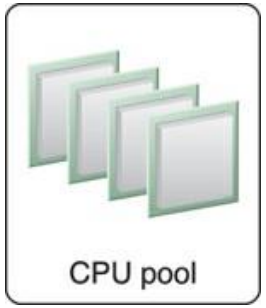
Virtual server pools are usually configured using one of several available templates chosen by the cloud consumer during provisioning.

For example, a cloud consumer can set up a pool of mid-tier Windows servers with 4 GB of RAM or a pool of low-tier Ubuntu servers with 2 GB of RAM.

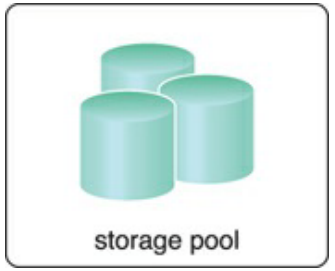
Cloud computing infrastructure

- **Resource Pooling Architecture:**

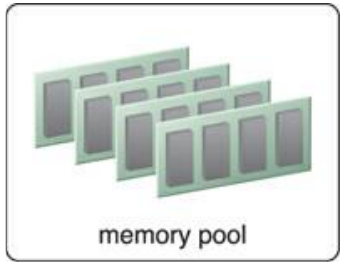




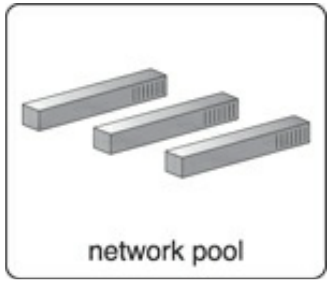
CPU pools are ready to be allocated to virtual servers, and are typically broken down into individual processing cores.



Storage pools, or cloud storage device pools, consist of file-based or block-based storage structures that contain empty and/or filled cloud storage devices.



Pools of physical RAM can be used in newly provisioned physical servers or to vertically scale physical servers.



Network pools (or interconnect pools) are composed of different preconfigured network connectivity devices.
For example, a pool of virtual firewall devices or physical network switches can be created for redundant connectivity, load balancing, or link aggregation.

Cloud computing infrastructure

Dynamic Scalability Architecture:

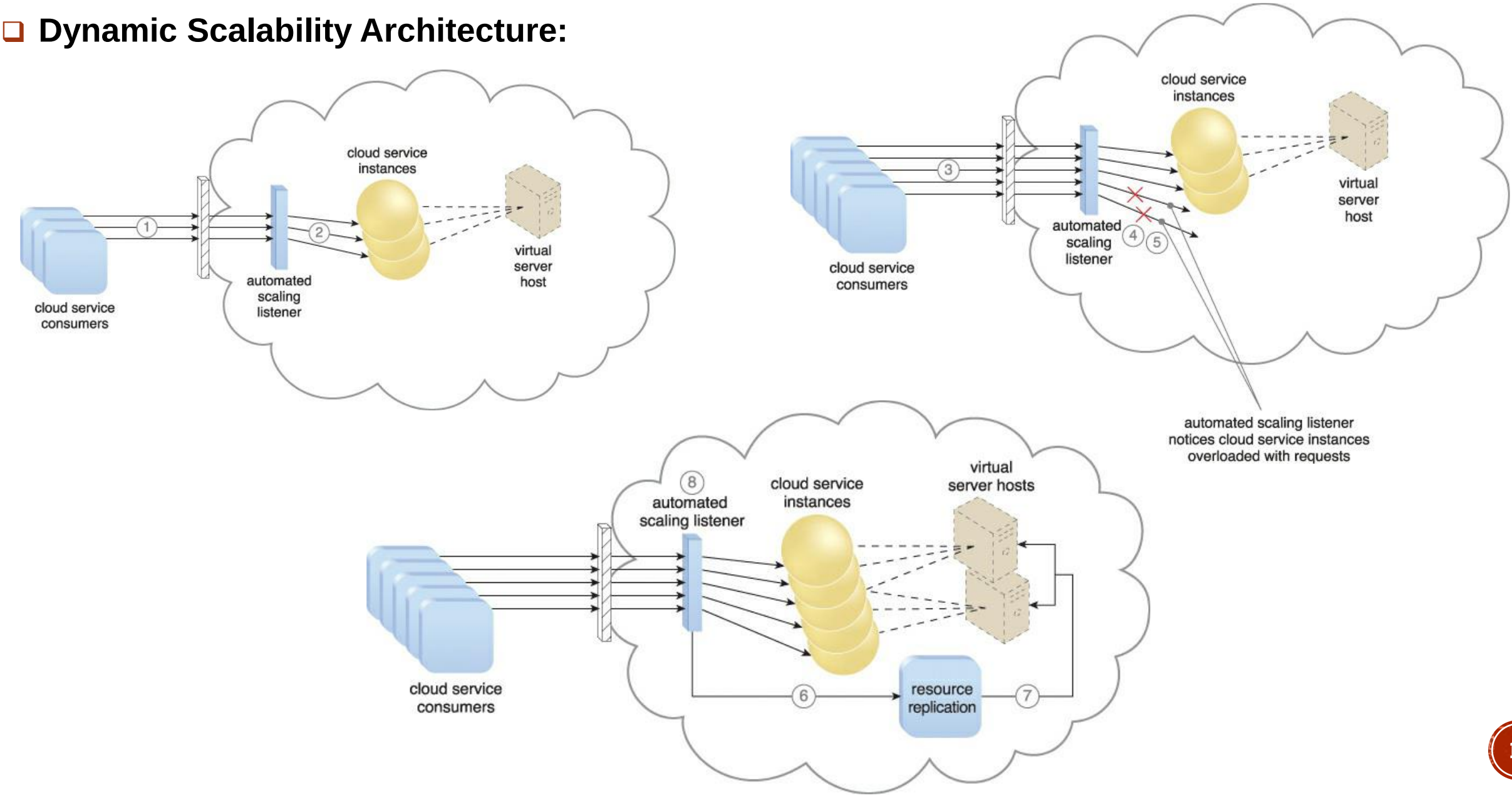
The dynamic scalability architecture is an architectural model based on a system of predefined scaling conditions that trigger the dynamic allocation of IT resources from resource pools.

Dynamic allocation enables variable utilization as dictated by usage demand fluctuations, since unnecessary IT resources are efficiently reclaimed without requiring manual interaction.

- 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, as per requirements and permissions.
- Dynamic Vertical Scaling – IT resource instances are scaled up and down when there is a need t adjust the processing capacity of a single IT resource. For example, a virtual server that is being overloaded can have its memory dynamically increased or it may have a processing core added.
- Dynamic Relocation – The IT resource is relocated to a host with more capacity. For example, a database may need to be moved from a tape-based SAN storage device with 4 GB per second I/O capacity to another disk-based SAN storage device with 8 GB per second I/O capacity.

Cloud computing infrastructure

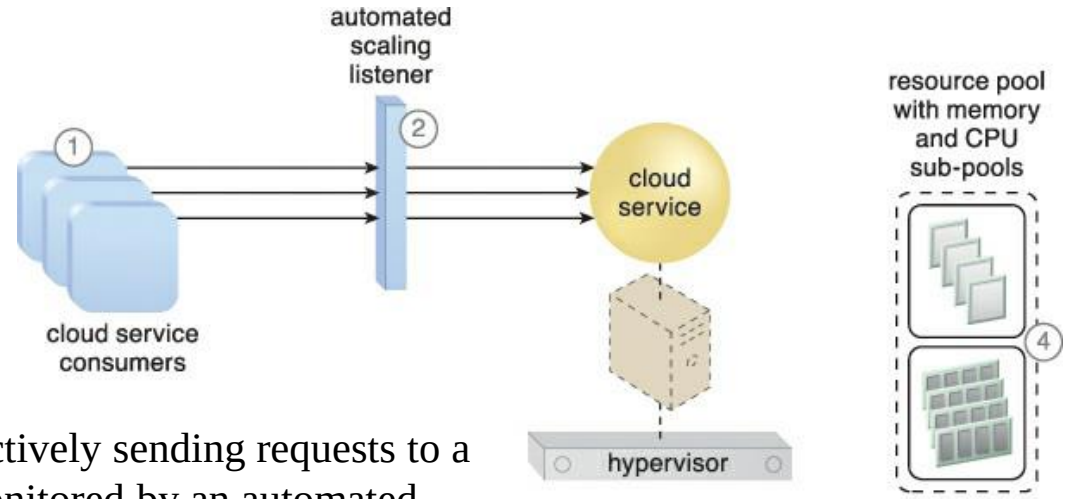
- **Dynamic Scalability Architecture:**



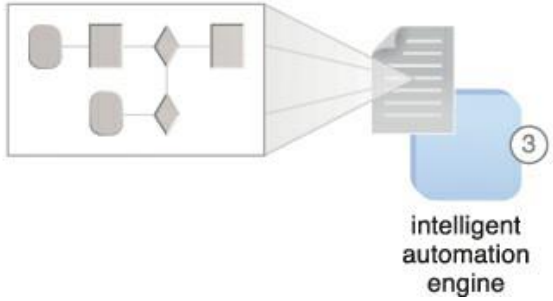
Cloud computing infrastructure

Elastic Resource Capacity Architecture:

The elastic resource capacity 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.



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

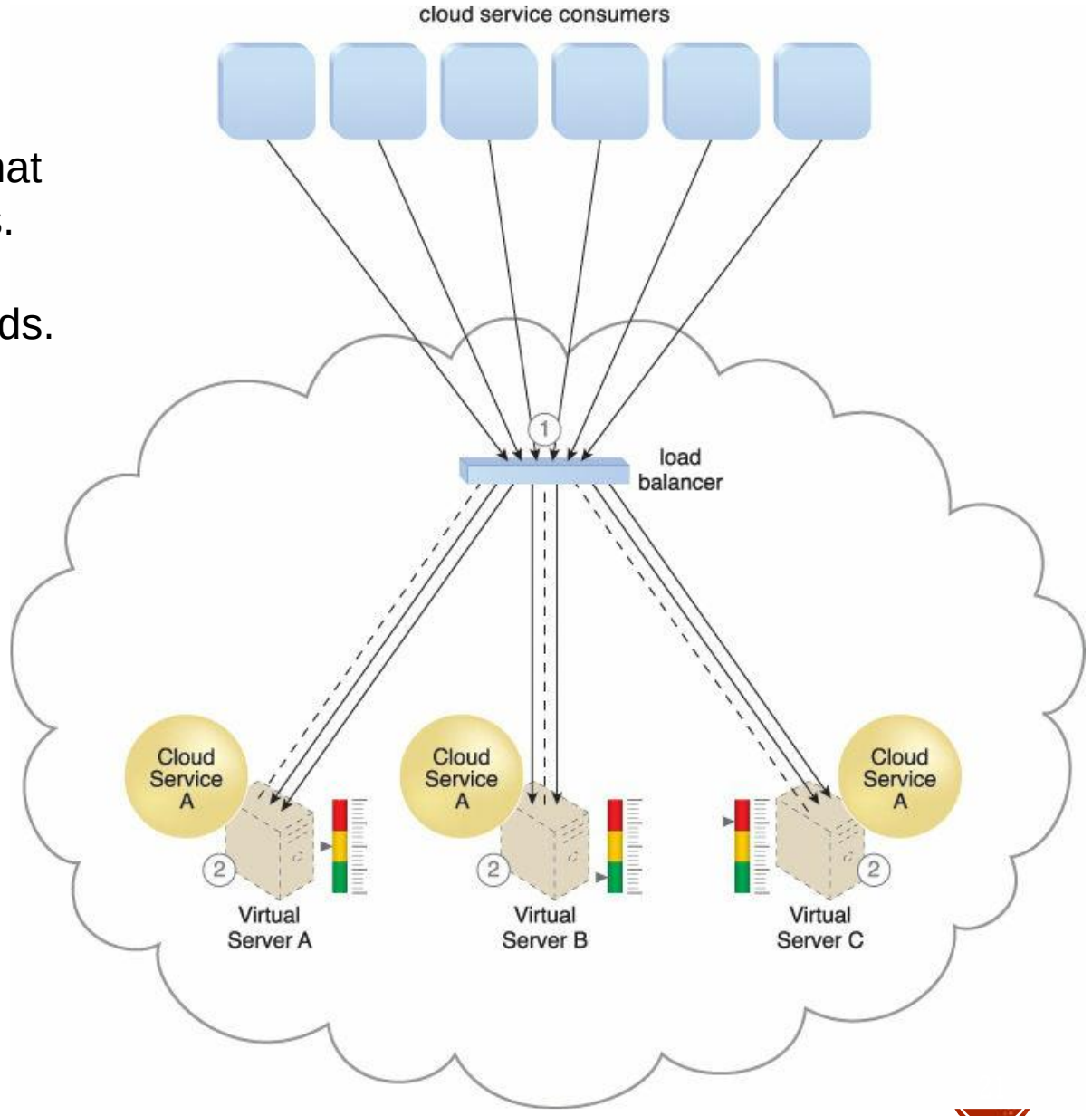


Cloud computing infrastructure

- **Service Load Balancing Architecture:**

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.

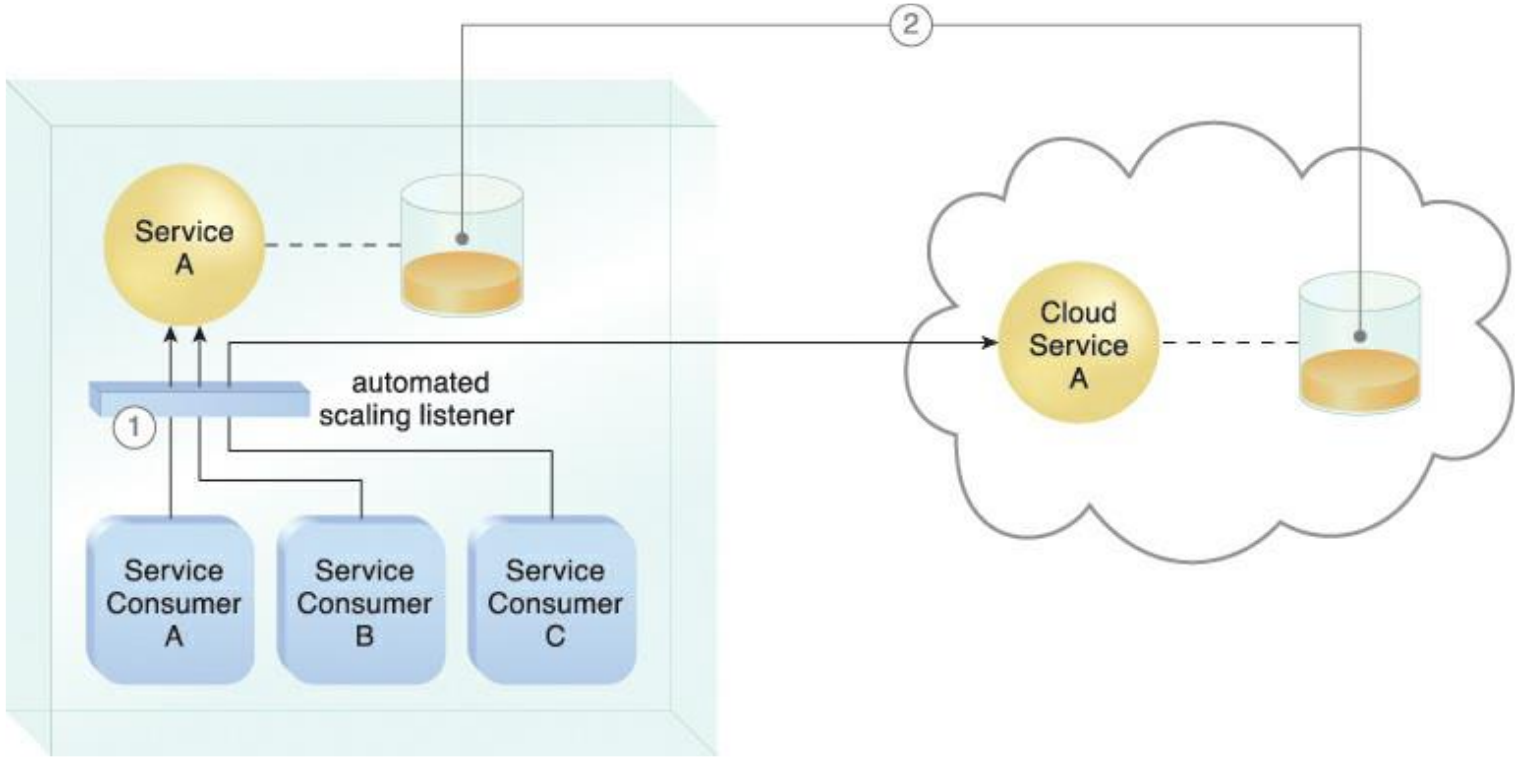
Depending on the anticipated workload and processing capacity of host server environments, multiple instances of each cloud service implementation can be generated as part of a resource pool that responds to fluctuating request volumes more efficiently.



Cloud computing infrastructure

Cloud Bursting Architecture:

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.



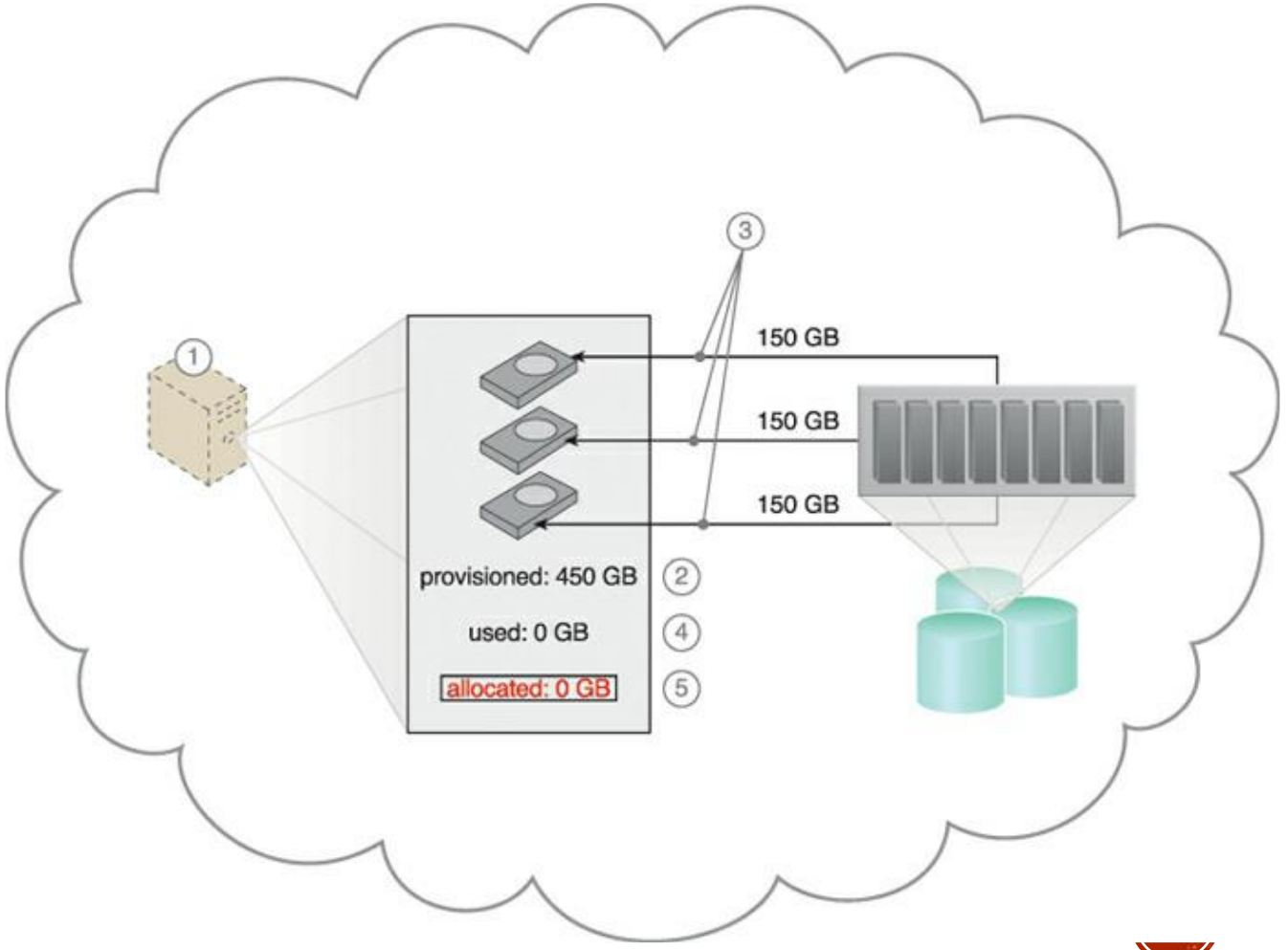
An automated scaling listener monitors the usage of on-premise Service A, and redirects Service Consumer C’s request to Service A’s redundant implementation in the cloud (Cloud Service A) once Service A’s usage threshold has been exceeded (1). A resource replication system is used to keep state management databases synchronized (2).

Cloud computing infrastructure

❑ Elastic Disk Provisioning Architecture:

Cloud consumers are commonly charged for cloud-based storage space based on fixed-disk storage allocation, meaning the charges are predetermined by disk capacity and not aligned with actual data storage consumption.

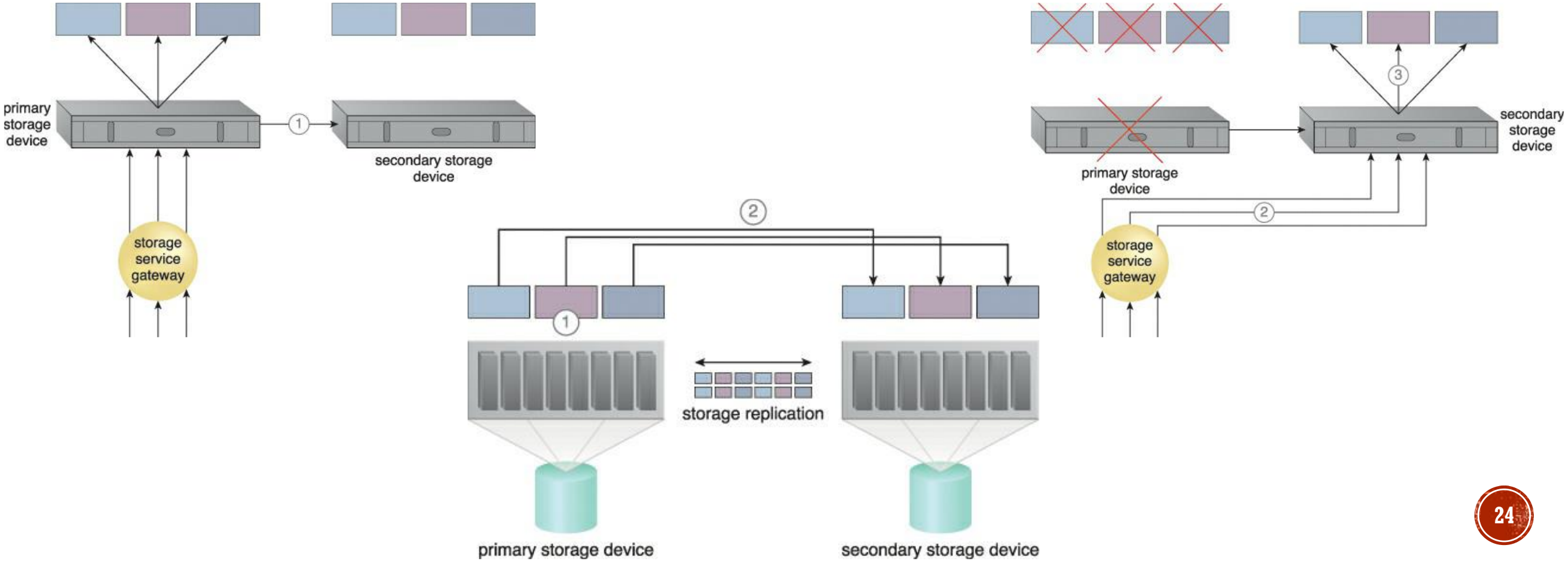
- The cloud consumer requests a virtual server with three hard disks, each with a capacity of 150 GB (1).
- The virtual server is provisioned according to the elastic disk provisioning architecture, with a total of 450 GB of disk space (2).
- The 450 GB is allocated to the virtual server by the cloud provider (3).
- The cloud consumer has not installed any software yet, meaning the actual used space is currently 0 GB (4).
- Because the 450 GB are already allocated and reserved for the cloud consumer, it will be charged for 450 GB of disk usage as of the point of allocation (5).



Cloud computing infrastructure

- **Redundant Storage Architecture:**

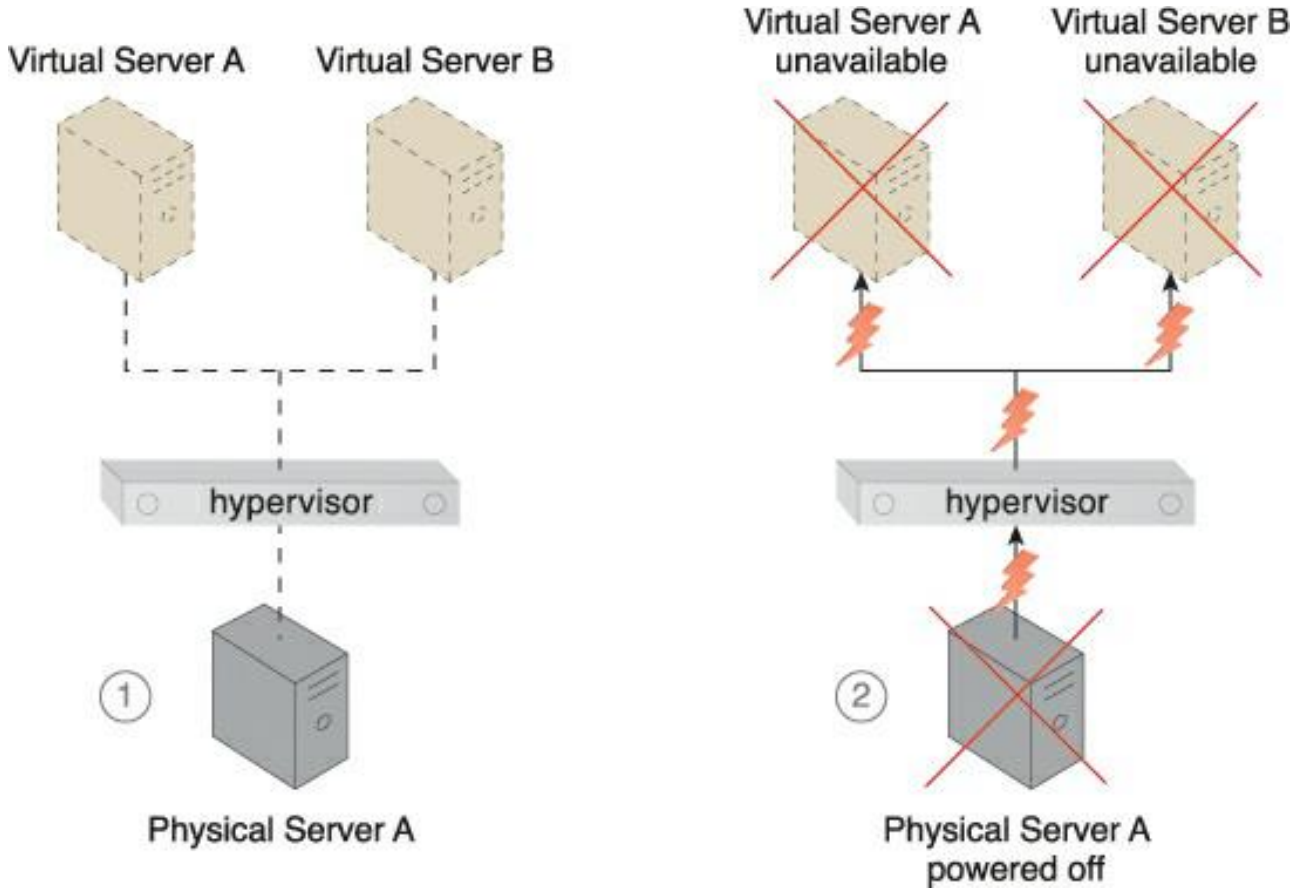
Cloud storage devices are occasionally subject to failure and disruptions that are caused by network connectivity issues, controller or general hardware failure, or security breaches. A compromised cloud storage device's reliability can have a ripple effect and cause impact failure across all of the services, applications, and infrastructure components in the cloud that are reliant on its availability.



Cloud computing infrastructure

Hypervisor Clustering Architecture:

Hypervisors can be responsible for creating and hosting multiple virtual servers. Because of this dependency, any failure conditions that affect a hypervisor can cascade to its virtual servers.

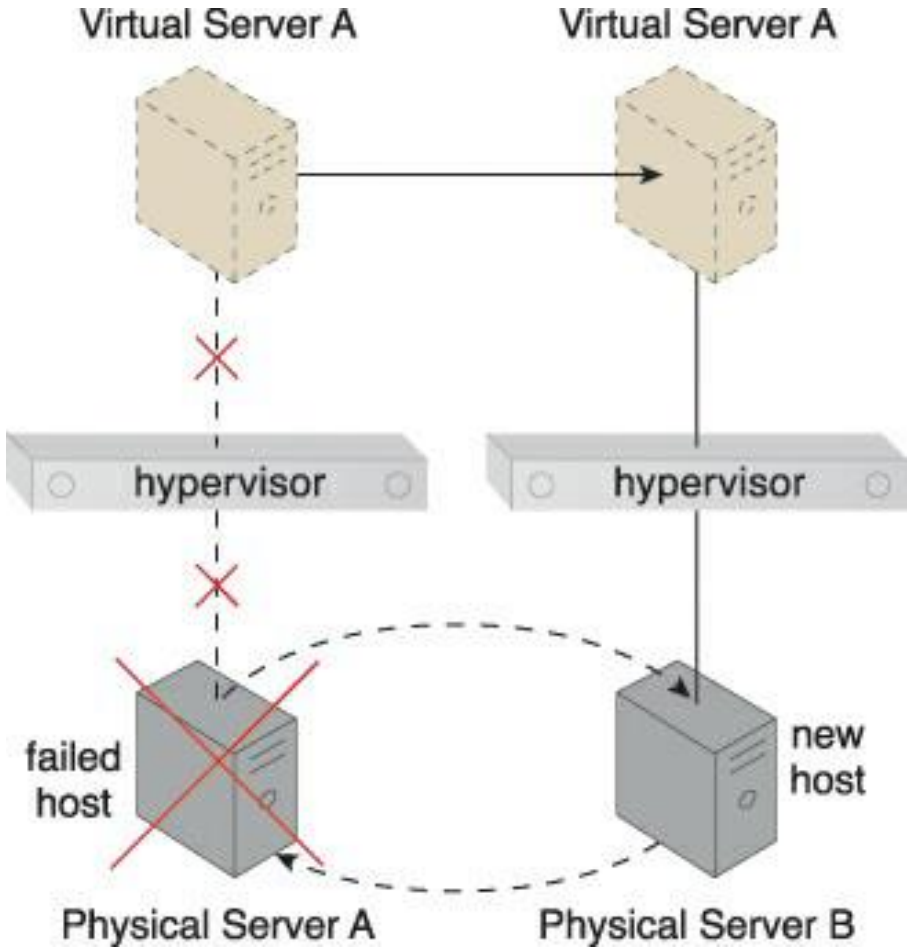


Physical Server A is hosting a hypervisor that hosts Virtual Servers A and B (1). When Physical Server A fails, the hypervisor and two virtual servers consequently fail as well (2).

Cloud computing infrastructure

- **Hypervisor Clustering Architecture:**

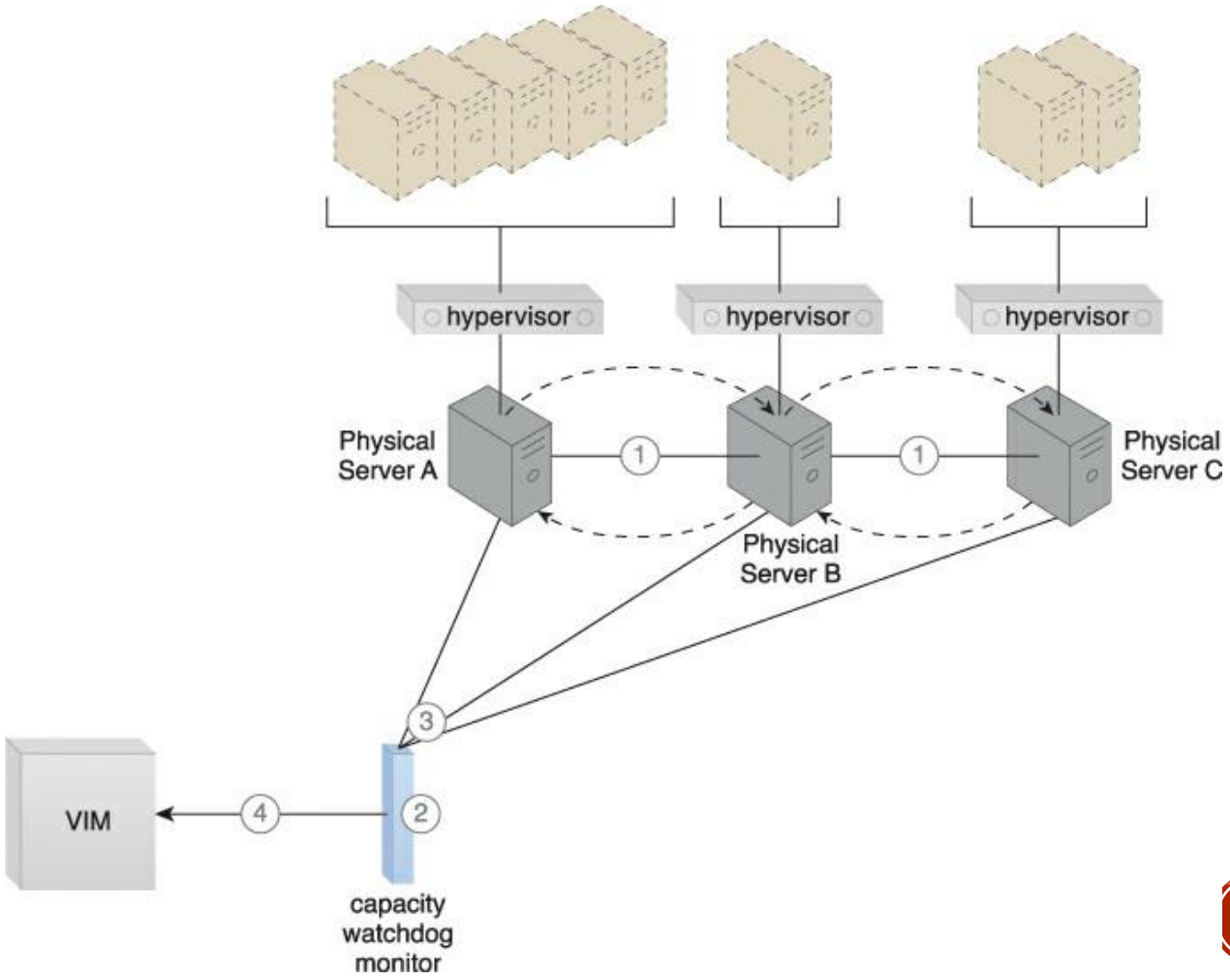
The hypervisor clustering architecture establishes a high-availability cluster of hypervisors across multiple physical servers. If a given hypervisor or its underlying physical server becomes unavailable, the hosted virtual servers can be moved to another physical server or hypervisor to maintain runtime operations.



Cloud computing infrastructure

- **Load Balanced Virtual Server Instances Architecture:**

Keeping cross-server workloads evenly balanced between physical servers whose operation and management are isolated can be challenging. A physical server can easily end up hosting more virtual servers or receive larger workloads than its neighboring physical servers.



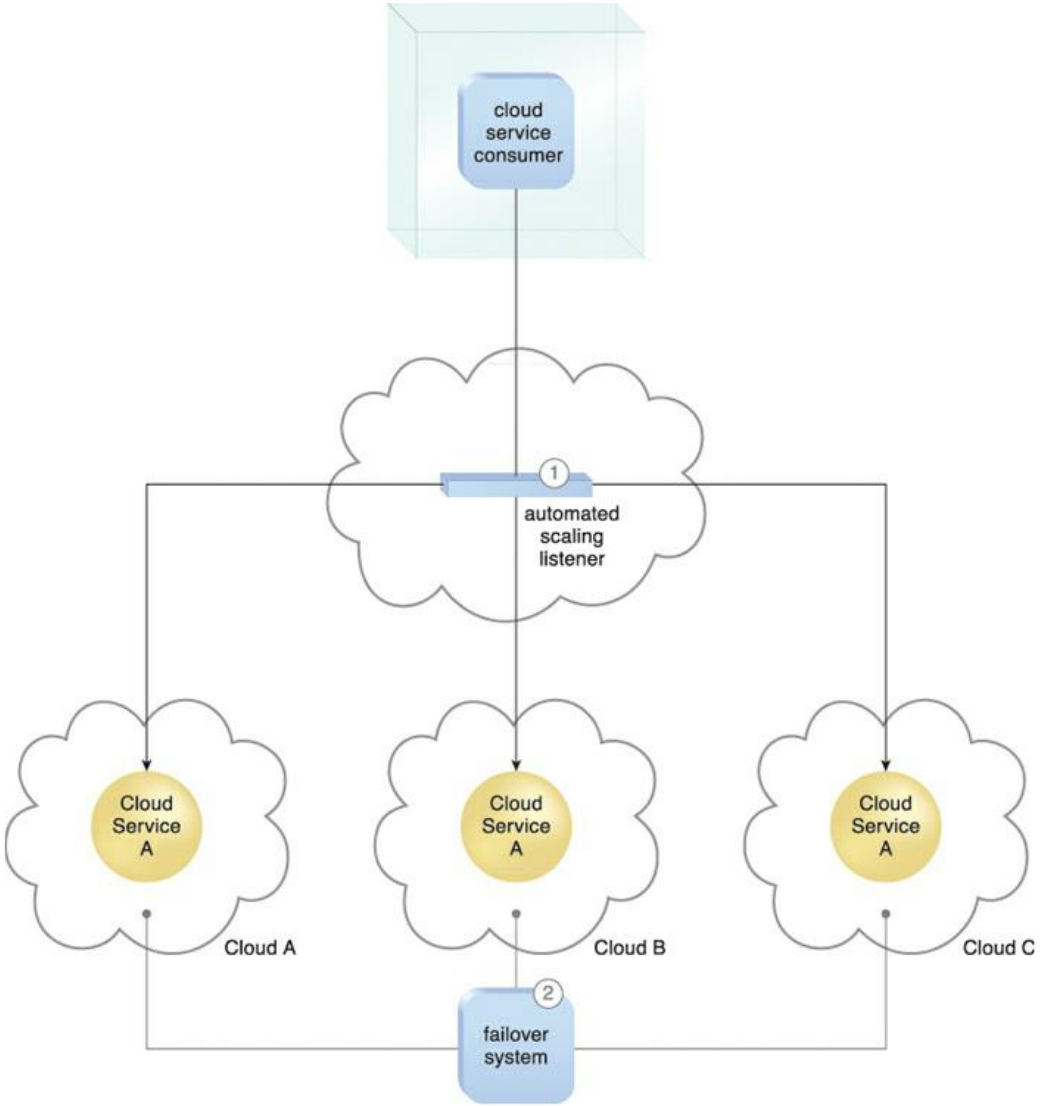
Cloud computing infrastructure

❑ Cloud Balancing Architecture:

The cloud balancing architecture establishes a specialized architectural model in which IT resources can be load balanced across multiple clouds.

The cross-cloud balancing of cloud service consumer requests can:

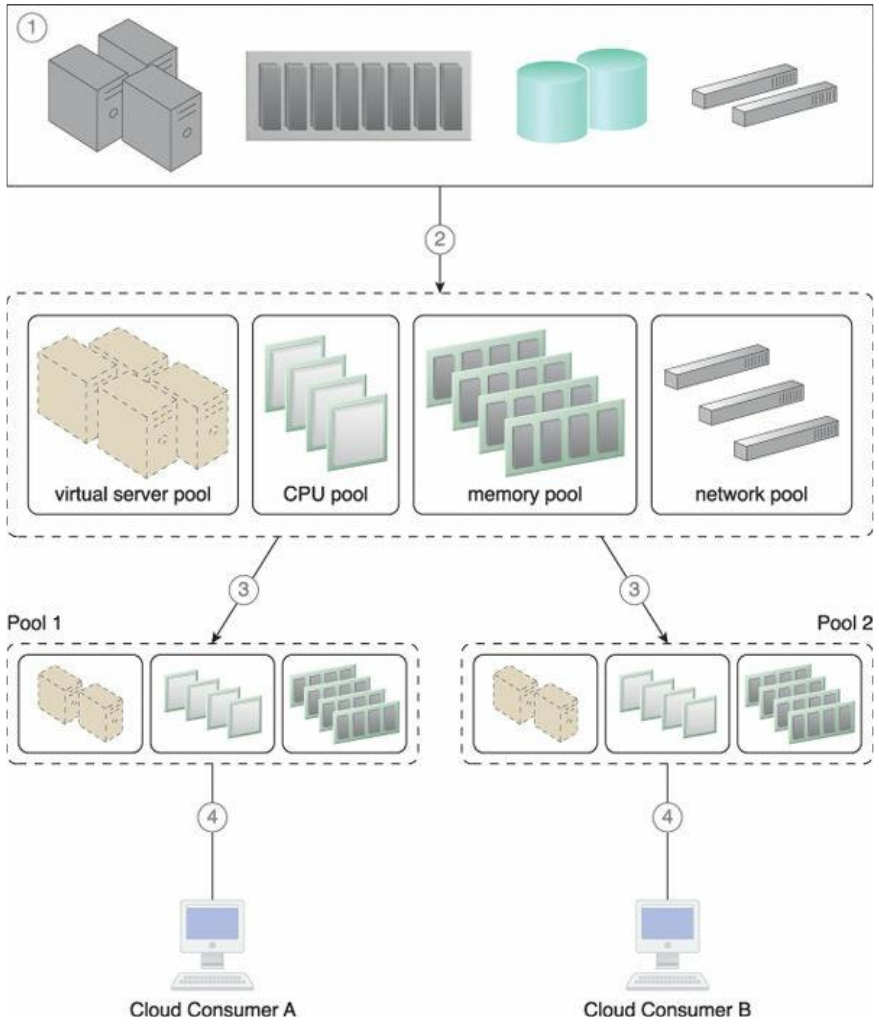
- improve the performance and scalability of IT resources
- increase the availability and reliability of IT resources
- improve load-balancing and IT resource optimization

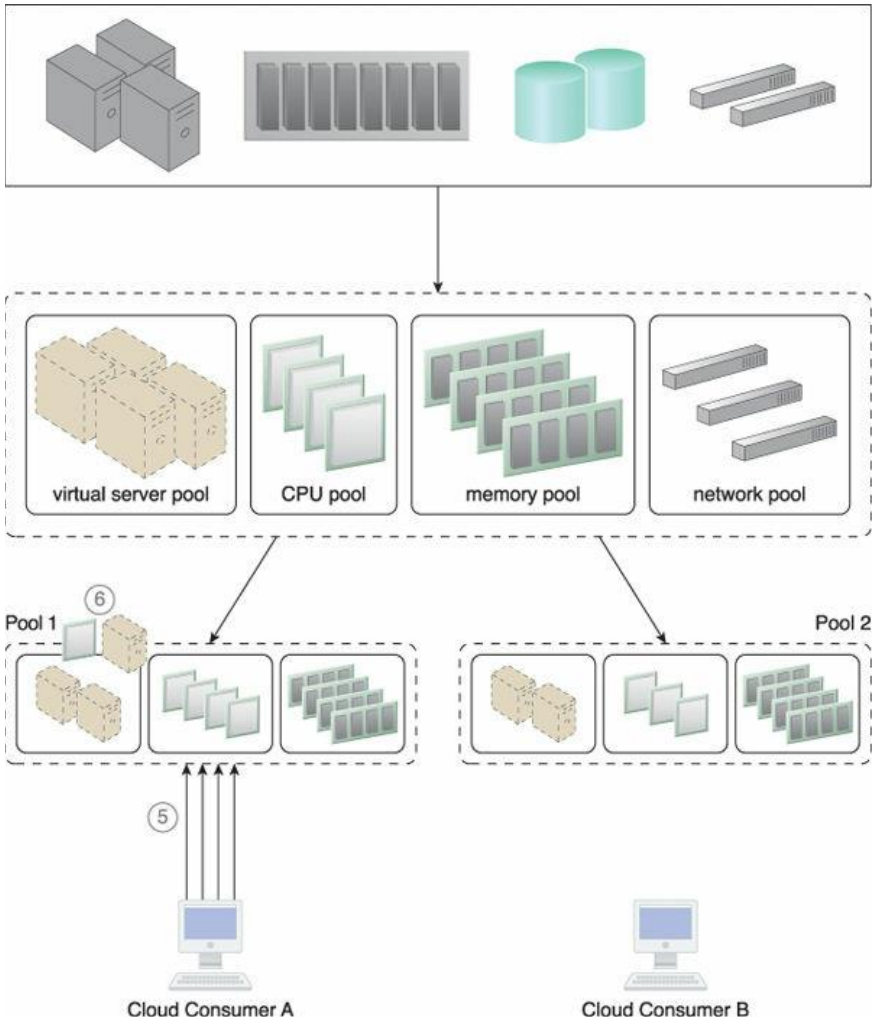


Cloud computing infrastructure

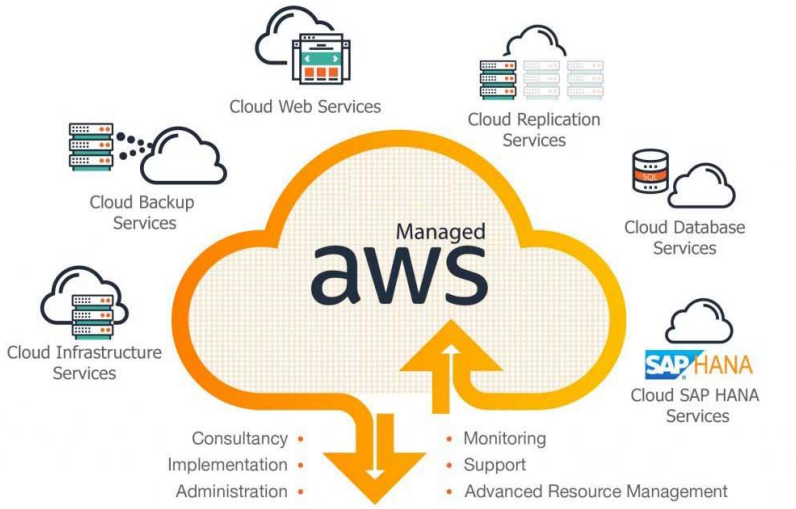
Resource Reservation Architecture:

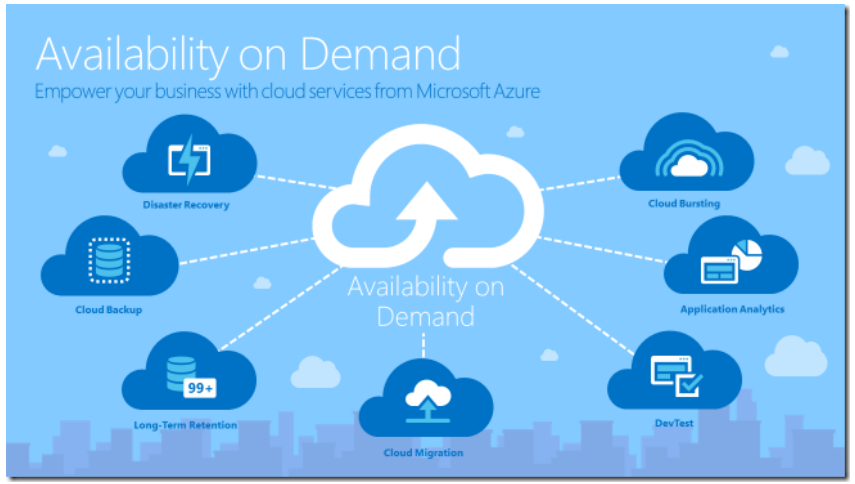
Depending on how IT resources are designed for shared usage and depending on their available levels of capacity, concurrent access can lead to a runtime exception condition called resource constraint.





Cloud Infrastructure and Applications





Cloud Infrastructure and Applications

- **Application architectures:**
  - Compute services
  - Storage services
  - Database services

  - Application services
  - Content delivery services
  - Analytics services
  - Deployment and Management services
  - Identity & Access management services

Cloud Infrastructure and Applications

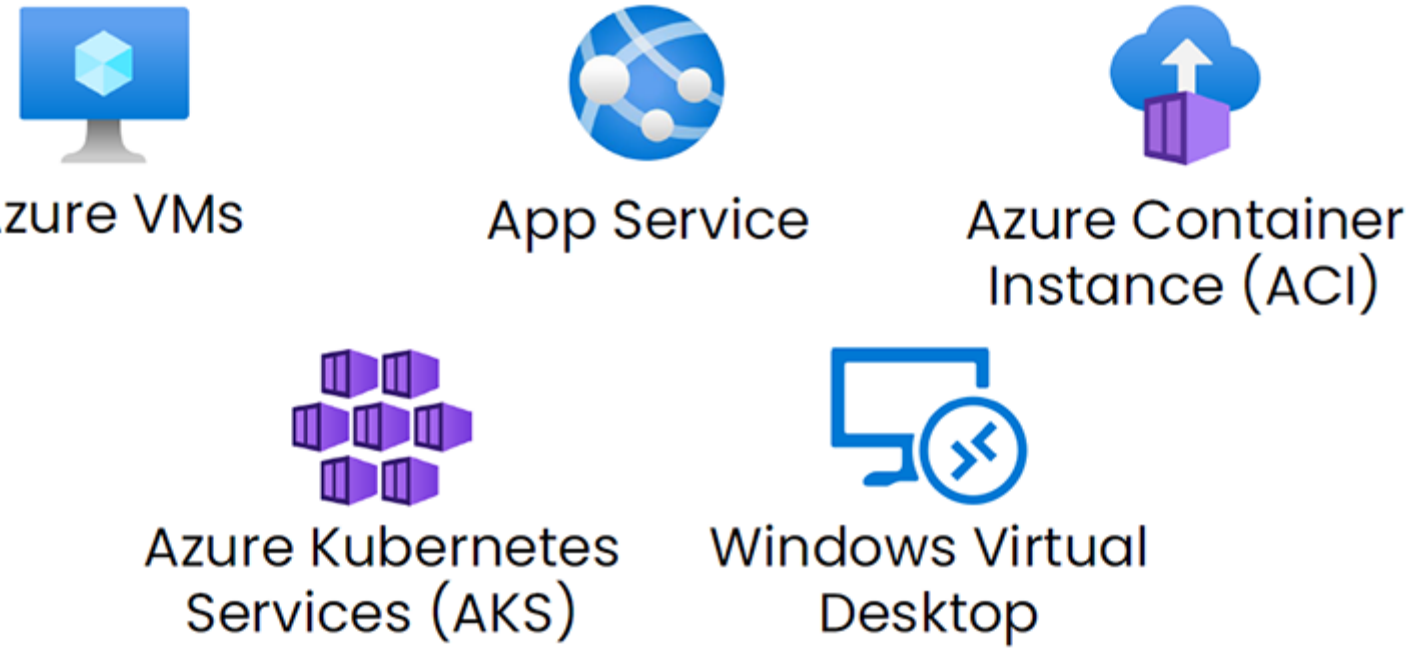
- **Compute services:**

.

Cloud Infrastructure and Applications

- **Compute services:**

.



Cloud Infrastructure and Applications

- **Compute services:**

.

Compute Engine

Run large-scale workloads on virtual machines hosted on Google's infrastructure

App Engine

A platform for building scalable web apps and mobile backends

Container Engine

Run Docker containers on Google's infrastructure, powered by Kubernetes

Cloud Run

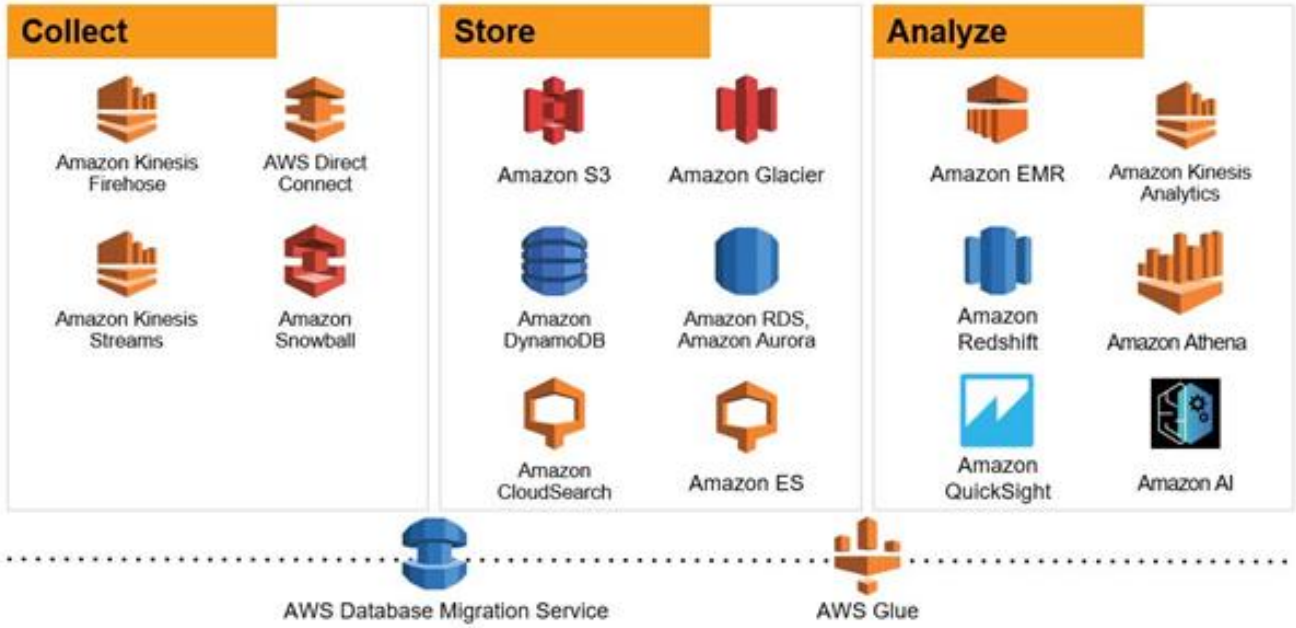
Fully managed compute platform for deploying and scaling containerized applications quickly and securely.

Cloud Functions

A serverless platform for building event-based microservices triggered by events in GCP

Cloud Infrastructure and Applications

- **Storage and Database services:**



Cloud Infrastructure and Applications

- **Storage and Database services:**

.


Queue Storage

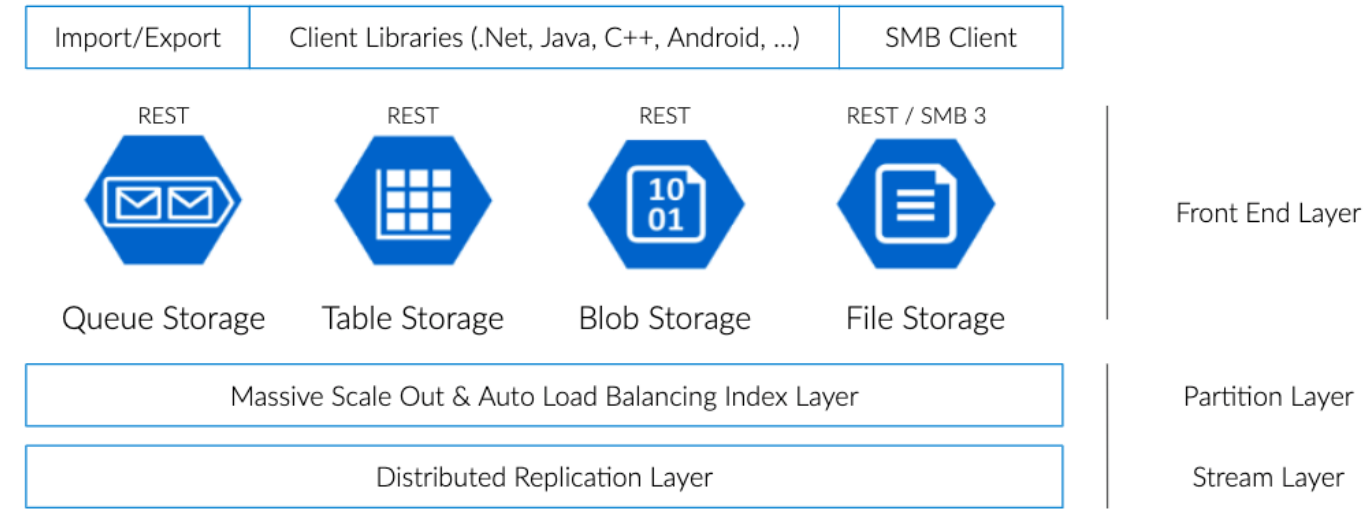

Table Storage

Blob Storage


File Storage


Disk Storage

Azure Storage Architecture



Cloud Infrastructure and Applications

- **Storage and Database services:**

.

Azure SQL

SQL virtual machines

Best for migrations and applications requiring OS-level access

Managed instances

Best for most lift-and-shift migrations to the cloud

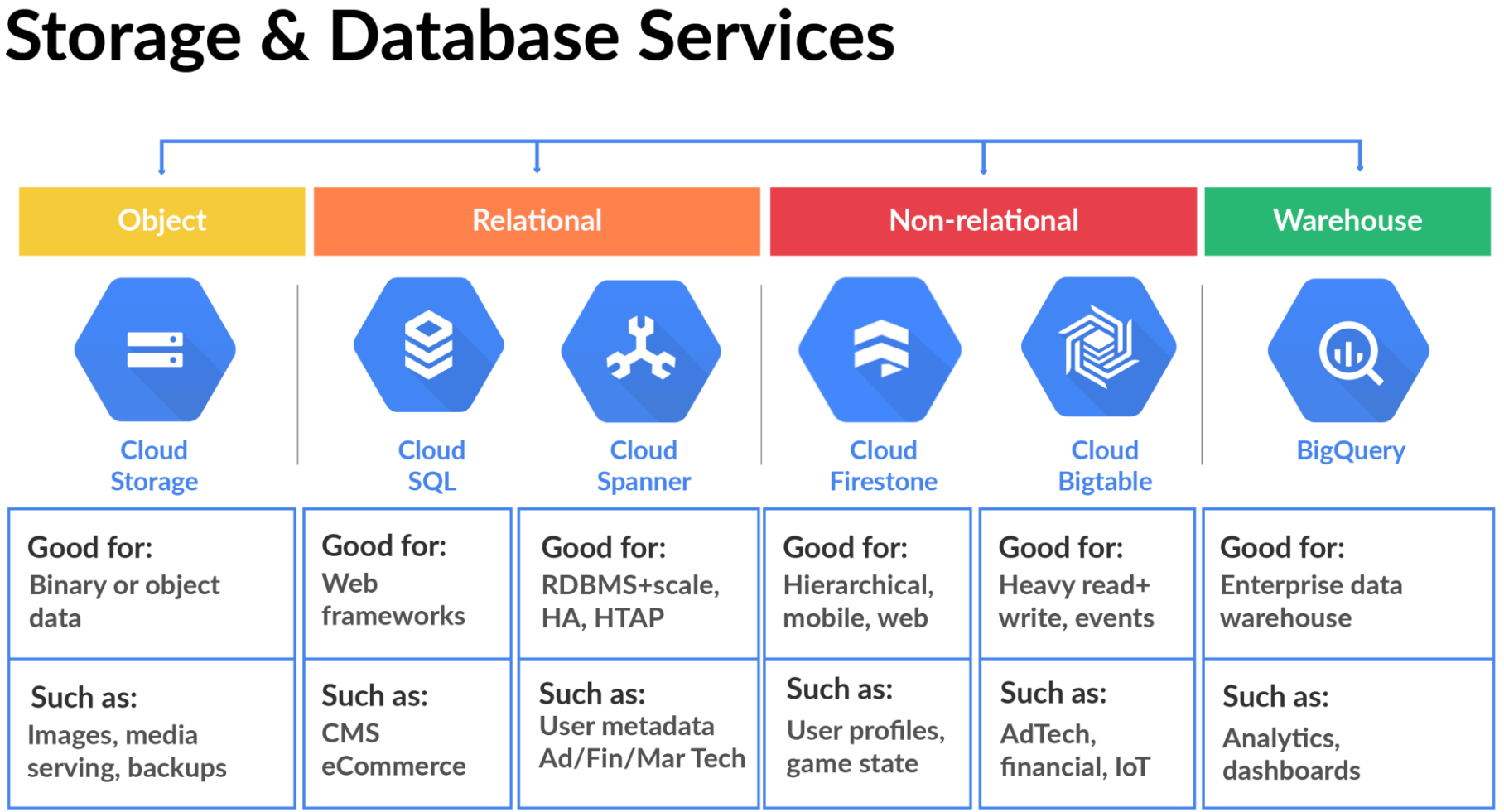
Databases

SQL virtual machine	Single instance	Instance pool	Single database	Elastic pool
• SQL Server and OS server access • Expansive SQL And OS version support • Automated manageability features for SQL Server	• SQL Server surface area (vast majority) • Native virtual network support • Fully managed service	• Pre-provision compute resources for migration • Enables cost-efficient migration. • Ability to host smaller instances (2Vcore) • Currently in public preview	• Hyperscale storage (up to 100TB) • Serverless compute • Fully managed service	• Resource sharing between multiple databases to price optimize • Simplified performance management for multiple databases • Fully managed service

Cloud Infrastructure and Applications

- **Storage and Database services:**

.

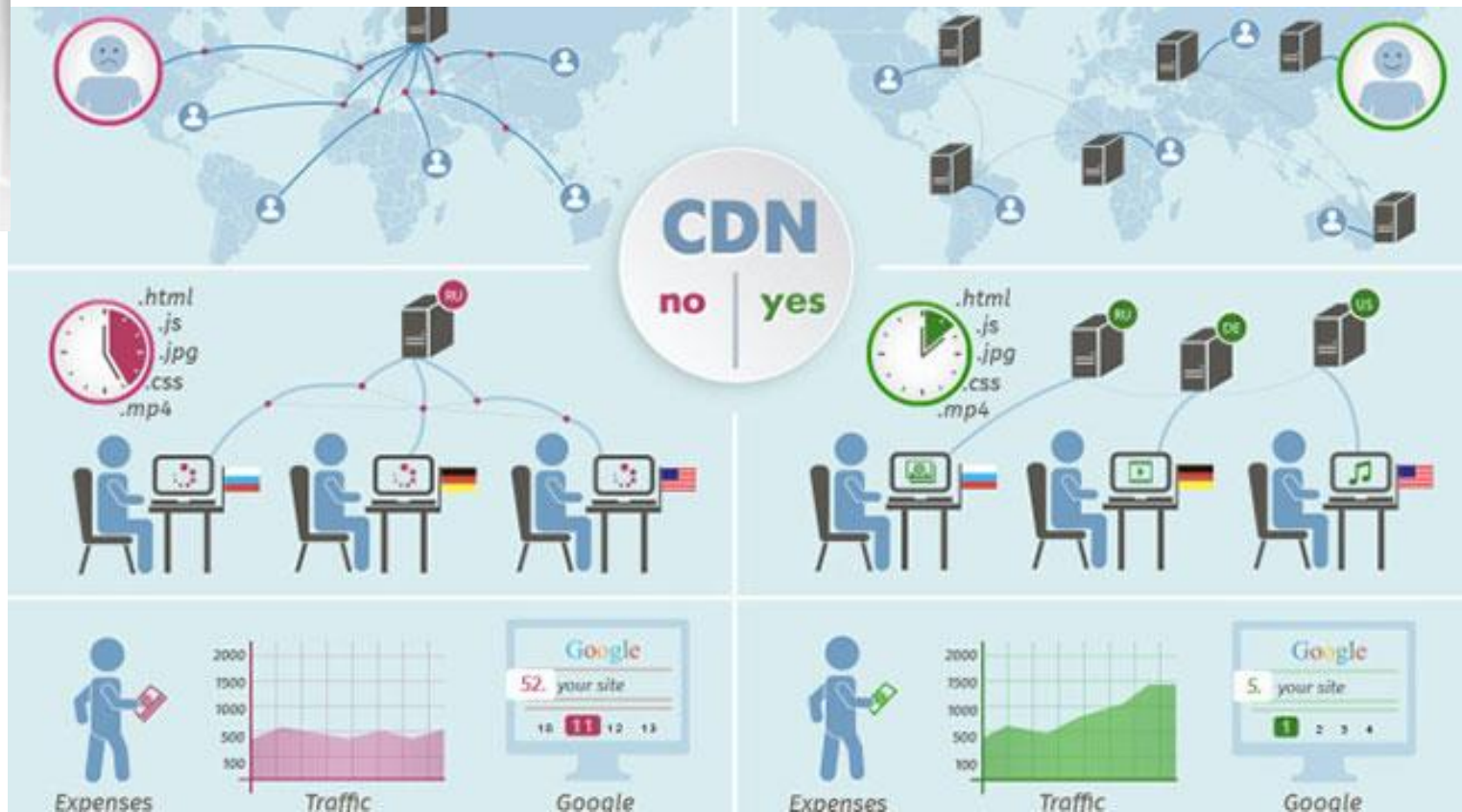


Cloud Infrastructure and Applications

- **Content delivery services:**

Cloud CDN works with the global external HTTP(S) load balancer or the global external HTTP(S) load balancer (classic) to deliver content to your users. The external HTTP(S) load balancer provides the frontend IP addresses and ports that receive requests and the backends that respond to the requests.

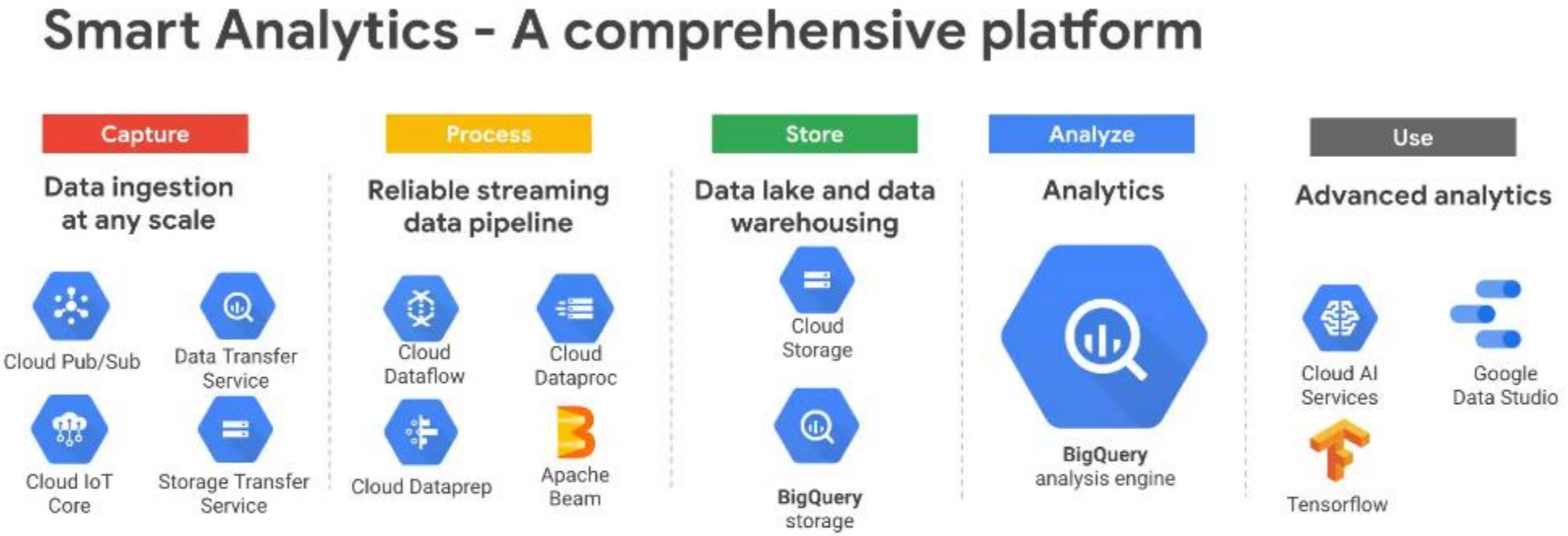
Cloud CDN (Content Delivery Network) uses Google's global edge network to serve content closer to users, which accelerates your websites and applications.



Cloud Infrastructure and Applications

- **Analytics services:**

.

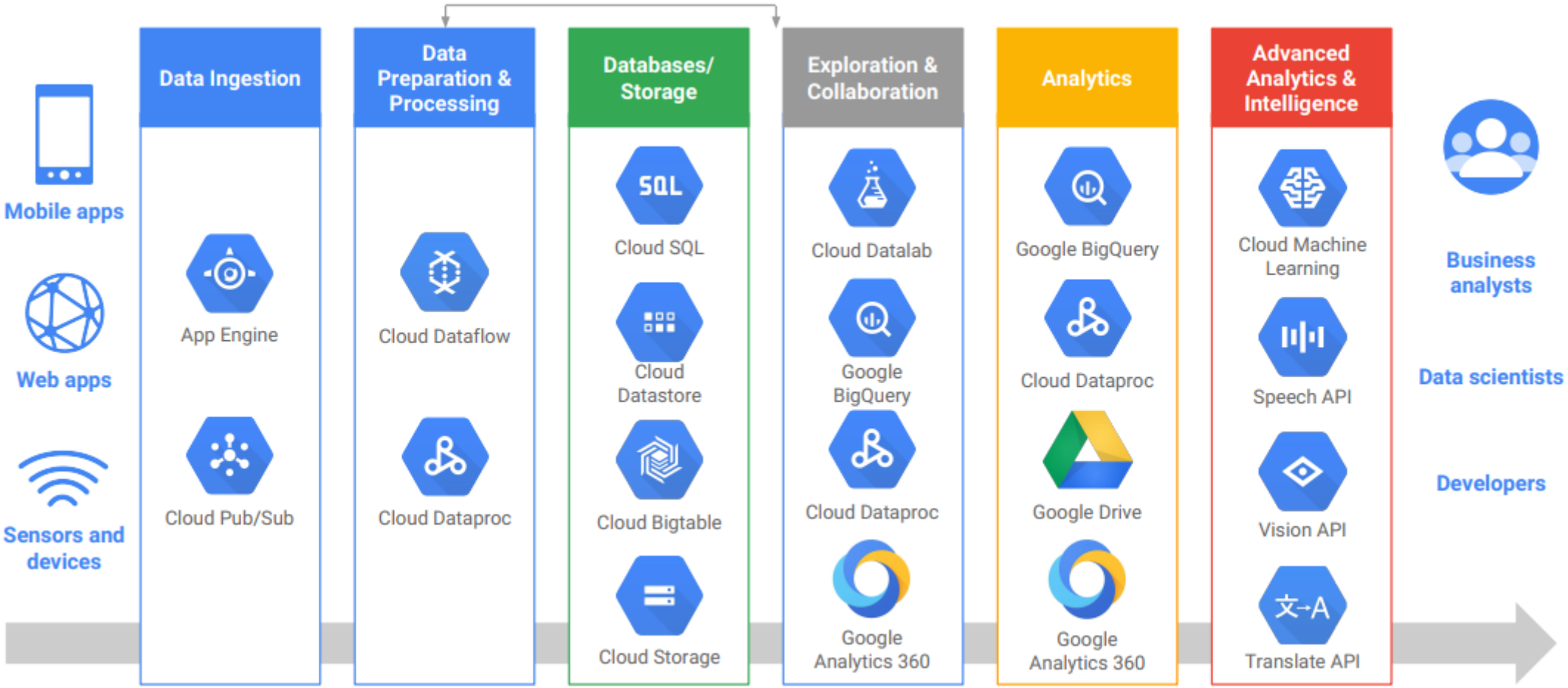


Google Cloud

Cloud Infrastructure and Applications

- **Analytics services:**

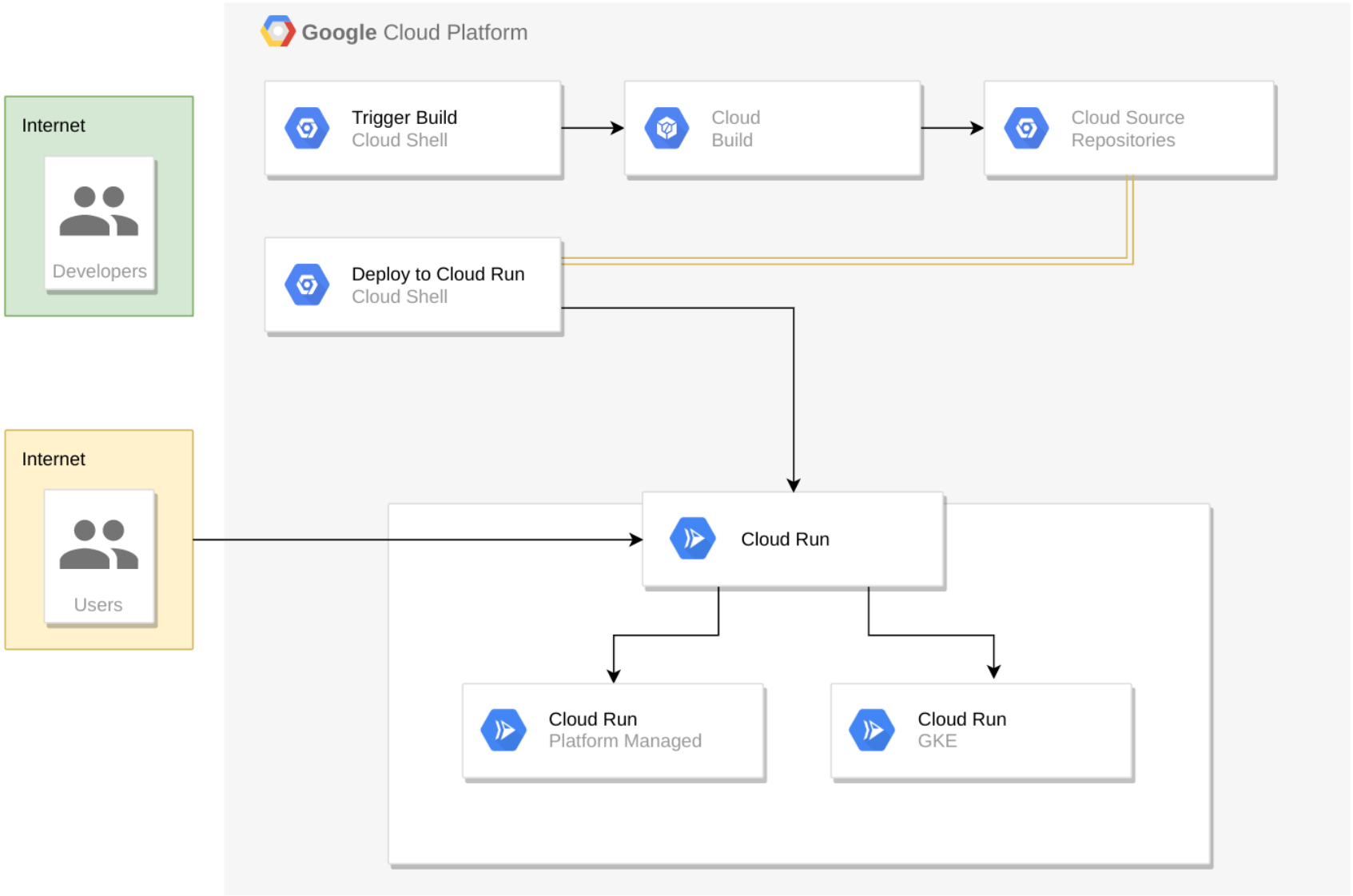
.



Cloud Infrastructure and Applications

- **Deployment and Management services:**

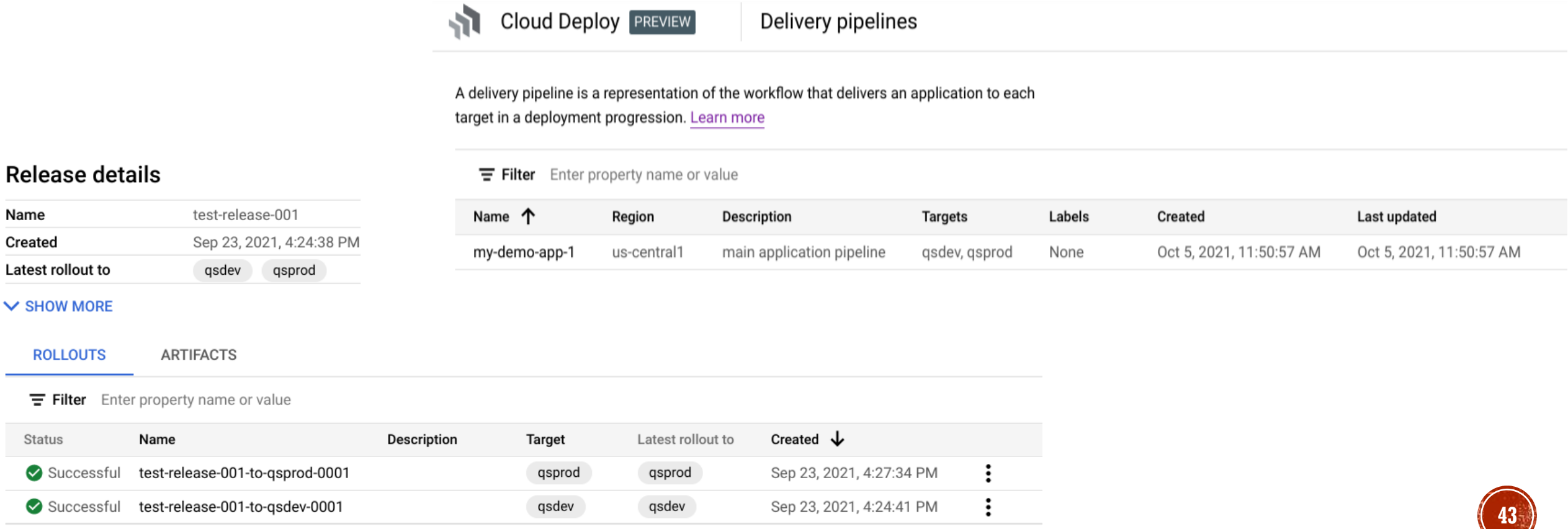
.



Cloud Infrastructure and Applications

❑ Deployment and Management services:

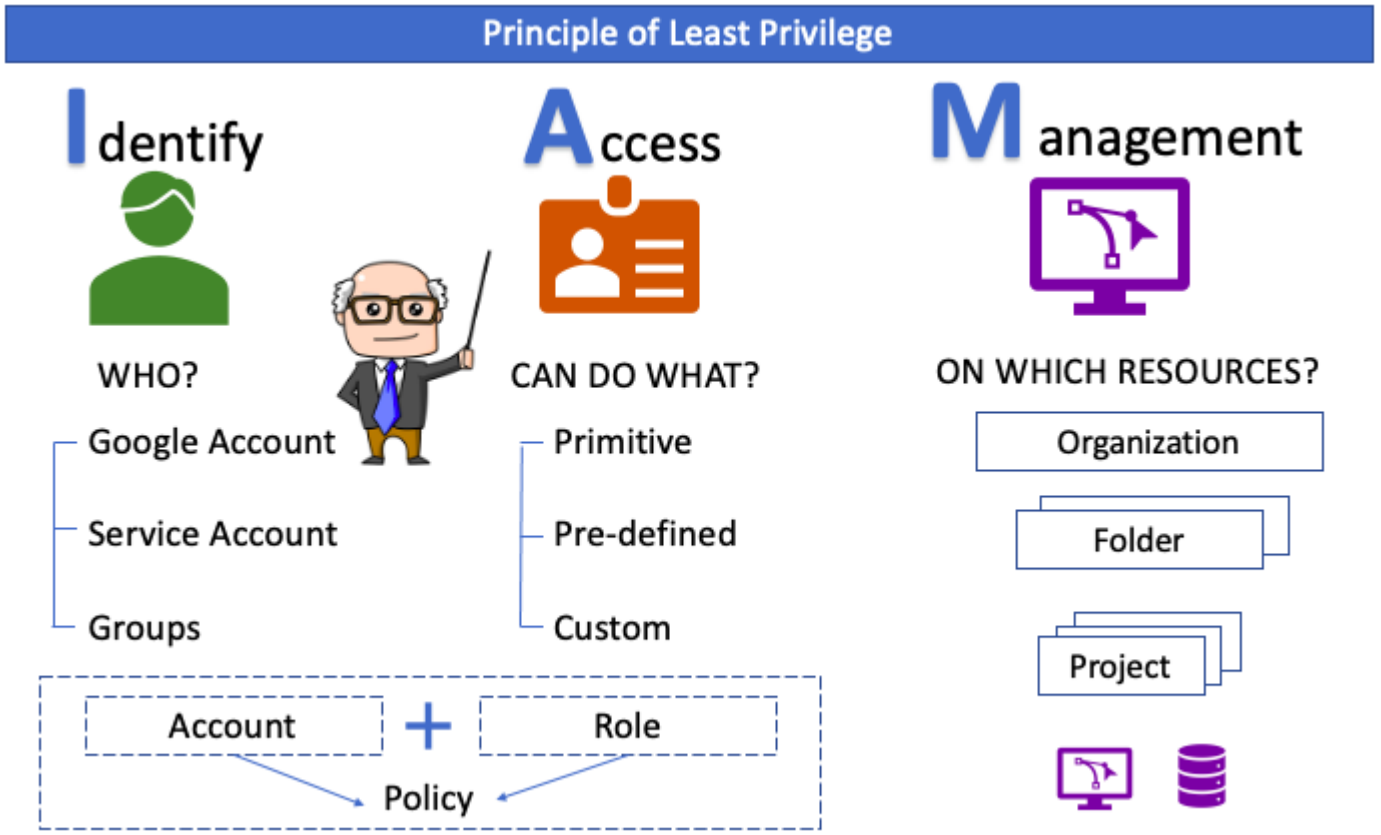
Google Cloud Deploy is a managed service that automates delivery of your applications to a series of target environments in a defined promotion sequence. When you want to deploy your updated application, you create a release, whose lifecycle is managed by a delivery pipeline.



Cloud Infrastructure and Applications

- **Identity & Access management services:**

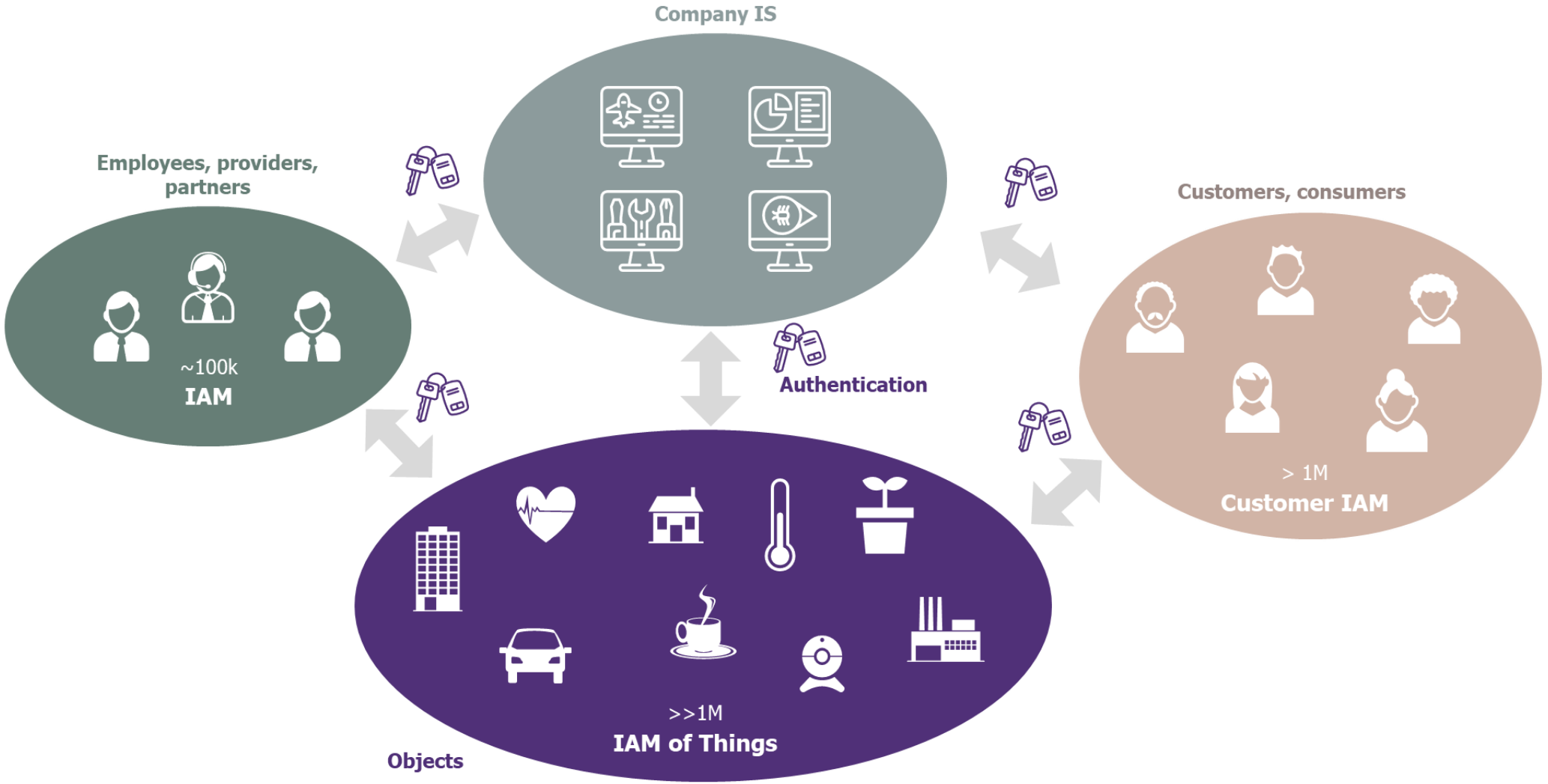
.



Cloud Infrastructure and Applications

- **Identity & Access management services:**

.



Cloud Infrastructure and Applications

- **Application services:**

https://cloud.google.com/solutions#industry-solutions

Cloud Infrastructure and Applications

- **Retail:**

Capture digital and omnichannel revenue growth

Become a customer-centric, data-driven retailer

Drive operational improvement

Carrefour

The Home Depot

wayfair

KOHL'S

Loblaws

shopify

Cloud Infrastructure and Applications

- **Consumer packaged goods:**

Unlock consumer growth with data-powered insights

Transform go-to-market in the omnichannel ecosystem

Drive connected, efficient, and sustainable operations

Unilever

Mondelēz International

Tyson

kao

Keurig DrPepper

P&G

Cloud Infrastructure and Applications

Financial services:

Drive business transformation across banking, capital markets, insurance, and payments to support data-driven innovation, customer expectations, and security and compliance needs with Google Cloud.

Grow your revenue through new products and innovation

Reduce your costs by reimagining operations

Manage risk and compliance

PayPal Bloomberg charles SCHWAB

citi REFINITIV KeyBank

Cloud Infrastructure and Applications

❑ Healthcare and life sciences:

Evolve the care paradigm, advance research at scale, and empower everyone in your organization to innovate and transform with Google Cloud's solutions for healthcare and life sciences.

Provide secure, continuous patient care

Empower caregivers with collaboration and productivity tools

Make data-driven clinical and operational decisions

BROAD INSTITUTE

CardinalHealth

McKESSON

moderna

NIH National Institute on Aging

Stanford University

Cloud Infrastructure and Applications

❑ Media and entertainment:

Modernize your content production and distribution operations while transforming audience experiences across the world.

Collaborate globally and produce great content efficiently

Improve audience engagement through data and AI

Distribute content globally

Cloud Infrastructure and Applications

❑ Telecommunications:

At Google Cloud, we're partnering with telecommunications companies around the world to help them drive transformation and accelerate 5G adoption and monetization.

Monetize 5G

Reimagine the customer journey with data

Bring operational efficiency to core IT and networks

docomo orange swisscom telenor kpn TIM

Cloud Infrastructure and Applications

❑ Gaming:

Build great player experiences and empower your game developers by minimizing infrastructure complexity and accelerating data insights with Google Cloud.

Build better games

Scale your game globally

Understand players more deeply with data

Activision Blizzard · King · unity · IMPROBABLE · netmarble · JAMCITY

Cloud Infrastructure and Applications

❑ Manufacturing:

Google Cloud helps manufacturers achieve their digital transformation goals with secure, data-driven solutions that reshape product development, factory-floor operations, and customer experiences.

Reimagine the digital customer journey

Turn every product into a smart product

Optimize supply chain and operations

AIRBUS ASML Ford GE APPLIANCES

Johnson & Johnson LG CNS ups SAMSUNG

Cloud Infrastructure and Applications

❑ Supply chain and logistics:

Make your supply chain resilient, sustainable, and transformative while improving customer experience with Google's digital supply chain solutions.

Deliver exceptional customer experience

Build a resilient and sustainable supply chain

Run your supply chain autonomously

Cloud Infrastructure and Applications

❑ Government:

Google Cloud helps government agencies improve citizen services, increase their operational effectiveness, and deliver proven innovation.

Improve operational effectiveness and service delivery

Deliver big data insights at scale with AI/ML

Enable the government workforce with collaboration tools

NIH National Institute on Aging

NYC Cyber Command

MANCHESTER CITY COUNCIL

Cloud Infrastructure and Applications

❑ Education:

We're committed to advancing learning for everyone. Explore our cloud solutions, teaching tools, and affordable devices that help transform classrooms, academic institutions, and edtech companies.

Build a secure, scalable infrastructure for your institution

Improve student success while optimizing costs

Accelerate groundbreaking research

GEORGETOWN UNIVERSITY

FIMM

Georgia State University

BROWN

Khan Academy