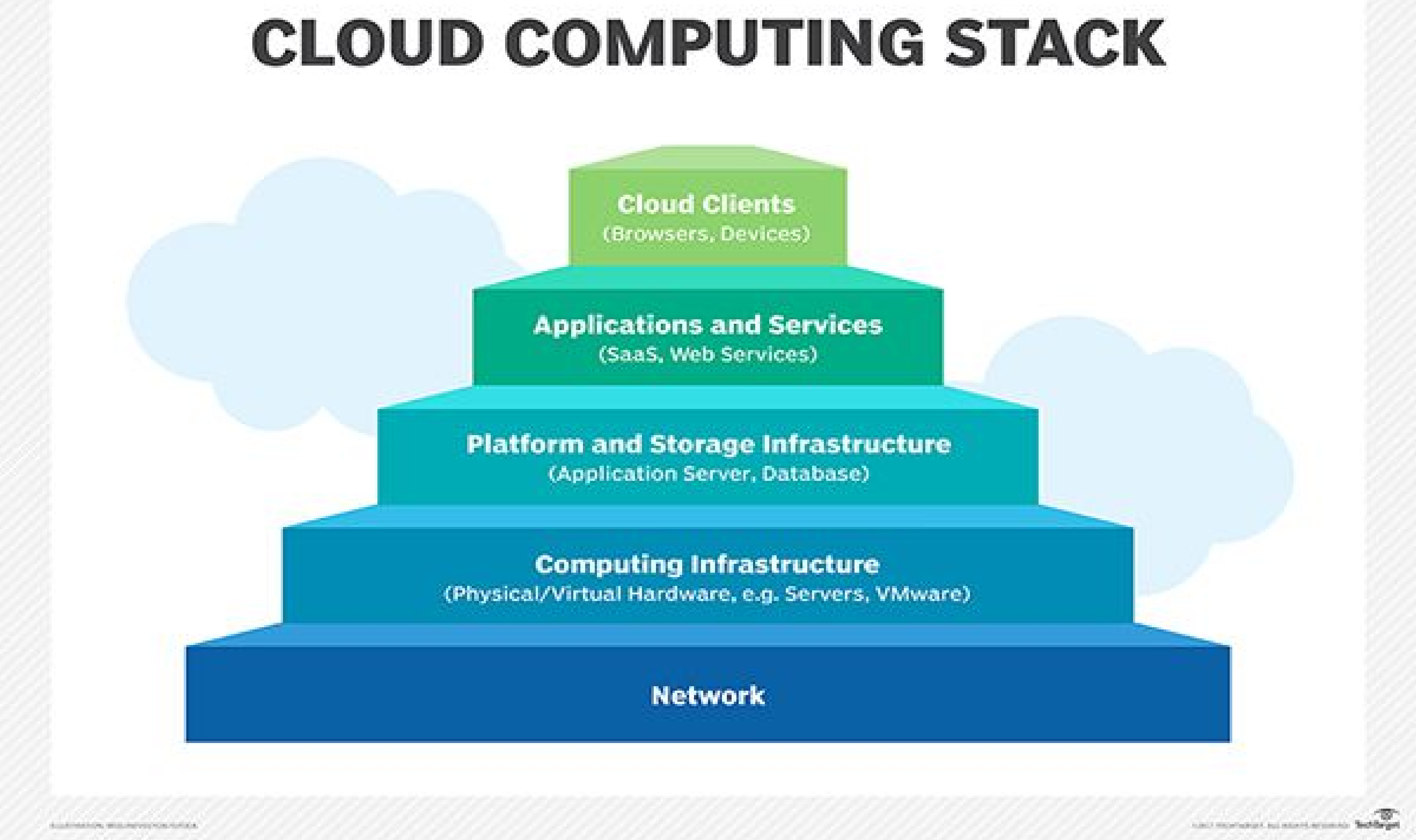
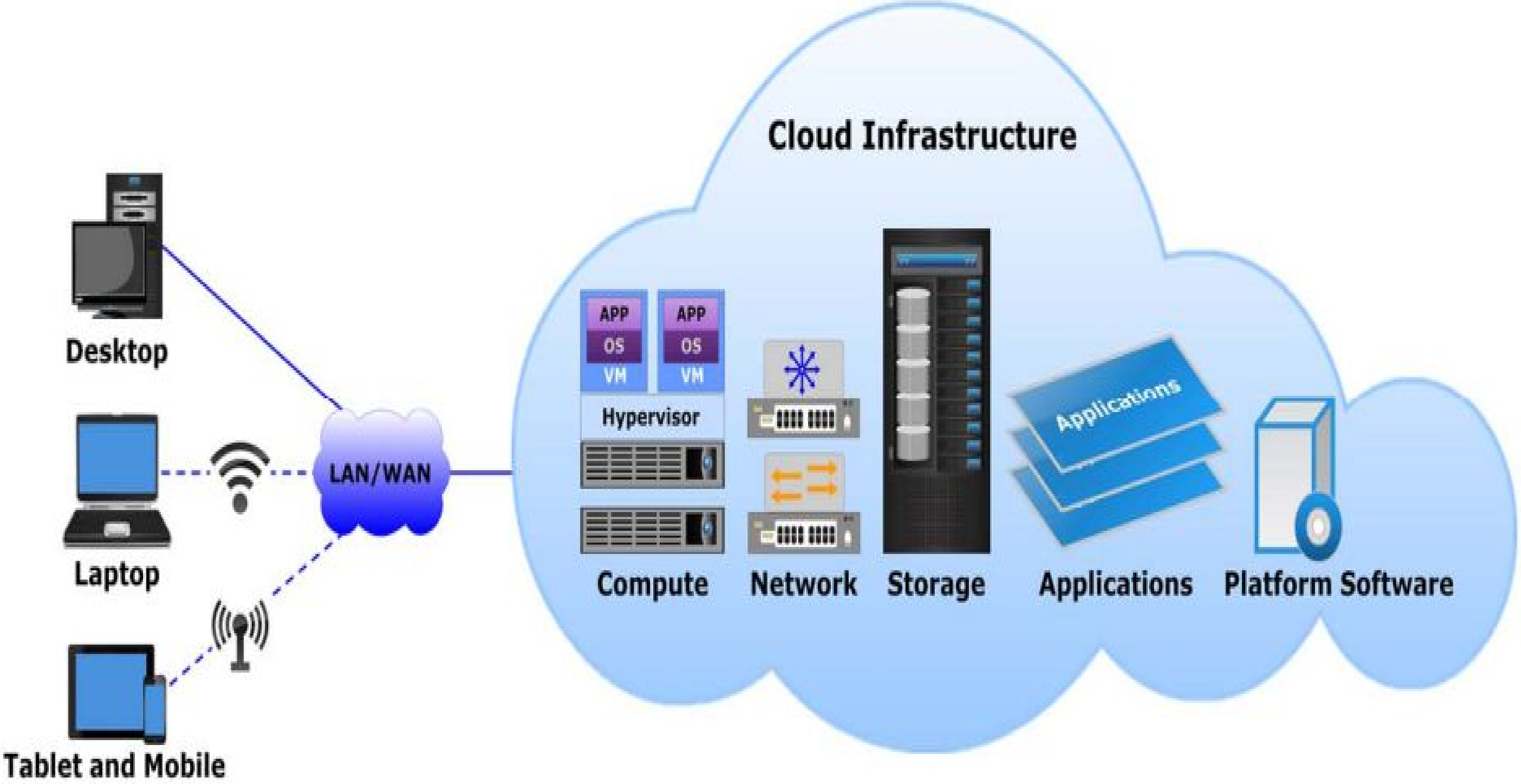
Cloud Infrastructure


CLOUD COMPUTING STACK
Cloud Clients
(Browsers, Devices)
Applications and Services
(SaaS, Web Services)
Platform and Storage Infrastructure
(Application Server, Database)
Computing Infrastructure
(Physical/Virtual Hardware, e.g. Servers, VMware)
Network
TechTarget

CHAPTER 02

Cloud Infrastructure

- Cloud infrastructure refers to the hardware and software components -- such as servers, storage, a network and virtualization software -- that are needed to support the computing requirements of a cloud computing model.


Cloud Infrastructure
Desktop
Laptop
Tablet and Mobile
LAN/WAN
APP
OS
VM
APP
OS
VM
Hypervisor
Applications
Compute
Network
Storage
Applications
Platform Software

Existing cloud infrastructure

- The cloud computing infrastructure at Amazon, Google, and Microsoft (as of mid 2012).
  - Amazon is a pioneer in Infrastructure-as-a-Service (IaaS).
  - Google's efforts are focused on Software-as-a-Service (SaaS) and Platform-as-a-Service (PaaS).
  - Microsoft is involved in PaaS.

Existing cloud infrastructure

- Private clouds are an alternative to public clouds. Open-source cloud computing platforms such as:
  - Eucalyptus,
  - OpenNebula,
  - Nimbus,
  - OpenStack

  can be used as a control infrastructure for a private cloud.

Cloud Computing at Amazon

- Amazon introduced a computing platform that has changed the face of computing.
- In mid-2000 Amazon introduced *Amazon Web Services* (AWS), based on the *IaaS* delivery model.
- In this model the cloud service provider offers an infrastructure consisting of compute and storage servers interconnected by high-speed networks that support a set of services to access these resources.
- An application developer is responsible for installing applications on a platform of his or her choice and managing the resources provided by the Amazon.

Cloud Computing at Amazon

- Businesses in 200 countries used AWS in 2012.
- A significant number of large corporations as well as start-ups take advantage of computing services supported by the *AWS* infrastructure.
- For example, one start-up reports that its monthly computing bills at Amazon are in the range of $100,000, whereas it would spend more than $2,000,000 to compute using its own infrastructure.

Cloud Computing at Amazon

- Businesses in 200 countries used AWS in 2012.
- A significant number of large corporations as well as start-ups take advantage of computing services supported by the *AWS* infrastructure.
- For example, one start-up reports that its monthly computing bills at Amazon are in the range of $100,000, whereas it would spend more than $2,000,000 to compute using its own infrastructure.

Cloud Computing at Amazon

- “Active in cloud, Amazon reshapes computing,” *New York Times*, August 28, 2012.

Amazon Web Services

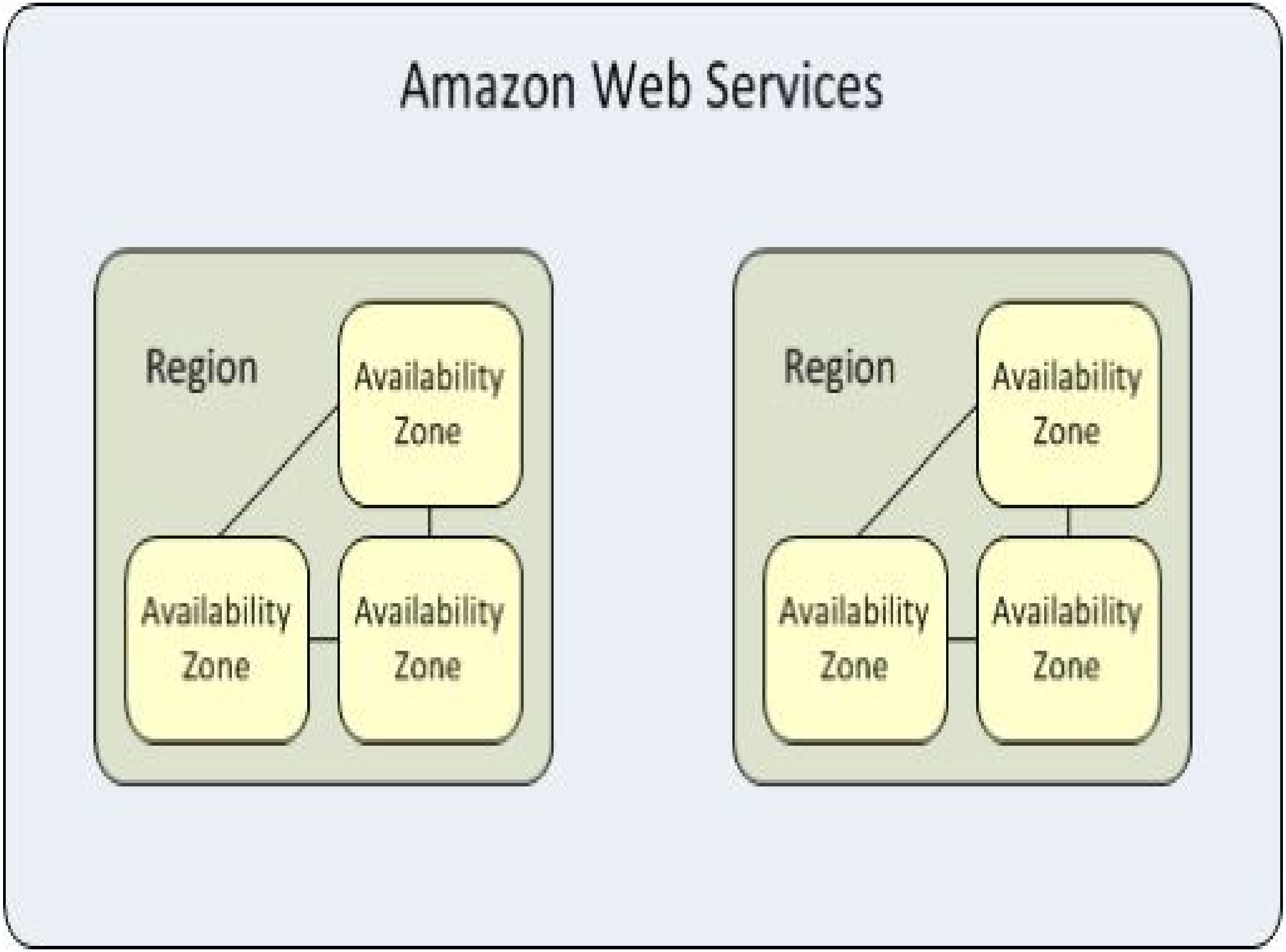
- AWS was one of the first companies to introduce a pay-as-you-go cloud computing model that scales to provide users with compute, storage or throughput as needed.
- AWS launched in 2006 from the internal infrastructure that Amazon.com built to handle its online retail operations.
- Amazon Web Services (AWS) is a secure cloud services platform, offering compute power, database storage, content delivery and other functionality to help businesses scale and grow.

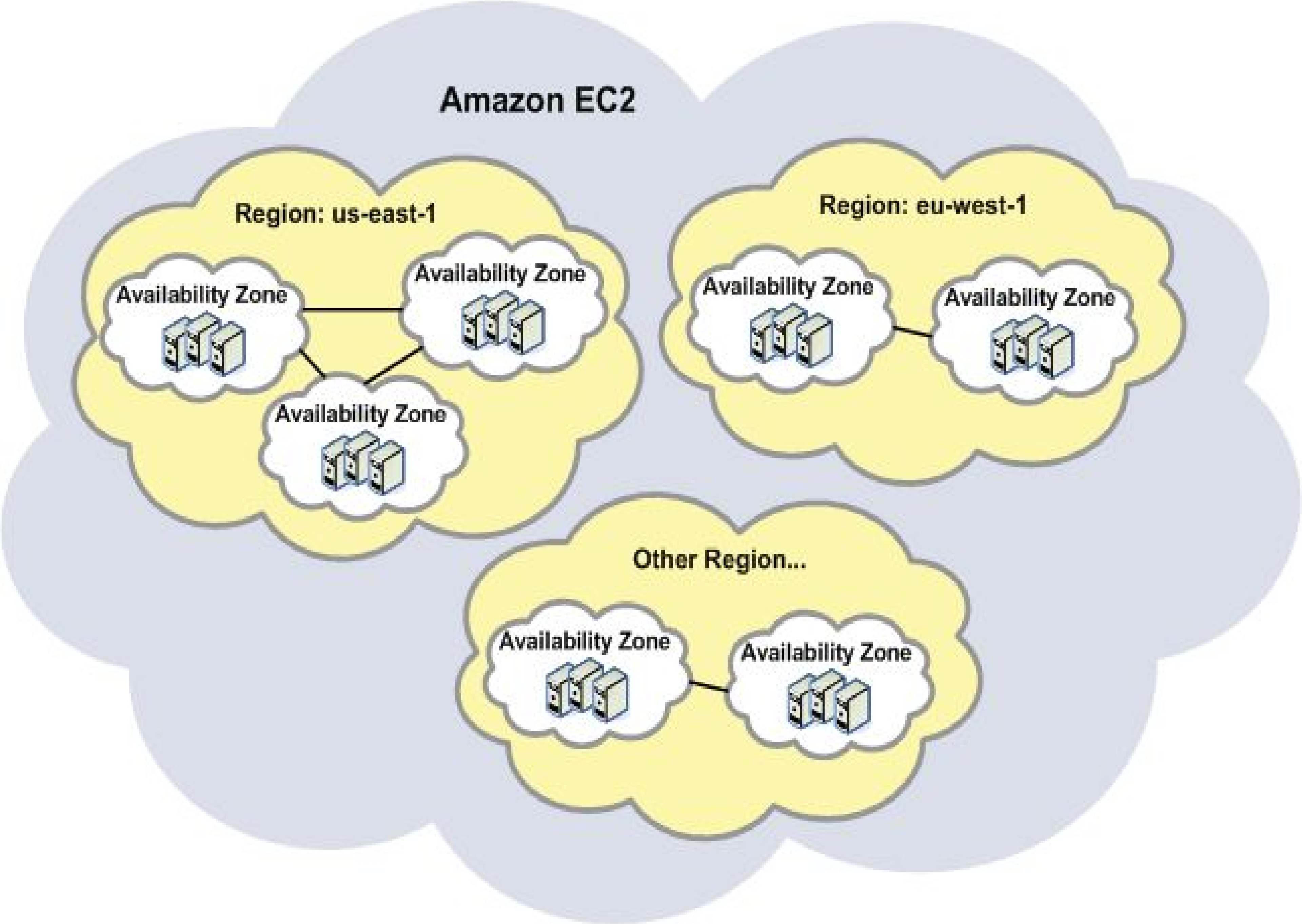
Amazon Web Services

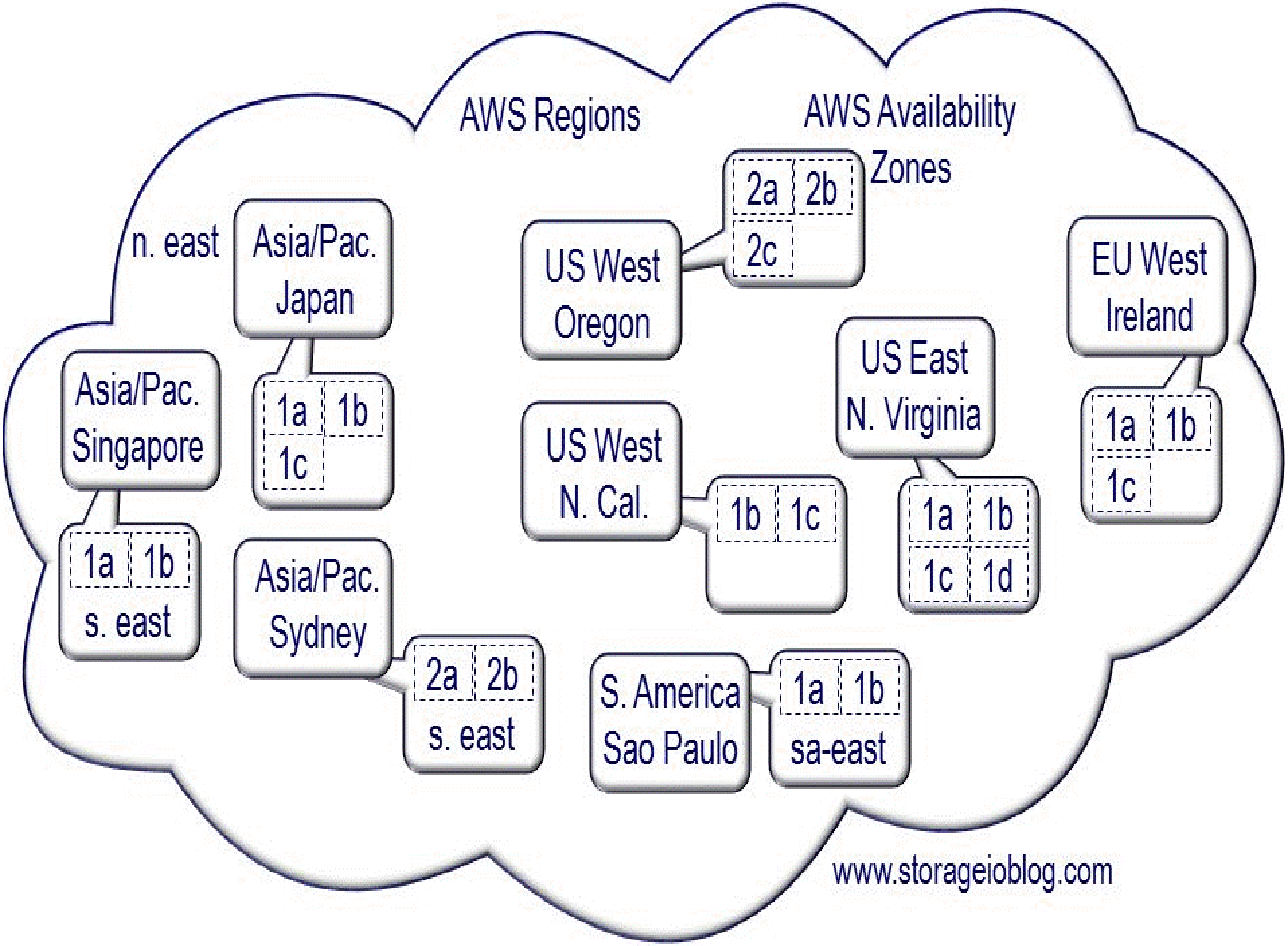
- It provides a mix of infrastructure as a service (IaaS), platform as a service (PaaS) and packaged software as a service (SaaS) offerings.
- The millions of customers are currently leveraging AWS cloud products and solutions to build sophisticated applications with increased flexibility, scalability and reliability.
- **Amazon Web Services (AWS**) is a subsidiary of Amazon.com that provides on-demand cloud computing platforms to individuals, companies and governments, on a paid subscription basis with a free-tier option available for 12 months.

AWS regions and availability zones

- Amazon offers cloud services through a network of data centers on several continents.
- In each *region* there are several availability zones interconnected by high-speed networks.
- An *availability zone* is a data center consisting of a large number of servers.
- Regions do not share resources and communicate through the Internet.


Amazon Web Services
Region
Availability Zone
Availability Zone
Availability Zone
Region
Availability Zone
Availability Zone
Availability Zone


Amazon EC2
Region: us-east-1
Availability Zone
Availability Zone
Availability Zone
Region: eu-west-1
Availability Zone
Availability Zone
Other Region...
Availability Zone
Availability Zone


AWS Regions
AWS Availability Zones
n. east
Asia/Pac. Japan
1a 1b
1c
Asia/Pac. Singapore
1a 1b
s. east
Asia/Pac. Sydney
2a 2b
s. east
US West Oregon
2a 2b
2c
US West N. Cal.
1b 1c
US East N. Virginia
1a 1b
1c 1d
EU West Ireland
1a 1b
1c
S. America Sao Paulo
1a 1b
sa-east
www.storageioblog.com

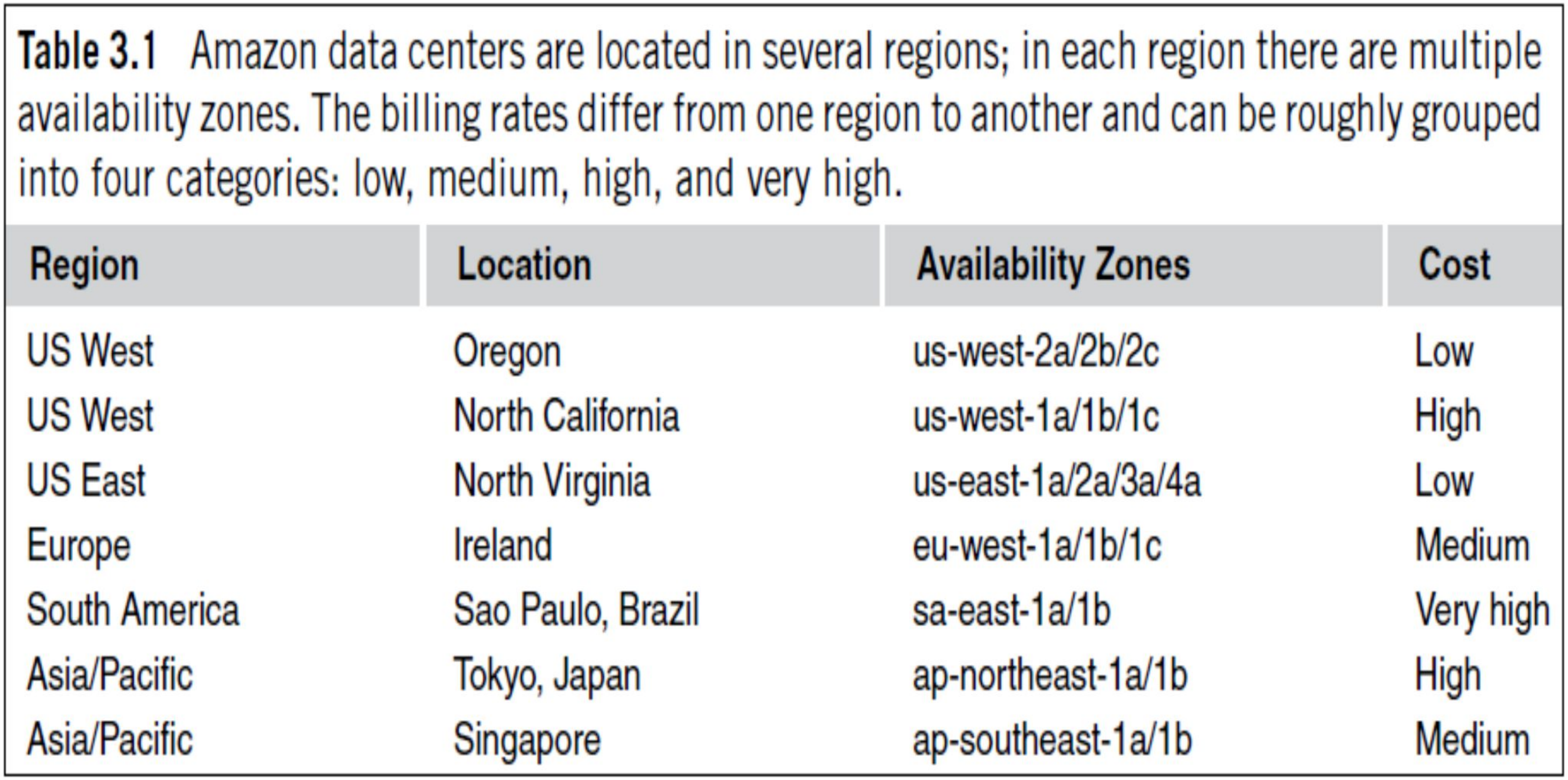
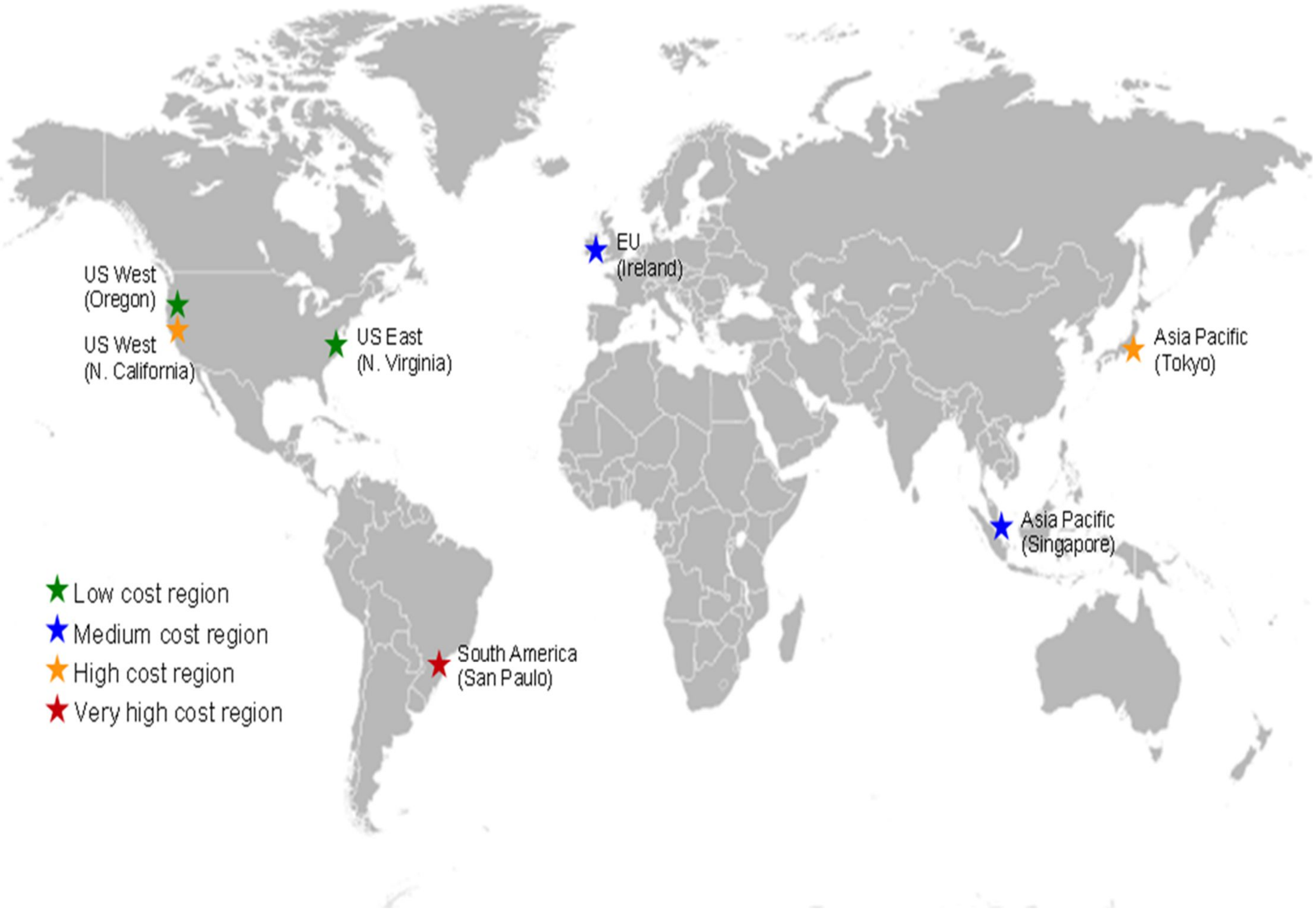


Table 3.1 Amazon data centers are located in several regions; in each region there are multiple availability zones. The billing rates differ from one region to another and can be roughly grouped into four categories: low, medium, high, and very high.

Region	Location	Availability Zones	Cost
US West	Oregon	us-west-2a/2b/2c	Low
US West	North California	us-west-1a/1b/1c	High
US East	North Virginia	us-east-1a/2a/3a/4a	Low
Europe	Ireland	eu-west-1a/1b/1c	Medium
South America	Sao Paulo, Brazil	sa-east-1a/1b	Very high
Asia/Pacific	Tokyo, Japan	ap-northeast-1a/1b	High
Asia/Pacific	Singapore	ap-southeast-1a/1b	Medium


US West
(Oregon)
US West
(N. California)
US East
(N. Virginia)
EU
(Ireland)
Asia Pacific
(Tokyo)
Asia Pacific
(Singapore)
South America
(San Paulo)
★ Low cost region
★ Medium cost region
★ High cost region
★ Very high cost region

AWS instances

- An instance is a virtual server with a well specified set of resources including: CPU cycles, main memory, secondary storage, communication and I/O bandwidth.
- The user chooses:
  - The region and the availability zone where this virtual server should be placed.
  - An instance type: CPU cycles, main memory, Secondary storage, Bandwidth and so on.

AWS instances

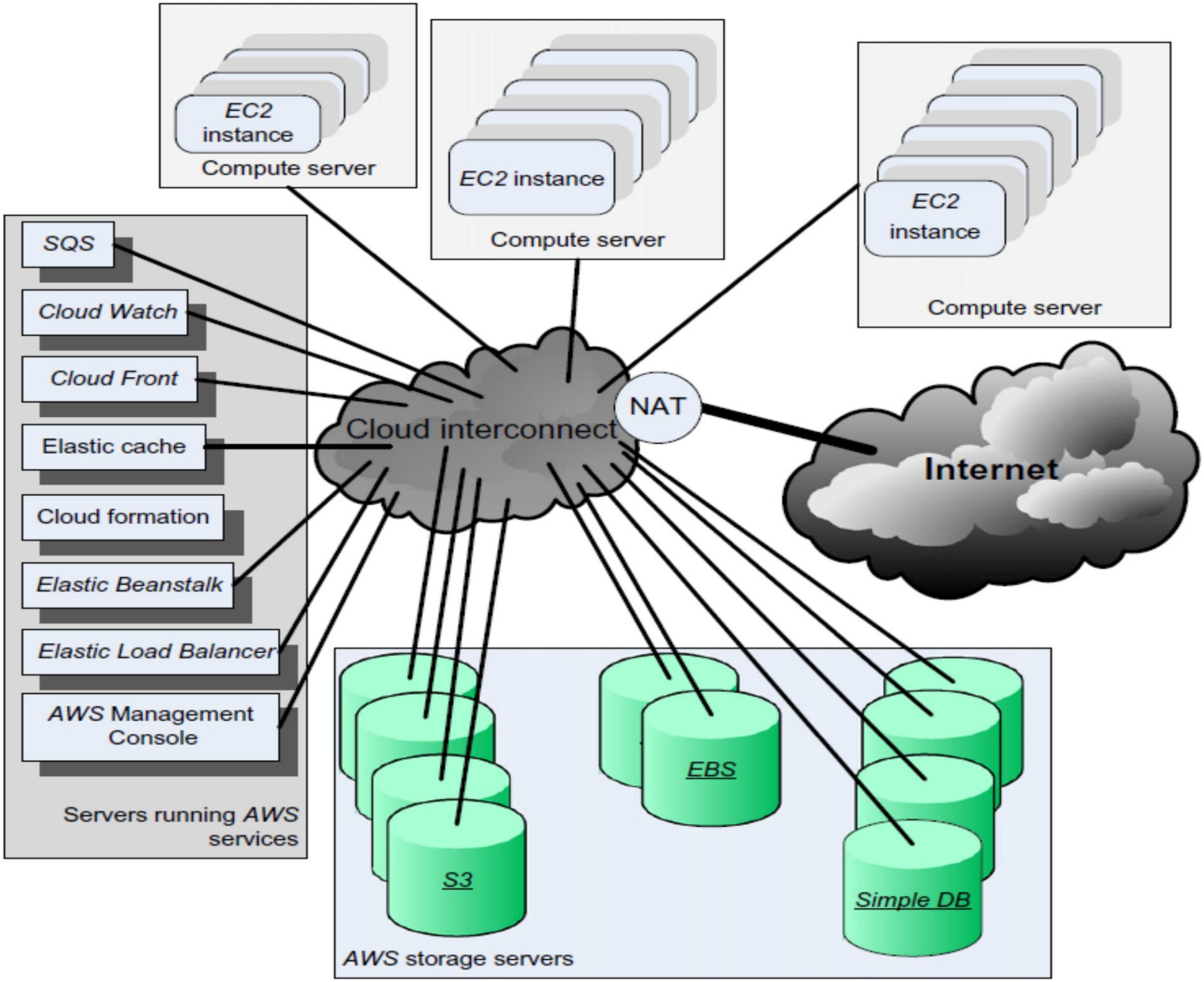
- When launched, an instance is provided with a DNS name; this name maps to a
  - private IP address → for internal communication within the internal Amazon network.
  - public IP address → for communication outside the internal Amazon network, e.g., for communication with the user that launched the instance.

AWS instances

- Network Address Translation (NAT) maps external IP addresses to internal ones.
- The public IP address is assigned for the lifetime of an instance and it is returned to the pool of available public IP addresses when the instance is either stopped or terminated.
- An instance can request an elastic IP address. The elastic IP address is a static public IP address allocated to an instance from the available pool of the availability zone.

AWS instances

- An elastic IP address is not released when the instance is stopped or terminated and must be released when no longer needed.

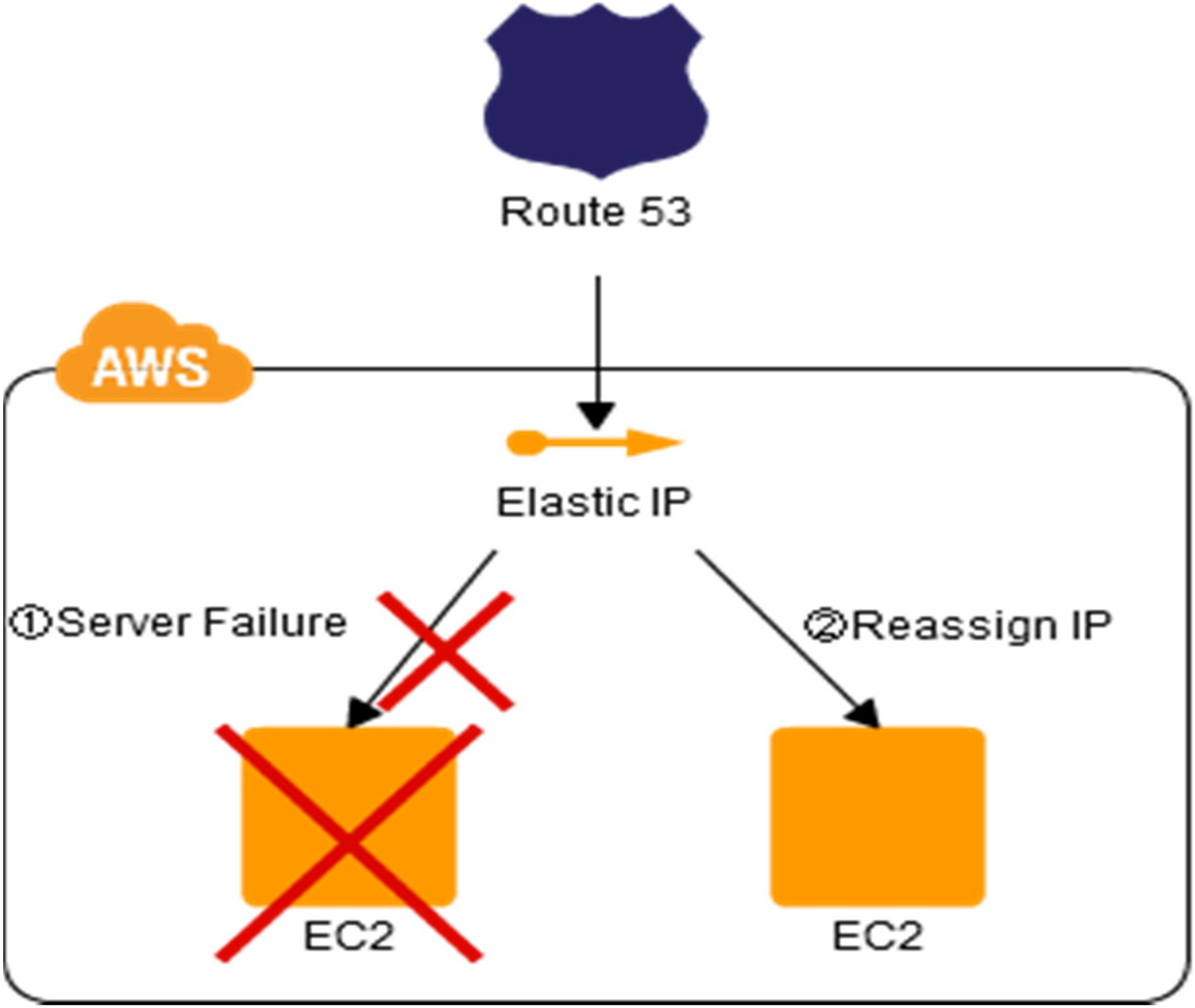

EC2 instance
Compute server
EC2 instance
Compute server
EC2 instance
Compute server
SQS
Cloud Watch
Cloud Front
Elastic cache
Cloud formation
Elastic Beanstalk
Elastic Load Balancer
AWS Management Console
Servers running AWS services
Cloud interconnect
NAT
Internet
S3
EBS
Simple DB
AWS storage servers

Elastic IP address

- An *Elastic IP address* is a static IPv4 address designed for dynamic cloud computing.
- An Elastic IP address is associated with your AWS account. With an Elastic IP address, you can mask the failure of an instance or software by rapidly remapping the address to another instance in your account.
- An Elastic IP is essentially tied to your AWS account. You can freely associate it with any AWS instance.

Elastic IP address

- The public IP is assigned to an instance when it is created.
- If you stop that instance, when you start it up you'll get another random public IP.
- Elastic IP is "permanent" in the sense that you own it and you associate it to a specific AWS instance ID.


Route 53
AWS
Elastic IP
①Server Failure
②Reassign IP
EC2
EC2

Route 53

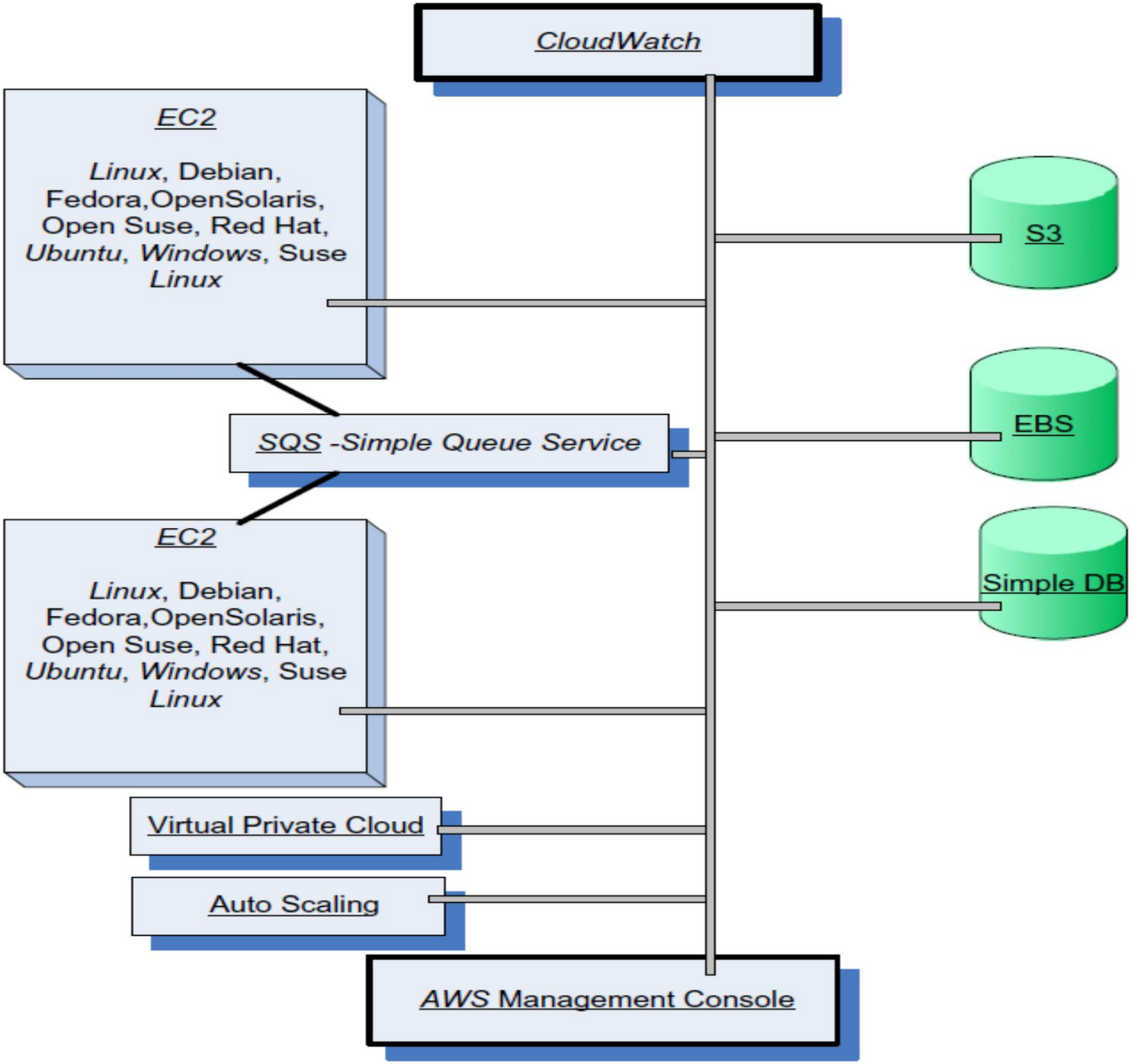
- **Amazon Route 53 (Route 53)** is a scalable and highly available Domain Name System (DNS).
- It is part of Amazon.com's cloud computing platform, Amazon Web Services (AWS).
- The name is a reference to TCP or UDP port 53,

AWS -Services

- AWS Management Console - allows users to access the services offered by AWS .
- Elastic Compute Cloud (EC2) - allows a user to launch a variety of operating systems.
- Simple Queuing Service (SQS) - allows multiple EC2 instances to communicate with one another.
- Simple Storage Service (S3), Simple DB, and Elastic Block Storage/Store (EBS) - storage services.

AWS - Services

- Cloud Watch - supports performance monitoring.
- Auto Scaling - supports elastic resource management.
- Virtual Private Cloud - provides a bridge between the existing IT infrastructure of an organization and the AWS cloud and allows direct migration of parallel applications. The VPC also allows existing management capabilities such as security services, firewalls, and intrusion detection systems to operate seamlessly within the cloud.


CloudWatch
EC2
Linux, Debian, Fedora,OpenSolaris, Open Suse, Red Hat, Ubuntu, Windows, Suse Linux
S3
SQS -Simple Queue Service
EBS
EC2
Linux, Debian, Fedora,OpenSolaris, Open Suse, Red Hat, Ubuntu, Windows, Suse Linux
Simple DB
Virtual Private Cloud
Auto Scaling
AWS Management Console

AWS Management Console

- The AWS Management Console is a browser-based GUI for Amazon Web Services (AWS).
- Through the console, a customer can manage their cloud computing, cloud storage and other resources running on the Amazon Web Services infrastructure.
- The AWS Console mobile app allows a user to perform operational tasks from a mobile device. The mobile app can be downloaded from the Amazon Appstore, Google Play or the Apple App Store.

AWS Management Console

- Is a web application for managing Amazon Web Services. It consists of list of various services to choose from. It also provides all information related to our account like billing.
- A web-based point and click interface to manage and monitor the Amazon infrastructure suite including (but not limited to) EC2, EBS, S3, SQS, Amazon Elastic MapReduce, and Amazon CloudFront.

Amazon CloudFront

- It is content delivery network (CDN).
- **Amazon CloudFront** is a web service that speeds up distribution of static and dynamic web content, for example, .html, .css, .php, image, and media files, to end users.
- CDNs serve a large portion of the Internet content today, including web objects (text, graphics and scripts), downloadable objects (media files, software, documents), applications (e-commerce, portals), live streaming media, on-demand streaming media, and social networks.

EC2 – Elastic Compute Cloud

- A web service that allows businesses to run application programs in the AWS public cloud.
- Is a Web service with a simple interface for launching instances of an application under several operating systems, such as several Linux distributions, Microsoft Windows Server 2003 and 2008, OpenSolaris, FreeBSD, and NetBSD.

EC2 – Elastic Compute Cloud

- An instance is created either from a predefined *Amazon Machine Image* (AMI) digitally signed and stored in *S3* or from a user-defined image.
- The image includes the operating system, the run-time environment, the libraries, and the application desired by the user.
- Each virtual machine or instance functions as a virtual private server.
- An instance specifies the maximum amount of resources available to an application, the interface for that instance, and the cost per hour.

EC2 – Elastic Compute Cloud

- Provides secure, resizable compute capacity in the cloud.
- EC2 automatically distributes the incoming application traffic among multiple instances using the *elastic load-balancing* facility
- An *EC2* instance is characterized by the resources it provides:
  - VC (Virtual Computers) – virtual systems running the instance.
  - CU (Compute Units) – measures the computing power of each system.
  - Memory.
  - I/O capabilities.

EC2 – Elastic Compute Cloud

- A user can:
  - Launch an instance from an existing AMI and terminate an instance;
  - start and stop an instance;
  - create a new image;
  - add tags to identify an image;
  - reboot an instance from an image.

Instance types

- Standard instances: micro (StdM), small (StdS), large (StdL), extra large (StdXL); small is the default.
- High memory instances: high-memory extra large (HmXL), high-memory double extra large (Hm2XL), and high-memory quadruple extra large (Hm4XL).
- High CPU instances: high-CPU extra large (HcpuXL).
- Cluster computing: cluster computing quadruple extra large (Cl4XL).

Nine instances supported by *EC2*.

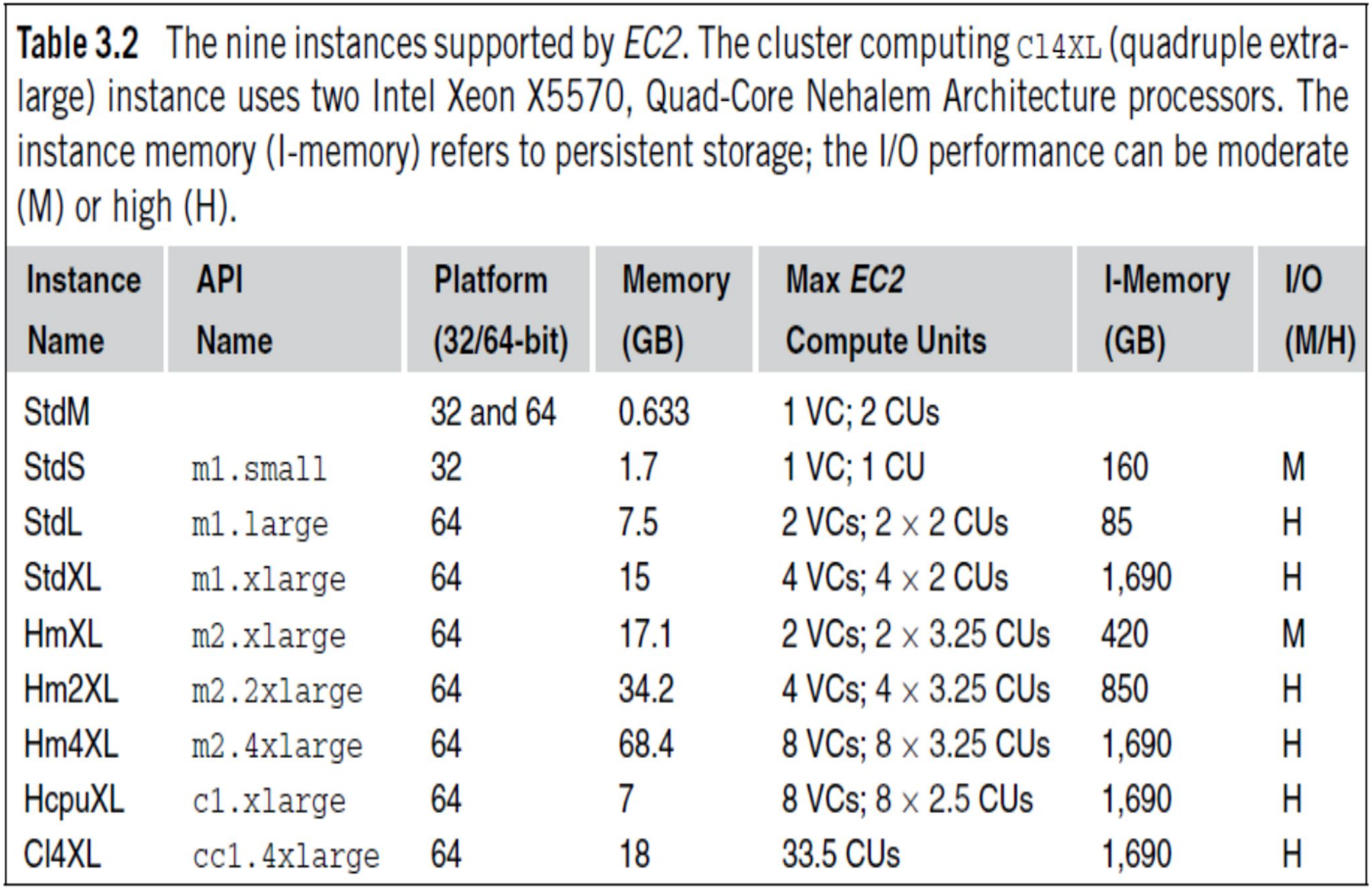


Table 3.2 The nine instances supported by *EC2*. The cluster computing C14XL (quadruple extra-large) instance uses two Intel Xeon X5570, Quad-Core Nehalem Architecture processors. The instance memory (I-memory) refers to persistent storage; the I/O performance can be moderate (M) or high (H).

Instance Name	API Name	Platform (32/64-bit)	Memory (GB)	Max *EC2* Compute Units	I-Memory (GB)	I/O (M/H)
StdM		32 and 64	0.633	1 VC; 2 CUs		
StdS	`m1.small`	32	1.7	1 VC; 1 CU	160	M
StdL	`m1.large`	64	7.5	2 VCs; 2 × 2 CUs	85	H
StdXL	`m1.xlarge`	64	15	4 VCs; 4 × 2 CUs	1,690	H
HmXL	`m2.xlarge`	64	17.1	2 VCs; 2 × 3.25 CUs	420	M
Hm2XL	`m2.2xlarge`	64	34.2	4 VCs; 4 × 3.25 CUs	850	H
Hm4XL	`m2.4xlarge`	64	68.4	8 VCs; 8 × 3.25 CUs	1,690	H
HcpuXL	`c1.xlarge`	64	7	8 VCs; 8 × 2.5 CUs	1,690	H
Cl4XL	`cc1.4xlarge`	64	18	33.5 CUs	1,690	H

Instance cost

- A main attraction of the Amazon cloud computing is the low cost.

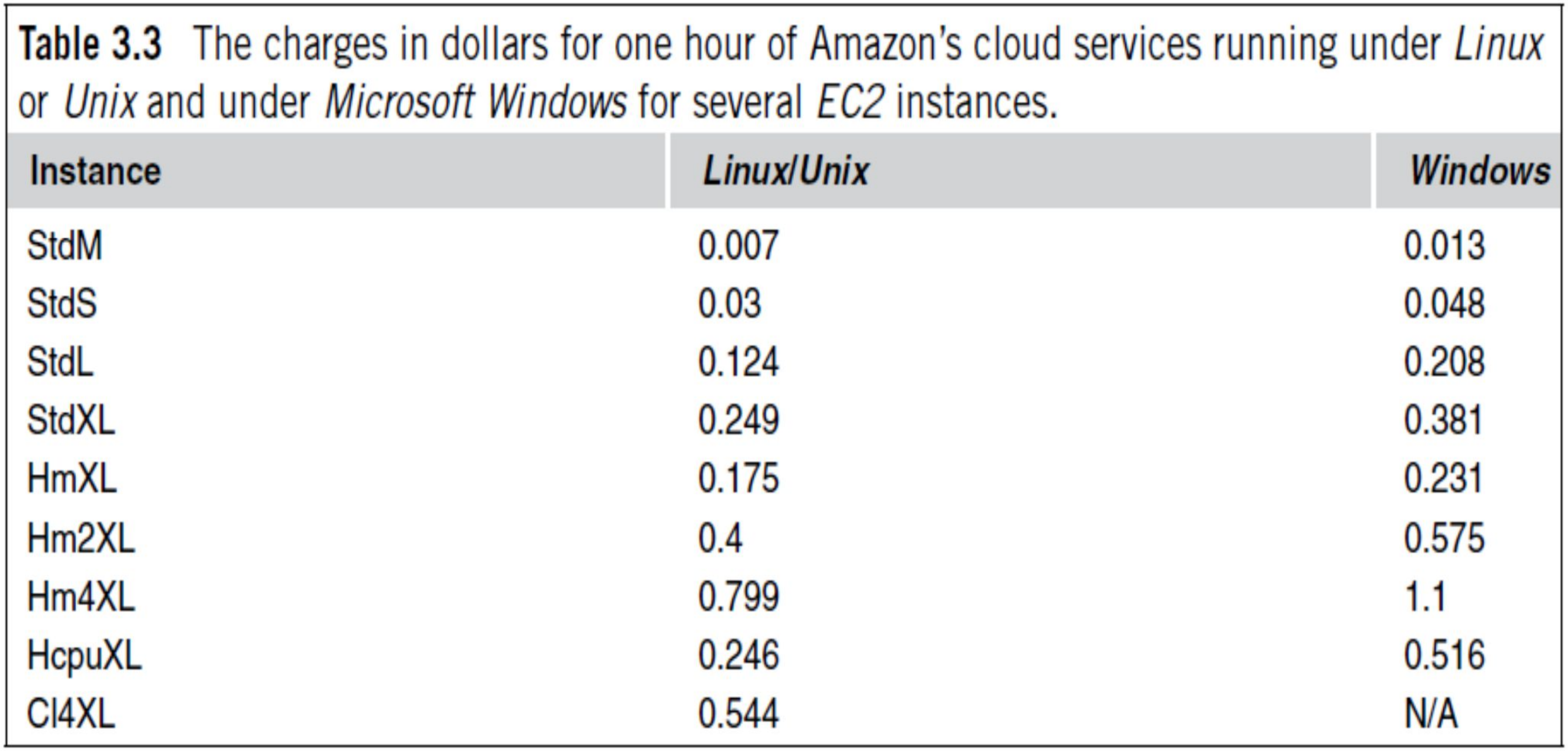


Table 3.3 The charges in dollars for one hour of Amazon's cloud services running under *Linux* or *Unix* and under *Microsoft Windows* for several *EC2* instances.

Instance	*Linux/Unix*	*Windows*
StdM	0.007	0.013
StdS	0.03	0.048
StdL	0.124	0.208
StdXL	0.249	0.381
HmXL	0.175	0.231
Hm2XL	0.4	0.575
Hm4XL	0.799	1.1
HcpuXL	0.246	0.516
Cl4XL	0.544	N/A

Simple Storage System (S3)

Amazon Simple
Storage Service
(S3)

Simple Storage System (S3)

- Amazon S3 is object storage built to store and retrieve any amount of data(any type of data) from anywhere – web sites ,mobile apps, corporate applications, and data from IoT sensors or devices.
- It is designed to deliver 99.999999999% durability, and stores data for millions of applications used by market leaders in every industry.
- Provides authentication mechanisms to ensure that data is kept secure; objects can be made public, and rights can be granted to other users.

Simple Storage System (S3)

- S3 stores data as objects within resources called **buckets**. The user can store as many objects as per requirement within the bucket, and can read, write and delete objects from the bucket.
- An application can handle an unlimited number of objects ranging in size from 1 byte to 5 TB.
- A bucket can be stored in a Region selected by the user.

Simple Storage System (S3)

- S3 supports PUT, GET, and DELETE primitives to manipulate objects but does not support primitives to copy, rename, or move an object from one bucket to another.
- The object names are global.
- S3 maintains for each object: the name, modification time, an access control list (Ex. 777), and up to 4 KB of user-defined metadata.

Simple Storage System (S3)

- S3 computes the MD5 of every object written and returns it in a field called ETag.
- A user is expected to compute the MD5 of an object stored or written and compare this with the ETag; if the two values do not match, then the object was corrupted during transmission or storage.
- MD5 (Message-Digest Algorithm) is a widely used cryptographic hash function; it produces a 128-bit hash value. It is used for checksums.

Elastic Block Store (EBS)

- Elastic Block Store (EBS) provides persistent block-level storage volumes for use with Amazon EC2 instances.
- Suitable for database applications, file systems, and applications using raw data devices.
- A volume appears to an application as a raw, unformatted and reliable physical disk.
- The size of the storage volumes ranges from 1GB to 1TB.

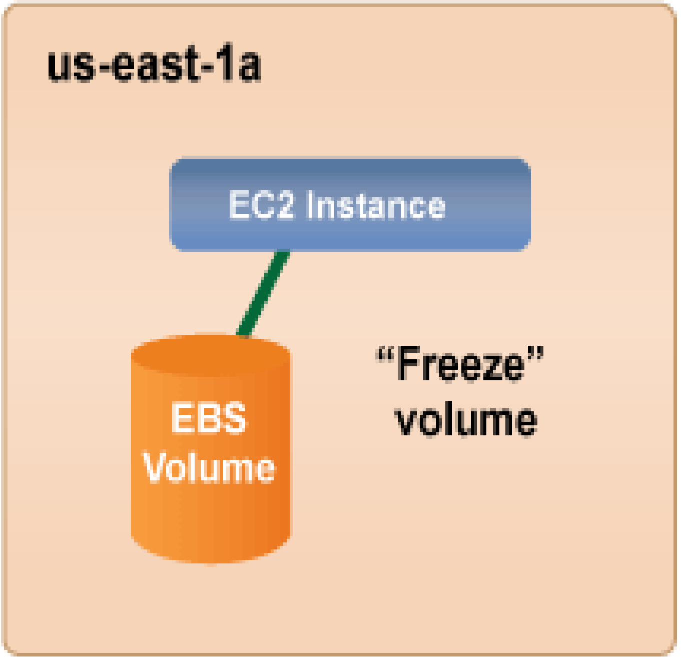
Elastic Block Store (EBS)

- The volumes are grouped together in availability zones and are automatically replicated in each zone to protect from component failure, offering high availability and durability.
- An EC2 instance may mount multiple volumes, but a volume cannot be shared among multiple instances.
- The EBS supports the creation of snapshots of the volumes attached to an instance and then uses them to restart an instance.

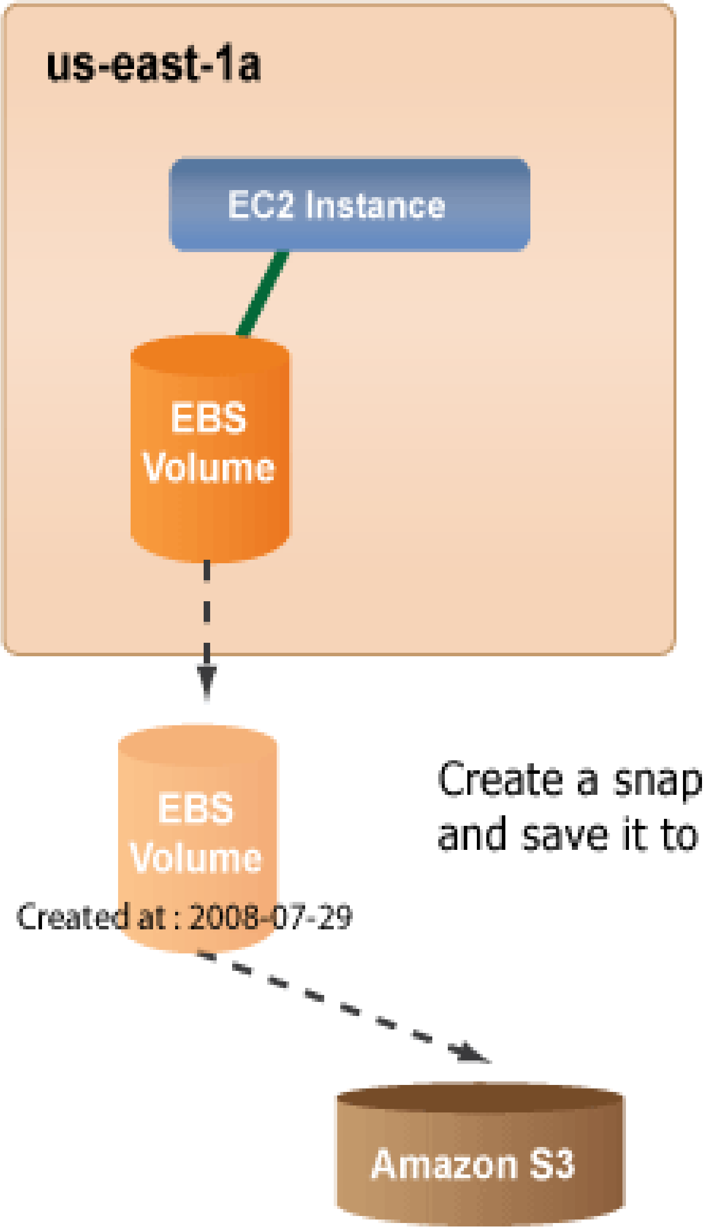
EBS Snapshots

- **Snapshots** are incremental backups, which means that only the blocks on the device that have changed after your most recent **snapshot** are saved.
- Used to back up the data on Amazon **EBS** volumes to Amazon S3 by taking point-in-time **snapshots**.

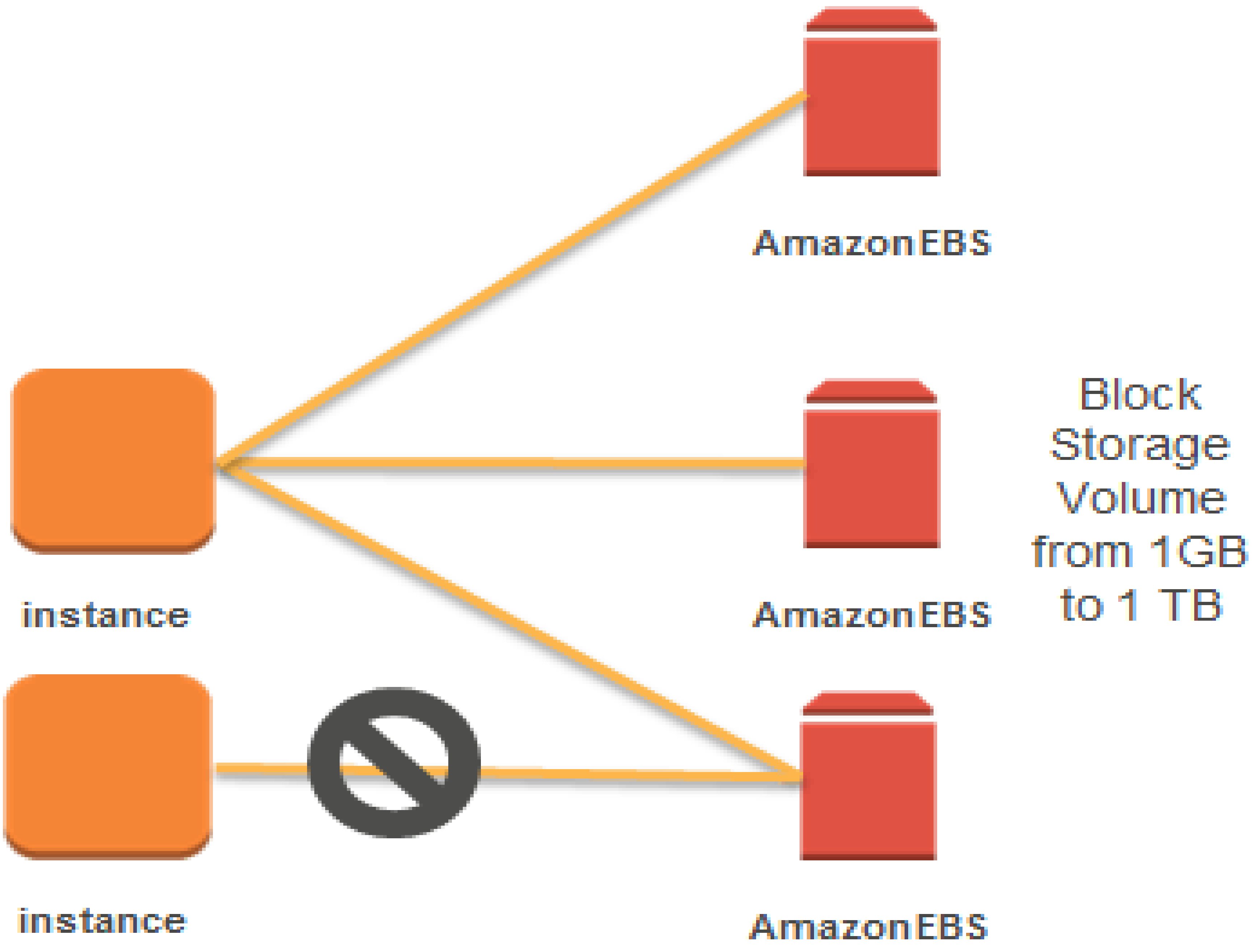
EBS snapshots



Before taking a snapshot, make sure that no more writes are being made to the volume in order to prevent data corruption and ensure that data remains synchronized.



Create a snapshot of a volume and save it to S3.


AmazonEBS
AmazonEBS
Block Storage Volume from 1GB to 1 TB
instance
instance
AmazonEBS

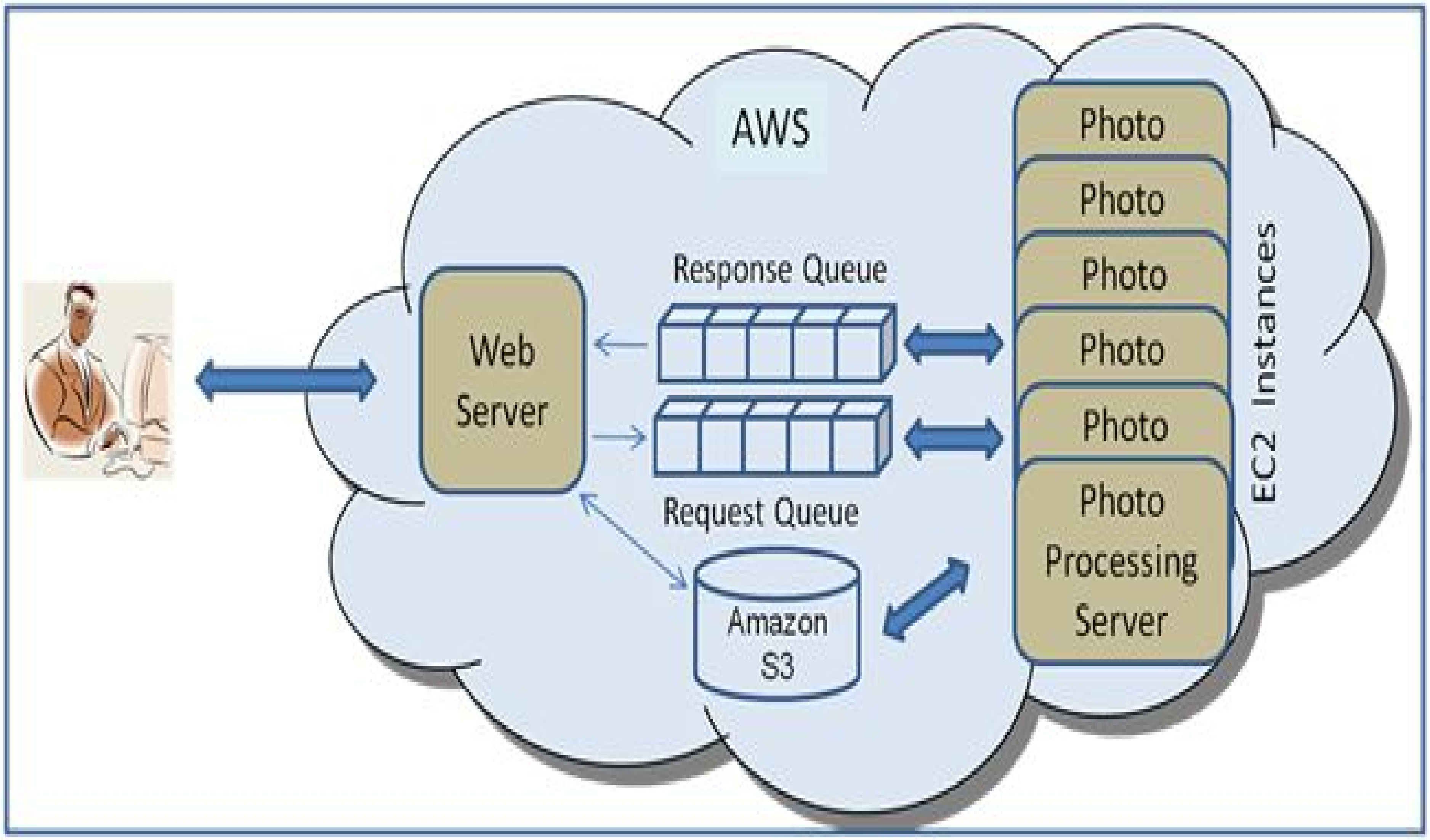
SimpleDB

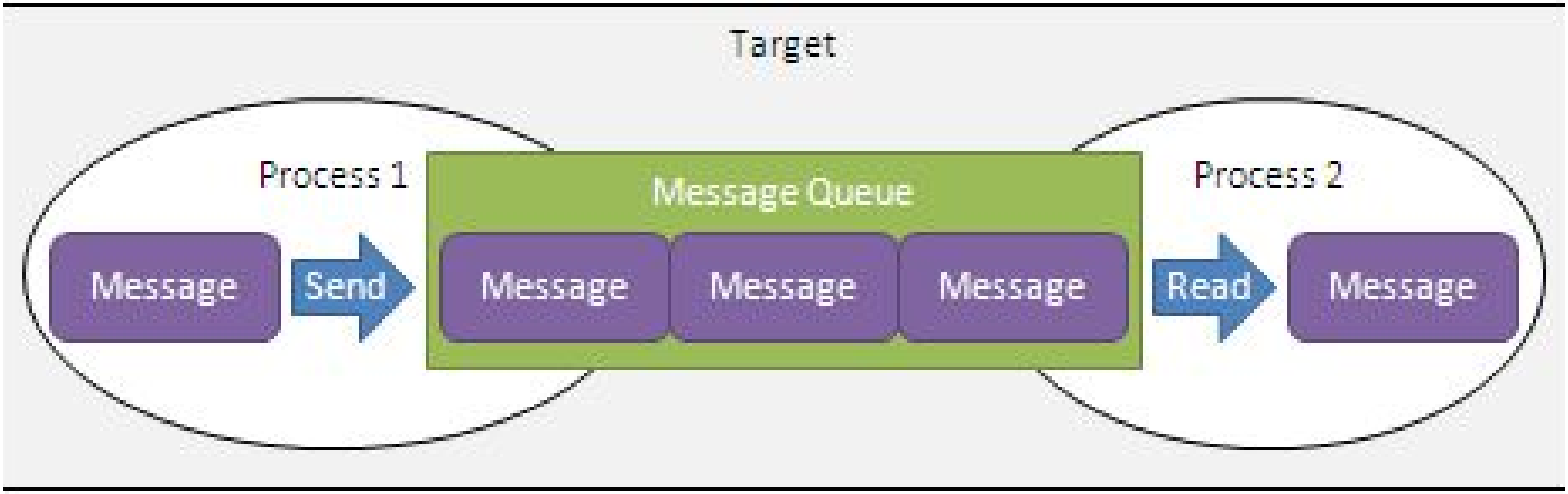
- *Simple DB* is a non-relational data store that allows developers to store and query data items via Web services requests.
- It supports store-and-query functions traditionally provided only by relational databases.
- Amazon Simple Database Service (SimpleDB), also known as a key value data store, is a highly available and flexible non-relational database that allows developers to request and store data.

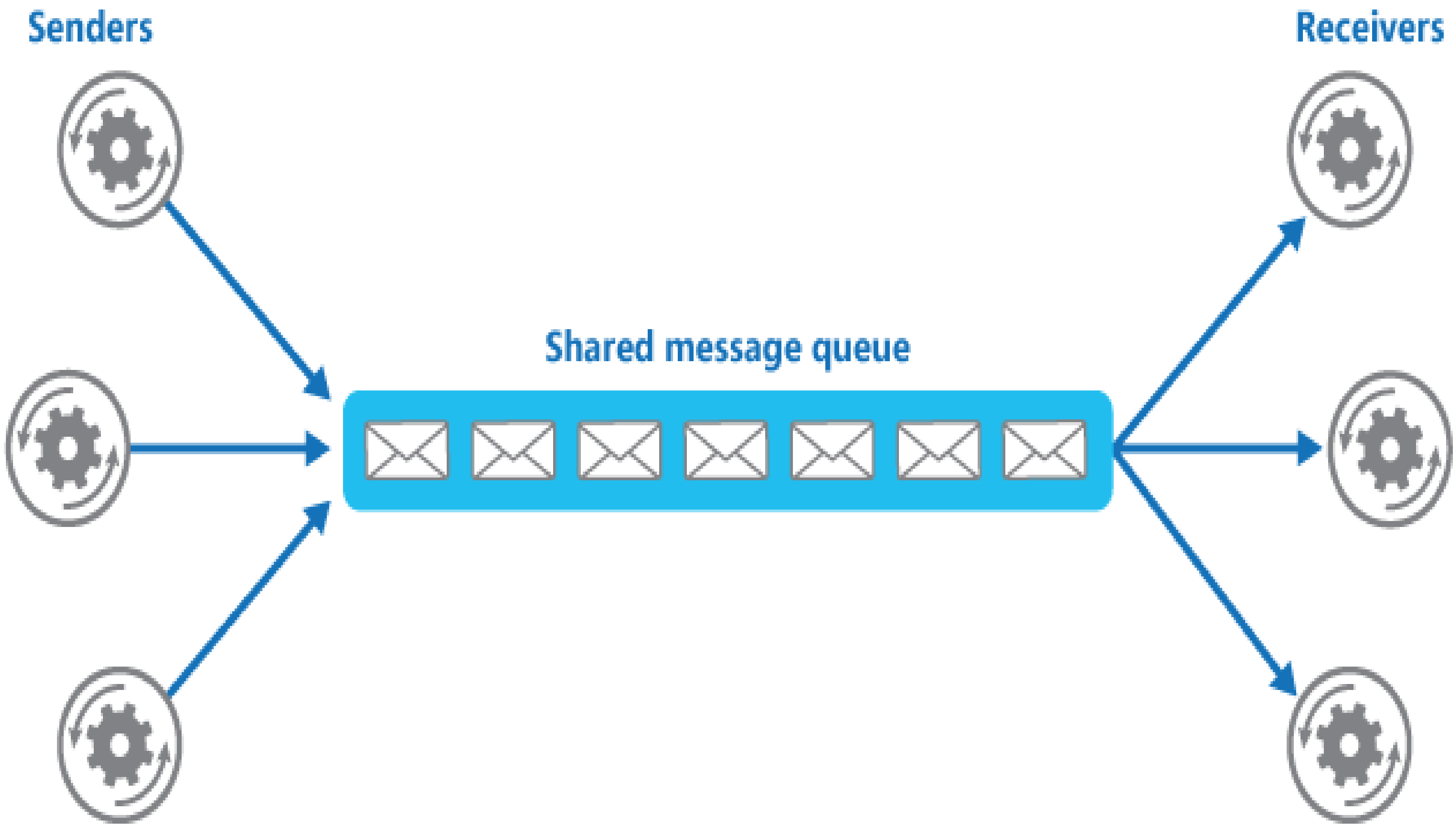
SimpleDB

- *Simple DB is* distributed database and creates multiple geographically distributed copies of each data item and supports high-performance Web applications.
- It manages automatically:
  - The infrastructure provisioning.
  - Hardware and software maintenance.
  - Replication and indexing of data items.
  - Performance tuning

SQS - Simple Queue Service




Target
Process 1
Message Queue
Process 2
Message
Send
Message
Message
Message
Read
Message


Senders
Shared message queue
Receivers

SQS - Simple Queue Service

- It is a hosted message queue.
- SQS is a system for supporting automated workflows; it allows multiple Amazon EC2 instances to coordinate their activities by sending and receiving SQS messages.
- Any computer connected to the Internet can add or read messages without any installed software or special firewall configurations.

SQS - Simple Queue Service

- Applications using *SQS* can run independently and asynchronously and do not need to be developed with the same technologies.
- Developers can access SQS through standards-based SOAP and Query interfaces.
- Queues can be shared with other *AWS* accounts and queue sharing can be restricted by IP address and time-of-day.

CloudWatch

- Amazon CloudWatch is a component of Amazon Web Services (AWS) that provides monitoring for AWS resources and the customer applications running on the Amazon infrastructure.
- Monitoring infrastructure used by application developers, users, and system administrators to collect and track metrics important for optimizing the performance of applications and for increasing the efficiency of resource utilization.

CloudWatch

- The CloudWatch interface provides current statistics that can be viewed in graph format.
- Users can set notifications (called “alarms”) to be sent when something being monitored surpasses a specified threshold. The app can also detect and shut down unused or underused EC2 instances.
- Used to gain system-wide visibility into resource utilization, application performance, and operational health. These insights are used to react and keep the application running smoothly.

CloudWatch

- When launching an Amazon Machine Image (AMI) the user can start the CloudWatch and specify the type of monitoring:
  - Basic Monitoring - free of charge; collects data at five-minute intervals.
  - Detailed Monitoring - subject to charge; collects data at one minute interval.

AWS services introduced in 2012

- Route 53 - low-latency DNS service used to manage user's DNS public records.
- Elastic MapReduce (EMR) - supports processing of large amounts of data using a hosted Hadoop running on EC2.
- Simple Workflow Service (SWF) - supports workflow management; allows scheduling, management of dependencies, and coordination of multiple EC2 instances.

AWS services introduced in 2012

- ElastiCache - enables web applications to retrieve data from a managed in-memory caching system rather than a much slower disk-based database.
- DynamoDB - scalable and low-latency fully managed NoSQL database service.
- CloudFront - web service for content delivery.

AWS services introduced in 2012

- Elastic Load Balancer - automatically distributes the incoming requests across multiple instances of the application.
- Elastic Beanstalk - handles automatically deployment, capacity provisioning, load balancing, auto-scaling, and application monitoring functions.

.

AWS services introduced in 2012

- CloudFormation
  - allows the creation of a stack describing the infrastructure for an application.
  - helps to model and set up Amazon Web Services resources so that one can spend less time managing those resources and more time focusing on applications that run in **AWS**.
  - The user creates a template, a text file formatted as in JavaScript Object Notation (JSON), describing the resources, the configuration values, and the interconnection among these resources.

Elastic Beanstalk

- It is a cloud deployment and provisioning service that automates the process of getting applications set up on the Amazon Web Services (AWS) infrastructure.
- AWS Elastic Beanstalk is an easy-to-use service for deploying and scaling web applications and services developed with Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker on familiar servers such as Apache, Nginx, Passenger, and IIS.

Elastic Beanstalk

- You simply upload your application, and AWS Elastic Beanstalk automatically handles the details of capacity provisioning, load balancing, auto-scaling, and application health monitoring.

- **AWS Elastic Beanstalk** is an orchestration service for deploying infrastructure which orchestrates various AWS services, including EC2, S3, Simple Notification Service (SNS), CloudWatch, autoscaling, and Elastic Load Balancers.

Elastic Beanstalk- **The management functions**

- Deploy a new application version (or rollback to a previous version).
- Access to the results reported by CloudWatch monitoring service.
- Email notifications when application status changes or application servers are added or removed.
- Access to server log files without needing to login to the application servers.

Steps to run an application

- Retrieve the user input from the front-end.
- Retrieve the disk image of a VM (Virtual Machine) from a repository.
- Locate a system and requests the VMM (Virtual Machine Monitor) running on that system to setup a VM.
- Invoke the Dynamic Host Configuration Protocol (DHCP) and the IP bridging software to set up MAC and IP addresses for the VM.

User interactions with AWS

- The AWS Management Console. The easiest way to access all services
- Toolkits are provided for several programming languages including Java, PHP, C#, and Objective-C and so on.
- REST/SOAP APIs for web services.

(REST stands for **RE**presentational **S**tate **T**ransfer)

Cloud computing: the Google perspective

- Google's effort is focused in the area of *Software-as-a-Service* (SaaS).
- It is estimated that the number of servers used by Google was close to 2.4 million in 2013 and was expected to reach close to 10 million in early 2018 [289].
- Google adheres to a bottom-up, engineer-driven, liberal licensing and user application development philosophy.

SaaS services offered by Google

- *Gmail* - hosts Emails on Google servers and provides a web interface to access the Email.
- *Google docs* - a web-based software for building text documents, spreadsheets and presentations.
- *Google Calendar* - a browser-based scheduler; supports multiple user calendars, calendar sharing, event search, display of daily/weekly/monthly views, and so on.

SaaS services offered by Google

- *Google Groups* - allows users to host discussion forums to create messages online or via Email.
- *Picasa* - a tool to upload, share, and edit images.
- *Google Maps* - web mapping service; offers street maps, a route planner, and an urban business locator for numerous countries around the world.

PaaS services offered by Google

- AppEngine - a developer platform hosted on the cloud.
    - Supports Python, Java, Node.js, Ruby, GO, PhP.
    - The database for code development can be accessed with GQL (Google Query Language) with a SQL-like syntax.
- Google Co-op - allows users to create customized search engines based on a set of facets/categories.
- Google Drive - online service for data storage [15GB].
- Google Base - allows users to load structured data from different sources to a central repository, that is a very large, self-describing, semi-structured, heterogeneous database.

Google Chromebook

- Purely Web-centric device running *Chrome OS.*
- The **Chromebook** is a new, faster computer. It starts in seconds, and offers thousands of apps. It has built-in virus protection, and backs up your stuff in the cloud. With automatic updates, it keeps getting better.
- Primarily used to perform a variety of tasks using the Google Chrome browser.

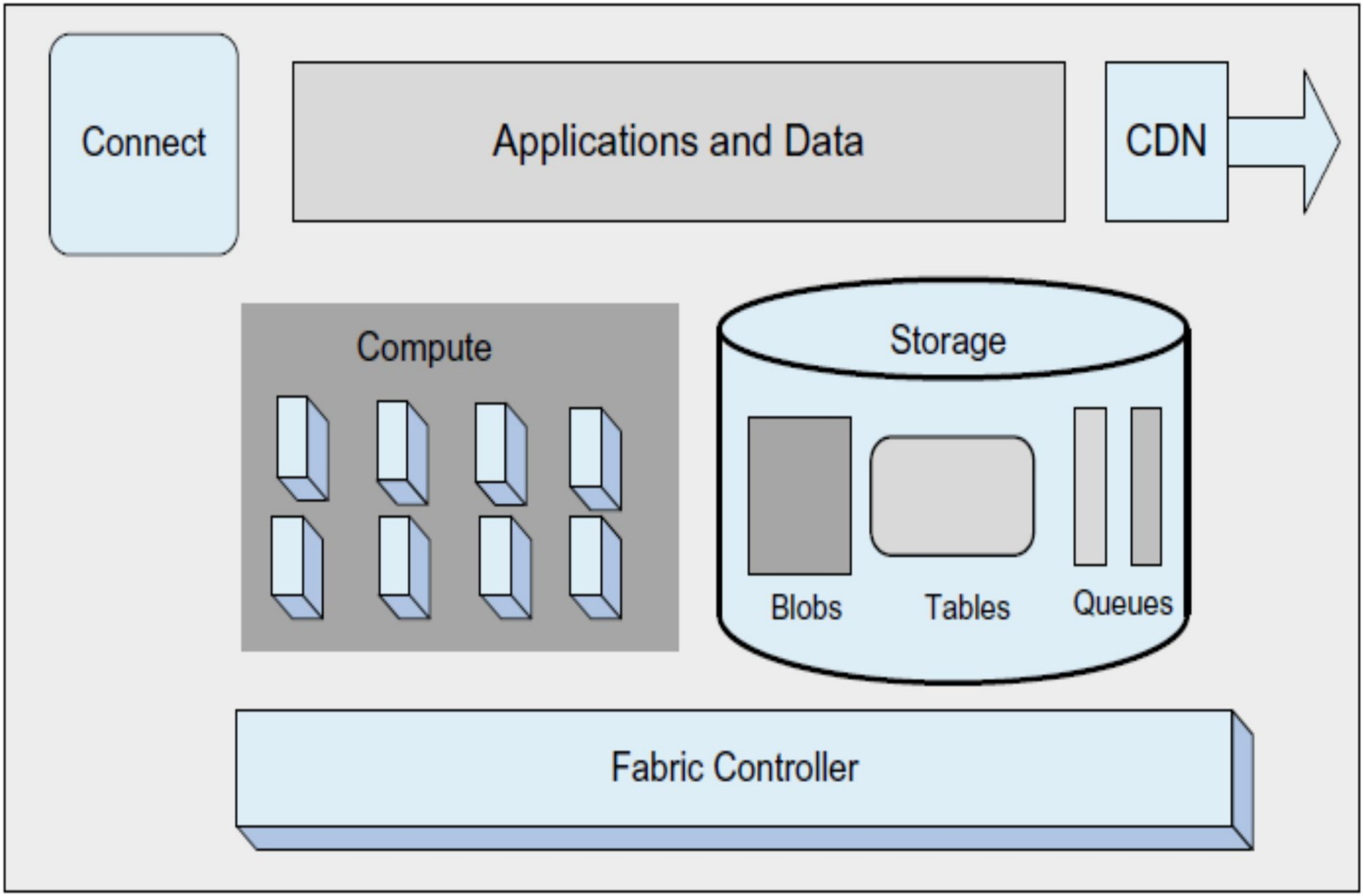
Microsoft Windows Azure and online services

- *Azure* and *Online Services* are, respectively, *PaaS* and *SaaS* cloud platforms from Microsoft.
- *Windows Azure is an operating system- has 3 components:*
  - *Compute - provides a computation environment.*
  - *Storage - for scalable storage.*
  - *Fabric Controller - deploys, manages, and monitors applications*.

Microsoft Windows Azure and online services

- *SQLAzure* is a cloud-based version of the SQL Server.
- *AzureAppFabric* (formerly .NET Services) is a collection of services for cloud applications.

Azure


Connect
Applications and Data
CDN
Compute
Storage
Blobs
Tables
Queues
Fabric Controller

AZURE

- The components of Windows Azure:
  - Compute, which runs cloud applications;
  - Storage, which uses blobs, tables, and queues to store data;
  - Fabric Controller, which deploys, manages, and monitors applications;
- CDN, which maintains cache copies of data; and Connect, which allows IP connections between the user systems and applications running on Windows Azure.

Open-source platforms for private clouds

- Private clouds provide a cost-effective alternative for very large organizations.
- A private cloud has essentially the same structural components as a commercial one: the servers, the network, virtual machines monitors (VMMs) running on individual systems, an archive containing disk images of virtual machines (VMs), a front end for communication with the user, and a cloud control infrastructure.

Open-source platforms for private clouds

- Eucalyptus - can be regarded as an open-source counterpart of Amazon's EC2.
- Open-Nebula - a private cloud with users actually logging into the head node to access cloud functions. The system is centralized and its default configuration uses the NFS file system.

Open-source platforms for private clouds

- Nimbus - a cloud solution for scientific applications based on Globus software (Grid computing software); inherits from Globus:
  - The image storage.
  - The credentials for user authentication.
  - The requirement that a running Nimbus process can **ssh** into all compute nodes.
- Ssh (Secure shell) : secure remote login

SSH (Secure shell)

- is a network protocol that allows data to be exchanged using a secure channel between two networked devices.
- SSH uses public-key cryptography to authenticate the remote computer and allow the remote computer to authenticate the user.
- It also allows remote control of a device.

Eucalyptus
(Elastic Utility Computing Architecture for Linking Your Programs To Useful Systems)

- *Eucalyptus (www.eucalyptus.com) can be regarded as an open-source counterpart of Amazon’s EC2,*
- *The systems supports several operating systems including CentOS 5 and 6, RHEL 5 and 6, Ubuntu 10.04 LTS, and 12.04 LTS.*

Eucalyptus (Components)

- *Virtual Machines* - run under several VMMs including Xen, KVM, and VMware.
- *Node Controller - Runs on every server or node designated to host a VM and controls the activities of the node. Reports to a cluster controller.*
- *Cluster Controller - Controls a number of servers. Interacts with the node controller on each server to schedule requests on that node. Cluster controllers are managed by the cloud controller*

Eucalyptus (Components)

- *Cloud Controller* - provides the cloud access to end-users, developers, and administrators.
- *Storage Controller* - provides persistent virtual hard drives to applications. It is the correspondent of EBS.
- *Storage Service (Walrus)* - provides persistent storage; similar to S3, it allows users to store objects in buckets.

Conclusions of the comparative analysis

- Eucalyptus is best suited for a large corporation with its own private cloud because it ensures a degree of protection from user malice and mistakes.
- OpenNebula is best suited for a testing environment with a few servers.
- Nimbus is more adequate for a scientific community.

OpenStack

- Is an open-source project started in 2009 at the National Aeronautics and Space Administration (NASA) in collaboration with Rackspace (www.rackspace.com) to develop a scalable cloud operating system for farms of servers using standard hardware.

OpenStack

- The current version of the system supports a wide range of features such as application programming interfaces (APIs) with rate limiting and authentication; live VM management to run, reboot, suspend, and terminate instances; role-based access control; and the ability to allocate, track, and limit resource utilization.
- The administrators and the users control their resources using an extensible Web application called the *Dashboard*.

Cloud storage diversity and vendor lock-in

- Risks when a large organization relies on a single cloud service provider:
  - Cloud services may be unavailable for a short or an extended period of time.
  - Permanent data loss in case of a catastrophic system failure.
- The provider(CSP) may increase the prices for service, and charge more for computing cycles, memory, storage space, and network bandwidth than other CSPs.

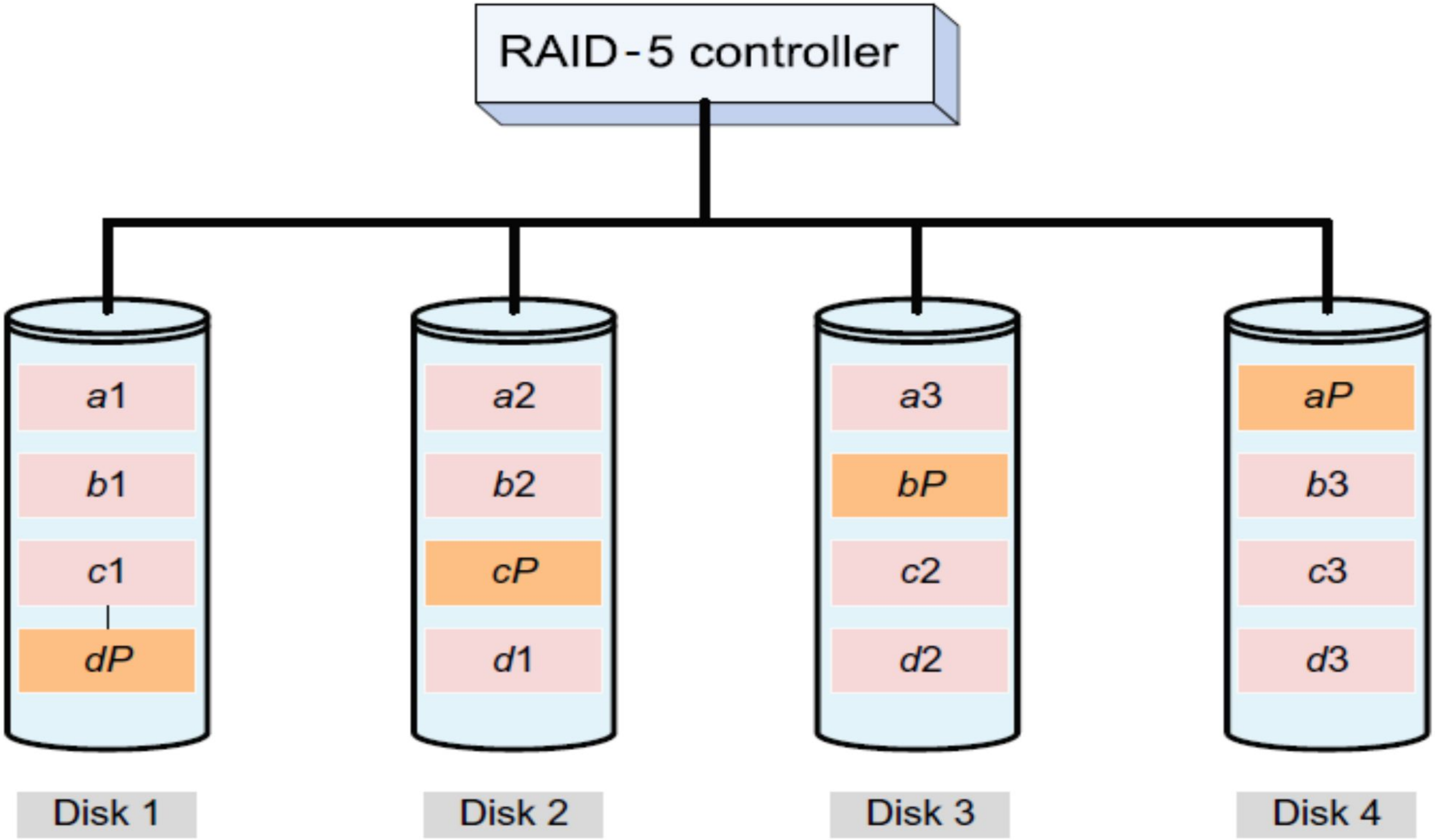
Cloud storage diversity and vendor lock-in

- Switching to another provider could be very costly due to the large volume of data to be transferred from the old to the new provider.
- Transferring terabytes or possibly petabytes of data over the network takes a fairly long time and incurs substantial charges for the network bandwidth.
- A solution to guarding against the problems posed by the vendor lock-in is to replicate the data to multiple cloud service providers, similar to data replication in RAID-5.

RAID-5

- A RAID-5 system uses block-level stripping with distributed parity over a disk array.
- The disk controller distributes the sequential blocks of data to the physical disks and computes a parity block by bit-wise XOR-ing of the data blocks.
- Example:
  - a1= a2 XOR a3 XOR aP
  - a2= a1 XOR a3 XOR aP
  - a3= a1 XOR a2 XOR aP
  - *aP* = *a*1 XOR *a*2XOR *a*3

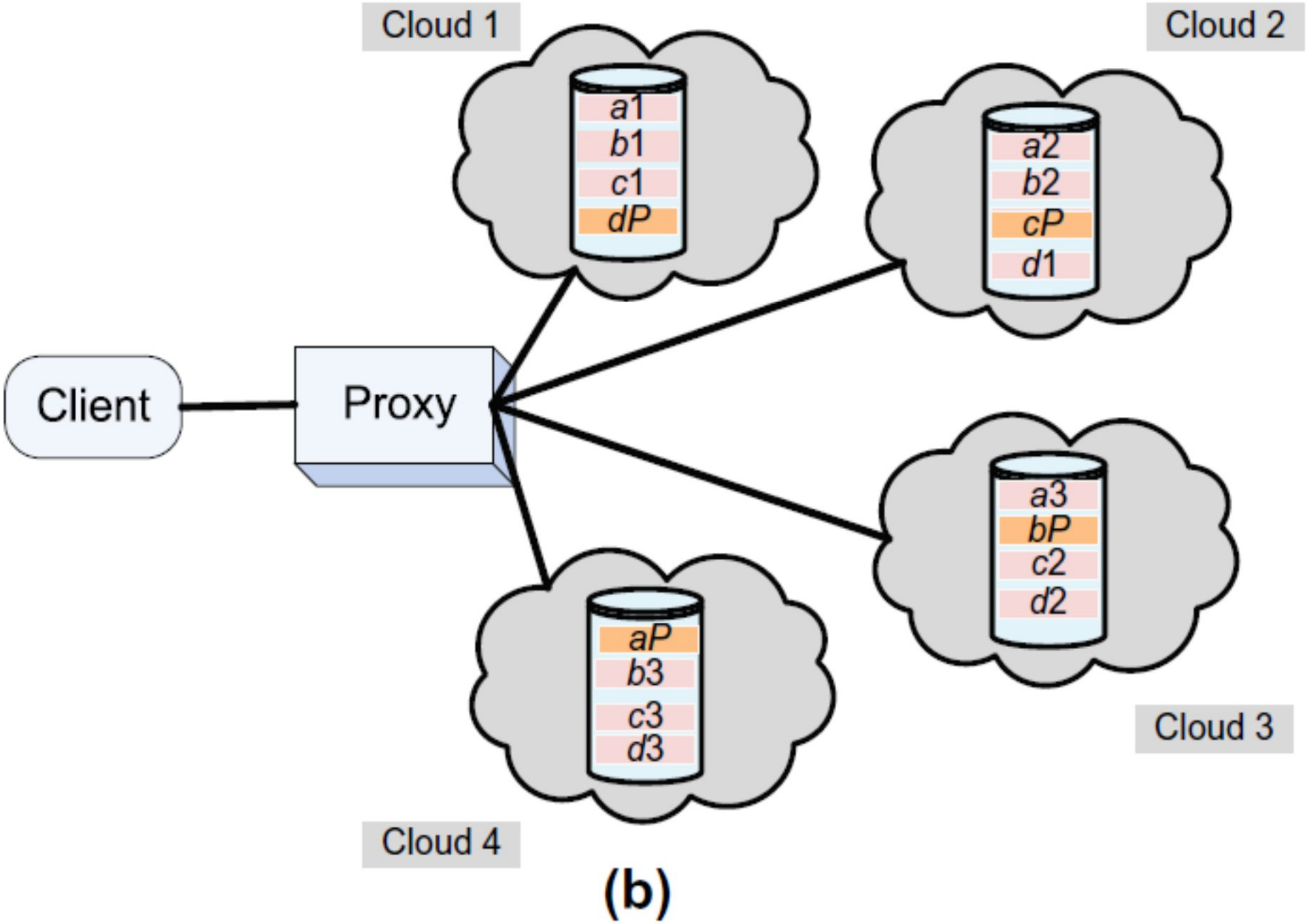
File 1 = {a1 , a2, a3}



(a) A (3, 4) RAID-5 configuration in which individual blocks are stripped over three disks and a parity block is added; the parity block is constructed by XOR-ing the data blocks (e.g., aP = a1XORa2XORa3). The parity blocks are distributed among the four disks: aP is on disk 4, bP on disk 3, cP on disk 2, and dP on disk 1.

RAID-5

- For example, if Disk 2 in Figure is lost, we still have all the blocks of the third file, *c*1,*c*2, and *c*3, and we can recover the missing blocks for the others as follows:
  - *a*2 = *(a*1*)* XOR *(aP)* XOR *(a*3*)*
  - *b*2 = *(b*1*)* XOR *(bP)* XOR *(b*3*)*
  - *d*1 = *(dP)* XOR *(d*2*)* XOR *(d*3*)*



A system strips data across four clouds; the proxy provides transparent access to data.

Cloud interoperability : the Intercloud

- An Intercloud ➔ a federation (“cloud of clouds,”) of clouds that cooperate to provide a better user experience.
- Is an Intercloud feasible?

Not likely at this time:

- There are no standards for either storage or processing.
- The clouds are based on different delivery models.
- The set of services supported by these delivery models is large and open; new services are offered every few months.
- CSPs (Cloud Service Providers) believe that they have a competitive advantage due to the uniqueness of the added value of their services.
- Security is a major concern for cloud users and an Intercloud could only create new threats.

Energy use and ecological impact of large-scale data centers

- The energy consumption of large-scale data centers and their costs for energy and for cooling are significant.
- In 2006, the 6,000 data centers in the U.S consumed 61x10^9 KWh of energy, 1.5% of all electricity consumption, at a cost of $4.5 billion.
- The energy consumed by the data centers was expected to double from 2006 to 2011 and peak instantaneous demand to increase from 7 GW to 12 GW.

Energy use and ecological impact of large-scale data centers

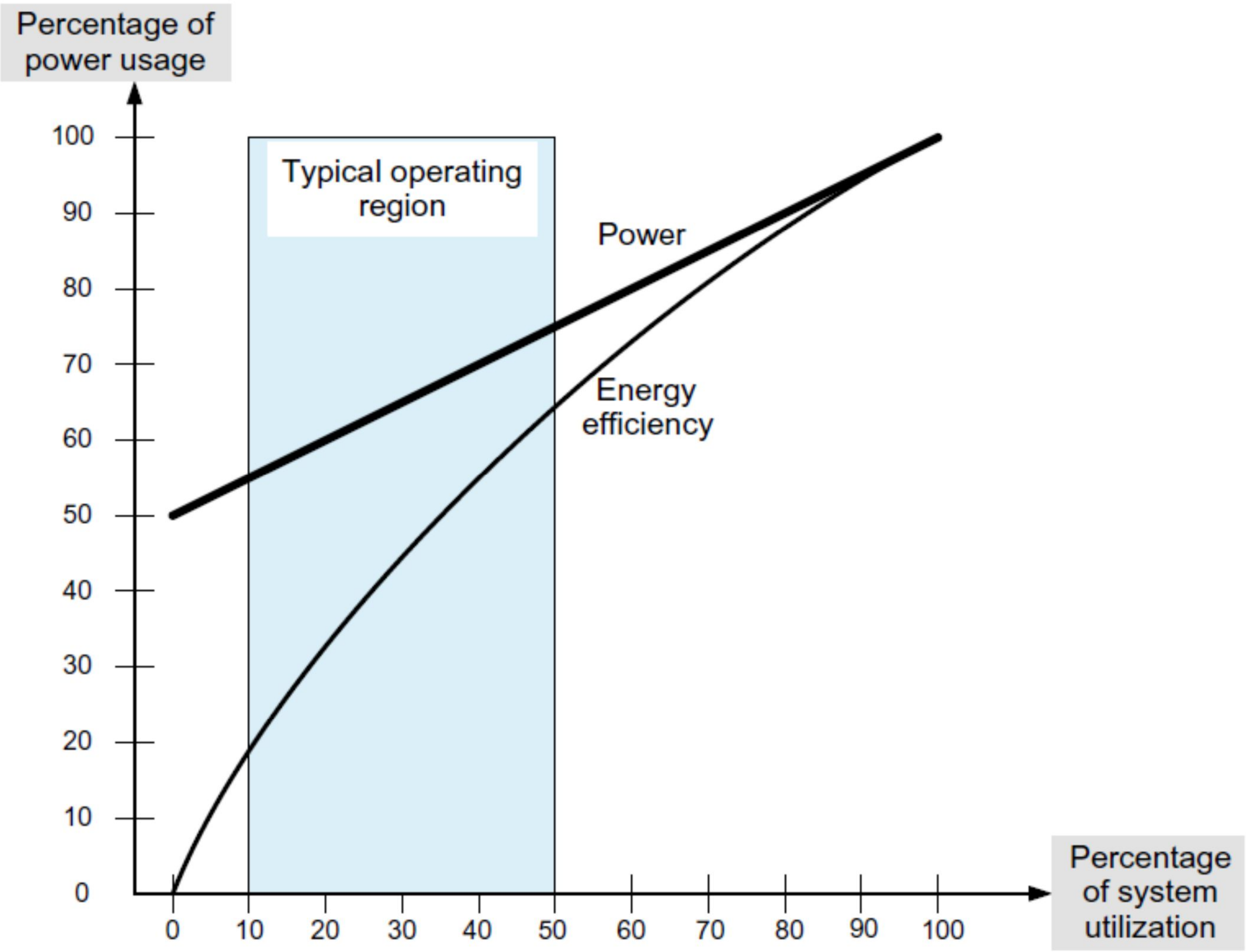
- The greenhouse gas emission due to the data centers is estimated to increase from 116 $x10^9$ tones of CO_2 in 2007 to 257 tones in 2020 due to increased consumer demand.
- The effort to reduce energy use is focused on the computing, networking, and storage activities of a data center.

Energy use and ecological impact of large-scale data centers

- Operating efficiency of a system is captured by the *performance per Watt of power*.
- The performance of supercomputers has increased 3.5 times faster than their operating efficiency – 7,000% versus 2,000% during the period 1998 – 2007.
- A typical Google cluster spends most of its time within the 10-50% CPU utilization range; there is a mismatch between server workload profile and server energy efficiency.

Energy-proportional systems

- An energy-proportional system consumes no power when idle, very little power under a light load and, gradually, more power as the load increases.
- By definition, an ideal energy-proportional system is always operating at 100% efficiency.
- Even when power requirements scale linearly with the load, the energy efficiency of a computing system is not a linear function of the load; even when idle, a system may use 50% of the power corresponding to the full load.


Percentage of power usage
100
90
80
70
60
50
40
30
20
10
0
Typical operating region
Power
Energy efficiency
0
10
20
30
40
50
60
70
80
90
100
Percentage of system utilization

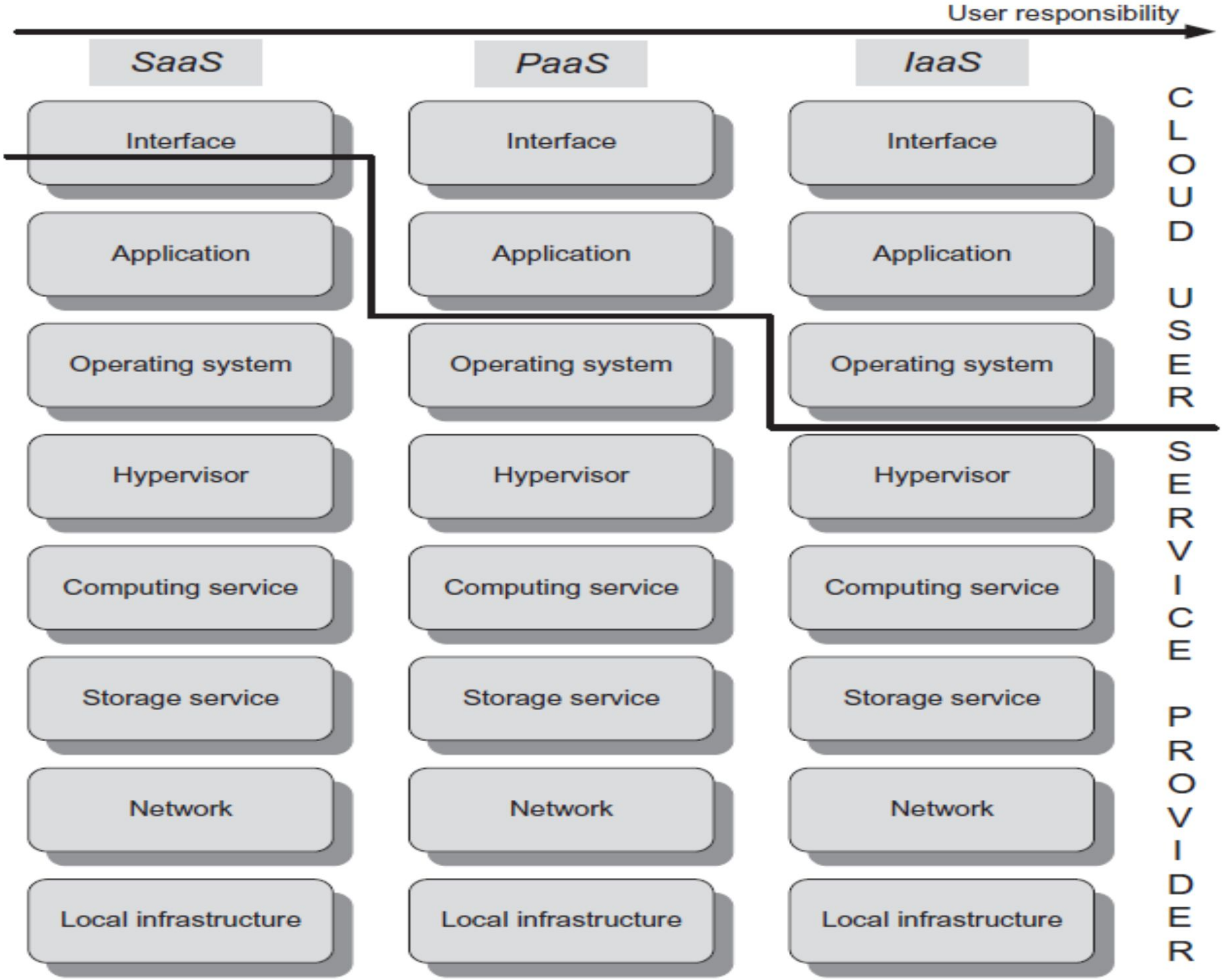
Service Level Agreement (SLA)

- It is a contract between a service provider and the end user that defines the level of service expected from the service provider.
- SLAs are output-based in that their purpose is to define what the customer will receive rather than how the service provider delivers the services.
- Particular aspects of the service – quality, availability, performance, Security / privacy of the data, responsibilities – are agreed between the service provider and the service user.

The objectives of the agreement are:

- Identify and define the customer’s needs and constraints including the level of resources, security, timing, and QoS.
- Provide a framework for understanding; a critical aspect of this framework is a clear definition of classes of service and the costs.
- Simplify complex issues; clarify the boundaries between the responsibilities of clients and CSP in case of failures.
- Reduce areas of conflict.
- Encourage dialog in the event of disputes.
- Eliminate unrealistic expectations.

Responsibility sharing between user and CSP



User experience

- The overall experience of a person using a product such as a website or computer application, especially in terms of how easy or pleasing it is to use.
- User experience (UX) focuses on having a deep understanding of users, what they need, what they value, their abilities, and also their limitations.
- **User Experience** (UX) refers to a person's emotions and attitudes about using a particular product, system or service.

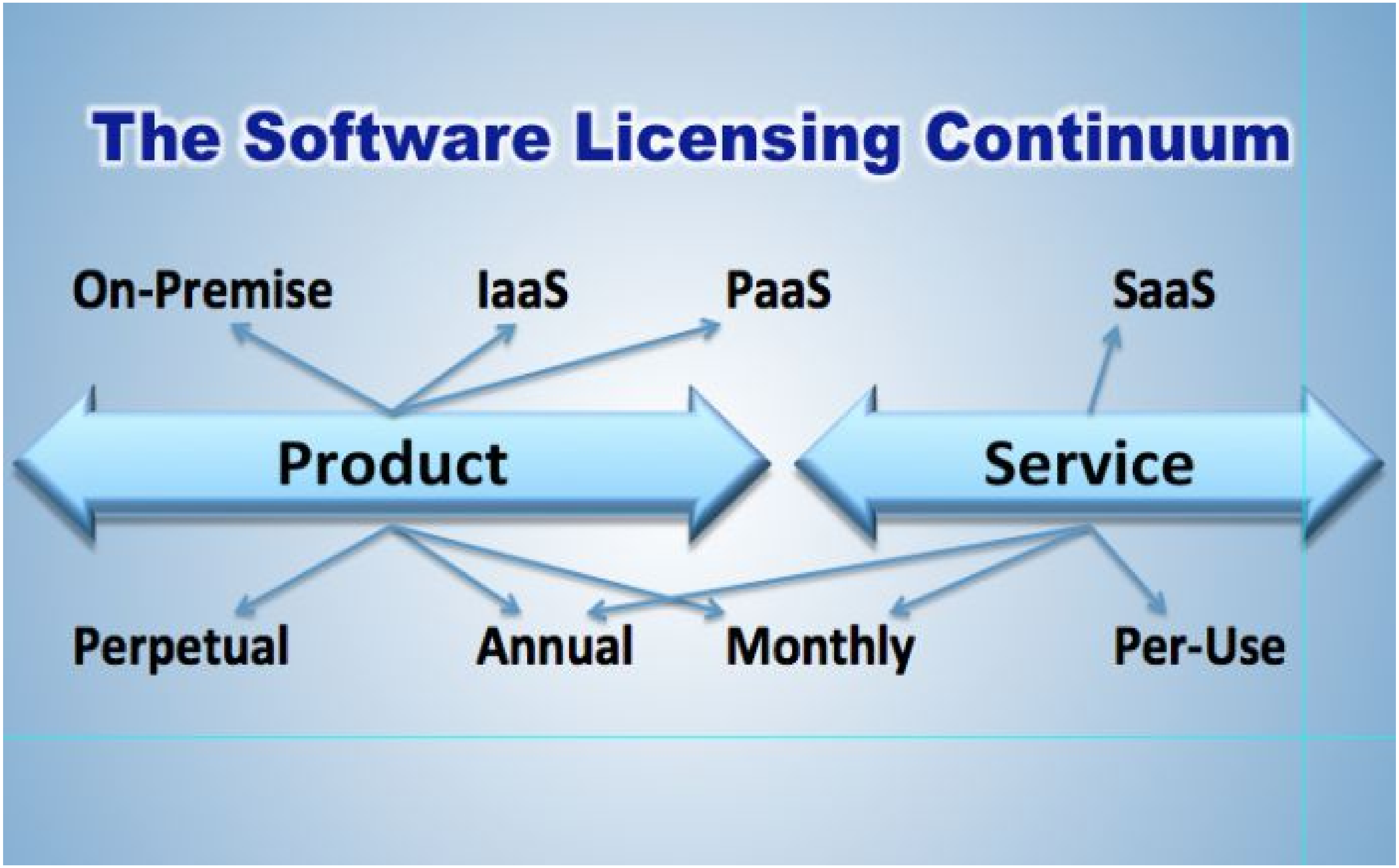
User experience

- The main user concerns are security threats, the dependence on fast Internet connections that forced version updates, data ownership, and user behavior monitoring.
- All users reported that trust in the cloud services is important.
- About half did not fully comprehend the cloud functions and its behavior,

Table 3.6 The reasons driving the decision to use public clouds.

Reason	Respondents Who Agree
Improved system reliability and availability	50%
Pay only for what you use	50%
Hardware savings	47%
Software license savings	46%
Lower labor costs	44%
Lower maintenance costs	42%
Reduced IT support needs	40%
Ability to take advantage of the latest functionality	40%
Less pressure on internal resources	39%
Solve problems related to updating/upgrading	39%
Rapid deployment	39%
Ability to scale up resources to meet needs	39%
Ability to focus on core competencies	38%
Take advantage of the improved economies of scale	37%
Reduced infrastructure management needs	37%
Lower energy costs	29%
Reduced space requirements	26%
Create new revenue streams	23%

Software Licensing

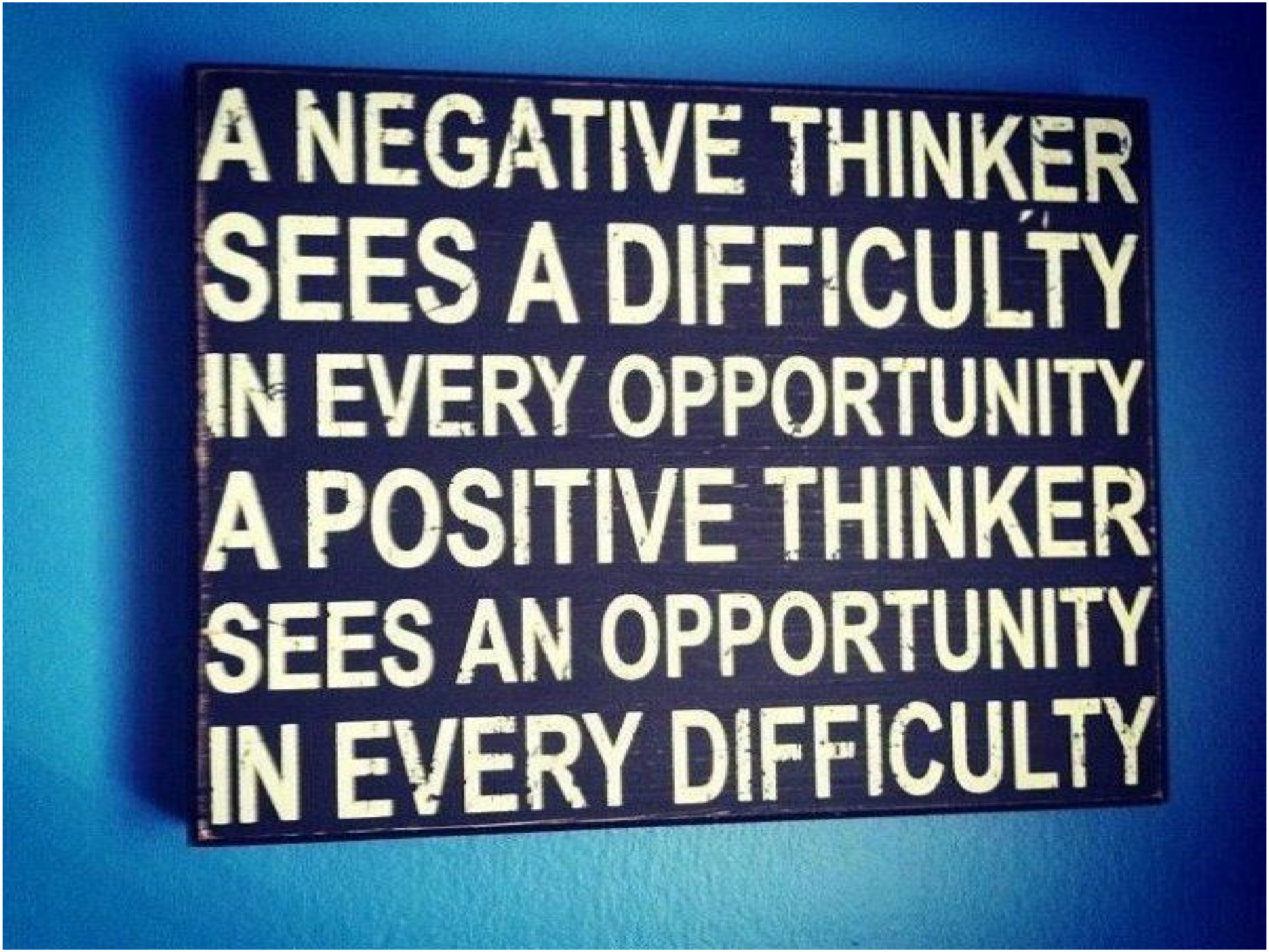


Software Licensing

- Software licensing for cloud computing is an enduring problem without a universally accepted solution at this time.
- The license management technology is based on the old model of computing centers with licenses given on the basis of named users or as site licenses.
- A **site license** is a type of software **license** that allows the user to install a software package in several computers simultaneously, such as at a particular **site** (facility) or across a corporation.

Software Licensing Challenges

- Software licenses based on physical hardware
- Virtual CPU capacity
- Software license tracking



A NEGATIVE THINKER
SEES A DIFFICULTY
IN EVERY OPPORTUNITY
A POSITIVE THINKER
SEES AN OPPORTUNITY
IN EVERY DIFFICULTY