CLOUD COMPUTING
Introduction

Zeinab Zali
Isfahan University Of Technology

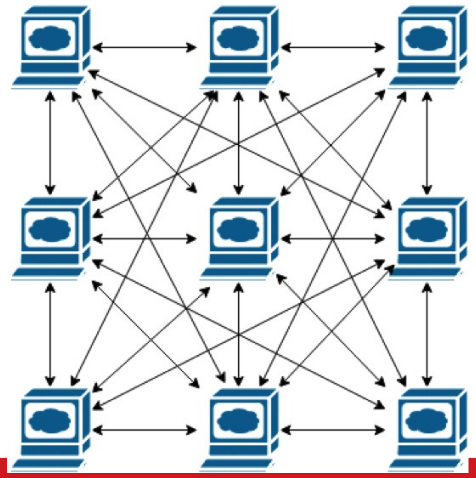
Syllabus and Plan

- https://docs.google.com/document/d/1qVa405gM2bwoFBN6XnZCphmmIzs0BzTp93Ew_XEAGWY/edit?usp=sharing

What is your expectations from this course?

What is a cloud?

- A cluster?
- A supercomputer?
- A data store?
- A Distributed system?
- None of the above?
- All of the above?

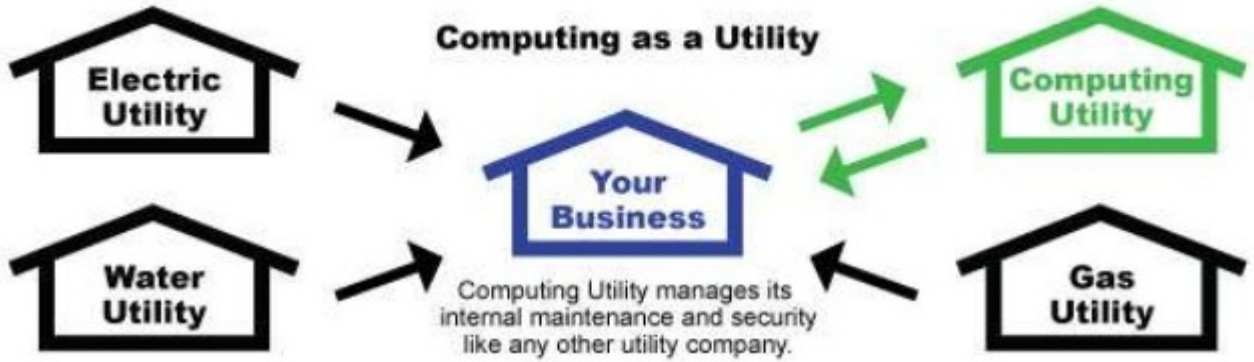


Computing utility

- Conceptually, computing can be viewed as another utility, like electricity, water, or gas, accessible to every household
- Computer clouds are the utilities providing computing services
  - the hardware and the software resources are concentrated in large data centers.
  - The users of computing services pay as they consume computing, storage, and communication resources

Computing utility

- Plug your thin client into the computing utility and play your favorite Intensive Compute & Communicate Application

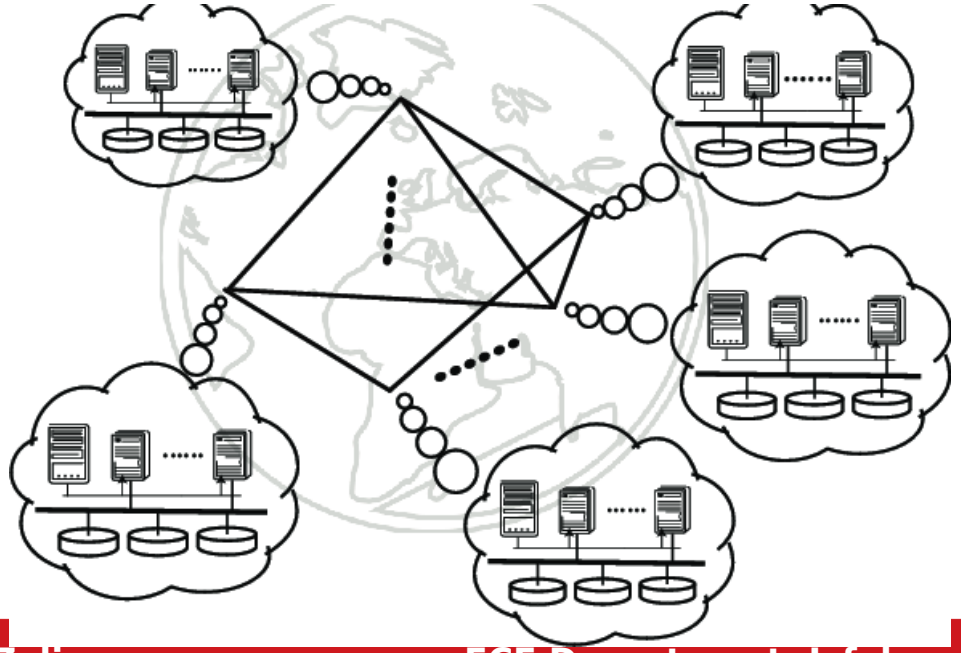


So what is a cloud?

- Clouds = lots of storage + computing cycles nearby
- A single-site cloud:
  - Compute nodes (grouped into racks)
  - Switches, connecting the racks
  - A network topology, e.g., hierarchical
  - Storage (back-end) nodes connected to the network
  - Front-end for submitting jobs and receiving client requests Software services

So what is a cloud?

- A geographically distributed cloud consists of
  - Multiple such sites
  - Each site perhaps with a different structure and services



Cloud vs Distributed Systems

- Distributed computing is when multiple autonomous machines communicate through a central network to perform a common goal.
- Cloud Computing is more about computing as a service, that is given to a computer over a network
- The cloud computing provides hardware, software and other infrastructure resources over the internet while the distributed computing divides a single task among multiple computers that are connected via a network

New Features in today clouds

- Massive scale.
- On-demand access: Pay-as-you-go, no upfront commitment
- Data-intensive Nature: What was MBs has now become Tbs, PBs and XBs.
  - Daily logs, Web data, etc.
- New Cloud Programming Paradigms: MapReduce/Hadoop, and many others.
  - High in accessibility and ease of programmability
  - Lots of open-source

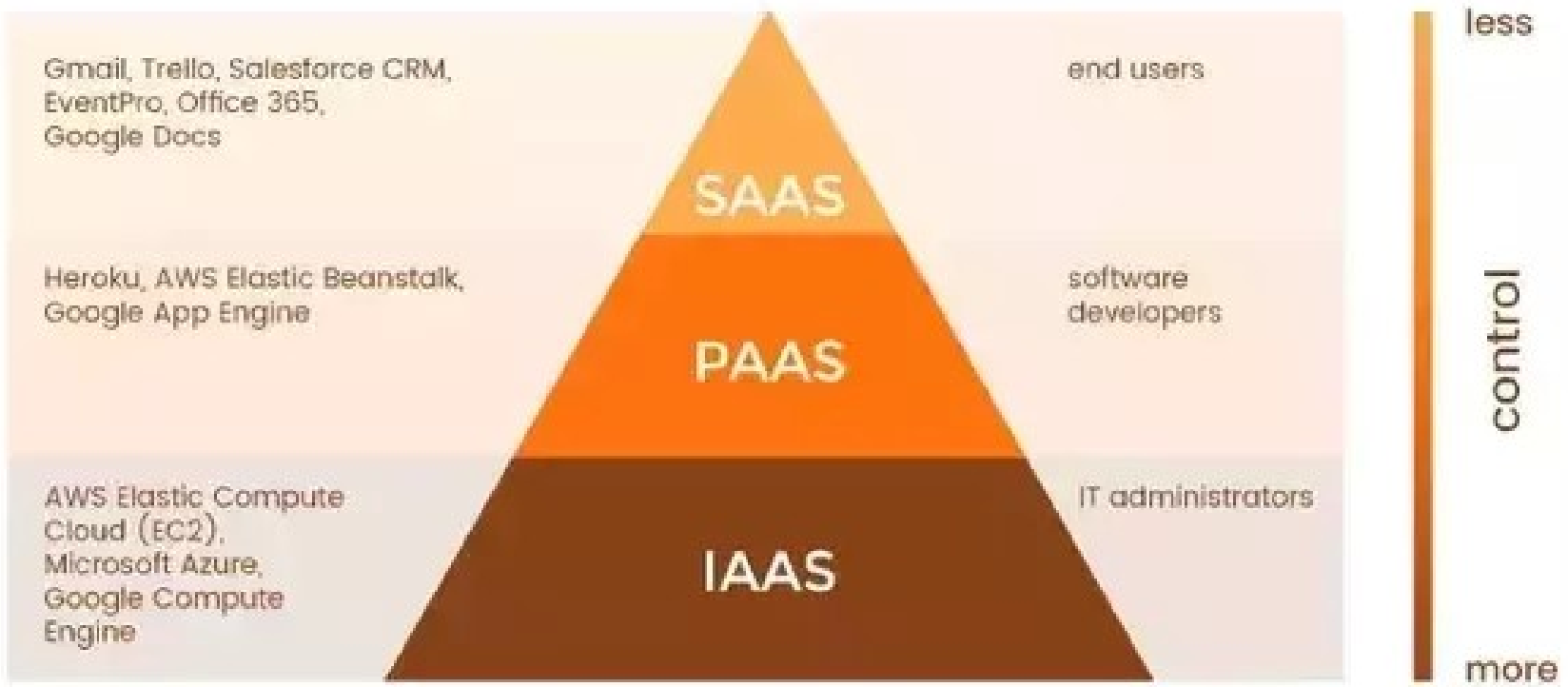
Massive scale Data Centers

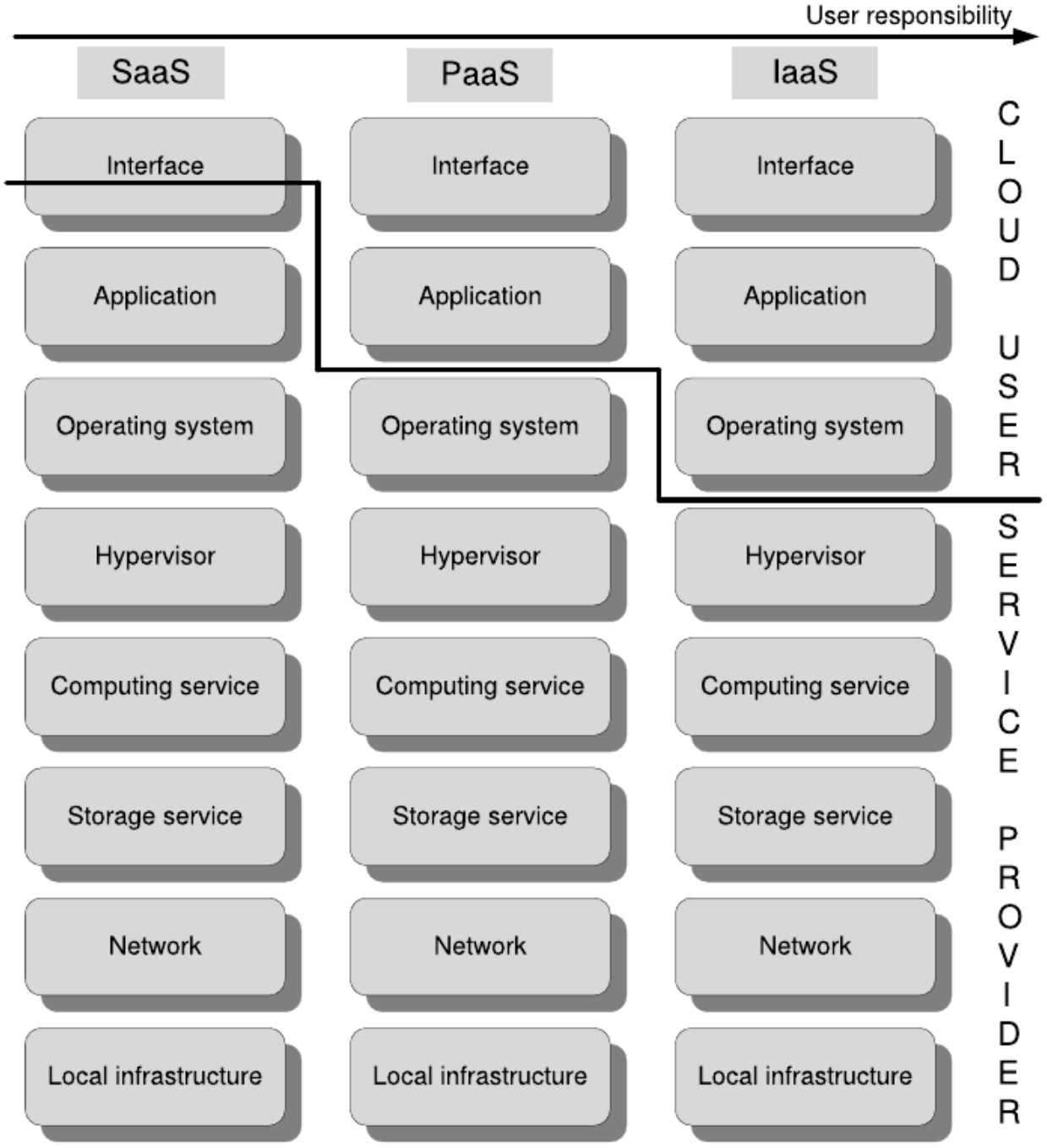
- Facebook
- Yahoo
- AWS EC2 [Randy Bias, 2009]
- eBay [2012]: 50K machines
- HP [2012]: 380K in 180 DCs
- Google: A lot

On-demand access

- renting a cab vs. (previously) renting a car, or buying one. Ex.:
  - Iaas (Infrastructure as a Service): You get access to flexible computing and storage infrastructure. Virtualization is one way of achieving this, Ex: AWS: EC2, S3
  - PaaS (Platform as a Service): You get access to flexible computing and storage infrastructure, coupled with a software platform, Ex: Google AppEngine, Microsoft Azure
  - SaaS (Software as a Service): You get access to software services, when you need them, Ex: Google docs

On-demand access




User responsibility
SaaS
PaaS
IaaS
Interface
Application
Operating system
Hypervisor
Computing service
Storage service
Network
Local infrastructure
CLOUD USER
SERVICE PROVIDER

Data-intensive Nature

- Computation-Intensive Computing Example areas:
  - MPI-based, high performance computing, grids, Typically run on supercomputers
- Data-Intensive
  - Typically store data at datacenters, use compute nodes nearby
  - the focus shifts from computation to the data:
    - CPU utilization no longer the most important resource metric, instead I/O is (disk and/or network)

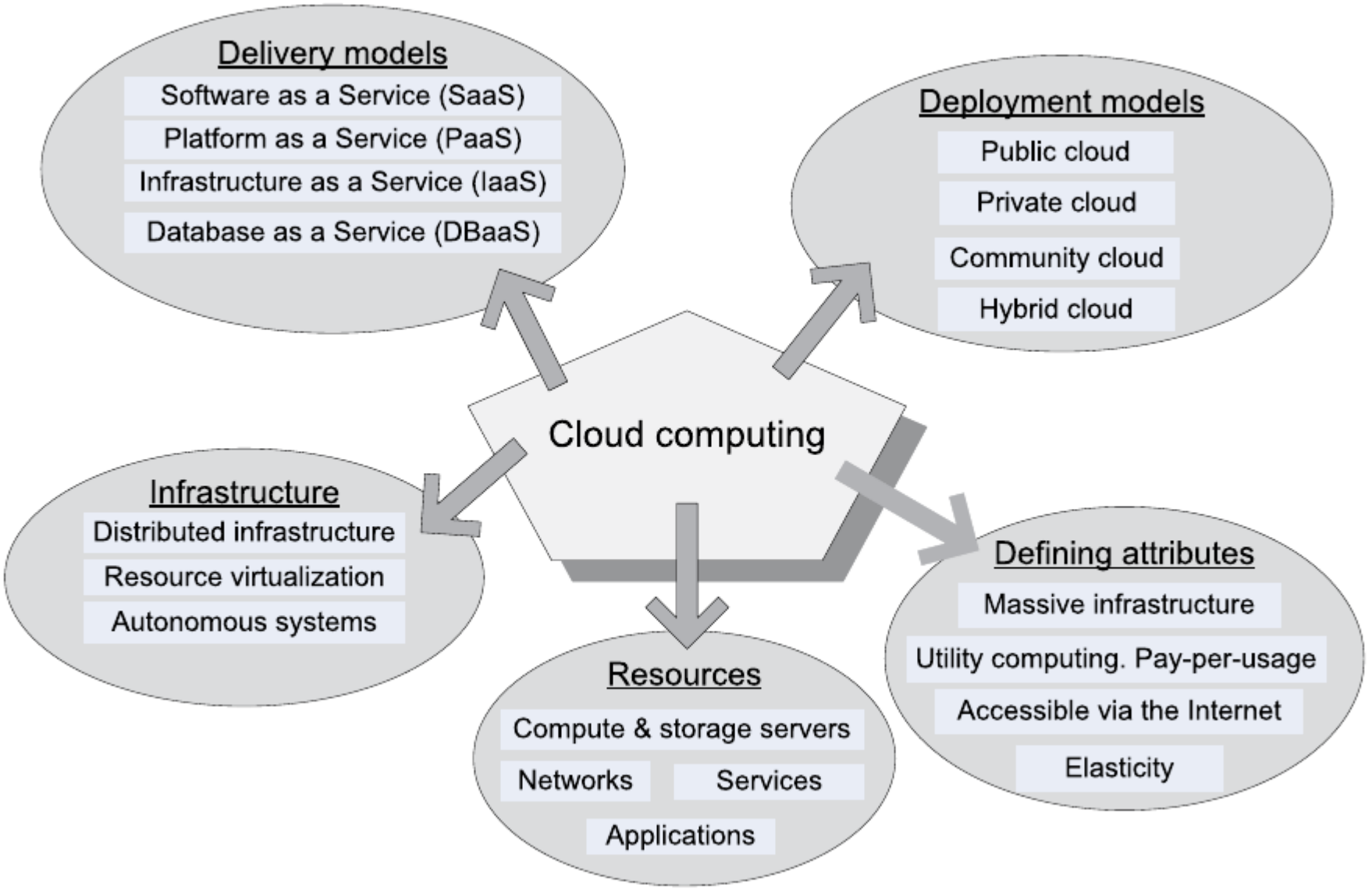
New Cloud Programming Paradigms

- Easy to write and run highly parallel programs in new cloud programming paradigms:
  - Google: Hadoop + MapReduce
  - Amazon: Elastic MapReduce service
  - Yahoo! (Hadoop + Pig)
  - Facebook (Hadoop + Hive)
  - NoSQL: MySQL is an industry standard, but Cassandra is 2400 times faster!

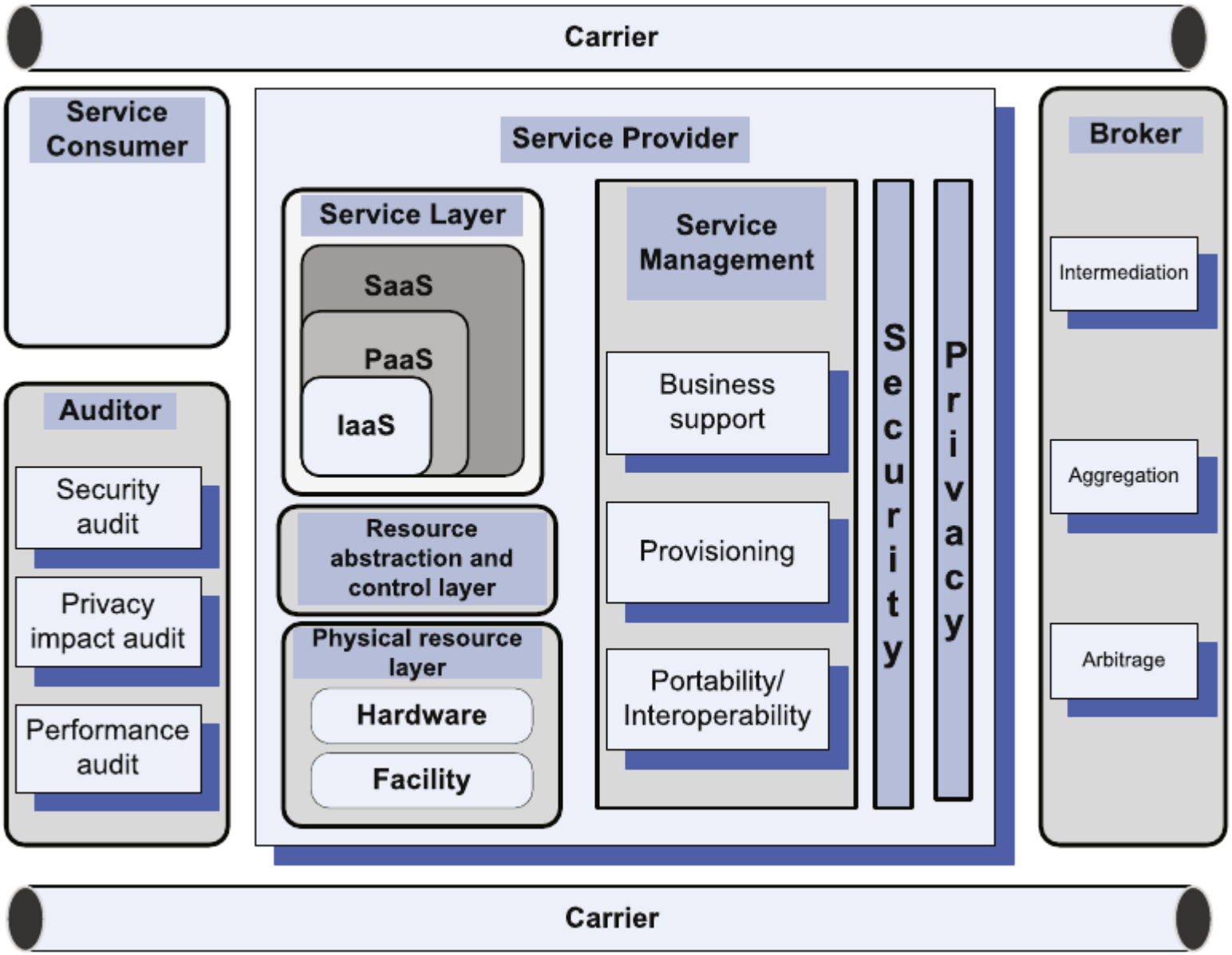
Cloud applications

- Data analytics
- data mining
- computational financing
- scientific and engineering applications
- gaming and social networking
- other computational and data-intensive activities

Cloud models and attributes



Cloud entities (I)



Cloud entities (II)

- Service consumer: entity that maintains a business relationship with, and uses service from service providers;
- Service provider: entity responsible for making a service available to service consumers;
- Carrier: the intermediary that provides connectivity and transport of cloud services between providers and consumers;
- Broker: an entity that manages the use, performance and delivery of cloud services, and negotiates relationships between providers and consumers

Cloud entities (III)

- Auditor – a party that can conduct independent assessment of cloud services, information system operations, performance and security of the cloud implementation.
  - An audit is a systematic evaluation of a cloud system by measuring how well it conforms to a set of established criteria.
  - For example, a security audit evaluates cloud security, a privacy-impact audit evaluates cloud privacy assurance, while a performance audit evaluates cloud performance.

Software as a Service

- The SaaS cloud infrastructure only runs applications developed by the service provider.
- ✔ A wide range of stationary and mobile devices allow a large population of clients to access the services provided by these applications using a thin client interface such as a web browser (e.g., web-based email).
- ✗ The users of services do not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities

Platform as a Service

- PaaS offers the capability to deploy consumer-created or acquired applications using programming languages and tools supported by the provider.
- ✔ The user has control over the deployed applications and, possibly, application hosting environment configurations
- ✗ The user does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage.

Infrastructure as a Service

- IaaS has the capability to provision processing, storage, networks, and other fundamental computing resources
- The consumer is able to deploy and run arbitrary software, which can include operating systems and applications
- The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of some networking components, e.g., host firewalls

IaaS applications and Services

- server hosting
- web servers
- Storage
- computing hardware
- Operating systems
- virtual instances
- Internet access
- bandwidth provisioning

Iaas and Elastic computing

- The term “elastic computing” refers to the ability of dynamically acquiring computing resources and supporting a variable workload
- It supports dynamic scaling of IaaS
  - it is based on a utility pricing model and variable cost
- IaaS cloud computing model is particularly useful when the demand is volatile and a new business needs computing resources and it does not want to invest in a computing infrastructure or when an organization is expanding rapidly

Two Cloud categories

- Private: are accessible only to company employees
- Public: provides service to any paying customer
  - Amazon S3: simple storage service
  - Amazon EC2:Elastic compute cloud
  - Google App engine/compute engine

Top 2020 cloud providers

- AWS
- Microsoft Azure
- Google Cloud

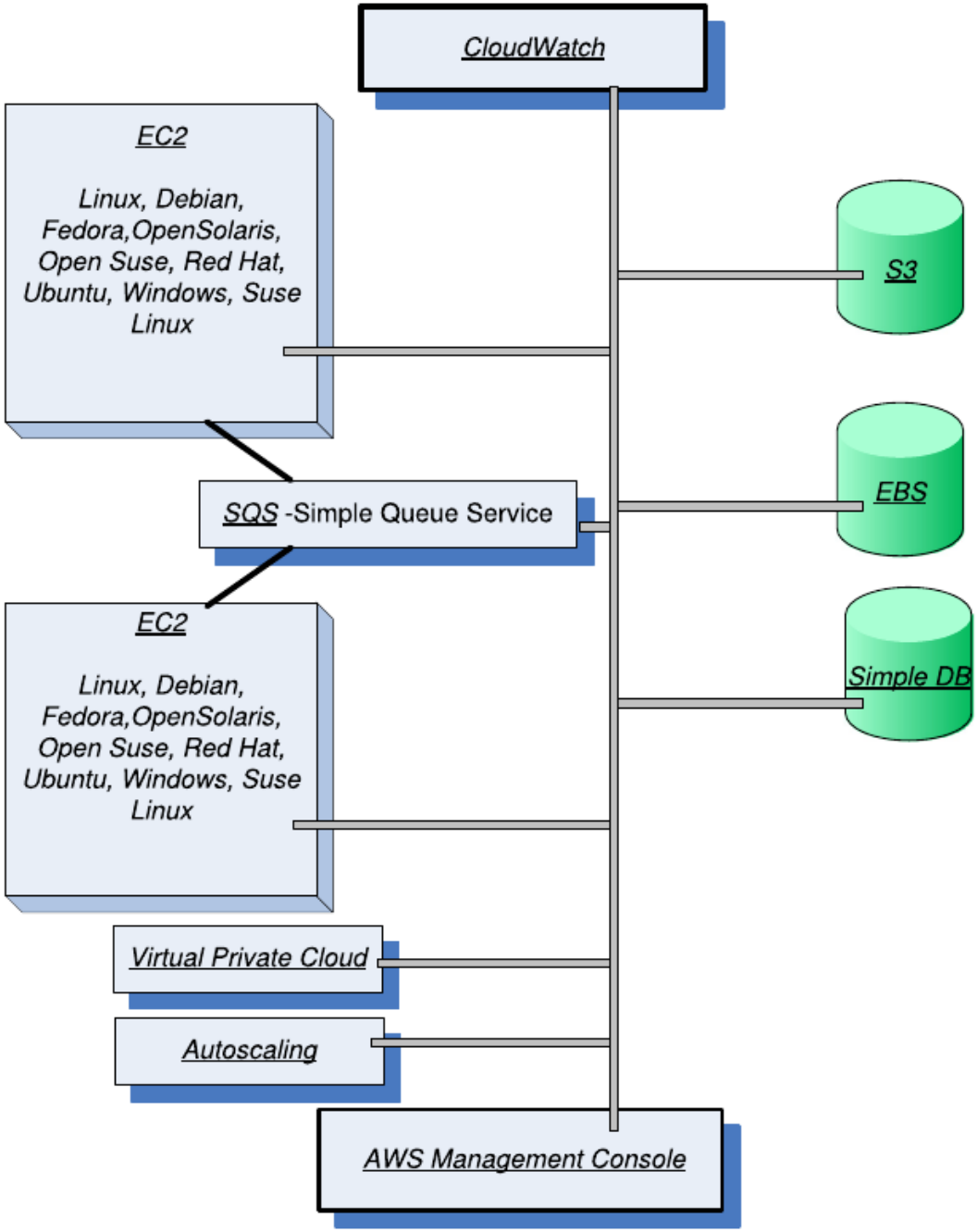
Amazon Web Service

Amazon Web Service

- Amazon first installed a powerful computing infrastructure to sustain its core business
  - selling online a variety of goods ranging from books and CDs to gourmet foods and home appliances.
- Then the company discovered that this infrastructure can be further extended to provide affordable and easy to use resources for enterprise computing, as well as computing for the masses.

AWS

- In mid 2006 Amazon introduced AWS based on the IaaS delivery model
- Amazon was the first provider of cloud computing



AWS regions

The AWS Cloud spans 69 Availability Zones
within 22 geographic regions around the world

AWS Elastic Compute Cloud (EC2)

- A web service with a simple interface for launching instances of an application under several operating systems
- An instance is a virtual server
  - the user chooses the region and the availability zone where this virtual server should be placed
  - The user selects from a limited menu of instance types the one which provides the resources needed by the application.
    - CPU cycles, main memory, secondary storage, communication and I/O bandwidth

EC2 Instances (I)

- 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

EC2 Instances (II)

- Instance creation in EC2 is based on the Xen virtualization strategy
- The instances can be placed in multiple locations in different Regions and Availability Zones.
- EC2 distributes automatically the incoming application traffic among multiple instances using the elastic load balancing facility

types of EC2 instances

- AWS offers several types of EC2 instances targeting different classes of applications:
  - T2: provide a baseline CPU performance and the ability to exceed the baseline.
  - M3 & M4: provide a balance of compute, memory, and network resources.
  - C4: use high performance processors and have the lowest price/compute.
  - R3: are optimized for memory-intensive applications.
  - G2: target graphics and general-purpose GPU applications.
  - I2, D2: are storage optimized and deliver high disk throughput.

Instance IP address

- An instance, when launched, is provided with a DNS name; this name maps to
  - A private IP address for internal communication within the internal EC2 communication network
  - A public IP address for communication outside the internal Amazon network
- Network Address Translation (NAT) maps external IP addresses to internal ones

AWS Storage Services

- Simple Storage System (S3) is a storage service designed to store large objects.
  - S3 supports a minimal set of functions: PUT, GET, and DELETE
- Elastic Block Store (EBS) provides persistent block level storage volumes for use with EC2 instances
  - A volume appears to an application as a raw, unformatted and reliable physical disk (1GB to 1TB)

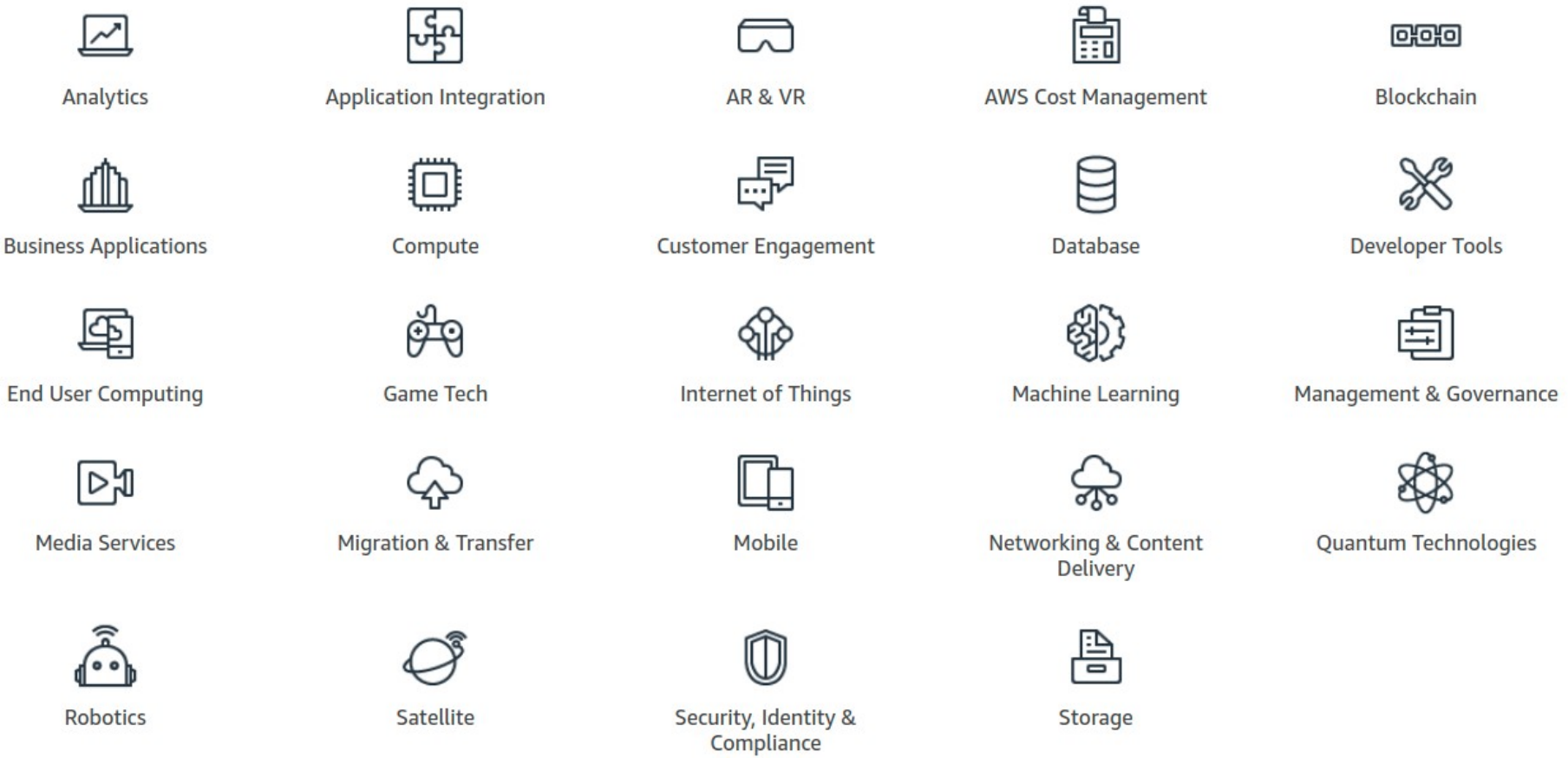
Other AWS services

- Simple DB: a non-relational data store that allows developers to store and query data items via web services requests
- Simple Queue Service: it allows multiple EC2 instances to coordinate their activities by sending and receiving SQS messages
- CloudWatch: monitoring infrastructure used by application developers, users, and system administrators to collect and track metrics important for optimizing the performance of applications

Other AWS services

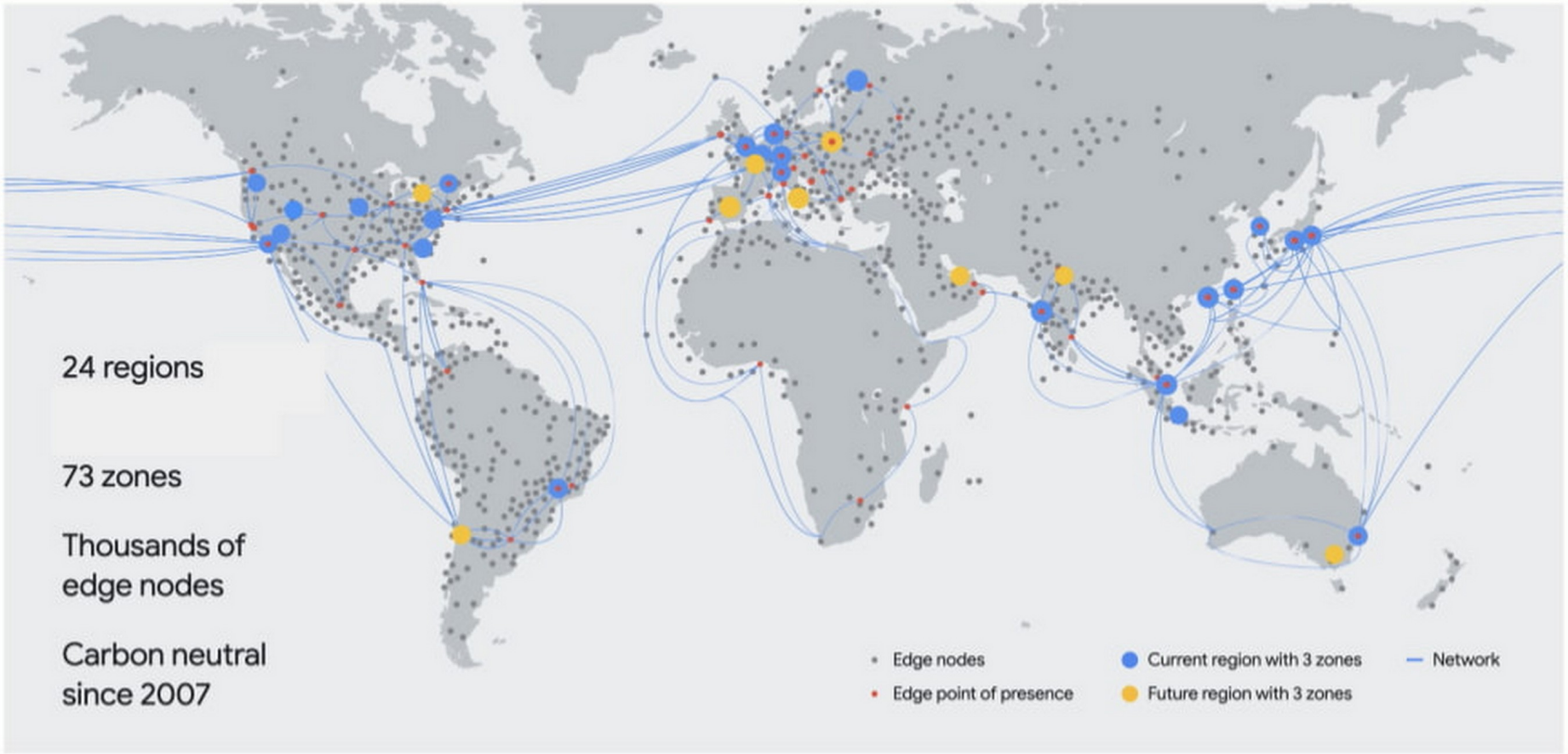
- Elastic MapReduce (EMR): a service supporting processing of large amounts of data using a hosted Hadoop running on EC2 and based on the MapReduce paradigm
- ElastiCache: a service enabling web applications to retrieve data from a managed in-memory caching system rather than a much slower disk-based database.
- Elastic Load Balancer: a cloud service to automatically distribute the incoming requests across multiple instances of the application

Other AWS services



Google Clouds

Google Cloud Platform (GCP)



Google Services (I)

- AppEngine: Google is a leader in the Platform-as-a-Service (PaaS) space
  - AppEngine (AE) is an infrastructure for building web and mobile applications and run these application on Google servers.
  - Initially, it supported only Python and support for Java was later added
- Compute Engine (CE): It supports the creation of VMs with configurations range
  - From micro instances to the ones with 32 vCPUs or 208 GB of memory.
  - Up to 64 TB of network storage

Google Services (II)

- Container Engine (CntE): a cluster manager and orchestration system for Docker containers built on the Kubernetes system
- Cloud Functions (CF): a lightweight, event-based, asynchronous system to create single-purpose functions that respond to cloud events.
  - CFs are written in Javascript and execute in a Node.js runtime environment.
- Cloud Load Balancing supports scalable load balancing on the Google Cloud Platform.

Google Services (III)

- Cloud Storage
- Big data services: Cloud SQL, Cloud Bigtable, Cloud datastore, Cloud Datalab
- Cloud Dataproc: Spark and Hadoop
- Cloud dataflow
- Cloud Pub/Sub
- Cloud Virtual Network (CVN)
- Cloud CDN is a low-latency, low-cost content delivery network.

Google Services (IV)

- Cloud Machine Learning is a managed service based on the TensorFlow model to build machine learning models, that work on any type of data.
- Cloud Natural Language API is a text analysis tool that can be used to extract information about people, places, events, and so on.
- Cloud Speech API allows developers to convert audio to text

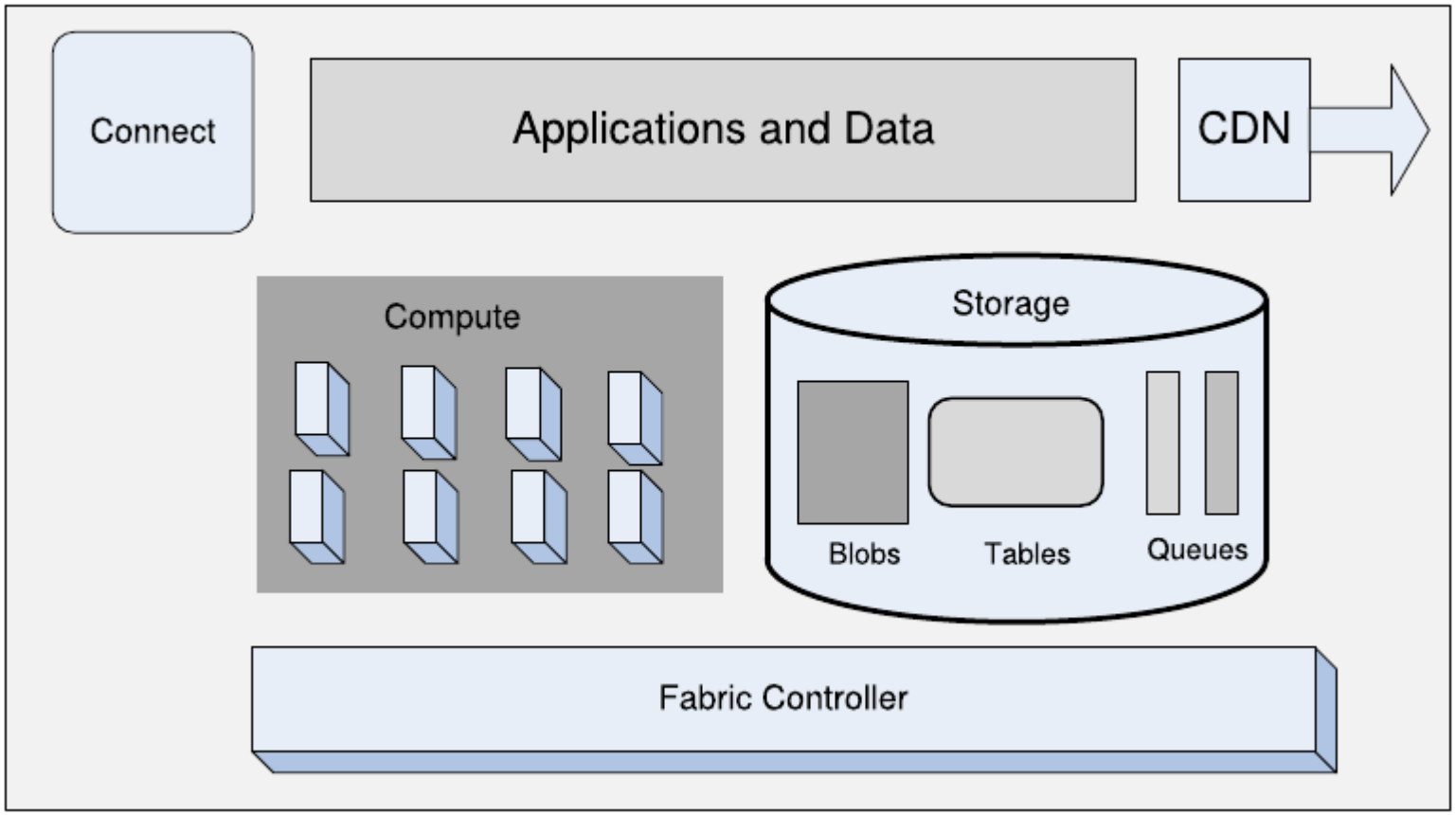
Google Services (V)

- Gmail
- Google Doc
- Google Calendar
- Google Drive

Microsoft Azure

Windows Azure components

- The API interface to Windows Azure is built on REST, HTTP and XML



Windows Azure components

- Compute: provides a computation environment
- Storage: It is for scalable storage
- Fabric Controller: deploys, manages, and monitors applications; it interconnects nodes consisting of servers, high-speed connections, and switches.
- Content Delivery Network (CDN) maintains cache copies of data to speedup computations.
- Connect subsystem supports IP connections between the users and their applications running on Windows Azure.

Windows Azure Services (I)

- The platform includes five services: Live Services, SQL Azure, AppFabric, SharePoint, and Dynamics CRM.
- A client library and tools are provided for developing cloud applications in Visual Studio
- multiple instances of a role for applications:
  - Web role instances used to create web applications;
  - Worker role instances used to run Window-based code;
  - VM role instances running user-provided Windows Server 2008 R2 images.

Windows Azure Services (II)

- 2017 book says: The Microsoft Azure platform currently does not provide or support any distributed parallel computing frameworks, such as MapReduce, Dryad or MPI, other than the support for implementing basic queue-based job scheduling
- 2018, the microsoft page says: Azure Batch is to run large-scale parallel and high-performance computing (HPC) batch jobs efficiently in Azure

Other Cloud concepts and issues

SERVICE-LEVEL AGREEMENTS

- A Service Level Agreement (SLA) is a negotiated contract between two parties, the customer and the service provider, some objectives are:
  - Identify and define the customer's needs and constraints including the level of resources, security, timing, and quality of service.
  - Simplify complex issues; for example, clarify the boundaries between the responsibilities of the clients and those of the provider of service in case of failures.

ENERGY USE AND ECOLOGICAL IMPACT OF CLOUD COMPUTIN

- The energy consumption of large-scale data centers and their costs for energy used for computing and networking and for cooling are significant now and are expected to increase substantially in the future
- In recent years the contribution of solar, wind, geothermal and other renewable energy sources has steadily increased (instead of fossil fuels such as coal and gas)
- Green Cloud: it involves the designing, engineering, and usage of computing devices in such a way that it can reduce any adverse impact on the environmental impact in the IT industry

Cloud Issues

- some of the risks related to relinquishing the control to third party services:
  - Unauthorized access
  - data corruption
  - infrastructure failure
  - service unavailability
- whenever a problem occurs it is difficult to identify the source and the entity causing it

Major challenges

- The challenges of parallel and distributed computing
  - Architecture and foundations
  - Concurrency
  - Load balancing
- resource sharing and resource virtualization
- Security
- Trust and privacy
- interoperability and standardization
  - a user should not be tied to a particular cloud service provider