INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

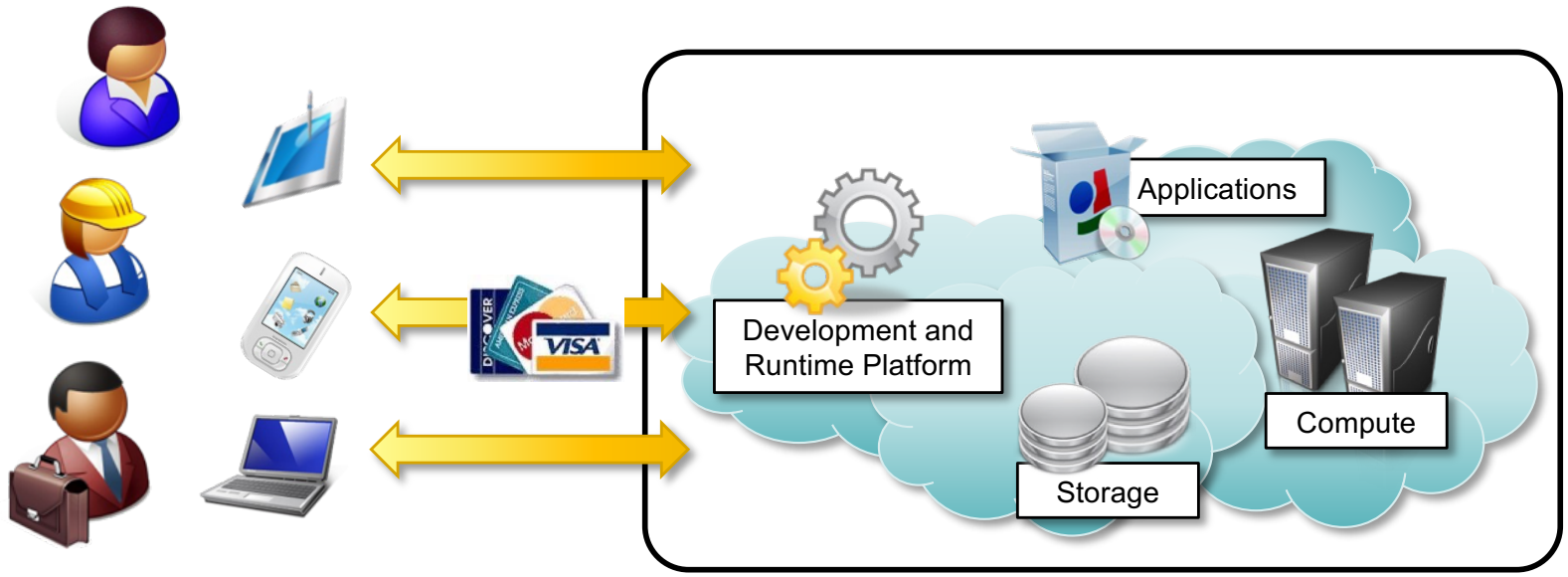
IaaS Clouds

Nikos Parlavantzas

Outline

- IaaS clouds
- Case study: Amazon Web Services

What is the cloud?



Everything as a Service

Essential characteristics

- On-demand self-service

- Broad network access

- Resource pooling

- Elasticity

- Metered service

Service models

IaaS Clouds

Infrastructure as a Service

- The provider delivers raw computing resources (typically virtualised)
  - Servers, storage, networking, …
- Consumers use these resources to deploy and run arbitrary software, including operating systems and applications

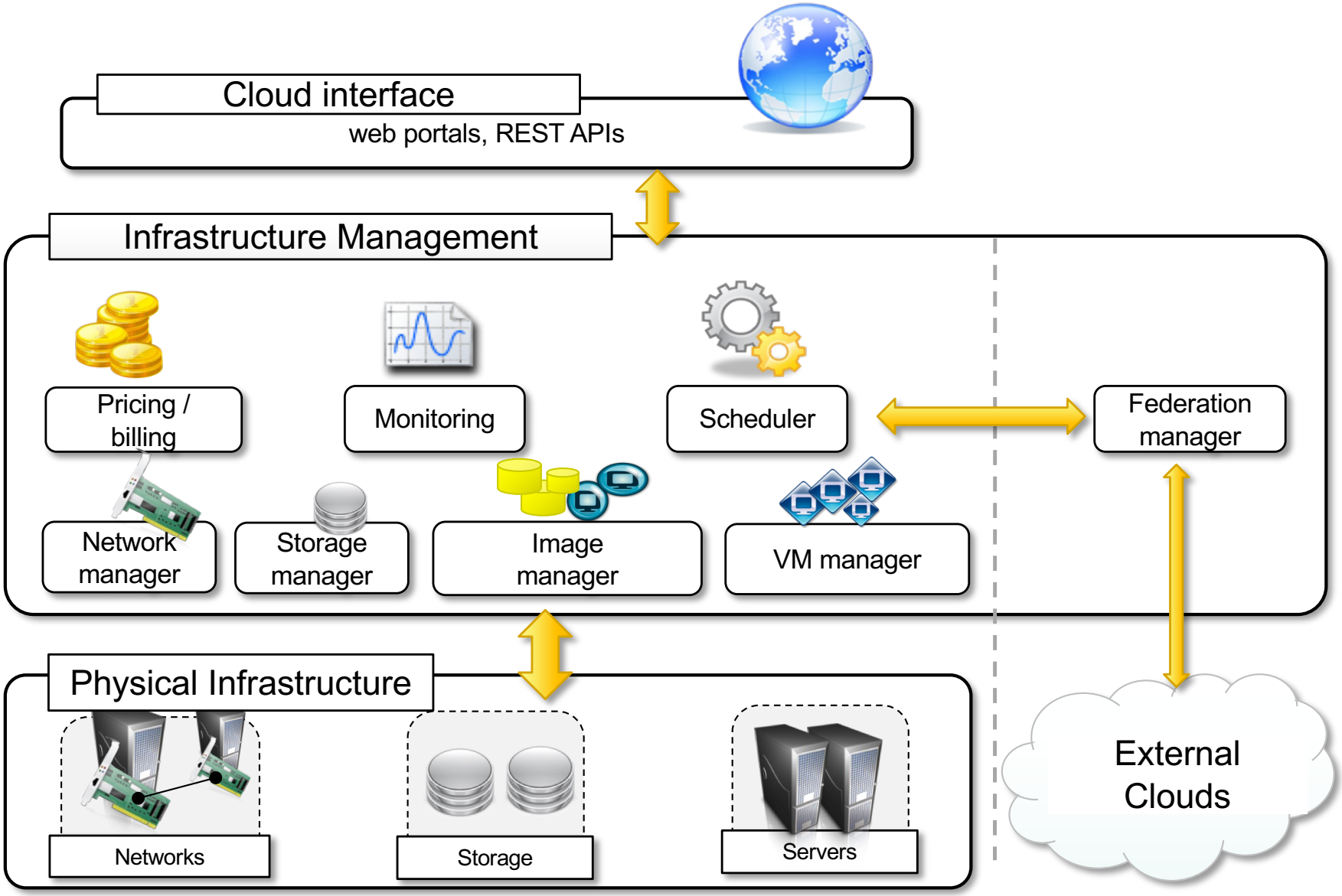
Common IaaS features

- Multiple types of VMs with different amounts of resources (e.g., virtual CPU, RAM, storage, network)
- Multiple storage options (e.g., block storage, object storage)
- Multiple geographical locations
- Load balancing, auto scaling, monitoring
- Virtual networks, content delivery networks
- Container support

Google Cloud Platform

Windows Azure

IaaS architecture



Physical Infrastructure

Physical infrastructure

- 10s-100s of thousands of servers

Data center costs

Amortized Cost*	Component	Sub-Components
	Servers	CPU, memory, disk
	Network	Switches, links, transit
	Infrastructure	UPS, cooling, generators
	Power draw	Electrical utility costs

*3 yr amortization for servers, 15 yr for infrastructure

The Cost of a Cloud: Research Problems in Data Center Networks. Sigcomm CCR 2009. Greenberg, Hamilton, Maltz, Patel.

Data center costs

Amortized Cost*	Component	Sub-Components
~45%	Servers	CPU, memory, disk
~15%	Network	Switches, links, transit
~25%	Infrastructure	UPS, cooling, generators
~15%	Power draw	Electrical utility costs

*3 yr amortization for servers, 15 yr for infrastructure

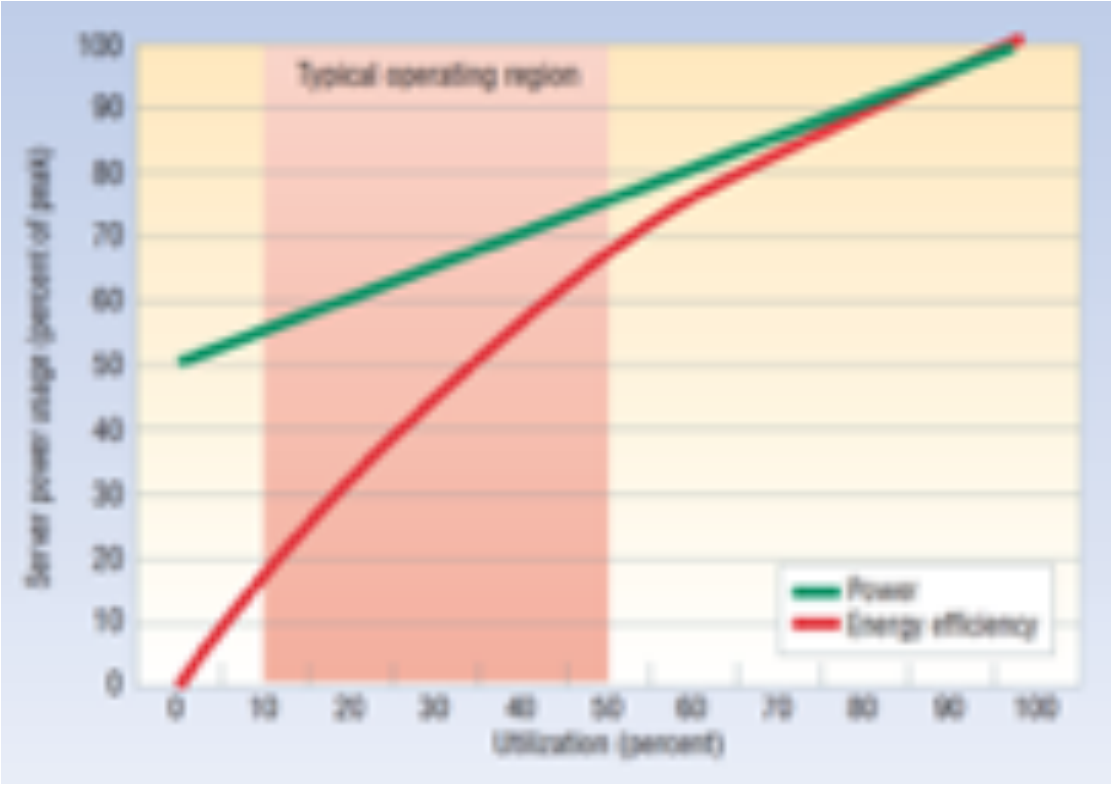
The Cost of a Cloud: Research Problems in Data Center Networks. Sigcomm CCR 2009. Greenberg, Hamilton, Maltz, Patel.

Data center challenges

- Increase server utilisation
  - Provide economic incentives to modulate consumption
    - e.g., dynamic pricing
  - Allow fine-grained resource allocation

Data center challenges

- Support energy proportionality



Barroso, L. A.; Hölzle, U. "The Case for Energy-Proportional Computing". Computer. 40 (12): 33–37

Data center challenges

- Increase reliability in an economical way
  - e.g., distributing state across data centers and allowing data centers to fail

Data center challenges

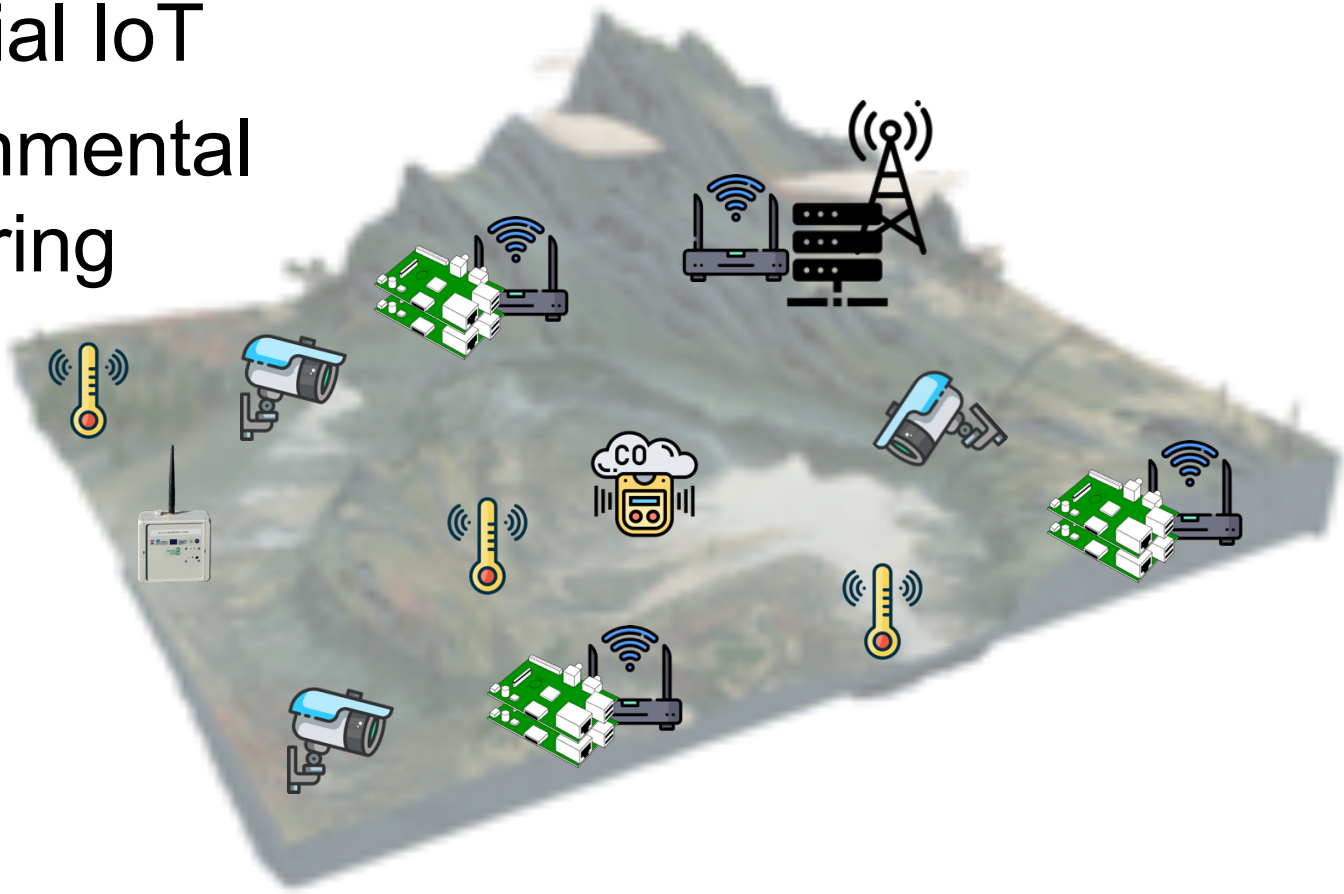
- Provide lower latency to end users
  - e.g., placing data centers close to users
  - cf., emergence of *Fog/Edge Computing*

Fog computing

- Extension of the traditional cloud computing model in which compute, storage, and network capabilities are distributed closer to users
- Drivers
  - Latency
  - Bandwidth
  - Privacy/security
  - Connectivity

Fog computing

- Use cases
  - Smart cities
  - Connected cars
  - Industrial IoT
  - Environmental monitoring



Fog computing

- Challenges
  - Resource heterogeneity
  - Workload dynamicity
  - Data management
  - Programming models
  - Economic models

Case study:
Amazon Web Services

Amazon AWS

- Grew out of Amazon's need to provision machines for its own business
- 2006 – S3 available in spring; EC2 in autumn
- 2008 – Elastic Block Store available
- 2009 – Relational Database Service
- 2012 – DynamoDB
- 2021 - **$17 billion** in profit (**74%** of Amazon operating profits)

Data centers

In Europe:
Frankfurt (3)
Ireland (3)
London (3)
Paris (3)
Stockholm (3)
Milan (3)
Zurich (3)
Spain(3)

Region (3 Availability Zones)

Services

- 200+ services accessed over the Internet
  - HTTP-based API
  - Command-line interface
  - Web-based user interface

Services

Notable services

- Elastic Compute Cloud (EC2)
- Elastic Block Store (EBS)
- Simple Storage Service (S3)
- Virtual Private Cloud (VPC)
- Simple Queue Service (SQS)
- EC2 Container Service (ECS)
- Amazon CloudFront

Amazon EC2

- Allows renting VMs (called *instances*) on a per second basis
- Bare-metal instances are also available

EC2 concepts

- Instance: an active VM with a specific resource capacity
- Amazon Machine Image (AMI): template for creating VMs (contains OS and other software and data)
  - EBS-backed AMI: the root device is stored on an EBS volume
  - Instance-store backed AMI: the root device is stored locally on host
- Availability zones, regions: determine instance location

EC2 instance types

- Offer different compute, memory, storage, and networking capacities

Instance	vCPU*	Mem (GiB)	Storage	Dedicated EBS Bandwidth (Mbps)	Network Performance
m4.large	2	8	EBS-only	450	Moderate
m4.xlarge	4	16	EBS-only	750	High
m4.2xlarge	8	32	EBS-only	1,000	High
m4.4xlarge	16	64	EBS-only	2,000	High
m4.10xlarge	40	160	EBS-only	4,000	10 Gigabit
m4.16xlarge	64	256	EBS-only	10,000	25 Gigabit

EC2 pricing

- On-demand instances
  - Per hour or per second charge
- Reserved instances
  - One-time fee and discounted hourly charge

EC2 pricing

- Spot instances
  - Excess capacity is offered at a fluctuating price
  - Users bid a maximum price (by default, the demand price) and run instances as long as the price is lower than bid

EC2 pricing

Elastic IPs

- IP addresses are normally *dynamic* (i.e., they do not persist when instances are powered off)
- **Elastic IP** addresses are *static* IP addresses that
  - belong to an AWS account
  - can be assigned and reassigned to running instances
- Elastic IP address are free *if* they are associated with a running instance
  - otherwise, hourly charged

Security groups

- Define a set of firewall rules for restricting the inbound and outbound traffic of instances

Inbound rules		
Protocol type	**Port number**	**Source IP**
TCP	22 (SSH)	203.0.113.1/32
TCP	80 (HTTP)	0.0.0.0/0
ICMP	All	0.0.0.0/0
Outbound rules		
Protocol type	**Port number**	**Destination IP**
All	All	0.0.0.0/0

Elastic Block Store (EBS)

- Persistent block storage volumes for EC2 instances
- Multiple volumes can be attached to one instance
- Automatic replication within an availability zone
- Snapshot support
- Pricing based on GB-month of provisioned storage and per million I/O requests

Using EC2

- Select AMI
- Choose instance type
- Choose availability zone
- Add EBS volumes
- Set security groups
- Attach elastic IP
- Set key pair
- Launch, stop, start, connect to instance, terminate instance, etc.

Simple Storage Service (S3)

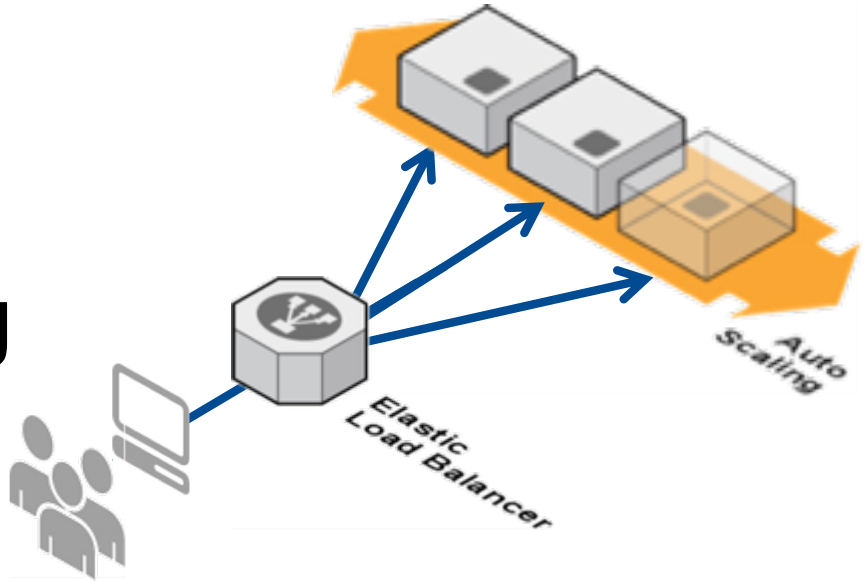
- Key-value store for large objects
- *Objects* are stored in *buckets* and retrieved via developer-assigned *keys*
  - *http://s3.amazonaws.com/<bucket>/<key>*
- Unlimited number of objects (of size up to 5TB)
- 99.999999999% durability and 99.99% availability
- Fine-grained access control

Simple Storage Service (S3)

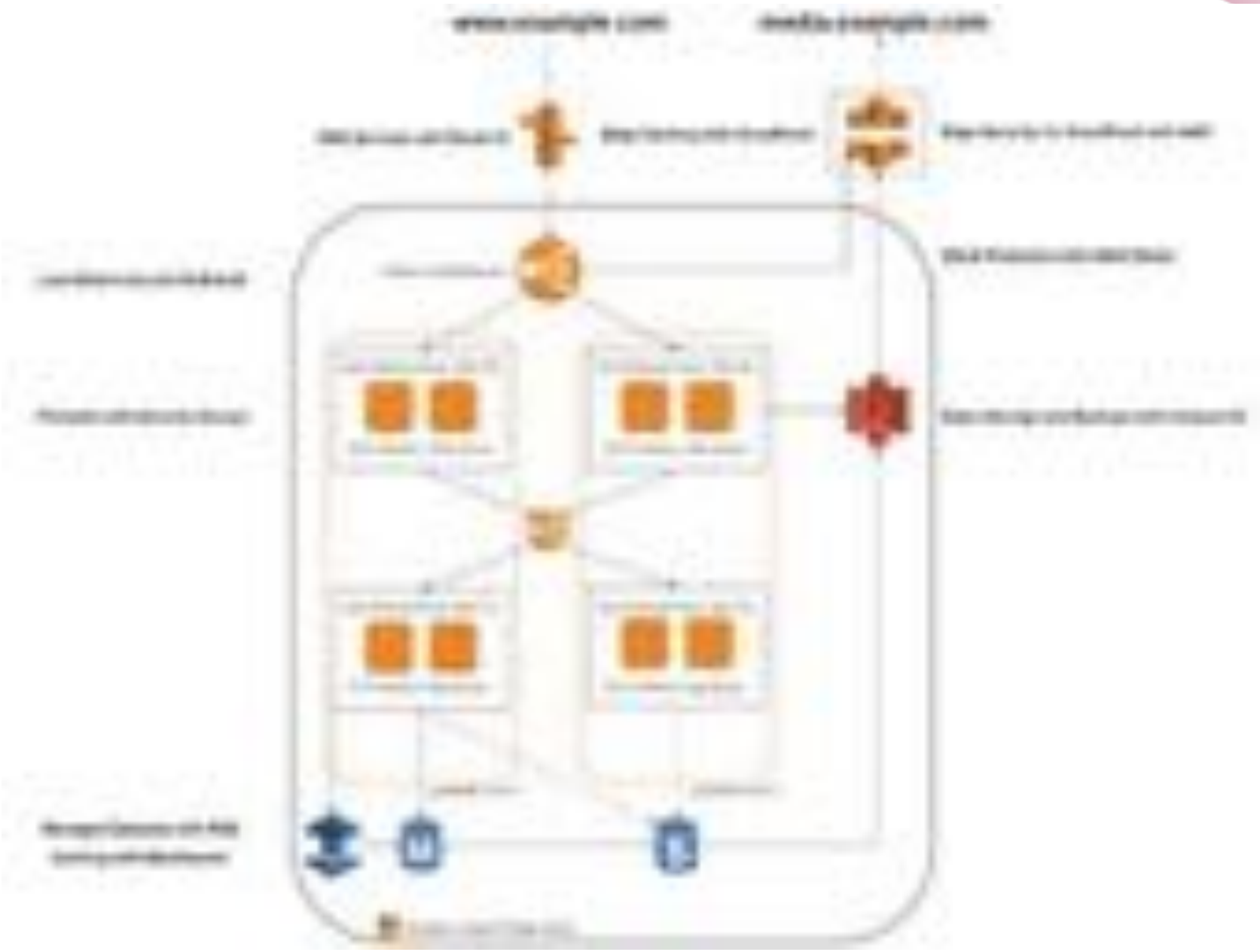
- Provides eventual consistency
- Useful for content storage and distribution, backup, archiving, …
- Pricing based on:
  - GBs used per month
  - Number and type of requests per month
  - GBs transferred out of S3 per month

Auto Scaling

- CloudWatch
  - Monitors metrics and sends alarms
- Elastic Load Balancing
  - Distributes incoming traffic across multiple instances
- Auto Scaling
  - Maintains availability and scales capacity according to rules



Web application hosting



https://docs.aws.amazon.com/whitepapers/latest/web-application-hosting-best-practices/an-aws-cloud-architecture-for-web-hosting.html

Summary

- IaaS is about offering computing resources (e.g., virtual machines, virtual disks, virtual networks, load balancers) as a service
- Amazon Web Services (AWS) is a representative IaaS offering
- Notable AWS services include EC2, EBS, and S3

References

- *Amazon Web Services*, http://aws.amazon.com
- *Mastering Cloud Computing: Foundations and Applications Programming,* R.Buyya, C. Vecchiola and S. Thamarai Selvi, Elsevier Science & Technology, 2013