Cloud e Datacenter Networking

Università degli Studi di Napoli Federico II

Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione DIETI

Laurea Magistrale in Ingegneria Informatica

Prof. Roberto Canonico

Amazon AWS: an introduction

Lesson outline

▶ **Amazon AWS suite**

▶ **Regions and Availability Zones**

▶ **EC2: AMI and Instance Types**

▶ **EC2: storage options**

▶ **EBS volumes**

▶ **S3 storage**

Amazon Web Services (AWS)

▶ Amazon started a commercial IaaS cloud computing service in 2006

▶ Amazon Web Services is today a rich suite of IaaS and PaaS cloud services

▶ The first IaaS service was Amazon Elastic Compute Cloud (EC2)

▶ Amazon offers their cloud services through a network of data centers on several continents

AWS regions and availability zones

▶ Amazon's AWS infrastructure is organized in geographical *Regions*

  ▶ Regions do not share resources and communicate through the Internet

  ▶ Each Amazon EC2 region is designed to be completely isolated from the others to achieve the greatest possible fault tolerance and stability

▶ Each region has multiple, isolated locations known as *Availability Zones*, interconnected by proprietary high-speed networks

  ▶ An *availability zone* corresponds to a data center facility

▶ If you distribute your instances across multiple Availability Zones and one instance fails, you can design your application so that an instance in another Availability Zone can handle requests
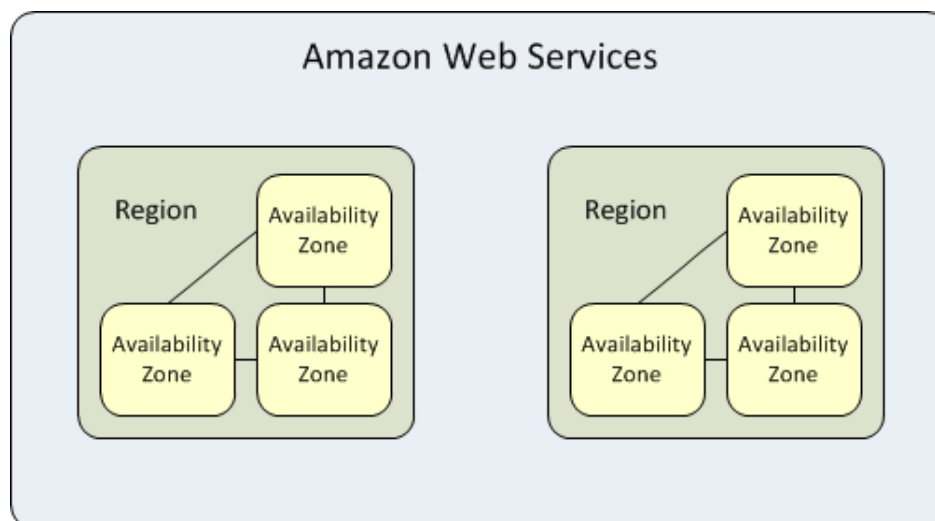
Amazon AWS global infrastructure

- **As of April 2016, ten Regions**
  - 3+1 in North America (the extra one is AWS GovCloud)
  - 1 in South America
  - 2 in Europe
  - 4+1 in Asia/ Australia (the extra one is in Beijing)

Code	Name
us-east-1	US East (N. Virginia)
us-west-2	US West (Oregon)
us-west-1	US West (N. California)
eu-west-1	EU (Ireland)
eu-central-1	EU (Frankfurt)
ap-southeast-1	Asia Pacific (Singapore)
ap-northeast-1	Asia Pacific (Tokyo)
ap-southeast-2	Asia Pacific (Sydney)
ap-northeast-2	Asia Pacific (Seoul)
sa-east-1	South America (São Paulo)
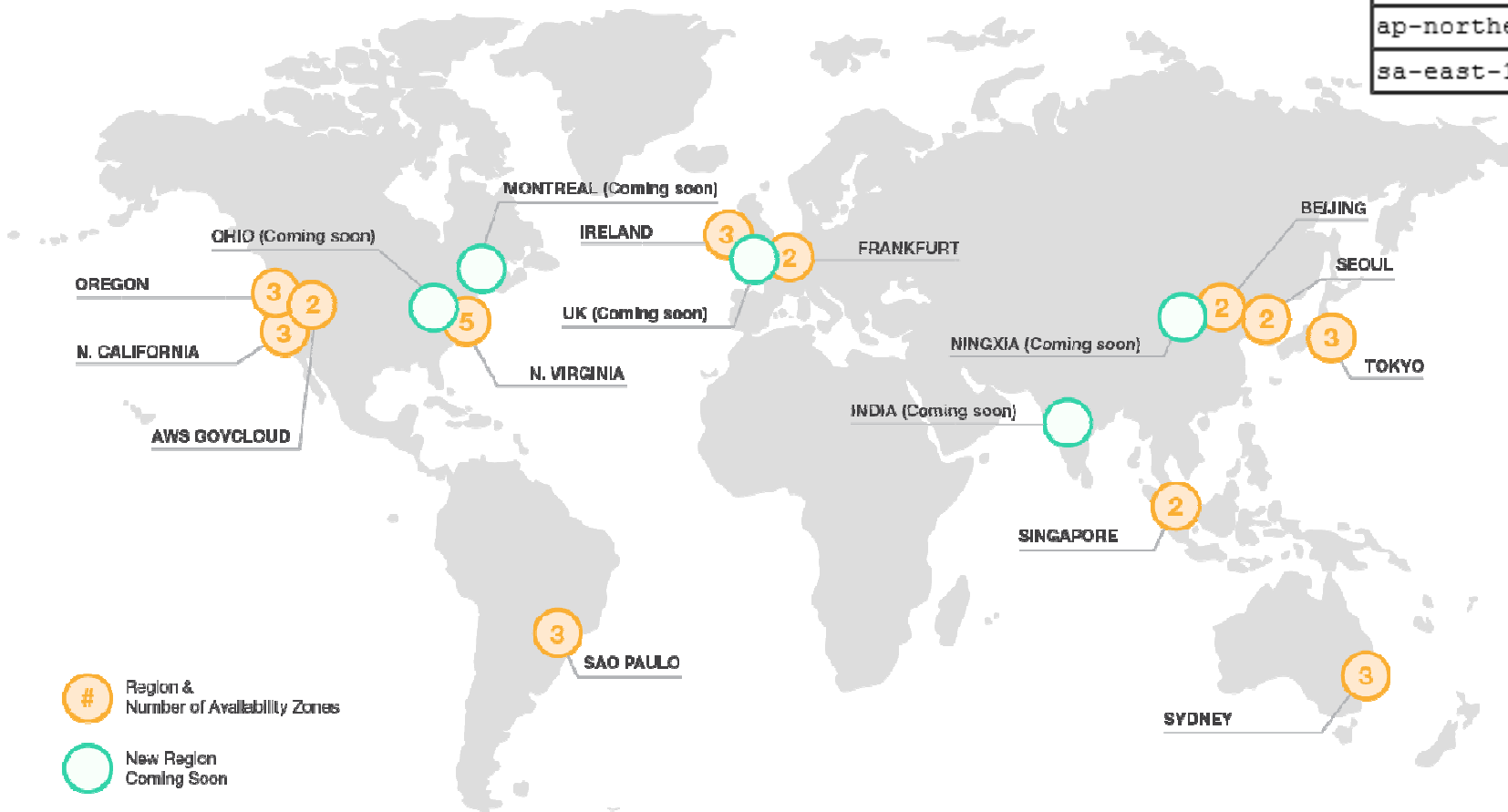
Amazon Elastic Compute Cloud service (EC2)

▸ EC2 allows users to rent virtual computers (*instances*) on which they run their own applications

▸ Instances may be Virtual Machines instantiated on a hypervisor or dedicated bare-metal servers

▸ A user can create, launch, and terminate instances as needed, paying by the hour for active servers - hence the term "elastic"

▸ Instances may be dynamically activated (and releases) through a web-service API

▸ EC2 encourages scalable deployment of applications

▸ EC2 provides users with control over the geographical location of instances that allows for latency optimization and high levels of redundancy

▸ Computational power of instances is expressed in ECUs (*Elastic Compute Units*)

　▸ Amazon's Definition of ECU notes "We use several benchmarks and tests to manage the consistency and predictability of the performance of an EC2 Compute Unit. One EC2 Compute Unit provides the equivalent CPU capacity of a 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor. This is also the equivalent to an early-2006 1.7 GHz Xeon processor referenced in our original documentation"

▸ Instances may be activated conforming to a predefined "menu" of configuration options termed "Instance Types"

EC2 instances

▶ An instance is a virtual server with a well specified set of resources including:

  ▶ CPU, 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 from a limited menu of instance types

  ▶ The instance template from which the instance should be derived (*Amazon Machine Image (AMI)*)

▶ When launched, an instance is provided with a DNS name; this name maps to a:

  ▶ *private IP address* for internal communication within the internal EC2 communication network

  ▶ *public IP address* for making the server reachable from the public Internet

▶ Network Address Translation (NAT) maps external IP addresses to internal ones

▶ The public IP address is assigned for the lifetime of an instance

▶ An instance can request an *elastic IP address*, rather than a public IP address

  ▶ The elastic IP address is a static public IP address allocated to an instance from the available pool of the availability zone

  ▶ An elastic IP address is not released when the instance is stopped or terminated and must be released when no longer needed (it is paid)

EC2 Instance Types

▸ **Instance Types are (virtual) hardware templates**

▸ **This is only a subset of available Instance Types**

Model	vCPU	CPU Credits / hour	Mem (GiB)	Storage (GB)
t2.micro	1	6	1	EBS Only
t2.small	1	12	2	EBS Only
t2.medium	2	24	4	EBS Only

Use Cases

Development environments, build servers, code repositories, low-traffic web applications, early product experiments, small databases.

| Model	vCPU	Mem (GiB)	SSD Storage (GB)
c3.large	2	3.75	2 x 16
c3.xlarge	4	7.5	2 x 40
c3.2xlarge	8	15	2 x 80
c3.4xlarge	16	30	2 x 160
c3.8xlarge	32	60	2 x 320

Use Cases

High performance front-end fleets, web-servers, on-demand batch processing, distributed analytics, high performance science and engineering applications, ad serving, batch processing, MMO gaming, video encoding, and distributed analytics.
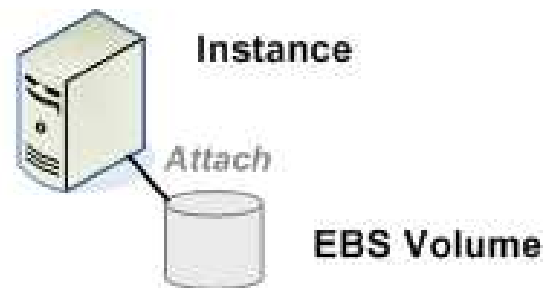
Amazon Machine Image (AMI)

▶ An *Amazon Machine Image (AMI)* is a template that contains a software configuration (for example, an operating system, an application server, and applications)

▶ From an AMI, users may launch *instances*, which are copies of the AMI running as a virtual server in the cloud

▶ Amazon publishes many Amazon Machine Images (AMIs) that contain common software configurations for public use

▶ In addition, third parties have published their own custom AMIs

▶ Users may launch multiple instances of an AMI, and create their own AMIs as well
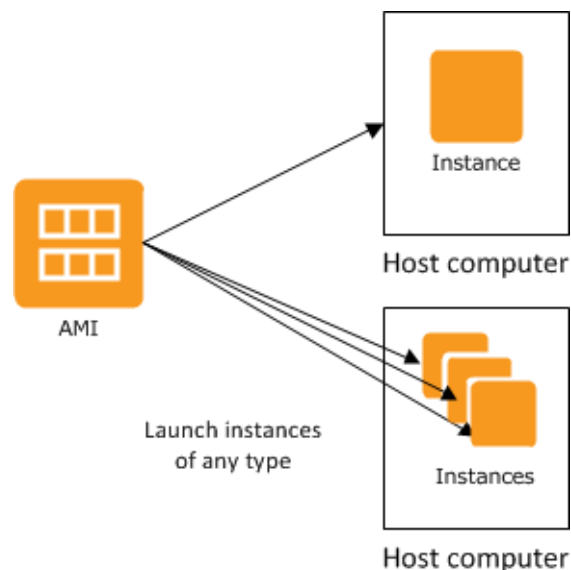
Storage options for EC2 instances: EBS

▸ To store data, Amazon EC2 offers the following storage options:

  ▸ Amazon Elastic Block Store (Amazon EBS)

  ▸ Amazon EC2 Instance Store

  ▸ Amazon Simple Storage Service (Amazon S3) (described separately)

▸ Amazon EBS provides the instances with persistent, block-level storage volumes

  ▸ Amazon EBS volumes are essentially virtual hard disks that you can attach to a running instance

▸ This is the recommended storage option for the majority of use cases

  ▸ Particularly suited for applications that require a database, file system,
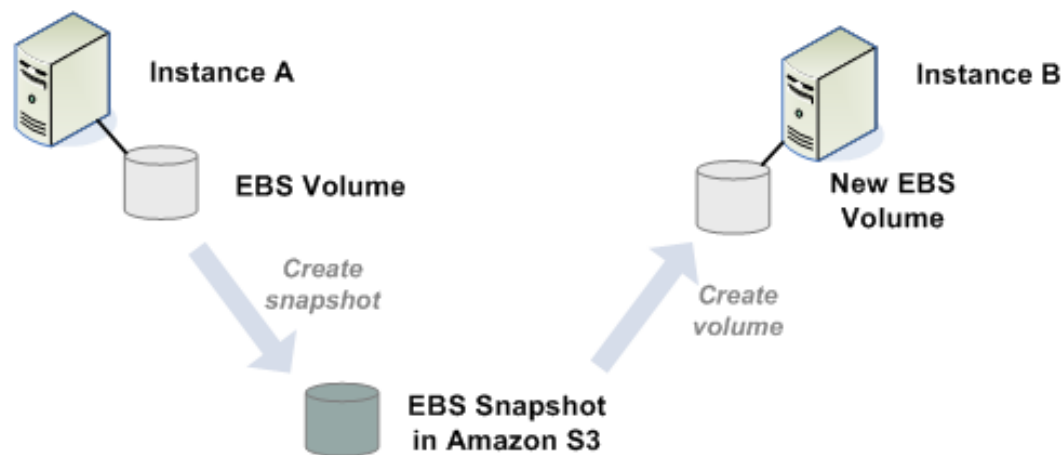   or access to raw block-level storage

Cloud and Datacenter Networking Course – Prof. Roberto Canonico – Università degli Studi di Napoli Federico II

10

EBS volumes: snapshot

▸ **To keep a back-up copy, you can create a snapshot of the volume**

　　▸ **EBS volume snapshots are stored in Amazon S3**
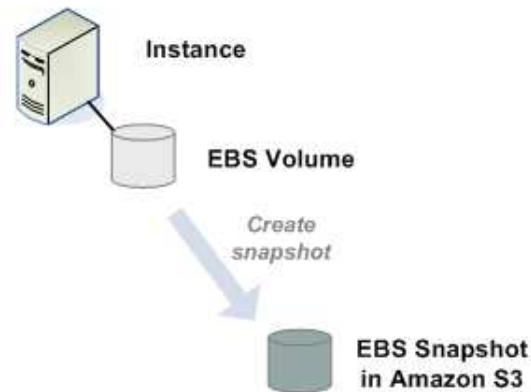
▸ **Users can create a new EBS volume from a snapshot, and attach it to another instance**

EBS volumes: attach/detach

▸ It is also possible to detach a volume from an instance and attach it to a different one


▸ Notice: EC2 instances can have an EBS volume as their root device ("EBS boot")

    ▸ Result: instance data persists independently from the lifetime of the instance

▶ An Instance Store provides temporary (*ephemeral*) block-level storage that does NOT persist if the instance is stopped or terminated

▶ This storage is located on disks that are <u>physically attached to the host computer</u>

▶ Instance store is ideal for temporary storage of information that changes frequently, such as buffers, caches, scratch data, and other temporary content, or for data that is replicated across a fleet of instances, such as a load-balanced pool of web servers

    ▶ This option can be considered when data persistence is not required

▶ If an instance reboots (intentionally or unintentionally), data in the instance store persists. However, data in the instance store is lost under the following circumstances:

    ▶ the underlying disk drive fails

   ▶ the instance stops or terminates
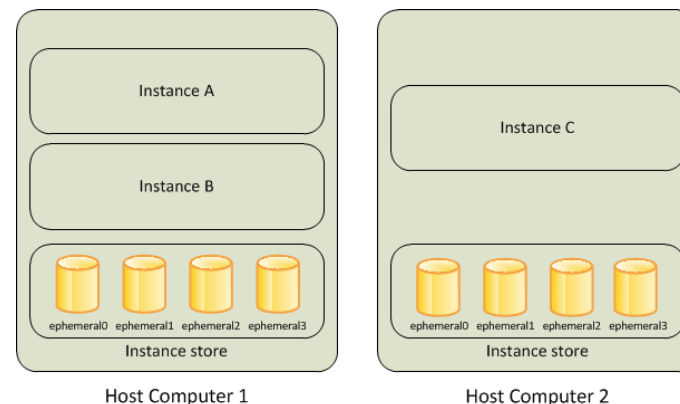
Amazon Simple Storage Service (S3)

▸ **Amazon Simple Storage Service** (*Amazon S3*) is *storage for the Internet*

▸ You can use Amazon S3 to store and retrieve any amount of data at any time, from anywhere on the web

▸ There are two main uses for S3:

  ▸ <u>Backup server</u>. You store your data through S3 and don't give anyone else access to it. Rather than buying your own disks, you rent disk space from Amazon

  ▸ <u>Data host</u>. You store your data on S3 and give others access to it. Amazon serves your data through HTTP. Rather than paying an ISP for bandwidth, you pay Amazon. Depending on your existing bandwidth costs this can save you a lot of money. Many of today's web startups use S3 to serve data files.

▸ You can accomplish these tasks using the AWS Management Console, which is a simple and intuitive web interface

▸ Amazon S3 stores data as **objects** within **buckets**

▸ An object consists of a file and metadata that describes that file

Sign Up for Amazon S3 → Create a Bucket → Add an Object to a Bucket → View an Object → Move an Object → Delete an Object and Bucket

▸ **The upper left quadrant of this chart shows the growth in data transfers into and out of the S3 object storage service from what looks like late 2010 through the middle of 2014 or so**

    ▸ **Y-axis measures petabytes per week**

    ▸ **132% growth growth in traffic generated by S3 in the last year**
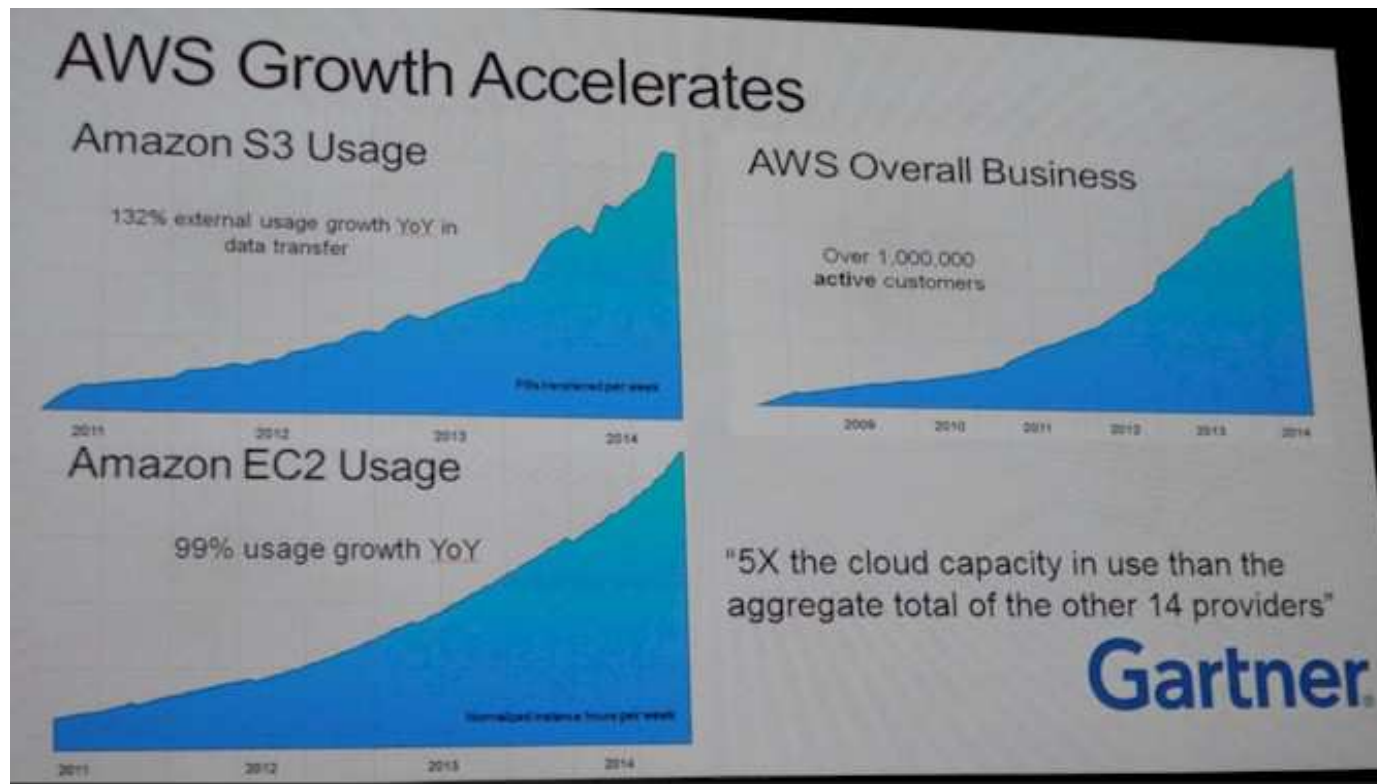
Amazon S3: buckets

▸ **Every file you upload to Amazon S3 is stored in a container called a bucket**

- ▸ A **bucket** is a container for objects and describes location, logging, accounting, and access control

- ▸ A bucket can hold any number of **objects**, which are files of up to 5TB

- ▸ Before you start working with Amazon S3 you have to create at least one bucket

- ▸ The bucket namespace is shared by all users of the system; therefore each bucket name should be unique

- ▸ You can create up to 100 buckets per account

- ▸ Buckets cannot be nested, you can not create a bucket within a bucket

▸ **Buckets may contain files and/or folders**

- ▸ Folders may be nested

S3 operational issues and their impact

▸ A huge number of websites store content and run services with Amazon S3

▸ It is estimated that Amazon S3 is used by more than 120,000 domains across the world

▸ On 28 February 2017 the S3 service provided by region US-EAST-1 experienced problems for a few hours

▸ This problem impacted several web sites and services

▸ In some cases entire sites were offline, but in others only parts of the services were affected

▸ A number of smart home devices that rely on S3 for parts of their functionality also experienced problems

S3 operational issue   Close

Details   Affected resources

We are investigating increased error rates for Amazon S3 requests in the US-EAST-1 Region.

We're continuing to work to remediate the availability issues for Amazon S3 in US-EAST-1. AWS services and customer applications depending on S3 will continue to experience high error rates as we are actively working to remediate the errors in Amazon S3.

Event data

Event	S3 operational issue
Status	Open
Region/AZ	us-east-1
Start time	February 28, 2017 at 6:37:00 PM UTC+1
End time	-
Event category	🔧 Issue