ĐIỆN TOÁN ĐÁM MÂY
Cloud Infrastructure

Nguyễn Thành Trung
trung.nguyenthanh@phenikaa-uni.edu.vn

Contents

- IaaS services from Amazon.
  - Regions and availability zones for Amazon Web Services.
  - Instances – attributes and cost.
  - A repertoire of Amazon Web Services.
- SaaS and PaaS services from Google.
- SaaS and PaaS services from Microsoft.
- Open-source platforms for private clouds.

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.
- 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.

Amazon Web Services (AWS)

- AWS → IaaS cloud computing services launched in 2006.

- Businesses in 200 countries used AWS in 2012.

- The infrastructure consists of computing and storage servers interconnected by high-speed networks and supports a set of services.

- An application developer:
    - Installs applications on a platform of his/her choice.
    - Manages resources allocated by Amazon.

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.

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

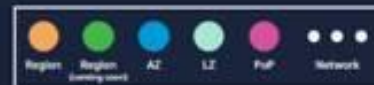- Regions do not share resources and communicate through the Internet.

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 from a limited menu of instance types.
  - 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 communication outside the internal Amazon network, e.g., for communication with the user that launched the instance.

- 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.
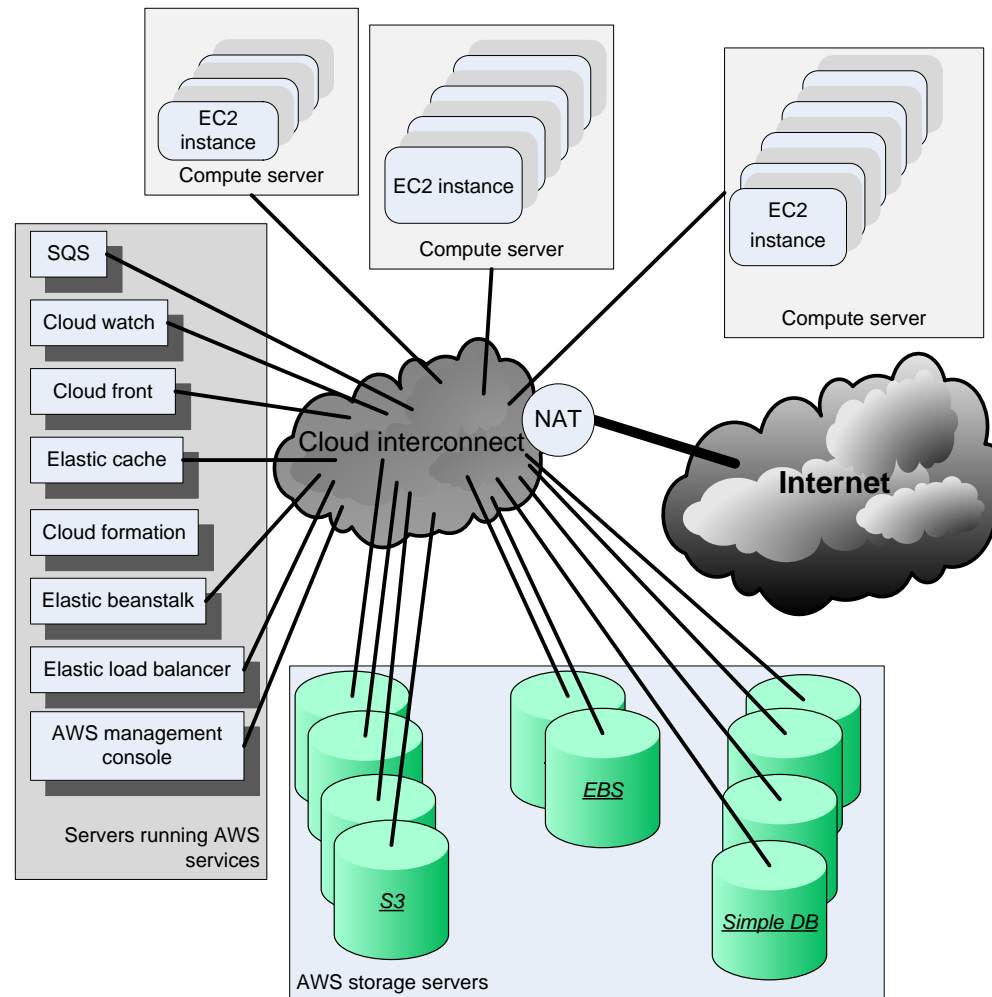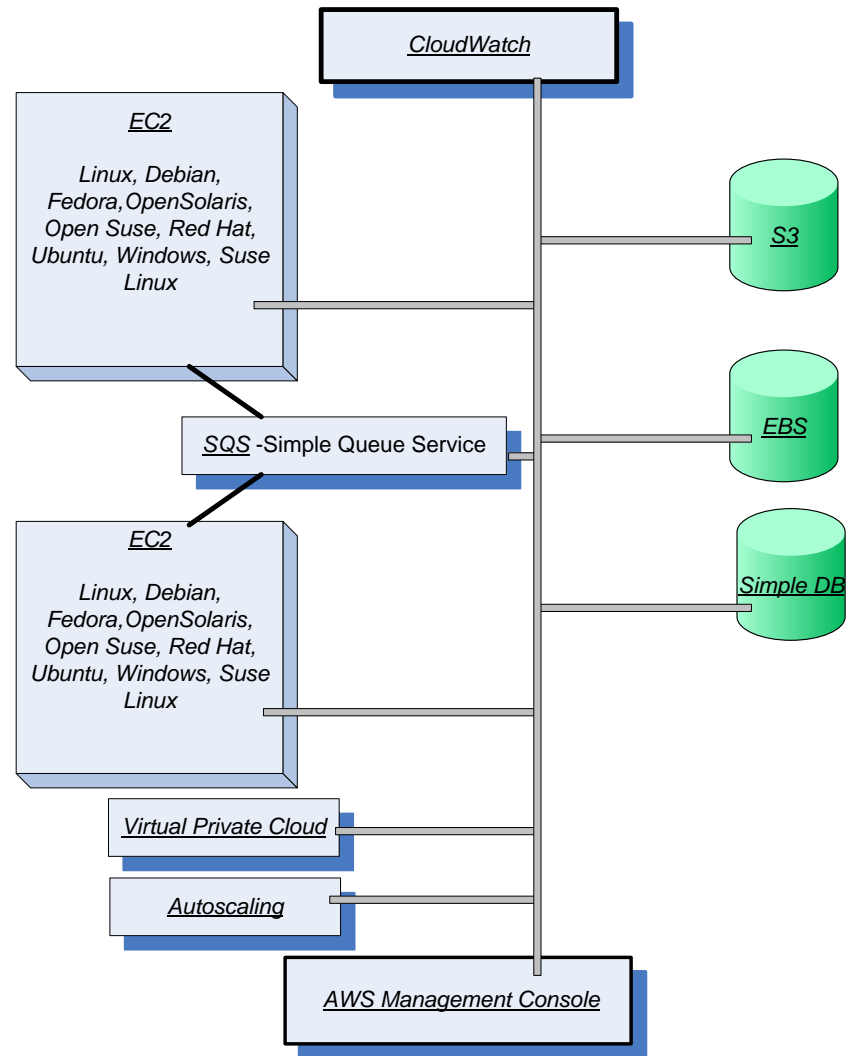
Steps to run an application

- 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, but not all options may be available.

- AWS SDK libraries and toolkits are provided for several programming languages including Java, PHP, C#, and Objective-C.

- Raw REST requests.

Examples of Amazon Web Services

- *AWS Management Console -* allows users to access the services offered by AWS.

- *Elastic Cloud Computing (EC2) -* allows a user to launch various operating systems.

- *Simple Queuing Service (SQS)* – allows multiple *EC2* instances to communicate with one another.

- *Simple Storage Service (S3), Simple DB*, and *Elastic Bloc Storage (EBS) -* storage services.

- *Cloud Watch -* supports performance monitoring.

- *Auto Scaling -* supports elastic resource management.

- *Virtual Private Cloud -* allows direct migration of parallel applications.

CloudWatch

EC2

Linux, Debian, Fedora,OpenSolaris, Open Suse, Red Hat, Ubuntu, Windows, Suse Linux

S3

EBS

SQS -Simple Queue Service

EC2

Linux, Debian, Fedora,OpenSolaris, Open Suse, Red Hat, Ubuntu, Windows, Suse Linux

Simple DB

Virtual Private Cloud

Autoscaling

AWS Management Console

EC2 – Elastic Cloud Computing

- *EC2* - web service for launching instances of an application under several operating systems, such as:
  - Several Linux distributions.
  - Microsoft Windows Server 2003 and 2008.
  - OpenSolaris.
  - FreeBSD.
  - NetBSD.

- A user can
  - Load an *EC2* instance with a custom application environment.
  - Manage network's access permissions.
  - Run the image using as many or as few systems as desired.

EC2 (cont'd)

- Import virtual machine (VM) images from the user environment to an instance through *VM import*.

- *EC2* instances boot from an AMI (Amazon Machine Image) digitally signed and stored in *S3*.

- Users can access:
  - Images provided by Amazon.
  - Customize an image and store it in *S3*.

- An *EC2* instance is characterized by the resources it provides:
  - VC (Virtual Computers) – virtual systems running the instance.
  - CU (Compute Units) – measure computing power of each system.
  - Memory.
  - I/O capabilities.

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).

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 Amazon cloud computing is the low cost.

PAGE CONTENT

On-Demand Pricing

Data Transfer

Data Transfer within the same AWS Region

EBS-Optimized Instances

Elastic IP Addresses

Carrier IP Addresses

Elastic Load Balancing

On-Demand Capacity Reservations

T2/T3/T4g Unlimited Mode Pricing

**Amazon CloudWatch**

Amazon Elastic Block Store

Viewing 489 of 489 available instances

🔍 | < **1** 2 3 4 5 6 7 … 25 >

Instance name ▲	On-Demand hourly rate ▽	vCPU ▽	Memory ▽	Storage ▽	Network performance ▽
a1.medium	$0.0255	1	2 GiB	EBS Only	Up to 10 Gigabit
a1.large	$0.051	2	4 GiB	EBS Only	Up to 10 Gigabit
a1.xlarge	$0.102	4	8 GiB	EBS Only	Up to 10 Gigabit
a1.2xlarge	$0.204	8	16 GiB	EBS Only	Up to 10 Gigabit
a1.4xlarge	$0.408	16	32 GiB	EBS Only	Up to 10 Gigabit
a1.metal	$0.408	16	32 GiB	EBS Only	Up to 10 Gigabit
t4g.nano	$0.0042	2	0.5 GiB	EBS Only	Up to 5 Gigabit
t4g.micro	$0.0084	2	1 GiB	EBS Only	Up to 5 Gigabit

Amazon EC2 instance lifecycle
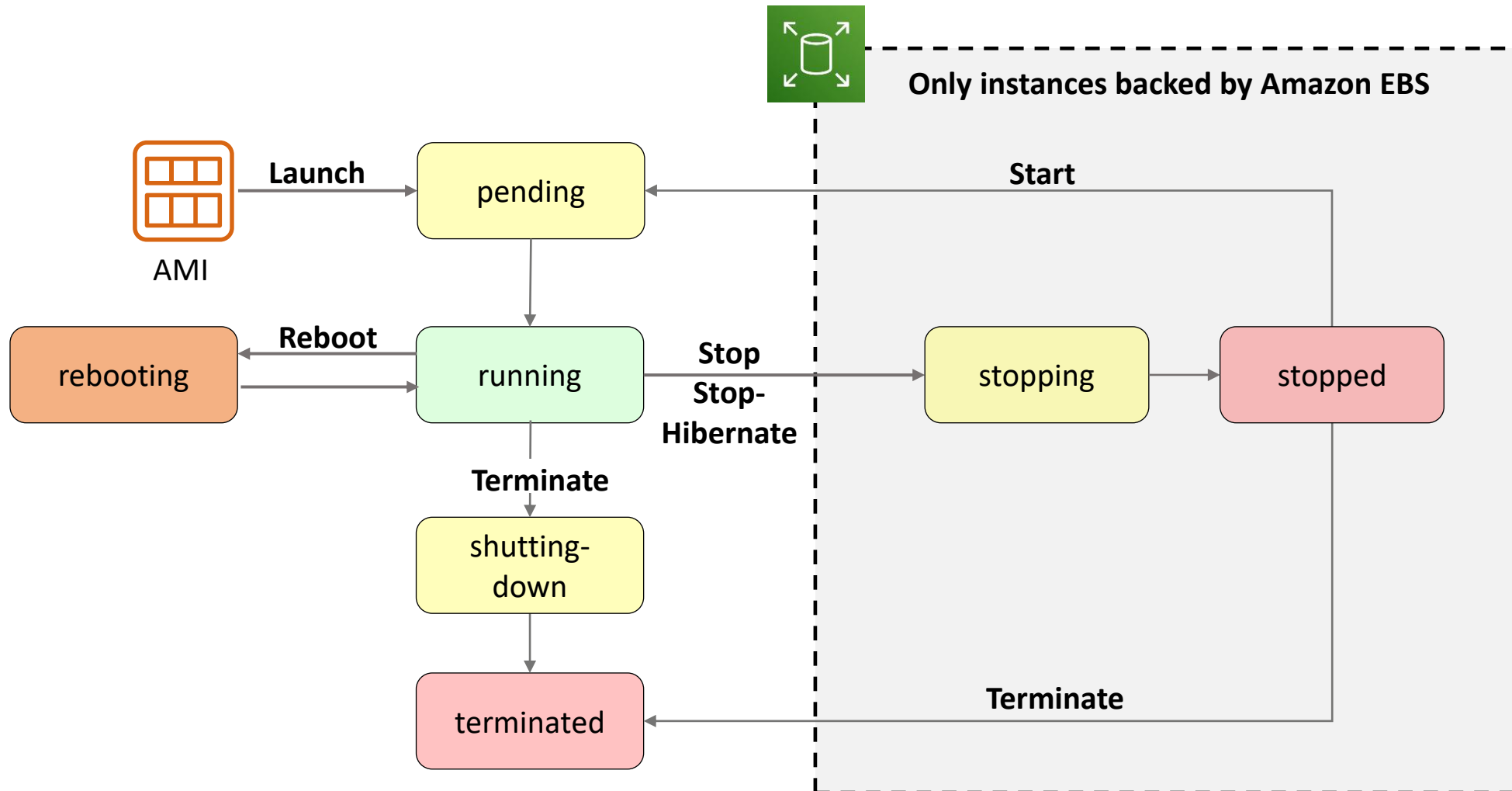
S3 – Simple Storage System

- Service designed to store large objects; an application can handle an unlimited number of objects ranging in size from 1 byte to 5 TB.

- An object is stored in a <u>bucket</u> and retrieved via a unique, developer-assigned key; a bucket can be stored in a Region selected by the user.

- Supports a minimal set of functions: write, read, and delete; it does not support primitives to copy, to rename, or to 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, and up to 4 KB of user-defined metadata.

- Authentication mechanisms ensure that data is kept secure.

- Objects can be made public, and rights can be granted to other users.

- *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.

Elastic Block Store (EBS)

- Provides persistent block level storage volumes for use with *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 range 1 GB -1 TB.

- An *EC2* instance may mount multiple volumes, but a volume cannot be shared among multiple instances.

- EBS supports the creation of snapshots of the volumes attached to an instance and then uses them to restart the instance.

- The volumes are grouped together in Availability Zones and are automatically replicated in each zone.

SimpleDB

- Non-relational data store. Supports store and query functions traditionally provided only by relational databases.

- Supports high performance Web applications; users can store and query data items via Web services requests.

- Creates multiple geographically distributed copies of each data item.

- It manages automatically:
  - The infrastructure provisioning.
  - Hardware and software maintenance.
  - Replication and indexing of data items.
  - Performance tuning.

SQS - Simple Queue Service

- Hosted message queues are accessed through standard SOAP and Query interfaces.

- Supports automated workflows - *EC2* instances can coordinate by sending and receiving SQS messages.

- Applications using SQS can run independently and asynchronously and do not need to be developed with the same technologies.

- A received message is "locked'' during processing; if processing fails, the lock expires, and the message is available again.

- Queue sharing can be restricted by IP address and time-of-day.

CloudWatch

- 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.

- Without installing any software a user can monitor either seven or eight pre-selected metrics and then view graphs and statistics for these metrics.

- 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 for up to seven metrics.
  - 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.

- *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.

AWS services introduced in 2012 (cont'd)

- *CloudFront* - web service for content delivery.

- *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.

- *CloudFormation* - allows the creation of a stack describing the infrastructure for an application.

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.

- *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.
  - Initially supported Python, Java was added later.
  - 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* - an online service for data storage.
- *Google Base* - allows users to load structured data from different sources to a central repository, a very large, self-describing, semi-structured, heterogeneous database.

PaaS and SaaS services from Microsoft

- *Windows Azure* - an operating system; has 3 components:
  - Compute - provides a computation environment.
  - Storage - for scalable storage.
  - Fabric Controller - deploys, manages, and monitors applications.

- *SQL Azure* - a cloud-based version of the SQL Server.

- *Azure AppFabric*, formerly .NET Services - a collection of services for cloud applications.

Azure
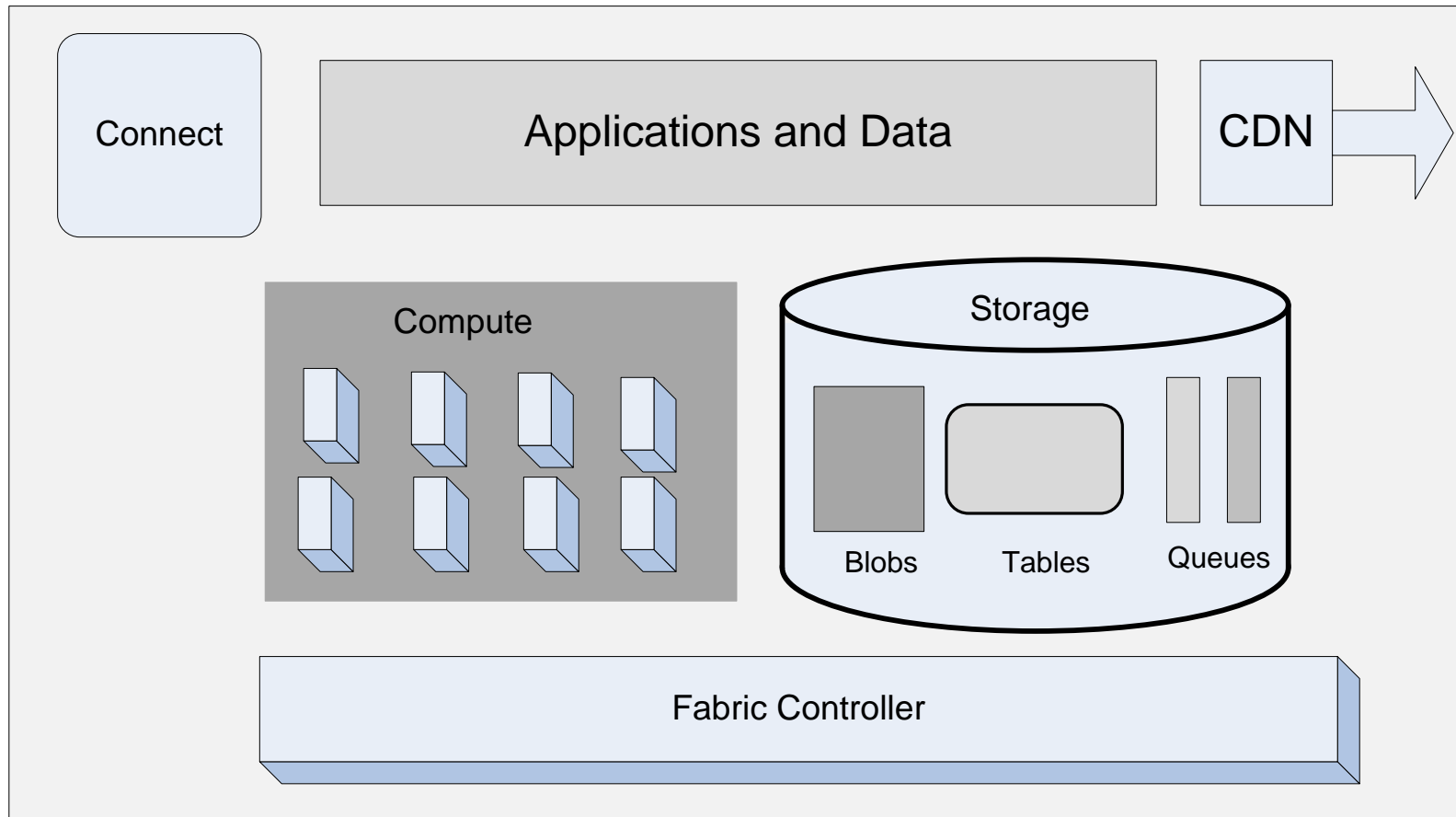
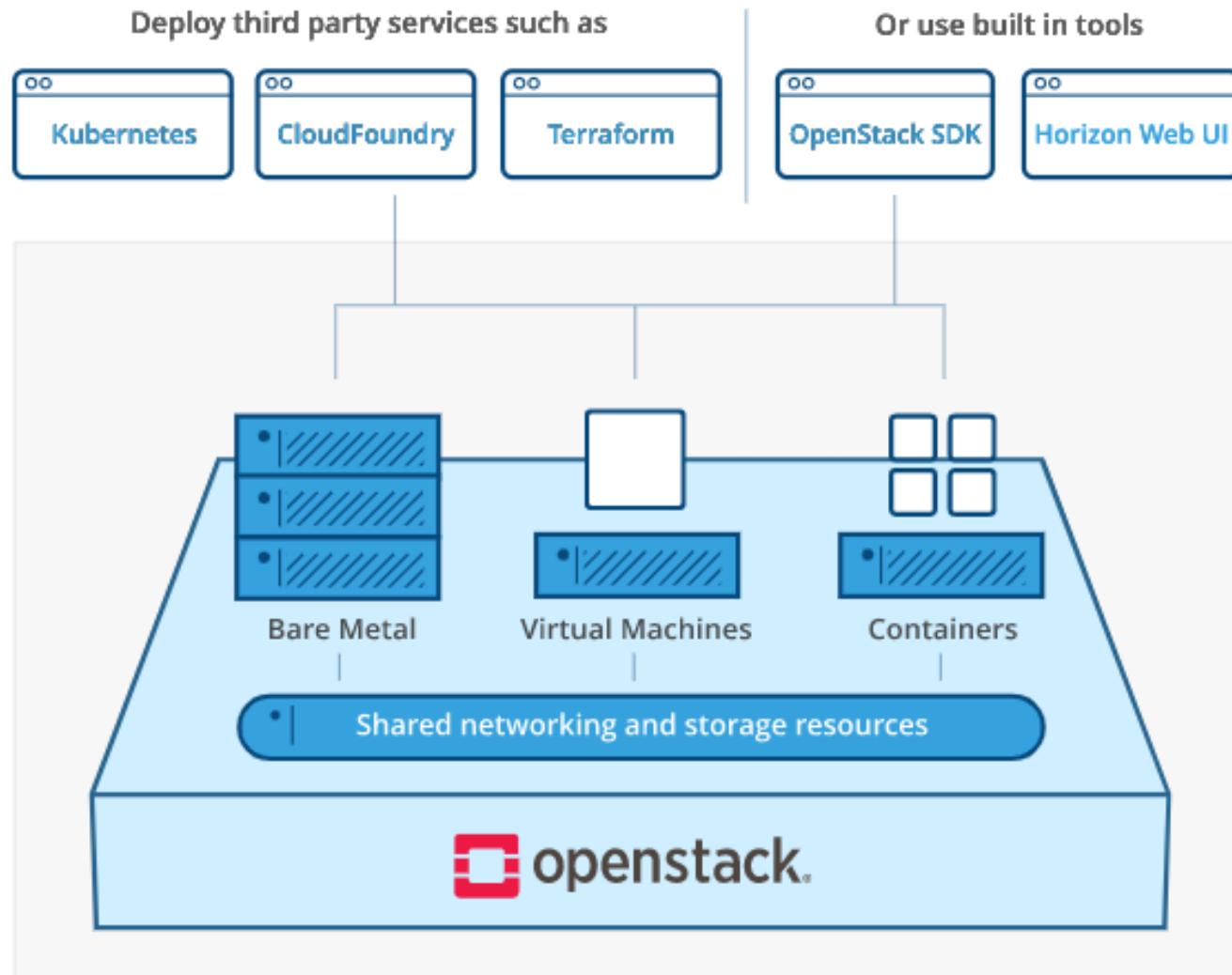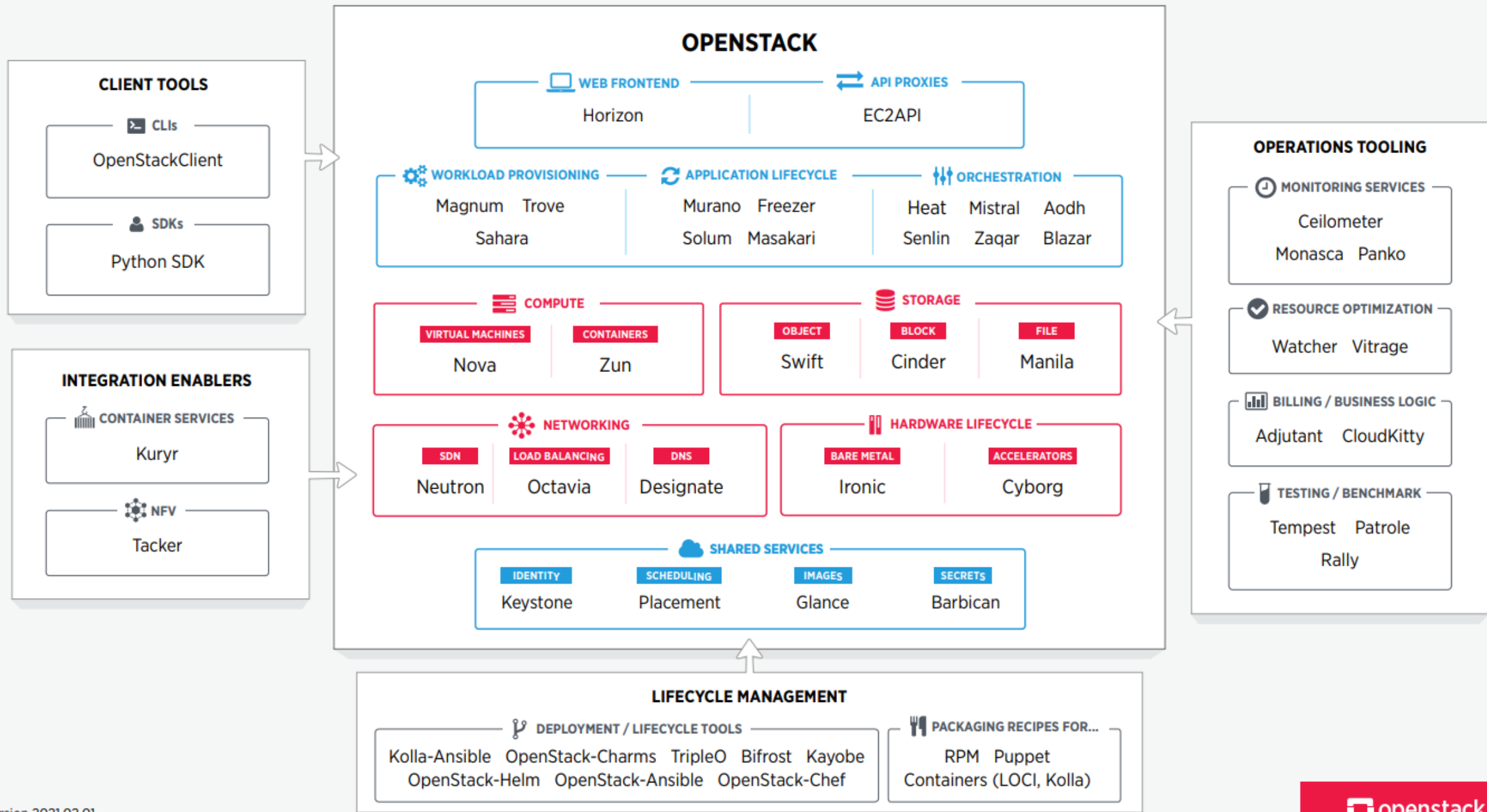# Open-source platforms for clouds

- *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.

- *Openstack* - Cloud Infrastructure for Virtual Machines, Bare Metal, and Containers

Openstack

OPENSTACK

WEB FRONTEND
Horizon

API PROXIES
EC2API

WORKLOAD PROVISIONING
Magnum Trove
Sahara

APPLICATION LIFECYCLE
Murano Freezer
Solum Masakari

ORCHESTRATION
Heat Mistral Aodh
Senlin Zaqar Blazar

COMPUTE

VIRTUAL MACHINES
Nova

CONTAINERS
Zun

STORAGE

OBJECT
Swift

BLOCK
Cinder

FILE
Manila

NETWORKING

SDN
Neutron

LOAD BALANCING
Octavia

DNS
Designate

HARDWARE LIFECYCLE

BARE METAL
Ironic

ACCELERATORS
Cyborg

SHARED SERVICES

IDENTITY
Keystone

SCHEDULING
Placement

IMAGES
Glance

SECRETS
Barbican

CLIENT TOOLS

CLIs
OpenStackClient

SDKs
Python SDK

INTEGRATION ENABLERS

CONTAINER SERVICES
Kuryr

NFV
Tacker

OPERATIONS TOOLING

MONITORING SERVICES
Ceilometer
Monasca Panko

RESOURCE OPTIMIZATION
Watcher Vitrage

BILLING / BUSINESS LOGIC
Adjutant CloudKitty

TESTING / BENCHMARK
Tempest Patrole
Rally

LIFECYCLE MANAGEMENT

DEPLOYMENT / LIFECYCLE TOOLS
Kolla-Ansible OpenStack-Charms TripleO Bifrost Kayobe
OpenStack-Helm OpenStack-Ansible OpenStack-Chef

PACKAGING RECIPES FOR...
RPM Puppet
Containers (LOCI, Kolla)

Version 2021.02.01

openstack.

Complete the knowledge check