Cloud Computing

Chapter 9 in Unix and Linux System Administration Handbook

Haakon André Reme-Ness

HVL

Haakon.Andre.Reme-Ness@hvl.no

February 5, 2025

Cloud Computing

- Practice of leasing computer resources from a pool of shared capacity
- Provision resources on demand and charge by consumptions

Advantages:

- Faster time to market
- Greater flexibility
- Lower capital and operating expenses

Cloud Computing[1]

A realisation of "utility computing" [1]

Utility: package of system resources

- computation
- storage
- networking ...

[1] First conceived by John McCarthy

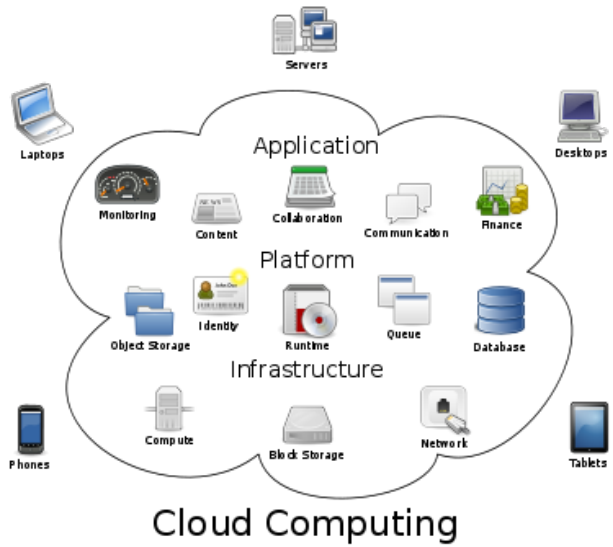
What makes Cloud Computing possible?

Some major advances in technologies:

- Reliable allocation of CPU, memory, storage and network resources on demand by virtualisation software
- Robust security layers
- Standardised hardware components
- A reliable global network connects everything

Cloud Computing

- On-demand network access to computing resources.
- Resources can rapidly be provisioned and released with minimal management effort.
- Autoscaling features
- Customer can increase capacity or add resources without investing in new infrastructure, training new personnel or licensing new software.
- Except for Microsoft Azure, Linux is at the heart of all cloud solutions.



Cloud Computing

Cloud Computing – Towards a definition

From A Break in the Clouds: Towards a Cloud Definition

– Luis M. Vaquero, Luis Rodero-Merino, Juan Caceres, Maik Lindner:

> *Clouds are a large pool of easily usable and accessible virtualised resources (such as hardware, development platforms and/or services). These resources can be dynamically reconfigured to adjust to a variable load (scale), allowing also for an optimum resource utilisation. This pool of resources is typically exploited by a pay-per-use model in which guarantees are offered by the Infrastructure Provider by means of customised SLAs.* [2]

[2] SLA: Service Level Agreement: A legal contract between the service providers and customers.

Types of Clouds – Public

vendor: controls all the physical hardware and affords access to systems over the Internet

user: no hardware installation and maintenance, but less control over the features and characteristics of the platform

E.g., Amazon Web Services (AWS), Google Cloud Platform (GCP), DigitalOcean (DO)

Types of Clouds – Private

- Similar to public cloud
- Hosted within an organisation's own data centre, or managed by a vendor on behalf of a single customer
- Servers are single-tenant, i.e., not shared with other customers

  E.g., OpenStack

Types of Clouds – Hybrid

- Combination of public and private cloud
- Useful for:
  - Initial migration from local servers to a public cloud
  - Temporary additional capacity to handle peak loads
  - A variety of other organisation-specific scenarios

e.g., VMware vCloud Air

Cloud Platforms

Provider	Notable qualities
Amazon Web Services	900lb gorilla. Rapid innovation. Can be expensive. Complex.
DigitalOcean	Simple and reliable. Lovable API. Good for development.
Google Cloud Platform	Technically sophisticated and improving quickly. Emphasizes performance. Comprehensive big-data services.
IBM Softlayer	More like hosting than cloud. Has a global private network.
Microsoft Azure	A distant second in size. Has a history of outages. Possibly worth consideration for Microsoft shops.
OpenStack	Modular DIY open source platform for building private clouds. AWS-compatible APIs.
Rackspace	Public and private clouds running OpenStack. Offers managed services for AWS and Azure. Fanatical support.
VMware vCloud Air	Buzzword-laden service for public, private, and hybrid clouds. Uses VMware technology. Probably doomed.

Cloud Computing Service Categories

Mainly three categories:

Infrastructure as a service (IaaS):

– **Vendors** provide virtual private servers (VPSs), including computing power, memory, network and storage

– **Users** maintain everything above the hardware

Platform as a service (PaaS):

– **Vendors** provide computing platforms to customers

- A computing platform will usually be a virtual computer with OS and software

– **Users** submit the application[3] and maintain its code running on the platform

Software as a service (SaaS):

– **Vendors** provide access to applications (on-demand software, e.g., WordPress) to customers

– **Users** maintain neither the OS nor the application.

[3] wrapped in a format specified by the vendor

Cloud Computing Service Categories

Layer	Local[a]	IaaS	PaaS	SaaS
Application	✔	✔	✔	
Databases	✔	✔	✔	
Application runtime	✔	✔	✔	
Operating system	✔	✔		
Virtual network, storage, and servers	✔	✔		
Virtualization platform	✔			
Physical servers	✔			
Storage systems	✔			
Physical network	✔			
Power, space, and cooling	✔			

a. Local: local servers and network
IaaS: Infrastructure-as-a-Service (virtual servers)
PaaS: Platform-as-a-Service (e.g., Google App Engine)
Saas: Software-as-a-Service (e.g., most web-based services)

Cloud Computing Service Categories – examples

- **IaaS** – Amazon Web Services, Microsoft Azure, GoGrid, etc.
- **PaaS** – All examples are web development:
  - Heroku
    – Ruby, Java, Node.js, Scala, Clojure, Python and PHP.
  - Google App te Engine (GAE)
    – Python, Java, Groovy, JRuby, Scala, Clojure, Go and PHP.
  - AWS Elastic Beanstalk (AEB)
    – Ruby, PHP and Python, .NET, Java and Node.js.
  - Microsoft Azure
    – ASP.NET, PHP, Node.js and Python.
- **SaaS** – Office 365, Google Docs, Dropbox.

Cloud service fundamentals

Access to the cloud

- Web-based GUI: primary interface to the cloud
- APIs are also available to access the same underlying functionality as that of the web console
- Also command-line tools for automation and repeatability
- In general, users uses SSH to access UNIX and Linux systems running in the cloud

Regions and availability zones

- region:
  - a location where a cloud provider maintains data centers
  - usually named after the territory of intended service
- availability zones:
  - collections of data centres within a region
  - inter-zone communication is fast
  - zones are independent of one another regarding power and cooling
  - geographically dispersed to reduce the effect from a natural disaster
- fundamental to build highly available network services
  - multiregion deployments may allow higher availability, but more complex

Virtual private servers (VPS)

The cloud's flagship service

- virtual machine that runs on the provider's hardware
- sometimes called instances
- customers can create as many instances as they need, and run their preferred OS and applications
- configurable virtual machines
- pay for what is used
- created from "images" (a saved state of an OS), which contains at least a root filesystem and a boot loader

Networking

- Virtual networks with custom topologies can be created to isolate customers' systems from each other and from the Internet, by
  - Setting the address ranges of the networks, define subnets, configure routes, set firewall rules, and construct VPNs
- Systems on the Cloud without public addresses are not directly accessible from the Internet
  - Need to use e.g., VPN that connects to the Cloud network
- Or, customers can rent publicly routable address to make the servers on the Cloud accessible to the Internet
- Users in general have less control over virtual networks than over traditional networks

Storage

- A major part of cloud computing
- Charge by the amount of data stored
- Some important ways to store data in the cloud:
  - object stores:
    – contain collections of discrete objects (e.g., files)
    – can store virtually unlimited amount of data with high reliability but relatively slow performance
    – design for a read-mostly access pattern

    E.g., AWS S3 and Google Cloud Storage
  - block storage devices:
    – virtualised hard disks that can be attached to a virtual server.
    – can be moved among nodes

    E.g., AWS EBS and Google persistent disks
  - ephemeral storage:
    – local disk space on VPS, which is fast and capacious
    – data is lost when the VPS is deleted
    – best of temporary files

    E.g., store volumes on AWS and local SSDs on GCP

Identity and authorisation

- AWS is exceptionally strong in controlling access
- IAM (Identity and Access Management) from Amazon provides advanced authorisation features for specifying access control
  - Can define users and groups, as well as roles for systems
  - Provides an API for key management
- Azure uses Microsoft's Active Directory
- Google's access control service (called IAM also) is relatively coarse-grained and incomplete compared with AWS IAM

Automation

- Tool to facilitate orchestrating large collection of resources, like
  - creating new network
  - configuring firewall
  - launching several VPS

  E.g., AWS CloudFormation uses JSON or YAML for describing the details about desired resources and associated configuration

Serverless functions

- One of the most innovative features in the Cloud
- Aka cloud function services or functions-as-a-service
- A model of code execution that does not require long-lived infrastructure
- Functions execute in response to an event
  e.g., the arrival of a new HTTP request
- E.g., Lambda in AWS

Example: AWS

- Install it with **pip install awscli**
- Runs **aws configure** to set API credential and default region

```
$ aws configure
AWS Access Key ID: AKIAIOSFODNN7EXAMPLE
AWS Secret Access Key: wJalrXUtnFEMI/K7MDENG/bPxRfiCYEXAMPLEKEY
Default region name [us-east-1]: <return>
Default output format [None]: <return>
```

  - Generate the credentials in IAM web console
    - access key ID
    - secret access key

Example: AWS (creating an EC2 instance)

Use **aws ec2 run-instances** to create an EC2[4] instance. For example:

```
$ aws ec2 run-instances --image-id ami-d440a6e7
    --instance-type t2.nano --associate-public-ip-address
    --key-name admin-key
```

- The base system image:
  an Amazon-supplied version of CentOS 7,
  named **ami-d440a6e7**
  – ami: Amazon Machine Images
  – use **aws ec2 describe-images** to decode the image id
- The instance type: **t2.nano** (the smallest instance type)
  – one CPU core and 512 MiB of RAM
- **--associate-public-ip-address**
  – allows reaching the EC2 instance directly from the Internet;
  – otherwise, by default, only accessible from other systems within the same virtual private cloud (VPC)
- A preconfigured key pair is assigned to control SSH access

[4] EC2: Elastic Compute Cloud

Example: AWS (creating an EC2 instance – output)


```
{
    "OwnerId": "188238000000",
    "ReservationId": "r-83a02346",
    "Instances": [
        ...
        "PrivateIpAddress": "10.0.0.27",
        "InstanceId": "i-c4f60303",
        "ImageId": "ami-d440a6e7",
        "PrivateDnsName": "ip-10-0-0-27.us-west-2.compute.internal",
        "KeyName": "admin-key",
        "SecurityGroups": [
            {
                "GroupName": "default",
                "GroupId": "sg-9eb477fb"
            }
        ],
        "SubnetId": "subnet-ef67938a",
        "InstanceType": "t2.nano",
    ...
}
```


- "security group" refers to firewalls in EC2, `default` means no access

Example: AWS (Stopping and terminating instances)

- Can **stop** the instance for shutting down but retaining it for later use; or
- Can **terminate** it to delete the instance entirely
  – A terminated instance can never be resurrected

```
$ aws ec2 stop-instances --instance-id i-c4f60303
{
    "StoppingInstances": [
        {
            "InstanceId": "i-c4f60303",
            "CurrentState": {
                "Code": 64,
                "Name": "stopping"
            },
            "PreviousState": {
                "Code": 16,
                "Name": "running"
            }
        }
    ]
}
```

Cost

Cloud tariffs generally consist of the followings:

- The compute resources of virtual private servers, load balancers, and whatever consumes CPU cycles to run the customer's services
  – charge by the hour
- Internet data transfer, and traffic among zones and regions
  – Charge by the GiB/TiB transferred
- Storage
  – Charge by the GiB/TiB stored per month