Introduction to data centers

Paolo Giaccone

Notes for the class on "Switching technologies for data centers"

Politecnico di Torino

September 2021

Outline

Section 1

Cloud Computing

Scenarios for data centers

Applications

- cloud computing
- cloud storage
- web services

Consolidation of computation and network resources

Very large data centers

- 1 000 - 10 000 - 100 000 - 1 000 000 servers

Cloud computing

USA National Institute of Standards and Technologies (NIST) definition

"Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."

http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

Cloud computing services

Models

- SaaS
- PaaS
- IaaS

Software as Service (SaaS)

Provides on-demand applications over the Internet. No control on network, servers, OS, individual application capabilities, etc. Accessible usually through a web browser. The state is maintained when accessing the application from different devices.
E.g., Office 365, Google Docs, Gmail, Dropbox, iCloud.

Cloud computing services

Platform as Service (PaaS)

Provides platform layer resources, e.g. operating system support and software development frameworks, to develop and to deploy applications over the Internet.
E.g., Google App Engine (Go, Java, Python, PHP), Microsoft Windows Azure (C#, Visual Basic, C++)
E.g., Amazon Elastic Map Reduce (AWS EMS):

- the user loads the data, writes the code through the provided API, specifies number of nodes (degree of parallelism) and then runs

Infrastructure as Service (IaaS)

Provides on-demand infrastructural resources, usually in terms of Virtual Machines
E.g., Amazon Elastic Compute Cloud (EC2), Microsoft Windows Azure, Google Compute Engine

Virtual machines (VM)

Possible scenarios

- OS image (e.g., Linux distribution)
- LAMP image (Linux + Apache + MySql +PHP)

Implementation

10-100 VMs on the same server, with their own IP and MAC address

VM migration

Migrate the entire VM state to

- achieve resource consolidation by turning off unused servers
- achieve load-balancing / statistical multiplexing
- route requests to servers with better bandwidth towards the clients
- avoid heat hot-spots
- adapt to different power availability/costs

Section 2

5G Networks

5G challenges and technologies

Expected performance

- per-user throughput: 100 Mbit/s
- peak data-rate: 20 Gbit/s
- mobility: 500 km/h
- latency: 1 ms

(Some) Key enabling technologies

- application: cloud-native verticals
- network: Software Defined Network (SDN) and NFV (Network Function Virtualization)
  - cloud to control the network and support any network function
- communication: 5G NR (New Radio) as access technology

Cloud-native verticals

Definition

"Cloud-Native is the name of an approach to designing, building and running applications/virtual functions fully exploiting the cloud delivery model" (taken from 5G-PPP Software Network Working Group)

Why?

- speed: to move quickly and get ideas to market fast
- scale: to support more users, in more locations, with a broader range of devices
- margin: pay for resources only if needed

Legacy 4G verticals

Verticals based on "legacy" technologies will need to adapt or migrate towards newer technologies as 5G systems mature.

Multi-access Edge Computing (MEC)

MEC in 5G

To reduce the latency, cloud-computing capabilities are available at the edges of the cellular network (e.g., in the cellular base station)

Benefits

- no need to reach a remote cloud
  - reduce the latency
  - reduce the network congestion
- network operator becomes PaaS/IaaS provider

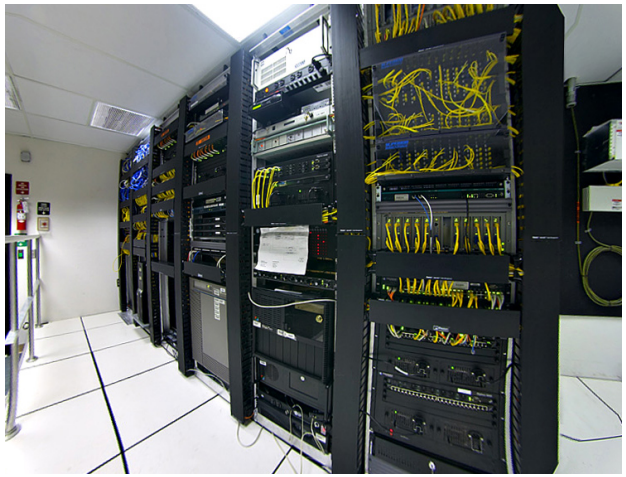
Section 3

Data Centers

Data center enables cloud computing

Data center

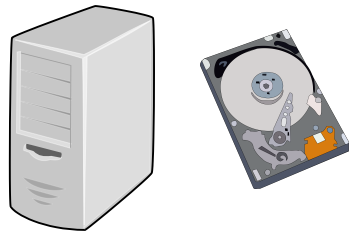
- Set of all the physical infrastructure required to support a cloud computing service
- The whole infrastructure is co-located either in a room, or in a building, or in a set of adjacent buildings



(Some) Basic ingredients

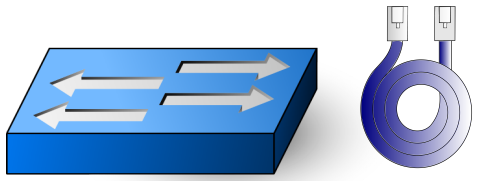
Computing and storage resources

- servers
- harddisk, SSD, ...



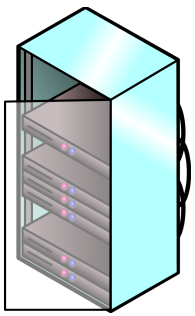
Communication resources

- Switches, routers
- Traffic balancers, firewalls, etc.
- Copper wires and fibre optics

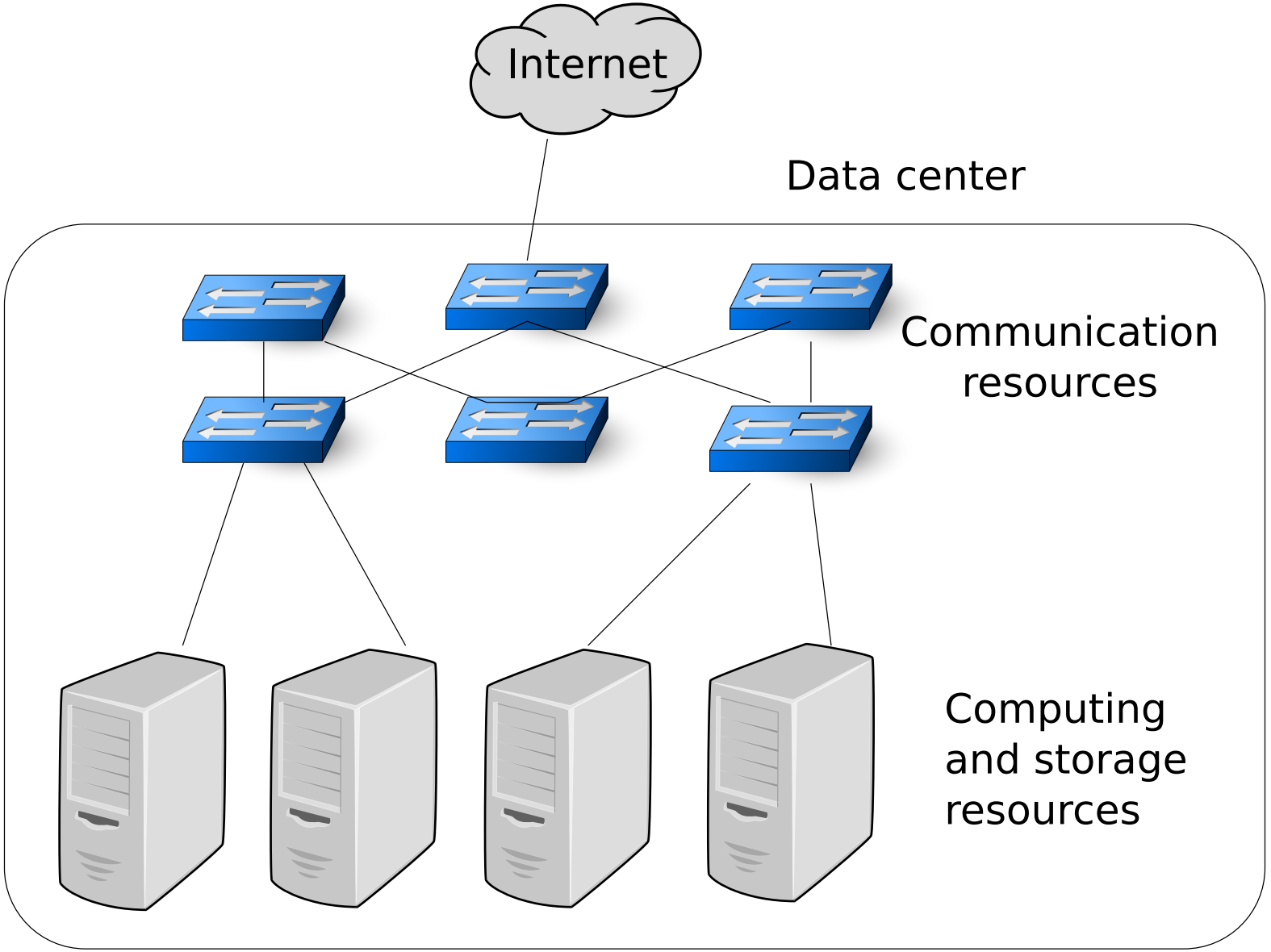


Infrastructure

- Racks
- Cooling and energy supply systems



Logical view of a data center



The design of the data center network

Design principles

- very scalable in order to support a very large number of servers
- minimum cost in terms of basic building blocks (e.g., switches)
- modular to reuse simple basic modules
- reliable and resilient
- may exploit novel/proprietary technologies and protocols not compatible with legacy Internet