Computing Infrastructures
Theory

Christian Rossi

Academic Year 2023-2024

Abstract

The course topics are:

- Hardware infrastructure of datacenters:
  - Basic components, rack structure, cooling.
  - Hard Disk Drive and Solid State Disks.
  - RAID architectures.
  - Hardware accelerators.
- Software infrastructure of datacenters:
  - Virtualization: basic concepts, technologies, hypervisors and containers.
  - Computing Architecture: Cloud, Edge and Fog Computing.
  - Infrastructure, platform and software-as-a-service.
- Methods:
  - Scalability and performance of datacenters: definitions, fundamental laws, queuing network theory basics.
  - Reliability and availability of datacenters: definitions, fundamental laws, reliability block diagrams.

Contents

CHAPTER 1

Introduction

1.1 Computing infrastructures

Definition (*Computing infrastructure*)**.** Computing infrastructure refers to the technological framework comprising hardware and software components designed to facilitate computation for other systems and services.

Data centers encompass servers tailored for diverse functions:

- *Processing Servers.*
- *Storage Servers.*
- *Communication Servers.*

1.1.1 Virtual machines

Virtual machines offer a comprehensive stack comprising an operating system, libraries, and applications. Applications rely on a guest operating system.

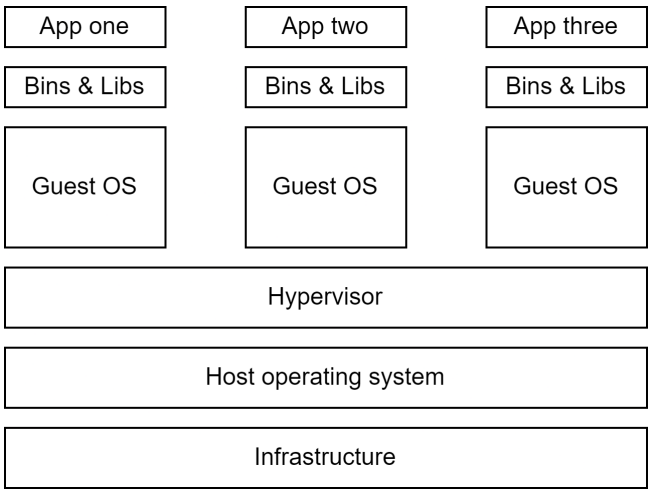


Figure 1.1: Virtual machine's structure

In this configuration, each operating system perceives the hardware as exclusively dedicated to itself. Virtualization results in significant power efficiency gains by consolidating the power consumption of individual machines, typically saving around 60%. Virtualization enables hot swaps, delivering two key advantages:

1. *Maintainability.*
2. *Availability*: overloaded machines can be supplemented by others.

1.1.2 Containers

Containers encapsulate applications along with their dependencies into a uniform unit for streamlined software development and deployment.

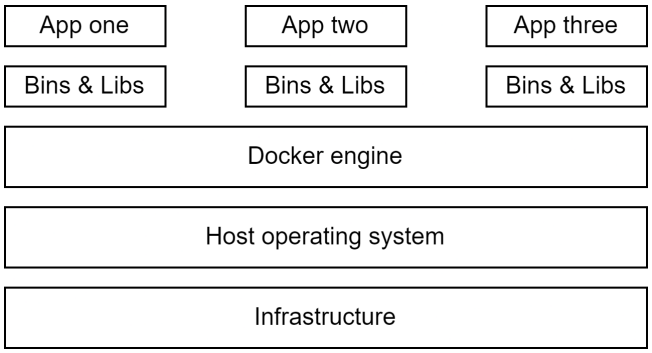


Figure 1.2: Container's structure

Containers are employed to execute specific services and offer a lighter alternative to virtual machines.

1.1.3 Summary

Data centers offer several advantages, including reduced IT costs, enhanced performance, automatic software updates, seemingly limitless storage capacity, improved data reliability, universal document accessibility, and freedom from device constraints. However, it also presents challenges such as the need for a stable internet connection, poor compatibility with slow connections, limited hardware capabilities, privacy and security concerns, increased power consumption, and delays in decision-making.

1.2 Edge computing systems

Edge computing is a distributed computing model in which data processing occurs as close as possible to where the data is generated, improving response times and saving on bandwidth. Processing data near the location where it is generated brings significant advantages in terms of processing latency, reduced data traffic, and increased resilience in case of data connection interruptions. Edge computing systems can be categorized as follows:

- *Cloud*: providing virtualized computing, storage, and network resources with highly elastic capacity.

- *Edge servers*: utilizing on-premises hardware resources for more computationally intensive data processing.
- *IoT and AI-enabled edge sensors*: enabling data acquisition and partial processing at the edge of the network.

Edge computing offers several advantages, including high computational capacity, distributed computing capabilities, enhanced privacy and security, and reduced latency in decision-making. However, it comes with drawbacks such as the requirement for a power connection and dependence on a connection with the Cloud.

1.2.1 Embedded PC

An embedded system refers to a computer system that comprises a computer processor, computer memory, and input/output peripheral devices, all serving a specific function within a larger mechanical or electronic system. Advantages of this approach include its pervasiveness in computing, high-performance units, availability of development boards, ease of programming similar to personal computers, and the support of a large community. On the other hand, it has disadvantages such as relatively high power consumption and the necessity for some hardware design work to be done.

1.2.2 Internet of things

The internet of things (IoT) encompasses devices equipped with sensors, processing capabilities, software, and other technologies. These devices are designed to connect and exchange data with other devices and systems over the internet or other communication networks. Advantages of IoT devices include their high pervasiveness, wireless connectivity, battery-powered operation, low costs, and their ability to sense and actuate. However, these devices also come with several disadvantages, such as their low computing ability, constraints on energy usage, limitations in memory (RAM/FLASH), and difficulties in programming them.

CHAPTER 2

Data centers

2.1 Introduction

Over the past few decades, there has been a significant shift in computing and storage, transitioning from PC-like clients to smaller, often mobile devices, coupled with expansive internet services. Concurrently, traditional enterprises are increasingly embracing cloud computing.

This shift offers several user experience improvements, including ease of management and ubiquitous access.

For vendors, Software-as-a-Service (SaaS) facilitates faster application development, making changes and improvements easier. Moreover, software fixes and enhancements are streamlined within data centers, rather than needing updates across millions of clients with diverse hardware and software configurations. Hardware deployment is simplified to a few well-tested configurations.

Server-side computing enables the swift introduction of new hardware devices, such as hardware accelerators or platforms, and supports many application services running at a low cost per user. Certain workloads demand substantial computing capability, making data centers a more natural fit compared to client-side computing.

2.1.1 Warehouse-scale computers

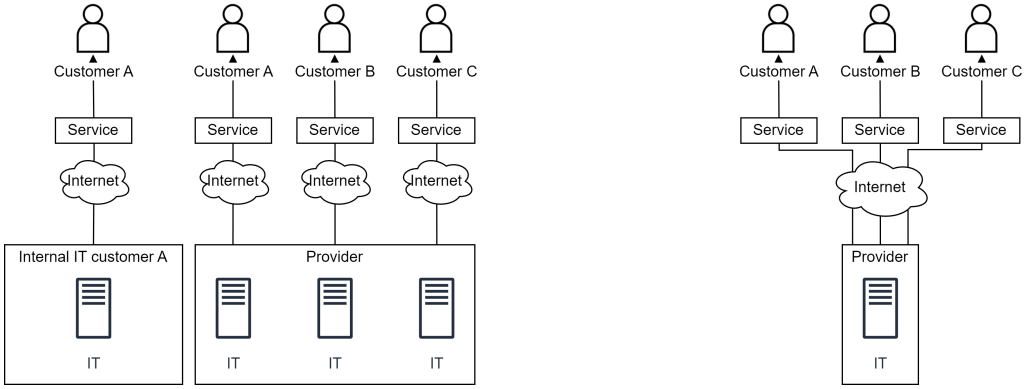
The rise of server-side computing and the widespread adoption of internet services have given rise to a new class of computing systems known as warehouse-scale computers (WSCs). In warehouse-scale computing, the program:

- Operates as an internet service.
- Can comprise tens or more individual programs.
- These programs interact to deliver complex end-user services like email, search, maps, or machine learning.

Data centers are facilities where numerous servers and communication units are housed together due to their shared environmental needs, physical security requirements, and for the sake of streamlined maintenance. Traditional data centers typically accommodate a considerable number of relatively small- or medium-sized applications. Each application operates on a

dedicated hardware infrastructure, isolated and safe guarded against other systems within the same facility. These applications typically do not communicate with one another. Moreover, these data centers host hardware and software for multiple organizational units or even different companies.

In contrast, warehouse-scale computers are owned by a single organization, employ a relatively uniform hardware and system software platform, and share a unified systems' management layer.



(a) Data center (b) Data warehouse

Figure 2.1: Structures of data centers and data warehouses

Warehouse-scale computers operate a reduced quantity of highly expansive applications, often internet services. Their shared resource management infrastructure affords considerable deployment flexibility. Designers are driven by the imperatives of homogeneity, single-organization control, and cost efficiency, prompting them to adopt innovative approaches in crafting WSCs.

Originally conceived for online data-intensive web workloads, warehouse-scale computers have expanded their capabilities to drive public cloud computing systems, such as those operated by Amazon, Google, and Microsoft. These public clouds do accommodate numerous small applications, resembling a traditional data center setup. However, all these applications leverage virtual machines or containers and access shared, large-scale services for functionalities like block or database storage and load balancing, aligning seamlessly with the WSC model.

The software operating on these systems is designed to run on clusters comprising hundreds to thousands of individual servers, far surpassing the scale of a single machine or a single rack. The machine itself constitutes this extensive cluster or aggregation of servers, necessitating its consideration as a single computing unit.

2.1.2 Geographical distribution of data centers

Frequently, multiple data centers serve as replicas of the same service, aiming to reduce user latency and enhance serving throughput. Requests are typically processed entirely within one data center.

Definition (*Geographic area*)**.** Geographic areas partition the world into sectors, each defined by geopolitical boundaries.

Within each geographic area, there are at least two computing regions. Customers perceive regions as a more detailed breakdown of the infrastructure. Notably, multiple data centers within the same region are not externally visible. The perimeter of each computing region

is defined by latency (with a round trip latency of two milliseconds), which is too far for synchronous replication but sufficient for disaster recovery.

Definition (*Availability zone*)**.** Availability zones represent more granular locations within a single computing region.

They enable customers to operate mission-critical applications with high availability and fault tolerance to datacenter failures by providing fault-isolated locations with redundant power, cooling, and networking. Application-level synchronous replication among availability zones is implemented, with a minimum of three zones being adequate for ensuring quorum.

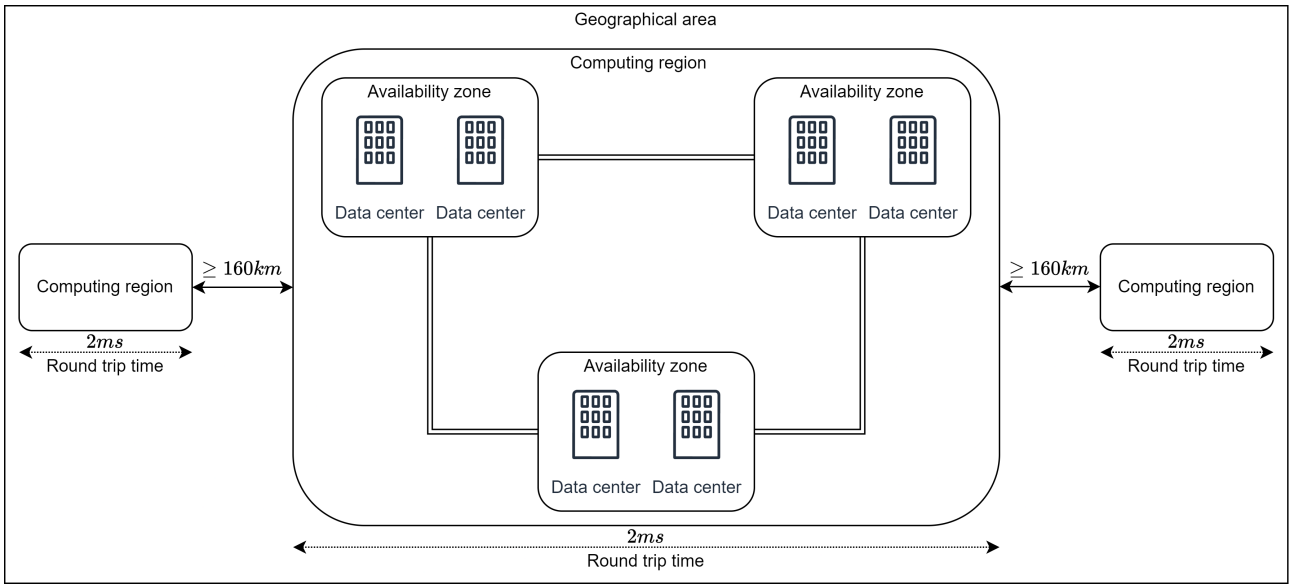


Figure 2.2: Geographical area structure