Compute: Background

In this section we will discuss fundamental concepts related to compute resources, their nature and their evolution throughout the years. Below is a brief overview of the topics to be covered:

- Compute services, and the hardware resources associated with it
- Virtualization, its purpose and financial impact
- Serverless, its advantages and disadvantages

If the reader believes themselves to already be familiar with the listed concepts, then they can feel free to skip this section and move on to the rest of the chapter.

Compute and related infrastructure

The resources that provide the computer system with processing power and are responsible for the running of applications are called compute resources. Whether it be on-prem or on the cloud, compute resources refer to three components:

- CPU (Central Processing Unit): Often called the "brain" of the computer, the CPU performs calculations and processes instructions required by applications. It handles all logical and arithmetic operations, controlling how quickly and efficiently tasks are executed.
- RAM (Random Access Memory): Already discussed in the storage chapter, RAM is the temporary holding space for information related to the programs currently being run on the computer.
- GPU (Graphics Processing Unit): Used to handle complex parallel-computing tasks like matrix and vector operations, heavily utilized in specific use cases such as gaming, cryptographic and AI operations.

Therefore, when we rent compute resources from a cloud provider what we are really renting is the CPU and the RAM, with us generally being charged based on the sophistication of the CPU Chip and the associated memory capacity.

Virtualization

Now, a layman might assume that separate physical servers would be allocated for every prospective renter but that is actually not the way cloud providers facilitate us with compute resources. In reality, cloud providers utilize **virtualization**, a technology that divides the underlying compute hardware into abstract partitions, and creating a **Virtual Machine (VM)** based off it, a virtual computer based on the partitioned resources.

Each VM has its own OS and computing resources, and operates independently of the other VMs, even if they were abstracted from the same underlying hardware enabling multiple users to share share a single physical server while maintaining the illusion of individual dedicated servers. Indeed, virtualization makes the leasing of compute resources a much more profitable venture for the cloud providers than it otherwise would be, allowing cloud providers to maximize the mileage they get out of their physical servers.

Serverless

Now, virtual machines have been around since the 1970s but there has been a major development in the field of managing compute and cloud computing in the last 10-15 years or so: **Serverless computing**.

Serverless allows us to run applications without having to manage any of the infrastructure ourselves, with the cloud providers handling the allocation of resources required by our workload. This allows us to focus only on the development of the application and not worry about operating the servers that it runs on. However, there are benefits to having control over the underlying infrastructure, it allows us to be in complete control of the security and configuration of the resources, a level of control taken away from us in the serverless faustian bargain.