



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

WIP Title: Dynamic resource allocation in the cloud for compute heavy tasks in a containerized environment

TESI DI LAUREA MAGISTRALE IN
COMPUTER SCIENCE AND ENGINEERING

Author: **Elia Ravella**

Student ID: 967243

Advisor: Prof. Raffaella Mirandola

Co-advisors: Name Surname, Name Surname

Academic Year: 2021-22

Abstract

Keywords: Cloud, Containers, Dynamic infrastructure

Contents

Abstract	i
Contents	iii
1 The Problem and the Current Available Solutions	1
1.1 Introduction	1
1.2 The Problem	1
1.3 Containerized Environment and High Performance Computing	2
1.4 State of the Industry	3
1.4.1 SLURM	4
1.4.2 Shifter	4
1.4.3 Kubernetes	4
1.4.4 Serverless Approach	5
2 Design and Testing Phase	7
2.1 MapNCloud Original Architecture	7
2.1.1 System Design	7
2.2 Problems Addressed	8
2.3 Testing and Validation	9
3 Implementation	11
3.1 Frontend	11
3.2 Backend	11
3.3 Database	11
3.4 Messaging Middleware	11
3.5 Computational Layer	11
3.5.1 Renderino	11

Bibliography

13

1 | The Problem and the Current Available Solutions

1.1. Introduction

1.2. The Problem

Scalability is one of the key design point that must be taken into consideration when developing a software. If a system cannot scale in power when the userbase or the load requested changes it slows down, making the response times growing for each request and compromising the overall performance of the application. The most intuitive approach to scalability, which is also the most common in cloud environments, is horizontal replication. With horizontal replication I mean (throughout all this thesis) the addition of identical software modules alongside the already existing ones to share the load; to do so, different incoming requests are routed to different modules when they arrive. The replicas being identical (and usually stateless) ensure that each request is carried out in the same way. The policies for deploying replicas can be either static or dynamic (based on the predicted load during the day or measuring the real time traffic incoming, for example).

This approach has seen a wide adoption in the industry and is the *de facto* standard to tackle scalability problems, especially in web environments. The horizontal replication approach gives applications the flexibility they need in reacting to the load that is applied, and is especially effective when the application is divided in submodules that can be individually scaled.

What happens when the requests that an application must serve change also in nature, and not only in volume? Horizontal replication works well when the load is mostly uniform (and for the majority of web applications, it is) and can be analyzed in a one dimensional fashion as "the number of requests". When requests set in motion heavy computational

pipelines, as image processing or complex mathematical problems, but the interface they are served is shared with all the other *light* requests¹ then a single request can weight, in terms of resources it needs to be carried out, very differently from the others. In this scenario, horizontal replication is harder to put in place effectively: if requests weight differently it is not possible to just share them equally among replicated servers² because in some cases a server will receive a much more higher share of heavy requests and be stuck executing them while other servers will be idling because they received only light requests. Horizontal replication, as it is implemented now, cannot face efficiently this scenario. The proposed solution uses dedicated ephemeral workers to execute *heavy* requests, and organises and schedules them with a ticketing middleware.

1.3. Containerized Environment and High Performance Computing

Containerization Virtualization is "the act of creating a virtual (rather than actual) version of something at the same abstraction level, including virtual computer hardware platforms, storage devices, and computer network resources"[2]. In cloud environments virtualization is the most used tool to provide isolated services, as compute capabilities, storage capabilities or networking. There are different kinds of virtualization, depending on the layer virtualized: some solutions just virtualize the hardware and let the user install a full fledged OS over it, other virtualizes all the technological stack from the hardware to the OS level, leaving to the user (in most of the cases, a developer) only the problem of developing the application s/he wants to ship, without the need to care about hardware limitations or operative systems settings.

One of the most used virtualization technique in the industry is containerization: when a software is said to be *containerized* it is packaged in a format that encapsulates also all its dependencies, configuration files and variables and OS settings. This package, the container, must be run through a container engine which provide the communication with the underlying software and OS to execute the software. Famous container engines are Docker, Podman, chroot or rkt. Containerized applications have several benefits over classic deployed ones:

- They do not rely on the machine or OS they are running, just the engine

¹this is not an impossible scenario: what a REST API exposes are a list of "light" HTTP requests that can trigger all kind of operations on the server they are executed on

²here *servers* is used to describe logical backend modules, not physical machines

- If the container has been built correctly, the configuration of the application is already done
- Packaging all their dependencies makes them independent from the other application installed on the system: two containerized applications running on the same engine could be using the same library but at different (and even incompatible) versions
- The container acts also as an isolation mechanism, that keeps the application from interacting with other systems: this improves the security of the container applications

If an application is containerized, moreover, it is very easy to deploy several instances of the same application working at the same time: this kind of deployments improve scalability and reliability of the application.

High Performance Computing With "high performance computing" is usually addressed the field in computer science that studies computational heavy problems and develops solutions to solve them via techniques as extreme parallelism, clusterization and high performance networking. In the early stages of this project I focused also on HPC in order to understand if a virtualization approach as the containers one is suitable for such kind of tasks. The problem was the additional layer of virtualization added by the container engine: since most of the times HPC software relies heavily on low-level procedure calls, the additional virtualized layer could degrade too much the performance. However, as stated in the paper "Exploring the support for high performance applications in the container runtime environment"[3] (which original aim was to compare native performance with containerized ones) the optimization of engines is reaching a level that can offer near-native performance for HPC-like loads, at least in not-extreme scenarios.

1.4. State of the Industry

There are several solutions available on the market that provides flexible infrastructure management and are built to automate the management of the infrastructure of an application. These systems can be seen as schedulers (so software components that organize when a task is executed and on which resources) with some additional features as the capability of actually *allocate* the resources needed or the automated management of the interfaces between resources and components.

1.4.1. SLURM

SLURM is "an open source, fault-tolerant, and highly scalable cluster management and job scheduling system for large and small Linux clusters"[1]. It aims to organize and schedule tasks on multiple nodes; these tasks can also be defined as OCI-compliant containers. SLURM was created to be executed on supercomputers or clusters of computers; in fact, SLURM focuses on communication between daemons and tasks through message passing framework as MPI and queue-managed resource access. Even if it can integrate containerized workloads, it is not suited to be deployed in a cloud environment rather than a computation center.

1.4.2. Shifter

Shifter is a simple scheduler which aims to utilize the container format in an HPC environment. It allows the user to specify the load in a docker image, then Shifter automates the conversion of that image to an HPC format and the scheduling of such task. Shifter is *not* an extension of Docker or the Docker engine, nor aims to automate the infrastructure, instead it just provides an additional interface (which is container compatible) to an already existing HPC platform.

1.4.3. Kubernetes

Kubernetes (often called "k8s") is a container orchestration system, and is the *de facto* standard for container orchestration. It provides an all-in-one system to manage containerized applications:

- It provides an abstraction over the container level (the Pod) that is used to define the service provided rather than the container itself
- It includes different ways to persist data and state across containers; this gives a kubernetes cluster the capability to hold an entire application, from data layer to presentation layer
- Kubernetes clusters embeds security and access control by enforcing the already existing isolation features of a containerized environment and by providing a set of tools that easily control the access to the cluster itself (the Ingress controllers)
- It allows developers to define the redundancy for every single service defined, so to set an "horizontal scaling width" beforehand to handle faults and heavy loads

Kubernetes has been designed for applications that are designed as microservices, and it

provides the tooling for administrate single services, replication, and scaling in such an environment. As already stated in (1.2) the level of abstraction provided by kubernetes is efficient when horizontal scaling must be automated, but also removes some of the controllability of a containerized system. Moreover, resource management within kubernetes deployment in the cloud is very difficult: all the needed resources (as computational resources, memory resources, storage and networking ones) must be available at any time so that kubernetes can manage them, since the available k8s deployments at the moment are coupled to the underlying virtual machine. As an example, on the major cloud service providers, during the selection of the resources to allocate to a k8s cluster it is explicitly required to select a virtual machine size which will be the host of the deployment; this allows cloud providers to fix the resources available to each deployment. This also forces said resources pool to be always active, and in the case of a heavy deployment this drives up the costs.

1.4.4. Serverless Approach

Functions-as-a-Service is a cloud computing execution model that allocates computing resources on demand. Popular FaaS services include AWS Lambda, Azure Functions and GCP Cloud Functions. This execution model can be seen as a "nanoservice" approach, where an application is further divided into single function calls rather than services, but it has a very operational connotation: the term "serverless" is usually used to describe a type of *application deployment* that minimizes code deployment time, since the whole infrastructure is completely abstracted and managed by the cloud provider. The problem with a serverless approach is the actual computing power available: let alone the high level of abstraction (which again removes some control over the infrastructure and the actual resources allocated) the currently available solutions all impose limits both in terms of power usage, measured in CPU allocated, and in runtime: this renders impossible to use such an approach in the scenario described.

2 | Design and Testing Phase

2.1. MapNCloud Original Architecture

This thesis project extends and develops the original Distributed Software Development project, which have been created during the first semester of the 2021 / 2022 academic year. The original project aim was to create a system which could offer photogrammetry services, so the rendering of a tridimensional model from a set of images of the subject, leveraging the container technology to encapsulate the rendering engine. In this section I describe the original architecture of that system and in the next ones I will explain how that architecture has been redesigned, adapted and extended in this thesis project.

2.1.1. System Design

The MapNCloud project was designed as a "three plus one" tiers application:

- a **presentation layer** or front end, which has been developed as a web site and offers a graphical interface to the system
- a **business layer** or back end, which embodies all the business logic of the application
- a **data layer**, the database which persists the data needed by the application
- a **computation layer**, which is a separated back end module that is tasked with the most computational heavy routines (as the photogrammetry pipelines)

Another module is present in the application, and it is the ticketing service. It is not considered a whole tier, in fact its only job is to manage the queues of tasks that the backend generates, and let the computational layer components access them in the right order with the right schedule.

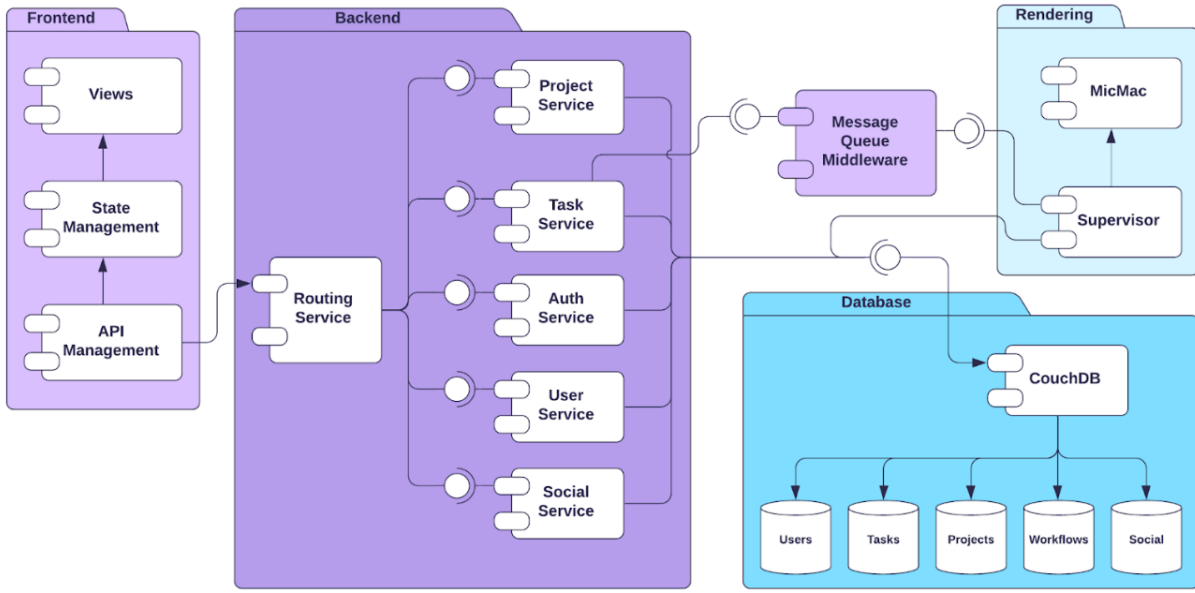


Figure 2.1: Original MapNCloud component diagram, with internal modules detailed.

Front End

The presentation layer of the application was developed as a web application; this exposes a graphical user interface for the services offered by the system. It is a JavaScript application (developed with the VueJS framework) which interacts with the backend via HTTP calls; this allows it to be considered, by the cloud provider, a static web application, which is easy to deploy and has a low resources demand.

The front end architecture follows the Model View ViewModel pattern, which separates the actual GUI elements from the logic of the user experience in View and ViewModel respectively, and lets the ViewModel reflect the changes on the Model; this is a common pattern to adopt as it provides a clear organization of functionalities and responsibilities among the internal software submodules.

This module of the application also uses a Continuous Integration / Continuous Deployment pipeline implemented with GitHub actions which deploys the changes directly in the hosting service.

2.2. Problems Addressed

1. database choice and API modification
2. queue monitoring
3. resizable backend containers

4. cloud provider integration

At the end of this section I will present the "final" design draft

2.3. Testing and Validation

HERE I will introduce the "diffusion analysis" to justify the test parameters

1. CouchDB testing
2. RabbitMQ testing
3. Cloud providers options, pros and cons
4. technological limitations (docker-compose, load balancers)

I will also present the real "final" Architecture that will be deployed here, with cloup provider's technological names and services

3 | Implementation

3.1. Frontend

3.2. Backend

3.3. Database

3.4. Messaging Middleware

3.5. Computational Layer

3.5.1. Renderino

Bibliography

- [1] Slurm documentation. URL <https://slurm.schedmd.com/containers.html>.
- [2] Virtualization — Wikipedia, the free encyclopedia. URL <https://en.wikipedia.org/w/index.php?title=Plagiarism&oldid=5139350>. [Online; accessed 28-July-2022].
- [3] J. P. Martin, A. Kandasamy, and K. Chandasekaran. Exploring the support for high performance applications in the container runtime environment. 2017.