



END OF STUDIES MASTER'S PROJECT

MASTER'S THESIS

Simulation of a Kubernetes Cluster with Validation in Real Conditions

Author

Théo LARUE

Evaluator

Surname NAME

FEBRUARY 24TH TO AUGUST 7TH 2020

Contents

1	Introduction / Abstract	2
2	Kubernetes	3
2.1	Kubernetes overview	3
2.2	HPC and Kubernetes	5
3	The scheduling problem	7
3.1	Batch scheduling	7
3.1.1	A general definition	7
3.1.2	Kubernetes batch schedulers	7
3.2	Simulating infrastructures	7
3.2.1	HPC simulators	7
3.2.2	Kubernetes simulation	7
3.3	The Batsim infrastructure simulator	8
3.3.1	Batsim concepts	8
3.3.2	Limitations	8
4	Problematic	9
4.1	Objectives	9
4.2	Translation	9
4.3	Synchronization	9
5	Implementation	10
5.1	Batkube architecture	10
5.2	API implementation	10
5.3	Time hijack	10
5.3.1	batsky-go	11
6	Evaluation	12
7	Conclusion	13

Introduction / Abstract

TODO

Kubernetes

2.1 Kubernetes overview

Cloud Native Computing

In the early stages of application development, organizations used to run their services on physical servers. With this direct approach came many challenges that needed to be coped with manually like resources allocation, maintainability or scalability. In an attempt to automate this process developers started using virtual machines which enabled them to run their services regardless of physical infrastructure while having a better control over resources allocation. This led to the concept of containers which takes the idea of encapsulated applications further.

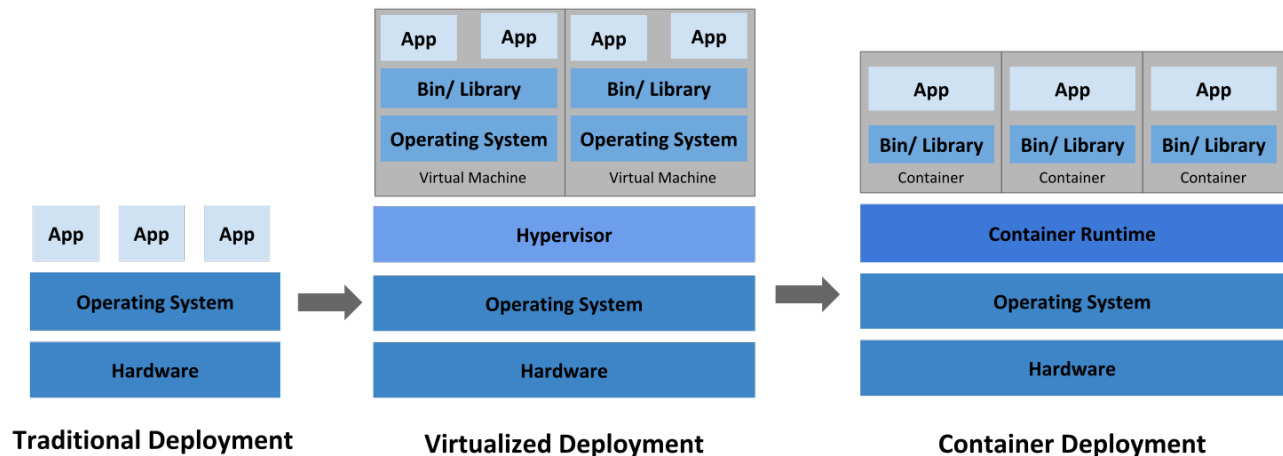


Figure 2.1: Evolution of application deployment.

Source: <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>

Containers can be thought of as lightweight virtual machines. Unlike the latter, containers share the same kernel with the host machine but still allow for a very controlled environment to run applications. There are many benefits to this : separating the development from deployment, portability, easy resource allocation, breaking large services into smaller micro-services or support of continuous integration tools (containers greatly facilitate integration tests).

The CNCF¹ (Cloud Native Computing Foundation) was founded in the intent of leveraging

¹<https://www.cncf.io/>

the container technology for an overall better web. In a general way, we now speak of these containerized and modular applications as cloud native computing :

“Cloud native technologies empower organizations to build and run scalable applications in modern, dynamic environments such as public, private, and hybrid clouds. Containers, service meshes, microservices, immutable infrastructure, and declarative APIs exemplify this approach.

These techniques enable loosely coupled systems that are resilient, manageable, and observable. Combined with robust automation, they allow engineers to make high-impact changes frequently and predictably with minimal toil.”²

Kubernetes³ is the implementation of this general idea and was announced at the same time as the CNCF. It aims at automating of the process of deploying, maintaining and scaling containerized applications. It is industry grade and is now the de-facto solution for container orchestration.

Kubernetes concepts

The basic processing unit of Kubernetes is called a **pod** which is composed of one or several containers and volumes⁴. In the cloud native context a pod most often hosts a service or micro-service.

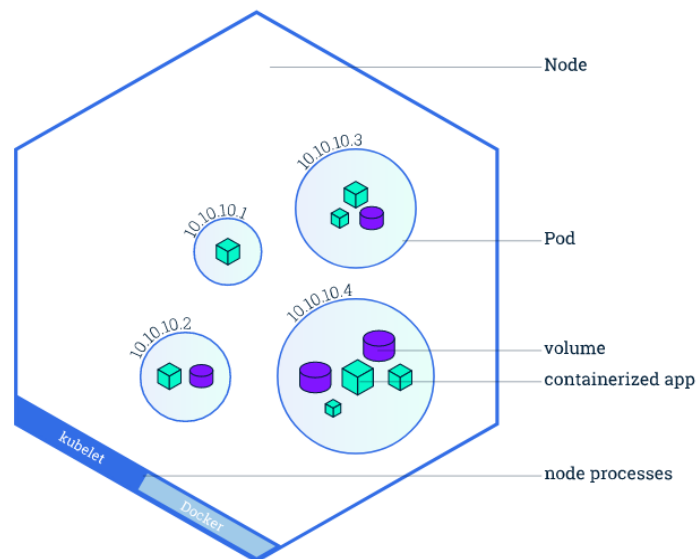


Figure 2.2: Node overview

Source: <https://kubernetes.io/docs/tutorials/kubernetes-basics/explore/explore-intro/>

Pods are bundled together in **nodes** (figure 2.2) which are either physical or virtual machines. They represent another barrier to pass through to access the outside world which can be useful to add layers of security or facilitate communication between pods. Nodes take the idea of containerisation further by encapsulating the already encapsulated services. Each node runs at least one pod and also one **kubelet** which is a process responsible for communicating with the rest of Kubernetes (or more precisely, with the master node which in turns communicates with the api

²<https://github.com/cncf/toc/blob/master/DEFINITION.md>

³<https://kubernetes.io/>

⁴A volume is some storage space on the host machine that can be linked to containers, so they can read persistent information or store data in the long term

server). A set of nodes is called a **cluster**. Each Kubernetes instance is responsible for running a cluster.

Kubernetes revolves its API server which is its central component (figure 2.3). The majority of operations between components go through this REST API like user interactions through `kubectl` or scheduling operations.

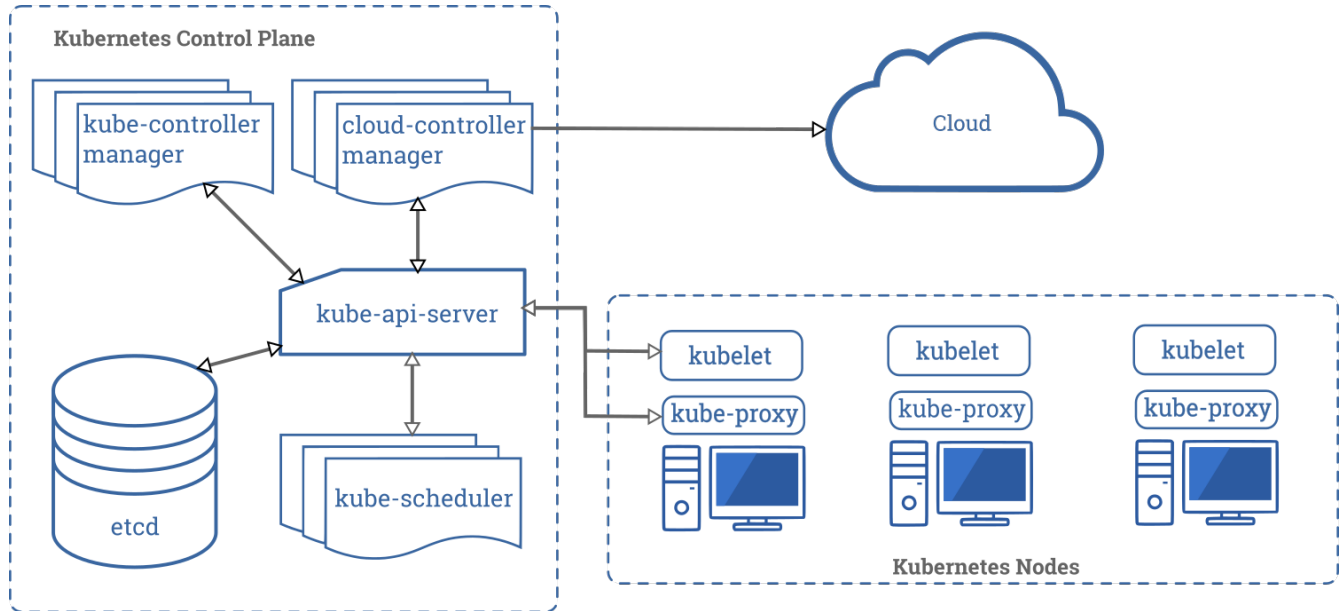


Figure 2.3: Components of Kubernetes

Source: <https://kubernetes.io/docs/concepts/overview/components/>

2.2 HPC and Kubernetes

The difference between HPC (High Performance Computing) and Kubernetes lies in the workloads they are intended to tackle.

A general definition of HPC would be : “*High Performance Computing (HPC) most generally refers to the practice of aggregating computing power in a way that delivers much higher performance than one could get out of a typical desktop computer in order to solve large problems in science, engineering, or business.*”⁵. HPC can either refer to “High Performance Computing” or “High Performance Computer” but it is generally clear which one it refers to, given the context.

A HPC workload is composed of numerous tasks hungry for computational resources which are executed in parallel on different machines (that we can refer to as compute resources or nodes). These tasks may be completely independant like when different users each submit a single task, or they may also be tightly coupled together as when a single user submits a job that is composed of several tasks than can be run in parallel. In that case, the whole system becomes very sensitive to latency as these tasks have got to communicate together. This is done through MPis (Message Passing Interfaces) which represent a large part of the HPC field.

⁵https://www.eni.uni.lu/university/high_performance_computing

Kubernetes is now the standard for AI and Machine Learning as shown by the many efforts at making this coupling an efficient environment[4][7][5], which brought an increasing interest for container driven HPC aswell and Kubernetes for HPC in particular. Batch schedulers such as kube-batch⁶ have been implemented for kube, and numerous HPC applications like slurm⁷ have been containerized aswell.

Indeed, containers have many advantages that HPC users can benefit from. Here are some notable ones:

- First off, research has shown that Kuberenetes offer similar performance to more standard bare metal HPC[2].
- Users will get the same environment everywhere making up for a uniform and standardized workplace.
- Portability : users could seamlessly hop from one infrastructure to another based on their needs and criteria like price, performance, and capabilities rather than compatibility.
- Encapsulation : HPC applications often rely on complex dependencies that can be easily concealed into containers.

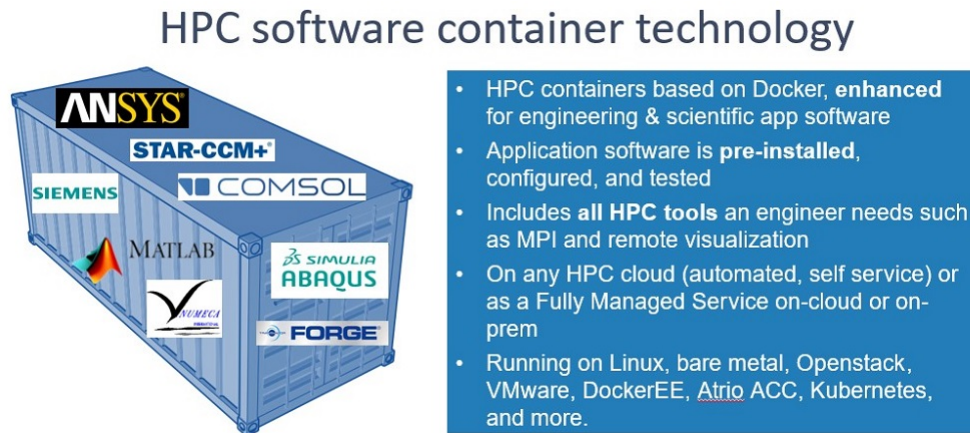


Figure 2.4: The container technology for HPC

Source: <https://www.hpcwire.com/2019/09/19/kubernetes-containers-and-hpc/>

Despite all those advantages, Kubernetes is not ready yet to be used in proper HPC environment because it lacks vital components like a proper batch job queuing system, and support for MPI applications. It cannot yet compete against the very well established HPC ecosystem, however efforts are being made in this direction. This master thesis is focused on a part of this problem : Kubernetes simulation for batch schedulers evaluation.

⁶<https://github.com/kubernetes-sigs/kube-batch>

⁷<https://slurm.schedmd.com/containers.html>

The scheduling problem

3.1 Batch scheduling

3.1.1 A general definition

3.1.2 Kubernetes batch schedulers

`kube-scheduler`

Not exactly a batch scheduler, it is the default scheduler for Kubernetes made by the CNCF

`kube-batch`

`Poseidon`

`bashScheduler` by rothgar

`random-scheduler` by Banzaicloud

`k8s-custom-scheduler` by IBM

`scheduler` by kelseyhightower

3.2 Simulating infrastructures

3.2.1 HPC simulators

3.2.2 Kubernetes simulation

This raises the question of scheduler development. Developing a scheduler implies being able to test its performances throughout the development process, however, testing in real conditions is time consuming and expensive. Organizations can either have enough resources to cover these costs, or test their scheduler against a simulation.

Kubernetes cluster simulations is an open problem and is the subject of this master project. Our approach relies on the Batsim[1] infrastructure simulator, which is itself built upon Sim-grid[6]. Batsim is currently mostly used to simulate HPC infrastructures but was designed to be able to simulate any kind of infrastructure and therefore is theoretically able to simulate any Kubernetes cluster, moreover, Kubernetes was designed to run services but is capable of handling High-Performance Computing[3]. This project aims at adapting Batsim so it can evaluate Kubernetes schedulers.

3.3 The Batsim infrastructure simulator

3.3.1 Batsim concepts

3.3.2 Limitations

Problematic

4.1 Objectives

The goal of this project is to design and implement Batkube, which will be an interface between Batsim and Kubernetes schedulers. With this interface, we want to compare Batsim results against data from a real Kubernetes cluster, given HPC workloads.

4.2 Translation

4.3 Synchronization

Implementation

5.1 Batkube architecture

TODO

5.2 API implementation

5.3 Time hijack

TODO

5.3.1 batsky-go

Algorithm 1: Requester loop

Input: req: request channel, res: result channel map

```
1 while Batkube is not ready do
2   | wait
3 requests = []request
4 while req is not empty do
5   | m = <- req /* Non blocking receive */
6   | requests = append(requests, m)
7 sendToBatkube(requests) /* Only requests with duration > 0 are actually sent.
   Batkube will always answer. */
8 now = responseFromBatkube()
9 for m in range requests do
10  | res[m.id] <-now /* The caller continues execution upon reception */
```

Algorithm 2: Time request (time.now())

Result: Current simulation time

Input: d: timer duration, req: request channel, res: response channel map

Output: now : simulation time

```
1 if requester loop is not running then
2   | go runRequesterLoop() /* There can only be one loop running at a time */
3 id = newUUID()
4 m = newRequestMessage(d, id) /* Requests are identified using uuids */
5 resChannel = newChannel()
6 res[id] = resChannel /* A channel is associated with each request */
7 req <- m /* The code blocks here until request is handled */
8 now = <-resChannel /* The code blocks here until response is sent by the
   requester loop */
9 return now
```

Evaluation

Conclusion

Bibliography

- [1] *Batsim docs*. URL: <https://batsim.readthedocs.io/en/latest/>.
- [2] A. M. Beltre et al. “Enabling HPC Workloads on Cloud Infrastructure Using Kubernetes Container Orchestration Mechanisms”. In: *2019 IEEE/ACM International Workshop on Containers and New Orchestration Paradigms for Isolated Environments in HPC (CANOPIE-HPC)*. 2019, pp. 11–20.
- [3] *Kubernetes Meets High-Performance Computing*. URL: <https://kubernetes.io/blog/2017/08/kubernetes-meets-high-performance/> (visited on 04/21/2020).
- [4] Mikyoung Lee, Sungho Shin, and Sa-Kwang Song. “Design on distributed deep learning platform with big data”. In: (2017).
- [5] Seetharami R. Seelam and Yubo Li. “Orchestrating Deep Learning Workloads on Distributed Infrastructure”. In: *Proceedings of the 1st Workshop on Distributed Infrastructures for Deep Learning*. DIDL ’17. Las Vegas, Nevada: Association for Computing Machinery, 2017, 9–10. ISBN: 9781450351690. DOI: 10.1145/3154842.3154845. URL: <https://doi.org/10.1145/3154842.3154845>.
- [6] *Simgrid official website*. URL: <https://simgrid.frama.io/>.
- [7] Boris Tvaroska. “Deep Learning Lifecycle Management with Kubernetes, REST, and Python”. In: Santa Clara, CA: USENIX Association, May 2019.