Development and evaluation of a Kubernetes cluster simulator based on Batsim

Presented by: Théo Larue
Supervised by: Olivier Richard & Michael Mercier

Université Grenoble Alpes

August 31, 2020

Table of contents

1. Introduction
2. Literature review
3. Integrating Kubernetes schedulers to Batsim
4. Study of the simulator
5. Discussion and future work

Introduction

Computer infrastructures

A distributed system is a system whose components are located on different networked computers, which communicate and coordinate their actions by passing messages to one another. [1]

[1] Andrew 2002

Computer infrastructures

A distributed system is a system whose components are located on different networked computers, which communicate and coordinate their actions by passing messages to one another. [1]

Many domains

Grid, Edge, HPC, Cloud, P2P, Volunteer.

[1] Andrew 2002

Studying distributed systems I

Why studying these infrastructures?

Studying distributed systems II

TODO One problem in particular: scheduling.

Different approches

How to study these infra?

Different approches

How to study these infra?

- Theoretical study.

Different approches

How to study these infra?

- Theoretical study.
- Real experiments.

Different approches

How to study these infra?

- Theoretical study.
- Real experiments.
- Emulation.

Different approches

How to study these infra?

- Theoretical study.
- Real experiments.
- Emulation.
- Simulation.

Contribution

TODO: Our contribution is Batkube, an interface between Batsim and Kubernetes schedulers

Literature review

Domain specific simulators

refs on domain specific simulators (summed up in a table?).
Explain briefly the concept behind some of them.

Software specific simulators

YARNSim, SLURM simulator

Publication specific simulators

“Publish and perish” - Milian Poquet

SimGrid

SimGrid: Versatile, scalable, accurate.
Cpu = a computation speed.
Storage = a seek time and a data transfert rate.
Network = a flow model, modeling bandwith sharing behaviors.
Simple models but thoroughly validated.

Batsim

Aimed at studying RJMS.
Strong decoupling decision process / simulator.

Related work

GridSim
Alea: modular, extensible.
Accasim: supports additional information (temperature, power consuption). Very efficient in terms of simulation time and memory usage.

Kubernetes

Explain containers real quick.
Container orchestration software, description based.

Kubernetes cluster simulation

k8s-cluster-simulator: open source, student project, delay jobs. Schedulers provided via a Go interface.

joySim: closed-source, fully fledged kubernetes cluster simulator, service oriented (mock nodes).

Integrating Kubernetes schedulers to Batsim

Technical challenges

Challenges to tackle

1. Integration with Kubernetes.

Technical challenges

Challenges to tackle

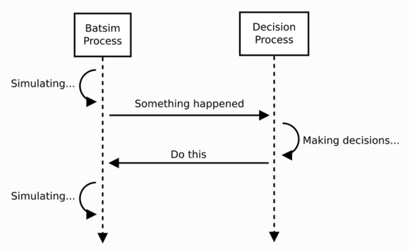
1. Integration with Kubernetes.
2. Intercepting scheduler time.

Technical challenges

Challenges to tackle

1. Integration with Kubernetes.
2. Intercepting scheduler time.
3. Time synchronization between Batsim and the scheduler.

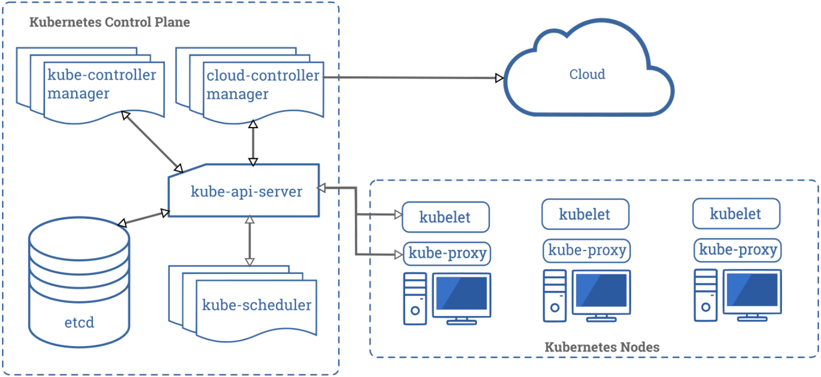
Batsim concepts



source `https://batsim.readthedocs.io`

Batsim events and protocol.
User defined workloads.
(insert json examples?)

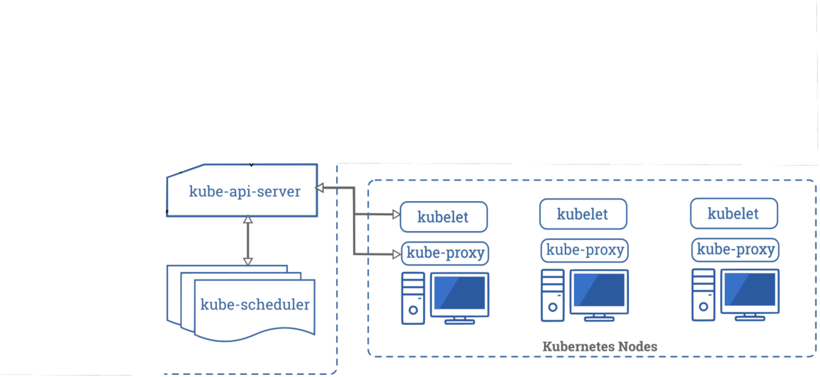
Kubernetes concepts



source: `https://kubernetes.io/docs/concepts/overview/components/`

Kubernetes components.

Kubernetes concepts



source: `https://kubernetes.io/docs/concepts/overview/components/`

Kubernetes components.

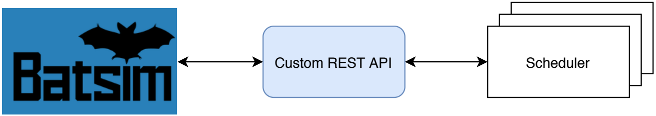
Different paradigms

Batsim: event based, simulation time.

Kubernetes scheduler: asynchronous calls to the API, machine time.

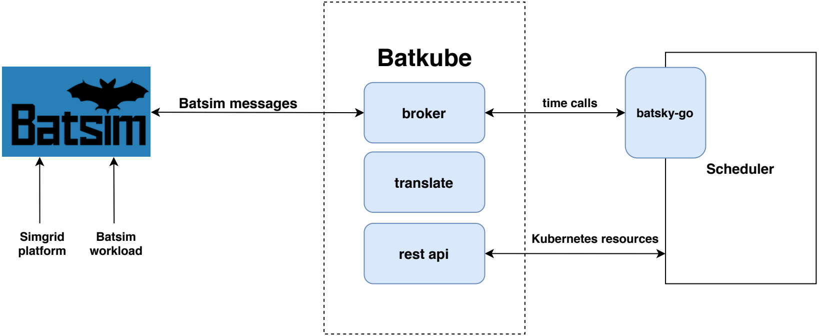
The goal is to make the scheduler event based and relying on simulation time for Batsim, and make Batsim a kube-api-server to the scheduler.

Batkube integration with Kubernetes



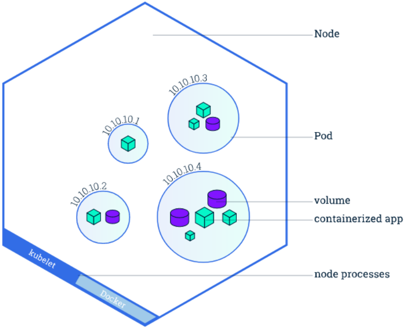
Reimplementation of a custom API.

Architeture of Batkube



Global architecture of Batkube.

Similar resources

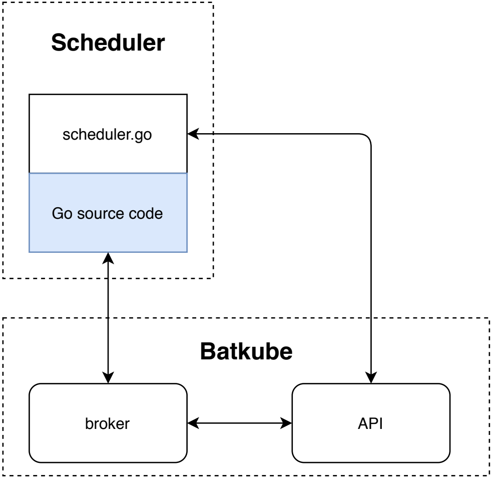


source: https://kubernetes.io/docs/tutorials/kubernetes-basics/explore/explore-intro/

Translation between Kubernetes and Batsim

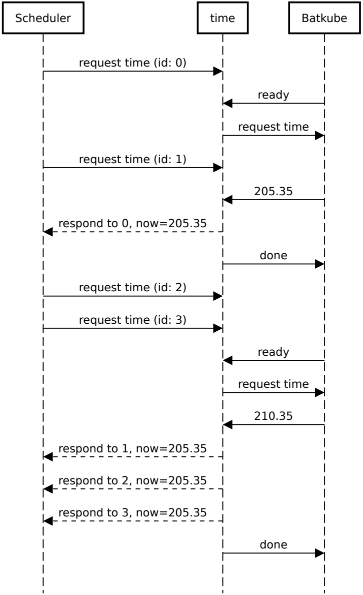
- A Pod = a job.
- A Node = a compute resource.

Time interception



Schedulers are patched to redirect their time.

batsky-go



Scheduler

time

Batkube

request time (id: 0)

ready

request time

request time (id: 1)

205.35

respond to 0, now=205.35

done

request time (id: 2)

request time (id: 3)

ready

request time

210.35

respond to 1, now=205.35

respond to 2, now=205.35

respond to 3, now=205.35

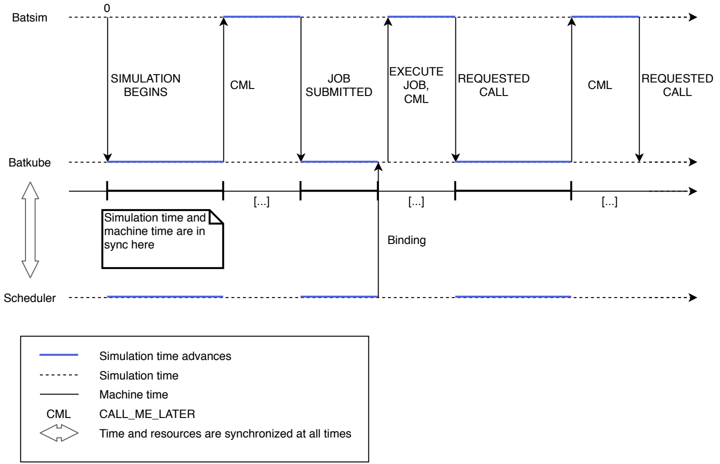
done

Exchanges between the scheduler, batsky-go (“time”) and Batsim

Time synchronization I

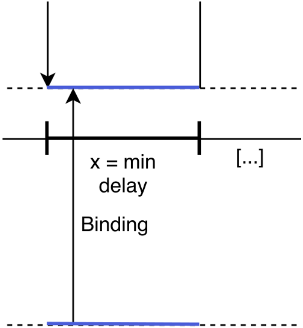
TODO: explain CML

Time synchronization II



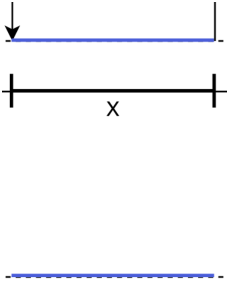
Time synchronization between Batsim and the scheduler

Parameters of the synchronization



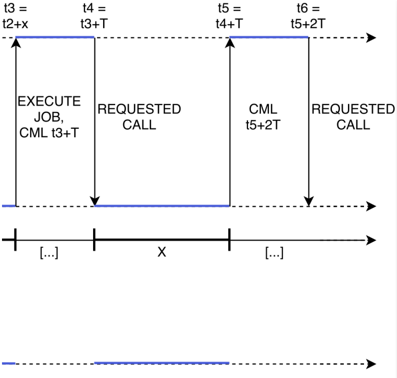
Minimum delay

Parameters of the synchronization



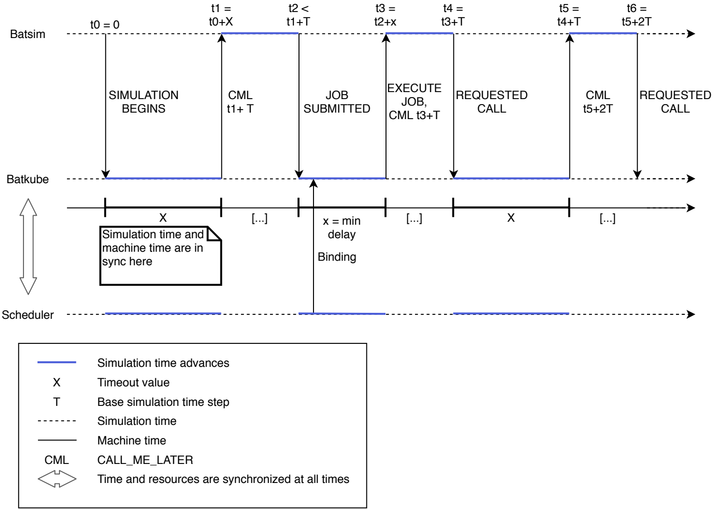
Timeout value

Parameters of the synchronization



Simulation time step ∈ [`base-simulation-timestep`, `max-simulation-timestep`]

Time synchronization breakdown



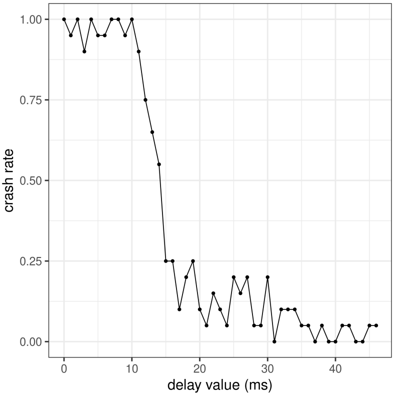
Time synchronization between Batsim and the scheduler

Study of the simulator

Studied workloads and platforms

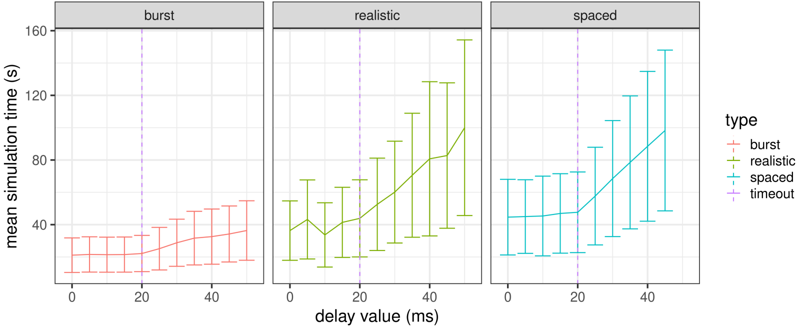
TODO

Minimum delay I

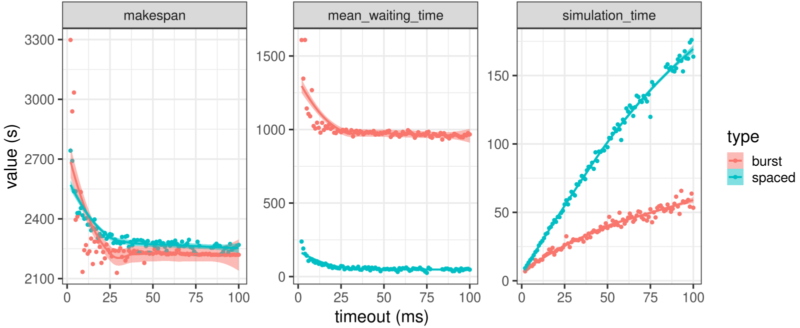


Note: inclure ce graphe?

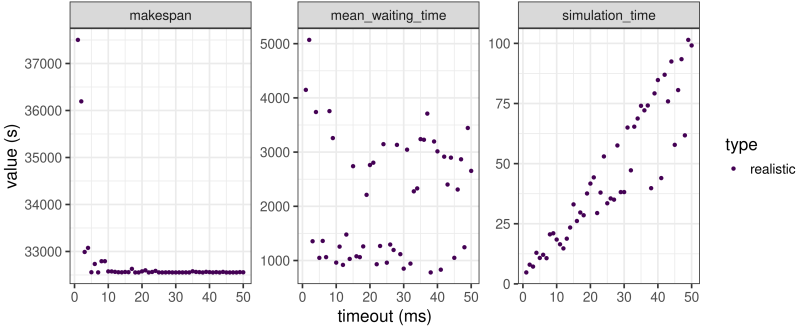
Minimum delay II



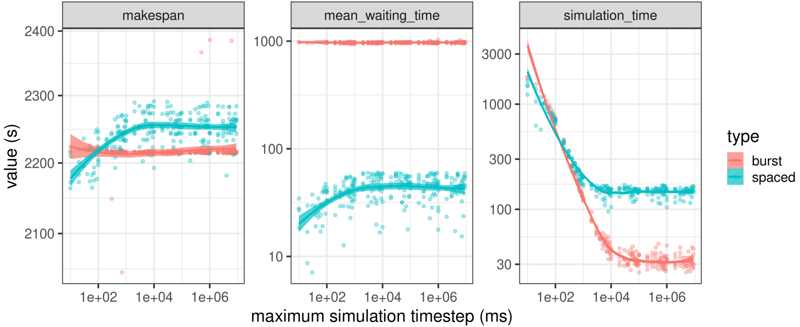
Timeout I


makespan
mean_waiting_time
simulation_time
value (s)
3300
3000
2700
2400
2100
1500
1000
500
0
150
100
50
0
0
25
50
75
100
timeout (ms)
type
burst
spaced

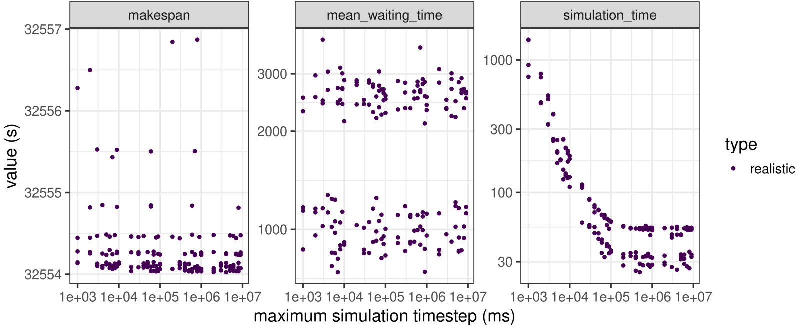
Timeout II


makespan
mean_waiting_time
simulation_time
value (s)
timeout (ms)
37000
36000
35000
34000
33000
5000
4000
3000
2000
1000
100
75
50
25
0
0
10
20
30
40
50
type
realistic

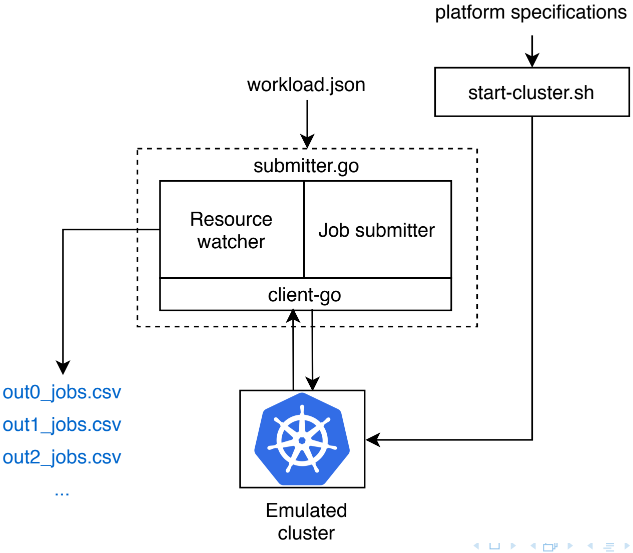
Maximum simulation timestep I



Maximum simulation timestep II



Experimentation on a real cluster



Deviation with reality

workload	makespan				mean waiting time			
	emulated		simulated		emulated		simulated	
	μ	σ	μ	σ	μ	σ	μ	σ
burst	2467	28.3	2215 (-252)	0.508	1077	10.6	970 (-107)	12.6
spaced	2468	5.14	2257 (-211)	16.9	146	1.67	48.1 (-97.9)	9.44
realistic	32556	-	32555 (-1)	1.30	2884	-	2020 (-864)	950

Discussion and future work

Capabilities of Batkube

- delay jobs
- cpu and memory requests
- can patch any kubernetes scheduler written in Go
- the api only supports the default scheduler

Limitations

- memory hungry (in fact, the scheduler is memory hungry)
- some problems with the scheduler
- not scalable

Perspectives for future work

- parallel jobs
- storage
- more complete api: support for more schedulers but also tools (monitoring tools)

References I

Andrew, S (2002). *Tanenbaum, Maarten van Steen." Distributed Systems. Principles and paradigms"*.