Investigating Placement Challenges in Edge Infrastructures through a Common Simulator (extended version)

Adwait Bauskar, Anderson da Silva, Adrien Lebre, Clement Mommessin, Pierre Neyron, Yanik Ngoko, Yoann Ricordel, Denis Trystram, Alexandre van Kempen

Investigating Placement Challenges in Edge Infrastructures through a Common Simulator

Adwait Bauskar, Anderson Da Silva, Adrien Lebre, Clément Mommessin, Pierre Neyron, Yanik Ngoko, Yoann Ricordel Denis Trystram, Alexandre Van Kempen,

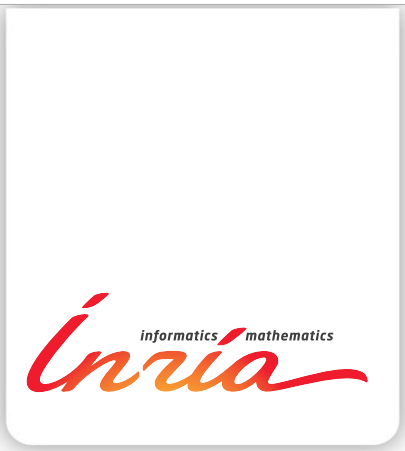
RESEARCH REPORT

N° 9282

May 2020

Project-Teams DATAMOVE, STACS

ISSN 0249-6399 ISRN INRIA/RR--9282--FR+ENG



Investigating Placement Challenges in Edge Infrastructures through a Common Simulator

Adwait Bauskar, Anderson Da Silva, Adrien Lebre, Clément Mommessin, Pierre Neyron, Yanik Ngoko, Yoann Ricordel Denis Trystram, Alexandre Van Kempen,

Project-Teams DATAMOVE, STACS

Research Report n° 9282 — May 2020 — 21 pages

Abstract: Scheduling computational jobs with data-sets dependencies is an important challenge of edge computing infrastructures. Although several strategies have been proposed, they have been evaluated through ad-hoc simulator extensions that are, when available, usually not maintained. This is a critical problem because it prevents researchers to –easily– perform fair comparisons between different proposals.

In this research report, we propose to address this limitation by presenting a simulation engine dedicated to the evaluation and comparison of scheduling and data movement policies for edge computing use-cases. Built upon the Batsim/SimGrid toolkit, our tool includes an injector that allows the simulator to replay a series of events captured in real infrastructures. It also includes a controller that supervises storage entities and data transfers during the simulation, and a plug-in system that allows researchers to add new models to cope with the diversity of edge computing devices.

We demonstrate the relevance of such a simulation toolkit by studying two scheduling strategies with four data movement policies on top of a simulated version of the Qarnot Computing platform, a production edge infrastructure based on smart heaters. We chose this use-case as it illustrates the heterogeneity as well as the uncertainties of edge infrastructures.

Our ultimate goal is to gather industry and academics around a common simulator so that efforts made by one group can be factorised by others.

Key-words: Edge, Simulation, Scheduling algorithms, Data movements

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Investigations sur les challenges de placement dans une infrastructure Edge à travers un simulateur commun

Résumé : Ordonnancer efficacement des travaux de calcul avec des dépendances de données est un des plus importants challenges des infrastructures de calcul edge. Bien que plusieurs stratégies ont été proposées, elles ont toutes été évaluées avec des extensions de simulateur ad-hoc qui sont, s'ils sont rendues disponibles, habituellement pas maintenus. C'est un problème critique parce que cela empêche les chercheurs de – facilement – conduire des évaluations équitables entre différentes stratégies proposées.

Dans ce rapport de recherche, nous proposons d'adresser cette limitation en présentant un simulateur dédié à l'évaluation et la comparaison de politiques d'ordonnancement et de mouvement de données pour le cas d'usage du calcul edge. Construit au dessus de Batsim/SimGrid, notre outil inclut un injecteur qui permet au simulateur de rejouer une série d'évènements capturés au sein d'une infrastructure réelle. Ce outil inclut aussi un contrôleur qui supervise les entités de stockage et les transferts de données durant la simulation, et un système de plug-in qui permet aux chercheurs d'ajouter des nouveaux modèles pour faire face à la diversité des appareils de calcul edge.

Nous démontrons la pertinence d'un tel outil de simulation en étudiant deux stratégies d'ordonnancement avec quatre politiques de mouvements de données sur une version simulée de la plateforme de Qarnot Computing, une infrastructure de production edge basée sur des radiateurs intelligents. Nous avons choisi ce cas d'usage car il illustre l'hétérogénéité et l'incertitude d'une infrastructure de l'edge.

Notre but ultime est de rassemble les industriels et les académiques autour d'un simulateur commun afin que les efforts faits par un groupe puissent être factorisés par d'autres.

Mots-clés : Edge, Simulation, Algorithmes d'ordonnancement, Mouvements de données

1 Introduction

The proliferation of Internet of Things (IoT) applications [7], as well as the advent of new technologies such as Mobile Edge computing [3], and Network Function Virtualisation [18] (NFV) have been accelerating the deployment of Cloud Computing-like capabilities at the edge of the Internet. Referred to as the Edge Computing [21] paradigm, the main objective is to perform on demand computations close to the place where the data are produced and analysed to mitigate data exchanges and to avoid too high latency penalties [28]. Among the open questions our community should address to favour the adoption of such infrastructures is the computation/data placement problem, i.e., *where to transfer data-sets according to their sources and schedule computations to satisfy specific criteria.* Although several works have been dealing with this question [4, 8, 12, 19, 22, 25, 26], it is difficult to understand how each proposal behaves in a different context and with respect to different objectives (scalability, reactivity, etc.). In addition to having been designed for specific use-cases, available solutions have been evaluated either using *ad hoc* simulators or through limited in-vivo (i.e., real-world) experiments. These methods are not accurate and not representative enough to, first, ensure their correctness on real platforms and, second, perform fair comparisons between them.

Similarly to what has been proposed for the Cloud Computing paradigm [16], we claim that a dedicated simulator toolkit to help researchers investigate Edge scheduling strategies should be released soon. Indeed, we claim that using placement simulators for Cloud Computing is not appropriate to study Edge challenges. Besides resource heterogeneity, network specifics (latency, throughput), and workloads, Edge Computing infrastructures differ from Cloud Computing platforms because of the uncertainties: connectivity between resources is intermittent, storage/computation resources are more heterogeneous and can join or leave the infrastructure at any time, for an unpredictable duration. In other words, a part of the infrastructure can be isolated or unavailable for minutes/hours preventing accessing some data-sets or assigning new computations.

In this article, we present several extensions we implemented on top of the Batsim/SimGrid toolkit [10, 13] to favour fair evaluations and comparisons between various scheduling and data placement strategies for Edge infrastructures. In particular, we developed an external module to allow injecting in the simulation any type of unforeseen events that could occur (e.g., a machine became unavailable at time t). We also implemented a Storage Controller to supervise all transfers of data-sets within the simulated platform. We chose to rely on Batsim/SimGrid instead of any other available Edge simulators [23, 27] for the following reasons:

- Batsim has been especially designed to test and compare resource management policies in distributed infrastructures. In other words, the design of Batsim enforces researchers to use the same abstractions and, thus, favours straightforward comparisons of different strategies, even if they have been implemented by different research groups;
- Batsim promotes separation of concerns and enables the decoupling between the core simulator and the scheduler. Moreover, Batsim provides APIs in different languages (including Python, C++ and Rust) that makes the development of a scheduling strategy accessible for a large number of researchers;
- The accuracy of the internal models (computation and network) of SimGrid has been already validated [11, 24] and extensively used [2];
- SimGrid provides a plug-in mechanism, which is of particular interest to deal with the diversity of Edge devices: it lets researchers add new models of specific Edge facilities without requiring intrusive modifications into the simulation engine.

By extending Batsim to the Edge paradigm, we target a tool that will enable researchers/engineers to re-evaluate major state-of-the-art load balancing strategies. In particular, we think about scheduling strategies that have been proposed in desktop computing platforms, volunteer computing and computational grids [5, 6] as these infrastructures have several characteristics in common with Edge platforms.

To demonstrate the relevance of our proposal, we discuss several simulations we performed for the *Qarnot Computing* [1] use-case. The infrastructure of *Qarnot Computing* is composed of 3,000 disk-less machines (smart heaters) distributed across several sites in France. Each computing resource can be used remotely as traditional Cloud Computing capabilities or locally to satisfy data processing requirements of IoT devices deployed in the vicinity of the computing resource. As such, the *Qarnot* platform is a good example of Edge infrastructure, with computing units and mixed local/global jobs with data-sets dependencies.

The strategies presented in this article are simple. They aim to illustrate what can be done without important efforts. More advanced strategies can be analysed in the same manner. We are, for instance, investigating more advanced strategies that consider pulling data-sets from other Edge resources rather than from the centralised storage system of the *Qarnot Computing* infrastructure.

The rest of the report is structured as follows. Section 2 gives an overview of the Batsim/SimGrid toolkit and the extensions we implemented. Section 3 presents the *Qarnot Computing* use-case and Section 4 describes how we simulated this case study. Section 5 discusses a first analysis of different scheduling strategies for the *Qarnot* platform. Section 6 presents the related work. Conclusion and future works are given in Section 7.

2 A Dedicated Simulator for Edge Platforms

Our proposal relies on extensions developed in the Batsim/SimGrid toolkit [13]. Released in 2016, Batsim delivers a high-level API on top of SimGrid [10] to ease the development and simulation of resource management algorithms. Thus, our proposal relies on tools already validated by our community.

2.1 Operational Components

We discuss in this section the role of the different components, namely SimGrid, Batsim, the *decision process*, and their interactions.

2.1.1 SimGrid

SimGrid [10] is a state-of-the-art simulation toolkit that enables the simulation of distributed systems. SimGrid's relevance in terms of performance and validity has been backed-up by many publications [2]. In addition to providing the program to be evaluated, performing simulations with SimGrid requires writing a platform specification and interfacing the program to simulate. SimGrid enables the description of complex platforms, such as hierarchical infrastructures composed of many interconnected devices with possibly highly heterogeneous profiles, such as the edge ones.

2.1.2 Batsim and the decision process

Batsim [13] is a simulator engine built on top of SimGrid. It proposes a specialised API to help researchers design and analyse jobs and I/O scheduling systems. Such systems are for

instance Batch Schedulers *a.k.a.,* Resource and Jobs Management Systems, in charge of managing resources in large-scale computing centres. Batsim allows researchers to simulate the behaviour of a computational platform in which workloads are executed according to the rules of a decision process. It uses a simple event-based communication interface: as soon as an event occurs, Batsim stops the simulation and reports what happened to the *decision process*.

The decision process embeds the actual scheduling code to be evaluated. In other words, to simulate a given scheduling algorithm, an experimenter has to implement this decision process. Comparing different algorithms consists in switching between different decision processes, which is easy in Batsim. Internally, the decision process (i) reacts to the simulation events received from Batsim, (ii) takes decisions according to the given scheduling algorithm, and (iii) drives the simulated platform by sending back its decisions to Batsim. Batsim and the decision process communicate via a language-agnostic synchronous protocol. In this work, we used Batsim's Python API to implement the decision process. For more details on Batsim and SimGrid mechanisms, we invite the reader to refer to Chapter 4 of Poquet's manuscript [20].

2.2 Extensions

To ease the study of scheduling and data placement strategies for Edge platforms, we have been working on a couple of extensions for Batsim. We present in this section those already available, namely the external events injector and the storage controller. Modifications made in Batsim and its Python API for this work are integrated in the main branch of the repositories[1]. Besides, we present the plug-in mechanism of SimGrid that researchers can leverage to provide models of particular Edge devices.

2.2.1 External events injector

To simulate the execution of an Edge infrastructure, which is by essence subject to very frequent unexpected or unpredictable changes, our simulator offers the opportunity to inject external events on demand. Those events impact the behaviour of the platform during the execution and thus the choices of the scheduling strategy. For example, one would be interested in studying the behaviour and resilience of a scheduling policy when a range of machines becomes unexpectedly unavailable for a period of time, due to a failure or action (e.g., from a local user) occurring at the edge.

An external event is represented as a JSON object composed of two mandatory fields: a *timestamp* that indicates when the event occurs, and the *type* of the event. Depending on the type of event, other fields can complement the event description, such as for instance the name of the unavailable resource, the new value of an environment parameter such as the network bandwidth, or anything of interest to the decision process.

Similarly to the workload submissions, external events are replayed thanks to the injector process of Batsim. For each external event file given as input to Batsim, with one aforementioned JSON object per line in the file, an *external events submitter* is created during the initialisation of Batsim. Each submitter parses the list of external events from the input file and iterates over the list to submit the external events to the main process of Batsim at the right simulation times. Then, the external event is processed by Batsim, the state of the platform is updated and the occurring external event is forwarded to the decision process.

This event injection mechanism is generic by concept: users can define their own types of event and associated fields, which will simply be forwarded to the decision process without requiring any modification in the code of Batsim.

[1] `https://gitlab.inria.fr/batsim/batsim` and `https://gitlab.inria.fr/batsim/pybatsim`

2.2.2 Storage Controller

The Storage Controller is a module included in Batsim's Python API to ease the management of storage entities and data-sets, and supervise data transfers during the simulation.

At the beginning of the simulation, the Storage Controller retrieves the list of *storage resources* of the platform and initialises one storage object per resource. These created storages are empty by default, but they can be filled on demand by the decision process by providing a single or a list of data-sets to be added to a storage. A data-set is represented by two fields, *id* and *size*, denoting the unique identifier of the data-set and its size in bytes. The Storage Controller exposes to the decision process an API to add data-sets to storages during the initialisation of the simulation. It also exposes functions to ask, for example, for the copy of a data-set from one storage to another, or to retrieve the list of all storages holding a copy of a given data-set during the simulation.

When a data-set should be copied from one storage to another, the Storage Controller creates a specific Batsim job for data transfers describing that a given amount of bytes should be transferred from the source to the destination storage resource. Once Batsim notifies that this job completed, the Storage Controller notifies back the decision process that the requested data transfer has completed.

A timestamp is saved for each data movement. In other words, there is a timestamp associated to each data-set in each storage. This timestamp corresponds to the last time the data-set has been requested on this storage.

When adding a new data-set to a storage, the Storage Controller makes sure that there is enough available space in the destination storage before starting the data transfer. In the case there is not enough space, an eviction policy is used to determine which data-sets should be removed to free space for the new data-set. The default policy in use is LRU (*Least Recently Used*), which removes the data-set with the smallest timestamp in the storage. However, this eviction policy can be easily overridden by end-users of our simulator without diving into the main code. When implementing their decision process, end-users should simply create a call that inherits from the Storage Controller and override the eviction method. This enables the evaluation of more advanced eviction policies that can impact the overall scheduling decisions.

Finally, the presence of a particular data-set on a storage can be enforced through the Storage Controller API by assigning meta-information on a data-set. This information can then be used by the eviction policy to prevent for instance the deletion of the data-set while being used by running jobs.

2.2.3 SimGrid plug-ins

When designing an Edge simulator, it is a nonsense to foresee all the models and devices that may compose the platform. There are just too many. However, we claim that leveraging generic models is not the right solution either and so a trade-off should be found. We propose to leverage the SimGrid plug-ins capability that facilitates the implementation of new models without requiring intrusive changes in the simulation engine. We underline that, unfortunately, there is no generic manner of exposing information that can be captured by new plug-ins to the scheduler. Hence, some modifications might be required to extend the communication protocol of Batsim and exchange information between a particular plug-in and the decision process. Examples of such modifications are explained for the case of the *Qarnot* platform in Section 4. This is the trade-off to be able to cope with the high heterogeneity of edge infrastructures while targeting accuracy of sub-models.

3 Case Study: the Qarnot Computing Platform

We present in this section the *Qarnot Computing* infrastructure we use to demonstrate the relevance of our simulation tool.

3.1 Infrastructure Overview

Qarnot Computing has been incorporated in 2010 to develop "a disruptive solution able to turn IT waste heat into a viable heating solution for buildings". The infrastructure is distributed in housing buildings, offices and warehouses across several geographical sites in France. As of writing this report, the whole platform is composed of about 1,000 computing devices hosting about 3,000 disk-less machines, and is growing quickly. On each of the 20 geographical sites, there is a NFS-based storage with a few TB of capacity that enables disk-less machines to manipulate data. In a typical configuration a computing machine has a 1 Gbps uplink to a common switch, which then has up to 40 Gbps uplink to the NFS server. The latency between a computing machine and the NFS server is of the order of 1 ms. The various deployment sites are connected to the Internet using either a public or enterprise ISP, with characteristics varying from 100 Mbps to 1 Gbps symmetric bandwidth to the Internet, and about 10 ms latency to French data centres used by *Qarnot* to host control and monitoring services, the central storage system, and gateways to its distributed infrastructure.

On a daily basis, *Qarnot* computing solution processes from a few hundred to several thousands of batch jobs and thousands of cores are provisioned for dedicated corporate customers, and up to tens of GB of data are replicated from central storage to Edge Computing sites.

3.2 Platform Organisation and Terminology

The jobs and resources manager of the *Qarnot* platform is based on a hierarchy of 3 levels, as shown in Figure 1: the *Qnode-*, the *QBox-* and the *QRad*-level. The QNode is a root node, a "global" server that takes placement decisions for the whole platform. It can be viewed as a load balancer for the platform. On the second level are the QBoxes, "local" servers in smart buildings that take scheduling decisions locally on their own computing nodes. Each QBox is in charge of a set of computing nodes, the QRads, which are composed of one or several disk-less computing units denoted by *QMobos*.

Moreover, a centralised storage server is present at the QNode-level, while each QBox has its own local storage disk. From a physical point of view, the QNode and the storage server are in the Cloud while QBoxes are distributed over smart buildings of several cities, while QRads among a building are distributed in different rooms.

The *Qarnot* platform receives two types of user requests: requests for computing and requests for heating. The computing requests describe the workload to be executed on the platform. They are made by users that first upload input data needed to execute their jobs (named *QTasks*) to the centralised server and upload a Docker image either to the centralised server or the Docker Hub. Then, they submit the QTasks to the QNode. A QTask can be decomposed as a bag of several *instances* that share the same Docker image and data dependencies, but with different command arguments. This can be used for example to process each frame of a given movie, with one frame or a range of frames per instance.

The heating requests are made by inhabitants that can turn on and off the smart heaters in their homes, or set a target temperature for rooms to be reached as soon as possible. Since the computing units in a smart heater are unavailable when cooling is necessary, and are available otherwise, such changes increases the heterogeneity challenges of an Edge infrastructure: the

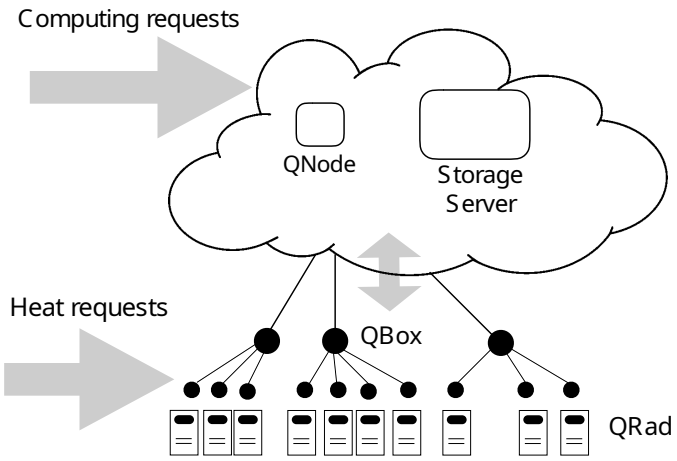


Figure 1: Scheme of the Qarnot platform.

computation capacity does not simply appear or disappear but also vary according to the heating needs.

3.3 Principle of the Actual Scheduling Policy

QTasks submitted to the platform are scheduled onto QMobos through two steps. The first step takes place at the QNode-level. The QNode greedily dispatches as much instances of the QTasks as possible on QBoxes, depending on the amount of QMobos available for computation on each QBox. The second step takes place at the QBox-level. Upon receiving instances of a QTask, the QBox will select and reserve a QMobo for each instance and fetch from the storage server each missing data dependency before starting the instances.

Notice that, at all times, a *Frequency Regulator* runs on each QRad to ensure that the ambient air is close to the target temperature set by the inhabitant, by regulating the frequencies of the QMobos and completely turning off a QRad if it is too warm. Moreover, whenever there is no computation performed on the QMobos while heating is required, "dummy" compute-intensive programs are executed to keep the QRad warm.

Modelling such an infrastructure to identify improvement opportunities and analyse new scheduling strategies is something possible with our Batsim extensions.

4 Simulated Platform

We detail in the following how we modelled and instantiated the *Qarnot* platform with our simulation toolkit, also depicted in Figure 2.

The schedulers of the QNode- and QBox-level were implemented using Batsim's Python API. The decision process is a Python process holding one instance of the QNode scheduler and the Storage Controller, and one instance of the QBox scheduler for each QBox of the simulated platform. Upon receiving a message from Batsim, the simulation events are directly forwarded to the correct destination scheduler or the Storage Controller that should handle it.

In the following we provide details about the descriptions of the platform, the workload and data-sets, the temperature modeling, and how we generated the input files for the simulation.

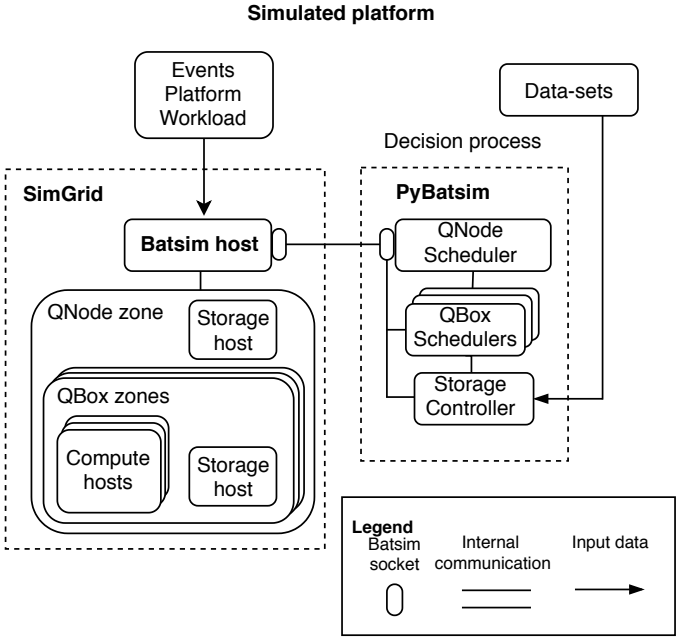


Figure 2: Scheme of the simulated Qarnot platform.

4.1 Platform Description

All details about the platform description are modelled by the XML platform file given to SimGrid at the beginning of the simulation.

Each QMobo is simulated as a single-core SimGrid *host* (representing a machine) as they are the only computing units of the platform. A host holds information such as its unique identifier, the list of speeds and corresponding power usage of its processor, and additional values such as the thermal coefficients required by the temperature plug-in. We keep the same hierarchical structure of the *Qarnot* platform: QMobos belonging to the same QRad are aggregated in the same SimGrid *zone* (representing a network). Similarly, all QRads of a same QBox are aggregated in the same zone, as well as all QBoxes of the QNode. The simulation of storage spaces is done by adding special hosts which carry the Batsim *storage* role. Thus, in each QBox zone, there is one additional storage host for the QBox disk. Similarly, there is one storage host in the QNode zone to represent the storage server.

4.2 Workload Description

The workload to simulate is represented by a JSON file fed to Batsim that contains a list of job and profile descriptions.

Job descriptions are defined by the user requests and contain: the id, the submission time, and the job profile to use. Profile descriptions represent how a job should be simulated, plus other specific information, and contain: the type of the job to simulate, the number of flops to compute, the job priority and the list of data-sets required as inputs.

Each instance of a given QTask in *Qarnot* can run independently from the others, so we transcribed each instance as one Batsim job and profile. Instances belonging to the same QTask have the same data-set dependencies and submission times.

4.3 Data-sets Description

The list of data-sets is described as a list of JSON objects (one per line). Each data-set is represented by the unique identifier of the data-set and its size in bytes. This file is read by the *decision process* and fed to the Storage Controller at the beginning of the simulation to initialise the state of the storage server.

4.4 Temperature Modeling

As temperature plays an important role in the platform and the scheduling decisions, we leveraged the plug-in mechanism of SimGrid to implement our own model. Built on top of the existing energy plug-in [15], our plug-in computes the temperature of a QRad and its ambient air from the energy consumption of the QMobos and other physical parameters, such as the thermal conductivity and mass of the QRad. For the simulation of heating requests, each change of a QRad target temperature is simulated as an external event injected in the simulation, with additional fields for the name of the QRad and the new target temperature. Besides, we take into account the outside temperature of the cities where the QRads are deployed. This value is measured on a one-hour basis and injected as another type of external event.

We modified Batsim to relay these external events to the SimGrid temperature plug-in, and we modified the communication protocol to periodically forward the ambient air temperature of each QRad to the scheduler.

The modifications that are specific to the Qarnot use-case, such as the handling of temperature events, are available in a separate branch of the repository[2]. The temperature plug-in is also available in a separate SimGrid repository of the authors[3].

4.5 Input Files Generation

A log extractor was built to generate all the input JSON files from real logs of the *Qarnot* platform, for a given time period (including the list of jobs to execute, the list of data-sets in the storage server and the list of external events). It is noteworthy that due to users' privacy reasons, we cannot provide access to the log extractor and the *Qarnot* logs used for the experiments. Since we want to simulate an exact time period, we added a special external event that enforces the simulation to stop at a particular time.

5 Simulations

Two kind of experiments have been performed to investigate the *Qarnot Computing* use case. The first aimed to compare the standard scheduling policy used in the real *Qarnot* platform with a policy based on locality of the data-sets. The second experiment enabled us to study the impact of replication policies for the data-sets that are uploaded on the platform (i.e., how they affect the scheduling decisions). The code of all evaluated schedulers is available in a dedicated branch of Batsim's Python API repository[4].

5.1 Data/Job Scheduling Policies

Along with the real *Standard Qarnot* scheduler that serves as a baseline for our experiments (see Section 3.3), we implemented a variant using the data-locality to take scheduling decisions

[2] `https://gitlab.inria.fr/batsim/batsim/tree/temperature-sbac-2020`
[3] `https://github.com/Mommessc/simgrid/tree/temperature-sbac-2020`
[4] `https://gitlab.inria.fr/batsim/pybatsim/tree/temperature-sbac-2020`

at the QNode-level, denoted by *LocalityBased*. Upon dispatching instances, LocalityBased gives priority to the QBoxes already having the data-set dependencies of the QTask on their storage disk. This variant aims at taking benefit from the data locality and reducing the data transfers.

To evaluate the impact of data placement on the scheduling decisions, we also implemented three variants of replication policies upon the submission of QTasks. The question we want to answer with these variants is whether replicating data-sets can achieve significant improvements, and at which cost? The first two variants, denoted by *Replicate3* and *Replicate10*, respectively replicates the data-dependencies of a submitted QTask on the 3 and 10 least loaded QBox disks among the 20 QBoxes in the platform, before applying the LocalityBased scheduling algorithm. These two variants aim at reducing the waiting time of the instances by providing more QBox candidates for the LocalityBased dispatcher. The last variant, denoted by *DataOnPlace*, instantaneously copies all data-set dependencies on all QBox disks upon the submission of a QTask. Even if it is unrealistic, this variant aims at visualising the behaviours of the standard scheduling policy without having any impact caused by the data transfers.

5.2 Simulated Workloads

We extracted 4 different simulation inputs corresponding to logs of the *Qarnot* platform for a 1-week period each. Since the simulation and the scheduling algorithms are deterministic, we ran one simulation with each combination of scheduler and workload. Each simulation took less than 20 minutes to run, with about 60% of the time spent in the decision process.

The considered workloads contained between 5,000 and 9,000 instances and between 40 and 60 different data-sets. In each workload, there was at least one data-set used by 50% of the instances, and up to 9 data-sets were used by 700 instances in workload 2 (as depicted in Figure 3). This information shows that using replication for data-sets should improve the quality of the schedules compared to standard scheduling decisions.

In our simulations, we compared the quality of the produced schedules using the waiting time of the instances, the total number of transfers that occurred, and the total data transferred in GB. For one instance, the waiting time denotes the difference between its starting and submission times.

5.3 Simulation Results

Figures 4 to 7 shows, respectively for each worklaod, the waiting time distribution for each scheduling strategy separated in 3 intervals for better clarity. Figure 8 shows the amount of manipulated data we observed through simulations for the four workloads. Note that, in the following, the analysis of the results is focused on workload 3.

5.3.1 Impact of data locality

As depicted on Figures 4 to 7, for each scheduler more than 60% of the instances waited less than one second before starting their execution. The last column shows that a few instances waited a long time before starting their execution (around 1,455 or 2,910 seconds). This is due to the long transfer time of one of their data dependencies that was as large as 36 GB, while other data-sets were smaller than 5 GB. Comparing the behaviour of Standard and LocalityBased, we do not observe a big difference in the distribution of the waiting times, except for the amount of instances that waited for 1,455 seconds. This is confirmed by the average value over all instances of 39 seconds for Standard and 34.6 for LocalityBased.

Regarding the amount of data manipulated Figure 8 shows, the results from the LocalityBased scheduler is as expected: dispatching instances on QBoxes already having the data-set

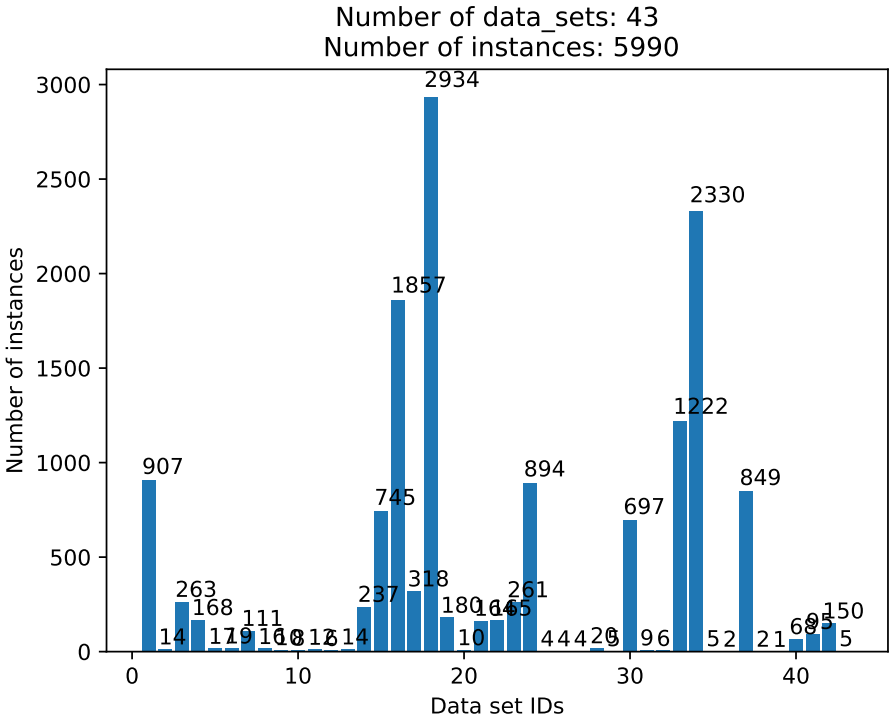


Figure 3: Number of instances using each dataset for the second workload.

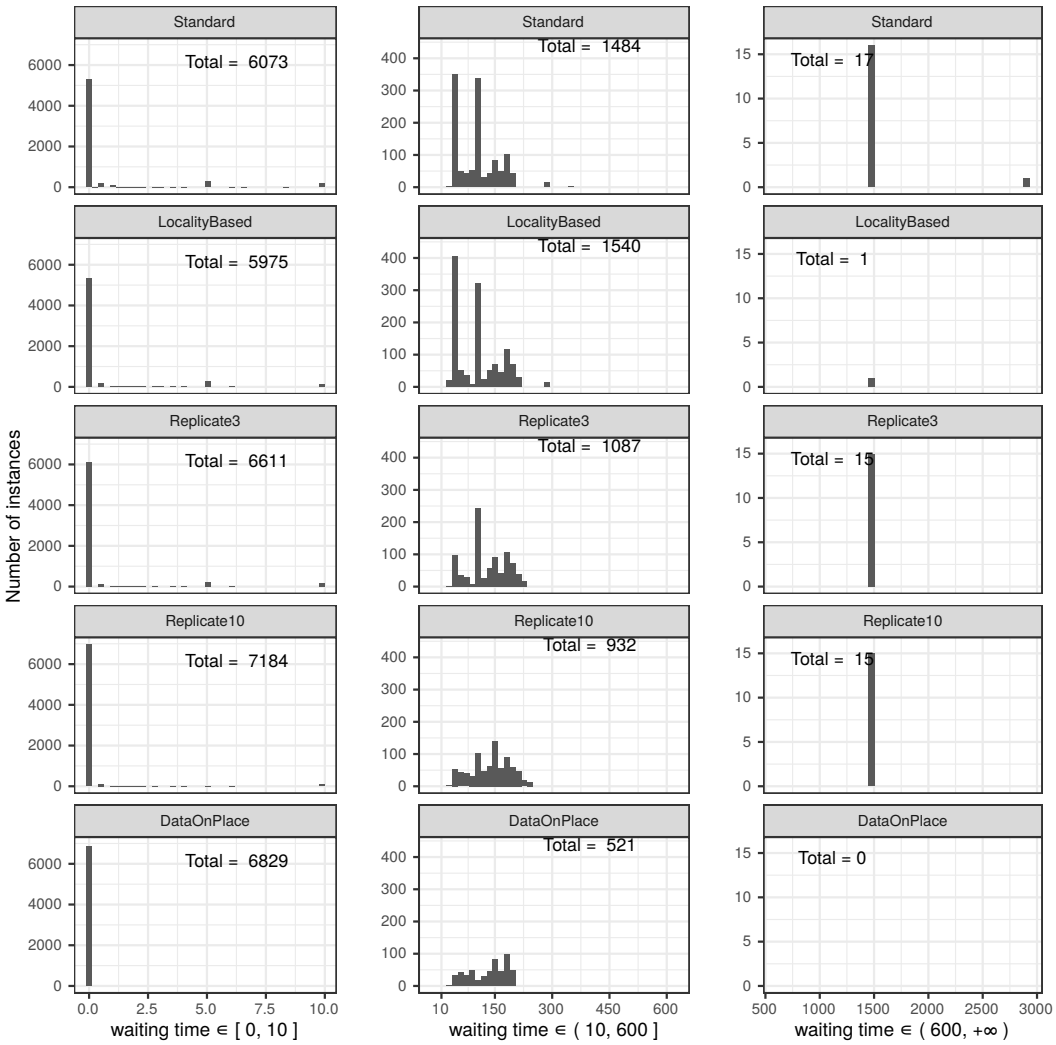


Figure 4: Waiting time distribution (in seconds) of all instances of the first workload.

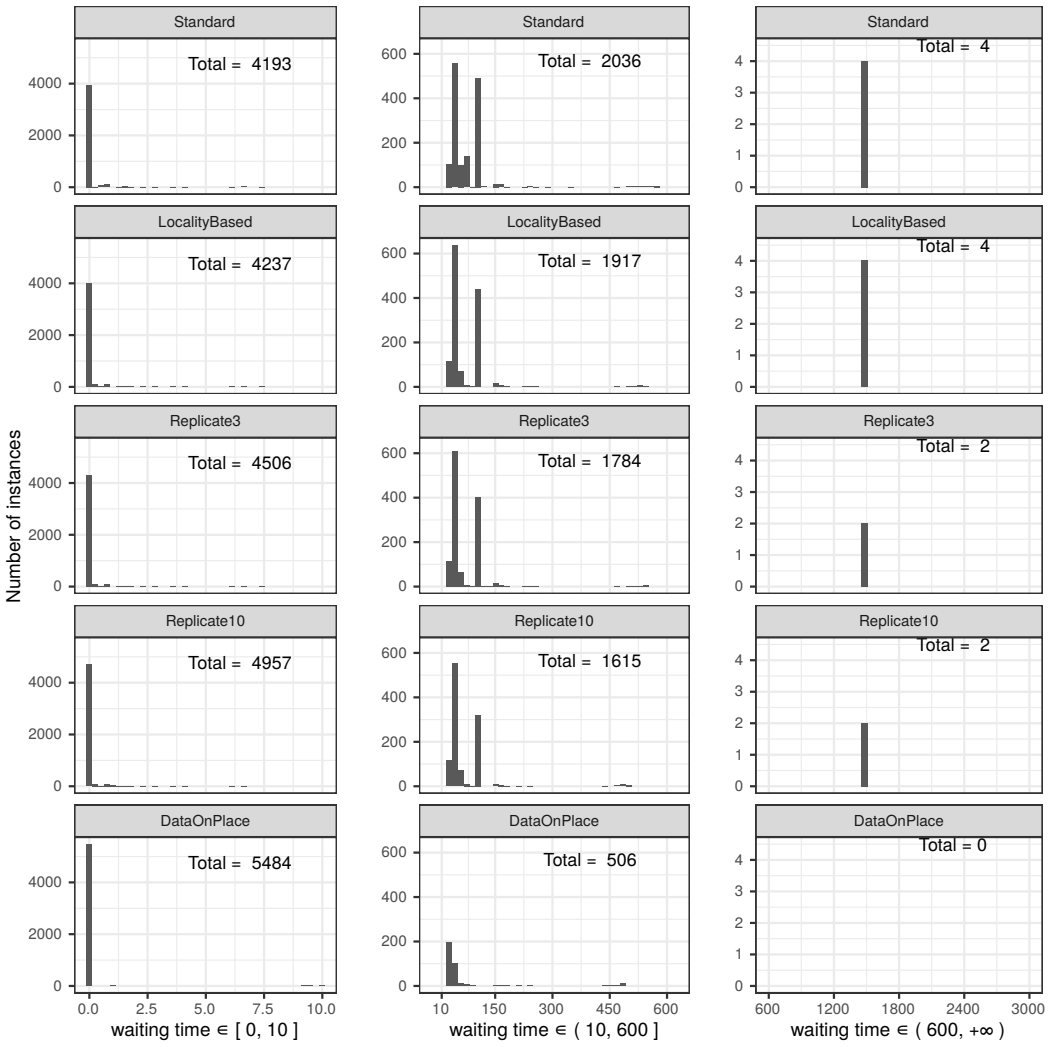


Figure 5: Waiting time distribution (in seconds) of all instances of the second workload.

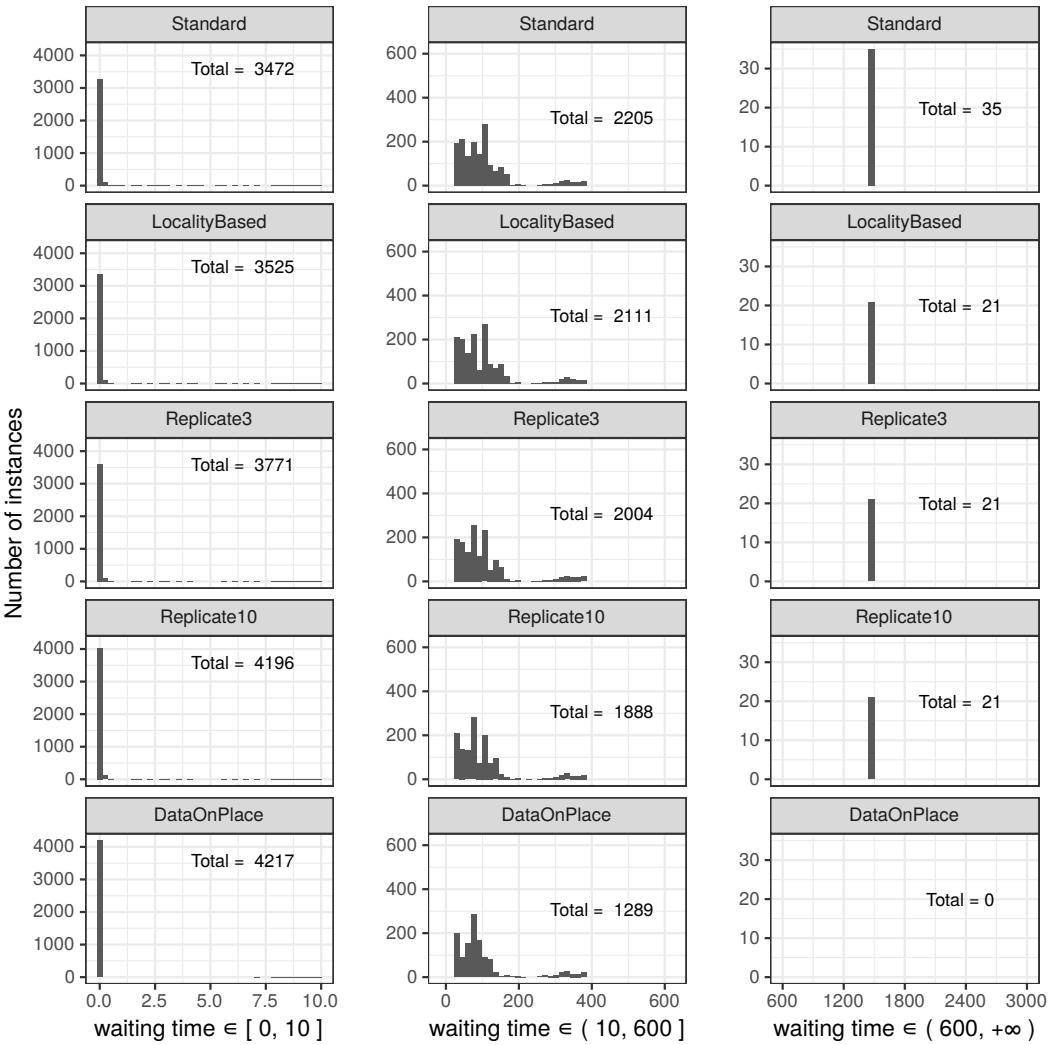


Figure 6: Waiting time distribution (in seconds) of all instances of the third workload.

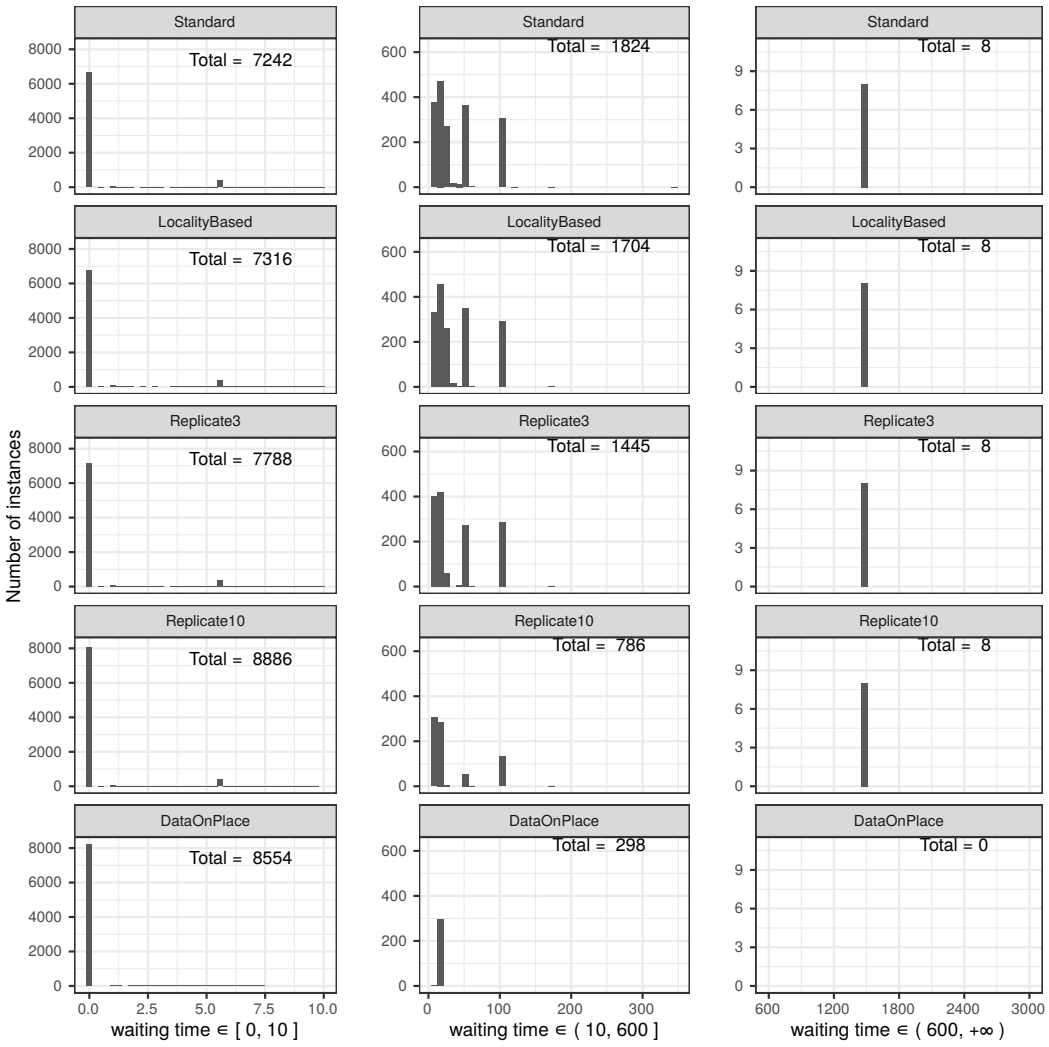


Figure 7: Waiting time distribution (in seconds) of all instances of the fourth workload.

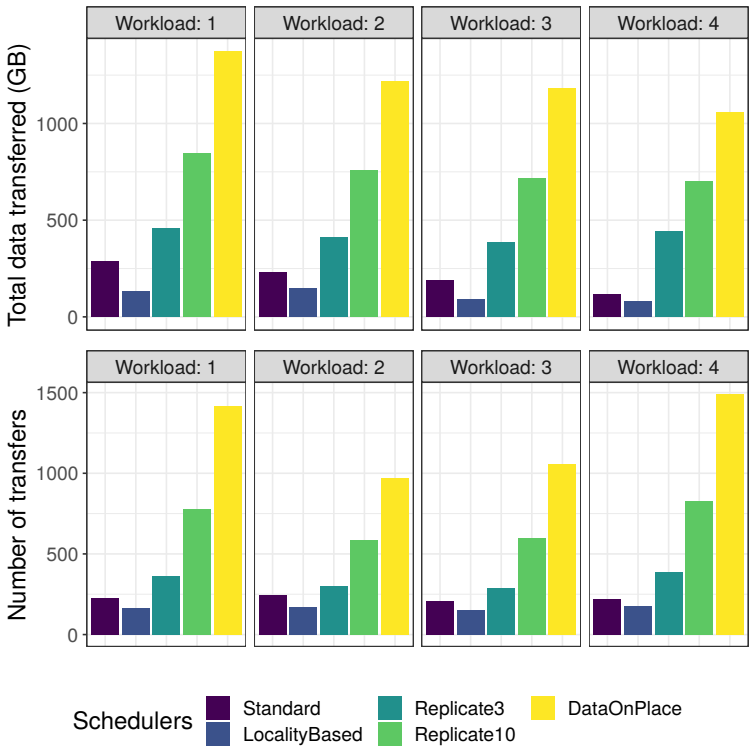


Figure 8: Number of transfers and total data transferred in GB.

dependencies on their disk permits to reduce the number of transfers by about 44%, and between 30 and 65% the total data transferred, compared to Standard.

To conclude, considering data locality decreases the amount of data transfer as expected but does not seem to be satisfactory enough to significantly improve waiting times.

5.3.2 Transferring data has a cost

Replicating data-sets permits to reduce the mean waiting times of the instances but at a cost of more data transfers, as depicted in Figure 8. More precisely for the 3rd workload, the mean waiting time of the instances decreases from 34.6 to 32.6, 28.6 and 22.2 seconds respectively for LocalityBased, Replicate3, Replicate10 and the unrealistic DataOnPlace strategies. While these results look encouraging, it is important to take into account the associated overhead in terms of data transfer: from 90 GB to 384 GB for Replicate3, 599 GB for Replicate10 and 1,056 GB for DataOnPlace. This respectively corresponds to an overhead in terms of data transfer of 4.3x, 6.7x and 11.7x.

Consequently, it is not clear whether replicating data-sets at a high ratio is a valid approach. On the first hand, taking into account only the data locality is not sufficient to have good waiting time performance (LocalityBased). On the second hand, it is crucial to control data-set exchanges as they have an impact on the overall performance. For instance, it may make sense to have a replication ratio that is dynamic according to the popularity of the data-set and the status of the platform. In other words, it is crucial to also consider the time spent in data transfer before taking scheduling decisions. This is critical as the size of data-sets should be increased with respect to IoT-based scenarios envisioned by *Qarnot Computing*. In this regard, we plan to extend the Storage Controller to estimate the transfer time of a data-set to a given storage entity at a certain time. This information is valuable for the schedulers to decide when triggering data transfers and on which QBoxes. Besides, we plan to leverage our proposal to evaluate whether exchanging data-sets directly between QBoxes can help us reduce the data transfer time.

Finally, we recall that our goal through this study was not to find the best scheduling algorithm but to illustrate the use of our simulation toolkit on a concrete scenario, and to demonstrate how such a simulator would help to drive the design of scheduling and data placement strategies. Capturing the aforementioned observations in the Qarnot Computing production platform would have been impossible.

6 Related Simulation Tools

We described in this report a novel simulation tool for easily designing and testing scheduling and data placement strategies on Edge Computing platforms. We motivated the effort of building a new simulator using adequate tools for modelling the processing and storage units and the network topology. We discuss briefly below the main competitors and argument for our simulator.

Some simulators have constraints that would prevent us to correctly simulate a platform such as the *Qarnot* one. For example, EmuFog [17] does not support hierarchical fog infrastructures, whereas *Qarnot* infrastructure is inherently hierarchical.

Other simulators such as iFogSim [14], EdgeCloudSim [23] and IOTSim [27], are simulation frameworks that enable to simulate Fog or Edge Computing infrastructures and execute simulated applications on top of it. These solutions are close to our work. However they have been built on top of the CloudSim toolkit [9]. Although widely used to study algorithms and applications, CloudSim is based on a top-down approach of cloud environments. This is efficient to deliver the right abstractions to the end-users but unfortunately lacks of validations of the low-level models. We believe it is an important issue as it may return invalid observations. Besides, the

Batsim/SimGrid toolkit is the only one that has been designed to study and compare scheduling challenges in an easy manner. In other simulators, including CloudSim, researchers have to implement a lot of business logic that is redundant each time they want to investigate a scheduling policy. Batsim/SimGrid delivers all this logic in a generic manner, making it more versatile and user-friendly for researchers/engineers.

7 Concluding Remarks and Future Steps

We presented in this report extensions we made to the Batsim/SimGrid framework to evaluate scheduling and data placement policies in Edge Computing infrastructures. Its integration into a simulator leads to a complete management system for Edge Computing platforms that focuses on the evaluation of scheduling strategies, taking into account both jobs and data.

While more extensions are still under development, the presented toolkit already enables researchers/engineers to easily evaluate existing load balancing and placement strategies. It may also serve at developing and testing new strategies thanks to its modular and clear interface.

To assess the interest of such simulator, we instantiated the toolkit to simulate the whole Edge platform of the *Qarnot Computing* company based on smart heaters. As a first use case, we investigated four scheduling strategies and compared them to the actual policy implemented in the *Qarnot* platform. We showed that replication of data-sets is an interesting approach to reduce job waiting times but requires additional investigations to determine how the replication ratio can be computed according to several metrics, such as data-set popularity, size, etc. To help researchers move forward on this question, we are currently extending the Storage Controller to monitor additional information such as the number of on-going data transfers. We are also discussing with the SimGrid team to see how we can leverage information about the current load of the links between resources to have estimations of bandwidth, latency or the time to transfer a particular data-set from a source to a destination.

Besides, we envision to design an automatic and probabilistic injector of machine and network failures based on statistical studies of the platform logs and learning techniques. Being able to model the dynamic of Edge infrastructures would be also an important added-value for our framework to capture side effects of such events on scheduling strategies.

Acknowledgements.

This work was supported by the ANR Greco project 16-CE25-0016-01 and by AUSPIN with the International Student Exchange Program from the University of São Paulo.

References

[1] Qarnot computing. https://www.qarnot.com

[2] SimGrid publications. http://simgrid.gforge.inria.fr/Publications.html

[3] Ahmed, A., Ahmed, E.: A survey on mobile edge computing. In: 2016 10th International Conference on Intelligent Systems and Control. pp. 1–8 (Jan 2016)

[4] Ait Salaht, F., Desprez, F., Lebre, A., Prud'Homme, C., Abderrahim, M.: Service Placement in Fog Computing Using Constraint Programming. In: SCC 2019 - IEEE International Conference on Services Computing. pp. 1–9. IEEE (Jul 2019)

[5] Allcock, B., Bester, J., Bresnahan, J., Chervenak, A.L., Foster, I., Kesselman, C., Meder, S., Nefedova, V., Quesnel, D., Tuecke, S.: Data management and transfer in high-performance computational grid environments. Parallel Computing **28**(5), 749–771 (2002)

[6] Anderson, D.P.: Boinc: A system for public-resource computing and storage. In: proceedings of the 5th IEEE/ACM International Workshop on Grid Computing. pp. 4–10. IEEE Computer Society (2004)

[7] Atzori, L., Iera, A., Morabito, G.: The internet of things: A survey. Computer networks **54**(15), 2787–2805 (2010)

[8] Brogi, A., Forti, S.: QoS-Aware Deployment of IoT Applications Through the Fog. IEEE Internet of Things Journal **4**(5), 1185–1192 (Oct 2017)

[9] Calheiros, R., Ranjan, R., Beloglazov, A., De Rose, C., Buyya, R.: Cloudsim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Software Practice and Experience **41**, 23–50 (01 2011)

[10] Casanova, H., Giersch, A., Legrand, A., Quinson, M., Suter, F.: Versatile, Scalable, and Accurate Simulation of Distributed Applications and Platforms. Journal of Parallel and Distributed Computing **74**(10), 2899–2917 (Jun 2014)

[11] Degomme, A., Legrand, A., Markomanolis, G., Quinson, M., Stillwell, M.L., Suter, F.: Simulating MPI applications: the SMPI approach. IEEE Transactions on Parallel and Distributed Systems **28**(8), 14 (Aug 2017)

[12] Donassolo, B., Fajjari, I., Legrand, A., Mertikopoulos, P.: Fog Based Framework for IoT Service Provisioning. In: IEEE CCNC (Jan 2019)

[13] Dutot, P.F., Mercier, M., Poquet, M., Richard, O.: Batsim: a Realistic Language-Independent Resources and Jobs Management Systems Simulator. In: 20th Workshop on Job Scheduling Strategies for Parallel Processing (May 2016)

[14] Gupta, H., Vahid Dastjerdi, A., Ghosh, S., Buyya, R.: ifogsim: A toolkit for modeling and simulation of resource management techniques in internet of things, edge and fog computing environments. Software: Practice and Experience (06 2016)

[15] Heinrich, F.C., Cornebize, T., Degomme, A., Legrand, A., Carpen-Amarie, A., Hunold, S., Orgerie, A.C., Quinson, M.: Predicting the Energy Consumption of MPI Applications at Scale Using a Single Node. In: Cluster 2017. IEEE (Sep 2017)

[16] Lebre, A., Pastor, J., Simonet, A., Südholt, M.: Putting the next 500 vm placement algorithms to the acid test: The infrastructure provider viewpoint. IEEE Transactions on Parallel and Distributed Systems **30**(1), 204–217 (Jan 2019)

[17] Mayer, R., Graser, L., Gupta, H., Saurez, E., Ramachandran, U.: Emufog: Extensible and scalable emulation of large-scale fog computing infrastructures. In: FWC. pp. 1–6. IEEE (2017)

[18] Mijumbi, R., Serrat, J., Gorricho, J.L., Bouten, N., De Turck, F., Boutaba, R.: Network function virtualization: State-of-the-art and research challenges. IEEE Communications Surveys & Tutorials **18**(1), 236–262 (2015)

[19] Naas, M.I., Parvedy, P.R., Boukhobza, J., Lemarchand, L.: iFogStor: An IoT Data Placement Strategy for Fog Infrastructure. In: ICFEC'17. pp. 97–104 (2017)

[20] Poquet, M.: Simulation approach for resource management. (Approche par la simulation pour la gestion de ressources). Ph.D. thesis, Grenoble Alpes University, France (2017), `https://tel.archives-ouvertes.fr/tel-01757245`

[21] Shi, W., Cao, J., Zhang, Q., Li, Y., Xu, L.: Edge computing: Vision and challenges. IEEE Internet of Things Journal **3**(5), 637–646 (Oct 2016)

[22] Skarlat, O., Nardelli, M., Schulte, S., Borkowski, M., Leitner, P.: Optimized IoT Service Placement in the Fog. SOC **11**(4), 427–443 (Dec 2017)

[23] Sonmez, C., Ozgovde, A., Ersoy, C.: Edgecloudsim: An environment for performance evaluation of edge computing systems. In: 2017 Second International Conference on Fog and Mobile Edge Computing (FMEC). pp. 39–44 (May 2017)

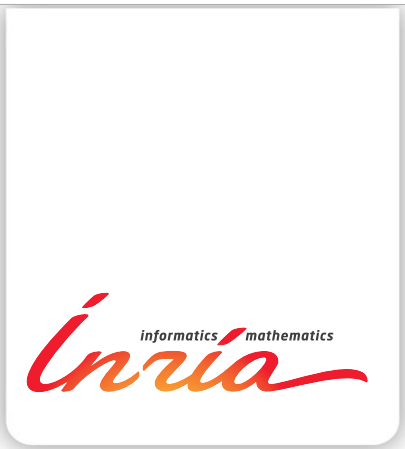
[24] Velho, P., Schnorr, L., Casanova, H., Legrand, A.: On the Validity of Flow-level TCP Network Models for Grid and Cloud Simulations. ACM Transactions on Modeling and Computer Simulation **23**(4) (Oct 2013)

[25] Xia, Y., Etchevers, X., Letondeur, L., Coupaye, T., Desprez, F.: Combining Hardware Nodes and Software Components Ordering-based Heuristics for Optimizing the Placement of Distributed IoT Applications in the Fog. In: Proc. of the ACM SAC. pp. 751–760 (2018)

[26] Yousefpour, A., Patil, A., Ishigaki, G., Jue, J.P., Kim, I., Wang, X., Cankaya, H.C., Zhang, Q., Xie, W.: QoS-aware Dynamic Fog Service Provisioning (2017)

[27] Zeng, X., Garg, S.K., Strazdins, P., Jayaraman, P.P., Georgakopoulos, D., Ranjan, R.: IOTSim: a Cloud based Simulator for Analysing IoT Applications. J. Syst. Archit. **72**(C), 93–107 (Jan 2017)

[28] Zhang, B., Mor, N., Kolb, J., Chan, D., Lutz, K., Allman, E., Wawrzynek, J., Lee, E., Kubiatowicz, J.: The Cloud is Not Enough: Saving IoT from the Cloud. In: HotStorage (2015)



RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399