Oversubscribing Micro-Clouds with Energy-aware Containers Scheduling

Seŕgio Mendes
INESC-ID Lisboa

José Simão
INESC-ID Lisboa
Instituto Superior de Engenharia de Lisboa (ISEL) / IPL
jsimao@cc.isel.ipl.pt

Luís Veiga
INESC-ID Lisboa
Instituto Superior Técnico / ULisboa
luis.veiga@inesc-id.pt

ABSTRACT

Cloud computation is being pushed to the edge of the network, towards Micro-clouds, to promote more energy efficiency and less latency when compared to heavy resourced centralized datacenters. This trend will enable new markets and providers to fill the current gap. There are however challenges in this design: (i) devices have less resources, leading to a frequent use of oversubscription (ii) lack of economic incentives to both provider and application owner to cope with less than full requests fulfilled. To support this trend, the virtualization layer of Micro-clouds is currently dominated by containers, which have a small memory footprint and strong isolation properties. We propose an extension to Docker Swarm, a widely used containers orchestrator, with an oversubscribing scheduling algorithm, based on improving resources utilization to levels where the energy efficiency is maximized. This solution improves CPU and memory utilization over Spread and Binpack (Docker Swarm strategies). Although we introduce a small overhead in scheduling times, our solution manages to allocate more requests, with a successful allocation rate of 83% against 57% of current solutions, measured on the scheduling of real CPU- and memory-intensive workloads (e.g. Video encoding, Key-value storages and a Deep-learning algorithm).

CCS CONCEPTS

• **Networks** → **Cloud computing**; • **Computer systems organization** → **Cloud computing**; • **Computing methodologies** → *Distributed computing methodologies*;

KEYWORDS

Oversubscription, Energy Efficienct, Containers orchestration

ACM Reference Format:
Seŕgio Mendes, José Simão, and Luís Veiga. 2019. Oversubscribing Micro-Clouds with Energy-aware Containers Scheduling. In *The 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19), April 8–12, 2019, Limassol, Cyprus.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3297280.3297295

SAC '19, April 8–12, 2019, Limassol, Cyprus

ACM ISBN 978-1-4503-5933-7/19/04…$15.00
https://doi.org/10.1145/3297280.3297295

1 INTRODUCTION

Cloud computing still relies nowadays on centralized architectures and public cloud providers, like Google, Amazon, Dropbox, to collect and analyze personal data. However, these providers make use of metadata and behavior, putting citizen-generated data in the hands of a few major actors, which use this data to monetize their apparently "free" infrastructures. With enormous amounts of sensitive data being collected and in need to be processed, user-centered devices rely on remote cloud storage and computational services. While this approach is reasonable for many applications, the ones dealing with sensitive data would benefit from using computation services in control of the user, or group of users, and closer to the source of the data.

Furthermore, the current approaches deepen the impact off overall energy consumed by these massive infrastructures, incurring high costs for Cloud Service Providers (CSPs) [16]. Recent trends push computation to the edge, with a paradigm known as Fog computing, which captures this idea of a cloud continuum [2, 17]. While it is clear that this distributed cloud infrastructure based on Micro-clouds at the edge of the network is part of the solution [10, 18], there are challenges at different levels, from the virtualization and scheduling approaches to the definition of an economic rational that can improve resource utilization and user satisfaction.

Containers [14] have been proposed as an alternative to VMs to virtualize resources. Containers are more lightweight than VMs, containing only the required application binaries to run a specific process and nothing more, not requiring a full guest Operating System (OS) instance. Since they are significantly more lightweight than VMs, a better resource utilization can be achieved using containers. The state-of-the-art regarding energy-aware strategies for cloud environments focus mostly on using VMs and not containers, with a few exceptions [4, 8]. The first has some limitations due to the of usage computationally intensive computations (through the use of X-means) which can be an overkill on real cloud environments. The second work approach can lead to hosts not serving any requests due to using a static amount of hosts for profiling and others for long duration requests. Therefore, if there are no requests to be profiled or there are only short duration requests, those hosts will not be used, wasting energy. This is also not considered on the current platforms for managing containers (e.g. Docker Swarm, Kubernetes). Their scheduling decisions, are not energy-aware.

In this paper we propose a scheduling algorithm that promotes energy efficiency in the context of cloud environments, managed by Docker containers, based on maximizing resource utilization according to levels of energy efficiency, without violating Service Level Agreements (SLAs). We have developed a prototype of the

solution in order to evaluate it in a realistic environment. This evaluation was performed according to a set of relevant metrics drawn from related work such as CPU and Memory utilization over time, comparing with relevant related systems.

2 RELATED WORK

As was mentioned on the previous section, our goal is to optimize energy efficiency where containers are concerned. In order to do so, some decisions have to be made, such as which scheduling strategy to use. These decisions were made based on our analysis of the related work.

The first important decision was deciding which container technology to use. The two most mature open-source solutions are **Docker**[1] and **Rocket**[2] (or *rkt*). Due to being daemon-less and not executing as root (as opposed to Docker which the daemon runs as root), Rocket provides more security guarantees than Docker. It is also simpler than Docker, since Docker provides significantly more different features such as Docker Compose[3], in comparison with Rocket. However this simplicity is also one of Rocket disadvantages, since these extra features Docker provides can be useful in different scenarios. Also, Rocket is still in the process of maturation while Docker is already a stable solution, already being deployed on production environments. For being more mature, we chose Docker as the container technology.

To schedule containers on cloud environments, there are three major orchestrator platforms: **Mesos** [5], **Kubernetes**[4] and **Docker Swarm**[5]. From our study we can conclude that Docker Swarm has the simplest architecture with just two entities, manager nodes and worker nodes, while Kubernetes has the more complex architecture having at least four separate entities. Regarding scheduling, Kubernetes has the simplest algorithm thanks to pods, which avoids the usage of filters (by Docker Swarm) and constraints (By Mesos) to co-relate similar containers. Docker Swarm is the less robust only replicating manager nodes while Mesos with Zookeeper and with health-checks provides a good reliability. Finally Docker Swarm uses the standard Docker API which simplifies the learning curve. None of the three solutions is significantly better than the others, in fact, they only differ on small aspects as could be seen by this brief analysis. We chose Docker Swarm because it has the closest architecture to the one we propose on the next section.

The last step is choosing a strategy for scheduling in an energy efficient manner. As was already mentioned on the previous section, we only managed to find two works that schedule containers in an energy efficient manner. However, VM strategies for scheduling VMs in an energy-efficient way can be leveraged for containers since both VMs and containers serve similar purposes. A panoply of strategies exist [9] but the most significant strategies are **VM Placement** [7], **Consolidation** [1], **Overbooking** [3, 15], **Brownout** [19] and **VM Sizing** [8]. There is no single strategy better than all the other and what should be used, depends on the environment and the goals. Some might even be used together, e.g., DVFS and VM Placement [6].

[1]https://www.docker.com/
[2]https://coreos.com/rkt/
[3]https://docs.docker.com/compose/overview/
[4]https://kubernetes.io/
[5]https://docs.docker.com/swarm/overview/

In this paper we opted for an overbooking strategy, which consists on allocating more resources beyond the hosts nominal capacity. The amount of resources that are wasted due to fixed size requests imposed by CSPs are a significant source of energy inefficiency, therefore creating an opportunity for increasing the energy efficiency by maximizing resource utilization. The *United States Data Center Energy Usage Report* [12] shows that approximately 30% of the servers on a data center are either idle or under-utilized, highlighting even further how an overbooking approach can be important to solve this problem by being able to allocate beyond the machine nominal capacity. In [13] this overbooking approach was applied to the cost model, incorporating a range-based non-linear reduction of utility, defined by the client, with impact in the price charged by each VM, although energy efficiency levels were not considered.

The other two works that perform energy-efficient scheduling with containers use different approaches. The authors in [8] use a **VM Sizing approach**. They propose finding efficient VM sizes for hosting containers in such way that the workload is executed with minimum wastage of resources. The challenge is therefore finding an optimal size such that applications have enough resources to be executed. **GenPack** [4] is a framework to schedule containers in cloud data centers, using principles from generational garbage collection (GC). It places containers in different groups, called generations, depending on the knowledge the system has about each container effective use of resources. All containers start in a generation called *nursery*, where an initial profiling is made, move to the young generation once the workload in properly understood, and finally reach the *old* generation for long running containers. Although there are improvements in energy-efficiency, GenPack assumes the cluster will always have resources for the incoming requests, and the profiling and migration steps have to potential do hinder execution time. The next section will describe the architecture of our solution, providing a high-level view of it.

3 ARCHITECTURE

At high level our system consists of two components, a manager and hosts, similarly to other Cloud scheduling platforms like Docker Swarm. It starts by a client submitting a request, indicating the request requirements. The request type refers to a **service** (does not have a finite execution time, e.g. a web server) or a **job** (if it has a finite execution time, e.g. calculating a factorial). The image refers to what the container is going to execute (e.g. an Apache web server). As for the classes, we provide four classes for the client to choose: (a) Class 1: No overbooking; (b) Class 2: 120% overbooking; (c) Class 3: 150% overbooking; (d) Class 4: 200% overbooking;

Class 1 requests do not tolerate overbooking. These requests must run on hosts that are not experiencing overbooking. As for the other classes, they tolerate $1 - (100/requestClassValue)$ overbooking. As an example, for a class 3 request: $1-(100/150) = 0.33(3)$, therefore these request classes can run on hosts that have up to 33% more resources allocated than its nominal capacity. After this process, the Manager receives this information and according to it, among all hosts, it selects the one which maximizes overall resource utilization, allocating the request to it.

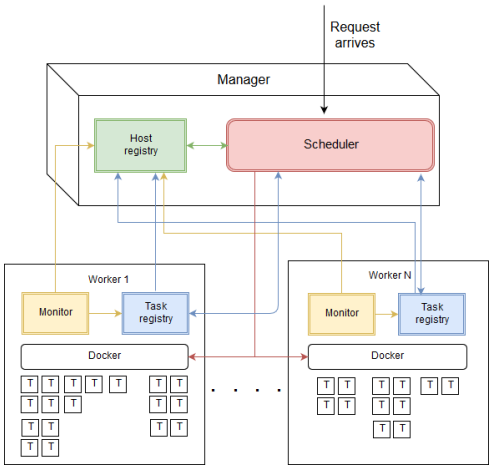

Figure 1: System Architecture

3.1 System Architecture

Fig.1 describes the architecture in more detail, the components inside the Manager and the hosts, and how they interact with each other. Next is provided a brief overview of each component.

Scheduler: The Scheduler is the first component the request interacts with. This component is Docker Swarm, which was extended to include our scheduling algorithm which is presented on the next Section.

Host Registry: This component maintains general information about the hosts (e.g. total resource utilization of each host) on these data structures. This component is also responsible for rescheduling tasks (sends them back to the Scheduler) that are terminated by the Scheduler and updating task information when a cut is performed.

The **cut** means that we are decreasing the resources assigned to that task. This is different from overbooking, because overbooking affects all the tasks on a host, while a cut affects a single task. This is useful for example, imagining that in a class 1 host there is 2GB RAM available and comes a class 4. request which requires 3 GB RAM, if we put the request there, it would increase the overbooking factor over 1 which is unacceptable on a class 1 host. But if we cut it (e.g. to 2GB RAM), then we can fit it there without bringing the overbooking factor over 1. The cut is equal to the overbooking that a class tolerates, so, for example, a class 2 task, would have its resources decreased by 16%. **Kills** refer to tasks that are terminated in order to allow lower level classes to be allocated to the host. The main purpose of resorting to terminate is to avoid the hosts to reach extremely high resource utilization levels which would reflect in a degradation of energy efficiency.

Task Registry: This component contains more specific information about each host (e.g. current tasks being served by the host). It is also responsible for terminating the tasks chosen by our algorithm.

Monitor: In order to make the best scheduling decisions, the Host Registry and the Task Registry must be constantly updated. For this purpose, the Monitor is responsible for measuring resource utilization on each host and each task, and sending updated information to the Host Registry and Task Registry.

To make decisions, besides the information regarding the request, the Scheduler requires additional information about the hosts. This is provided by the Host Registry and the Task Registry. Information on the Host Registry is the first to be considered, therefore being directly available at the Manager to avoid communication overheads. However, more specific information might be needed about what is running on each host. When that is the case, the Scheduler will request that information from the Task Registry of the host it requires that additional information.

Besides requesting information, the Scheduler also sends information to both registries. When a request is scheduled, the Scheduler informs the Host Registry to which host the request was scheduled and the corresponding request information (resources requirements, request type and request class). It also informs the Task Registry that a task was just created sending the same information. Upon receiving this information, each Registry will update its data structures accordingly. Next we will see how these data structures are built.

3.2 Key Data Structures

Our strategy for achieving better resource utilization and consequently, better energy efficiency, is based on the study performed by [11], which states that the energy consumed is proportional to the resource utilization and that energy efficiency starts degrading at high levels of resources utilization. Based on this, we decided to have three regions which map resource utilization (CPU and Memory) with energy efficiency:

- **Low Energy Efficiency (LEE)**: 0-50% total resources utilization;
- **Desired Energy Efficiency (DEE)**: 50-85% total resources utilization;
- **Energy Efficiency Degradation (EED):** >85% total resources utilization;

The LEE region refers to the region that has the lowest energy efficiency, due to under-utilized resources. We want to transit hosts on that region to the DEE region as quickly as possible, where an optimal energy efficiency is achieved. Our goal is to keep the hosts at region DEE, because heavily used resources (hosts at region EED) have a negative impact on the energy efficiency, increasing the energy consumption.

Host Registry: This component will maintain updated lists containing the hosts at each of these regions. For each region, we will have four lists, one for each overbooking class. What defines a host class is the lowest level class task currently running at that host. The region a host belongs depends on the current total resources utilization of the host. The total resources utilization is represented as *max{% of CPU utilization, % of memory utilization}*, since the highest of these two values is what is restraining more the utilization of the overall host resources. The overbooking factor is the *max {CPU shares allocated/Total CPU shares, Memory allocated/Total memory}*. Again, we use the max because it is what is the most restraining. As an example, if the overbooking factor is 1.3, it means we have 30% more resources allocated on that host than the total amount of resources of that host.

The lists on the regions LEE and DEE are ordered by descending order of total resources utilization and EED by ascending order. The

hosts on the LEE region are ordered by descending order, because the goal is to make the hosts leave this region of energy inefficiency, bringing them up to the the DEE region as quickly as possible. Therefore the scheduling algorithm will try to schedule the requests on the first elements of the list since they are closest to the DEE region. Since the DEE region is the desired region for hosts to be, we order its lists by descending order, to use a best fit approach, i.e. put as much requests on a host to maximize it but at same time avoid entering the EED region. The EED list will only be used to terminate as will be seen on the next section. The hosts on that region are experiencing high resource utilization, therefore we don't want them to be receiving more requests which would only aggravate their energy efficiency. What we want is to bring them down to the DEE region, therefore we order the lists by ascending order so that the first on the list is the closest to the DEE region.

Task Registry: As mentioned earlier, the Task Registry contains specific information about the tasks running on the host. Per host, there will exist four lists, one per overbooking class. The information of the Task Registry will only be used for the cut or terminate algorithm. Since the objective is to maximize resource utilization, priority is given to cut or terminate tasks that are using less resources. To achieve this, Task Registry data structures are ordered by ascending order of their total utilization resources. The next section presents the algorithms rely on these data structures.

3.3 Algorithms

There are three core algorithms: scheduling, cut and terminate. The first, tries to schedule the request, taking some restrictions into consideration. However, if the request does not fit with the first algorithm, there are two options, either cut or terminate tasks in order for the request to fit. The goal of the scheduling algorithm 1 is to try and schedule the request either in the LEE or in the DEE region.It starts by getting the hosts that are in the LEE region, then the hosts on the DEE region are appended to that list.We prioritize scheduling in the LEE region so that those hosts can leave that region of energy inefficiency. Since the lists are ordered by descending order of total resources utilization, as we saw the previous section, the first elements of the lists are always the best candidates in order to achieve the goals of the hosts on each region.

The hosts retrieved (line 2) must respect this condition: ***req.CLASS*** $\geq$ ***host Class*** and The hosts are aggregate by ascending order of the class. This is to try and aggregate class 1 requests so that they are not spread among the hosts, which would cause more energy inefficiency since no overbooking is allowed on class 1 hosts. If the request cannot be scheduled in any of those hosts, we must resort to cut or terminate. We first try to cut. We do not cut tasks on the region EED. Cutting a task and putting a request there, it would increase the overbooking on that host, worsening the decrease of energy efficiency that is already felt by hosts on that region. The algorithm gives priority to cutting the incoming request rather than the already running tasks, because cutting a task involves more overhead than cutting a request, due to the updates that have to be performed at the data structures. The following restrictions are due to the fact that when combining overbooking and cutting, class SLAs could be violated if these restrictions are not followed:

- Class 1 requests do not receive cuts;
- Class 2 requests can only receive a cut if they are assigned to a class 1 host;
- Class 3 requests can receive a full cut if they are assigned to a class 1 host. If they are assigned to class 2 host, they can only receive a cut equal to: 33% (class 3 value) - 16% (class 2 value), i.e. 17%. They cannot receive cuts for class 3 and 4 hosts;
- Class 4 requests can receive a full cut if they are assigned to a class 1 host. If they are assigned to class 2 host, they can only receive a cut equal to: 50% (class 4 value) - 16% (class 2 value), i.e. 34%. If the task is at a class 3 host then they can only receive a cut equal to: 50% (class 4 value)- 33% (class 3 value), i.e. 17%. They cannot receive cuts for class 4 hosts.

Algorithm 1 Scheduling algorithm

```
1:  function ScheduleRequest(request)
2:      listHostsLEE_DEE = getHostsLEE_DEE()
3:      for listHostsLEE_DEE as selectedHost do
4:          if requestFits(selectedHost, request) then
5:              allocateRequest(selectedHost, request)
6:              return
7:          end if
8:      end for
9:      listHostsLEE_DEE = getHostsLE_DEE()
10:     if cut(listHostsLEE_DEE, request) then
11:         allocateRequest(selectedHost, request)
12:         return
13:     end if
14:     listHostsEED_DEE = getHostsLEE_DEE()
15:     if terminate(listHostsEED_DEE, request) then
16:         allocateRequest(selectedHost, request)
17:         return
18:     end if
19:     warnClient()
20: end function
```

The first step is to try to fit the request by cutting it and checking if it fits. If it does fit, then the request is cut and allocated to that host. Otherwise, if the request class is higher than the host class, it continues to the next host because it is not worth to cut at this host. This is the case because, if the request class is higher than the host class, then it is likely that this host contains a majority of tasks that are below the request class therefore not being worth the time searching this host for tasks to cut. As mentioned before, the request can only be cut if the host class is lower than the request class. For the same reason, tasks can only be cut if their classes are lower than the host class. Therefore, if the host class is greater or equal than the request, only tasks whose class is higher than the request can be cut.

If it is not possible to cut tasks to fit the current request, the algorithm's last chance is to try terminate tasks in order to fit the request. Priority is given to terminating tasks on region EED, because by terminating tasks and assigning a new request to it, we could bring that host back to the DEE region. Since this is the scheduling algorithm last resort to fit a request, all the hosts on that region are considered regardless of their class. Tasks that are forced to terminate are rescheduled to other hosts. If after checking

all hosts the request does not fit in any, then it cannot be allocated and we warn the client.

4 IMPLEMENTATION

In this section we present lower level of abstraction, looking at how the system is setup and the components implemented. In order to start containers on remote hosts, Docker Swarm uses a discovery service. Docker Swarm provides a default discovery service but also supports different discovery services, such as key-value stores or DNS. The default discovery service requires constant communications with the Docker Hub[6], which is a slow process when compared to using a local discovery service without requiring external connections. We decided to use a key-value store discovery service for this purpose, Consul[7], for having a good integration with Docker Swarm.

The Host Registry is responsible for many different concurrent tasks, making it susceptible to bottlenecks and having inconsistencies within its data structures. The Task Registry is more light-weight, although it also deals with changes within its data structures. Both solutions that we found for these problems are applied at both registries in a similar way, therefore we present them both together at this section. However there are some differences that are highlighted when relevant.

Sorting: The constant insertions could result in bottlenecks and scalability problems since the data structures will grow very large in real cloud deployments. Therefore a quick, but simple insertion algorithm is required. Binary search is a common and simple algorithm used to find elements in a list with O(log N) complexity. We decided to adapt this algorithm to, instead of searching for an element, to search for an index position indicating the place we want to insert.

Data structures implementation: At the Host Registry each region will have 4 lists, one for each overbooking class. For a quick access, each region will be accessed through a map (e.g. names regions) where the key is a string with the region (LEE, DEE or EED) and the value is a struct (similar to C++ structs, there are no classes in Go [8], which is the language Docker Swarm is implemented) as follows:

```
1 struct {
2   classHosts map[string][ ]*Host
3 }
```

`ClassHosts` maps a host class (1, 2, 3 or 4) to a slice[9] of a Host struct. This struct contains all the information regarding a host (e.g. IP). These maps grant a very quick access to the hosts we want to access, useful for example, when the Scheduler asks for lists of hosts with restrictions about region and class. However, this approach is inefficient if we want to access a single host. To solve this problem we decided to create another map (e.g. named hosts) with the host IP as key (since it is unique) and as a value, we use a pointer to a Host struct, the same Host struct as above. To access a host cpu utilization and update it we can now simply use: `hosts[''193.146.164.10''].CpuUtilization=0.23`. Using this approach also increases free-locking accesses, consequently increasing overall performance.

Resources monitoring: Every 3 seconds samples are collected. After 30 seconds, we average all the samples collected during that interval and use those values (CPU and memory) to check if an update should be sent. In order for the update to be sent to the Host Registry, a condition must be verified. The difference (either CPU or memory) between the last update sent and the current measurement must be higher than a threshold. The threshold is defined at 10 *p.p.*

To collect resource usage information we use System Information Gatherer And Reporter (Sigar)[10]. It provides a simple and efficient way to access OS/hardware information. The rationale behind tasks resource monitoring is the same as of the host monitoring, except that the time between measurements is 45 seconds instead of 30. We increased the value because tasks resource utilization is not as volatile as the hosts resources utilization. We leverage Docker built-in command, stats[11], to get CPU and memory utilization of each task.

5 EVALUATION

This chapter describes the experiments carried out to evaluate the proposed solution against the two Docker Swarm scheduling algorithms, spread and binpack. We start by describing how the evaluation was carried out, followed by its results.

5.1 Setup

We present the evaluation based on a real deployment, with a small 6 hosts cluster, running Intel Core i7-2600K CPU @3.40HZ, 11926 MB RAM and HDD 7200RPM SATA 6GB/s 32MB cache. One host served as the Manager and the remaining hosts as workers, executing client's requests.

Due to the lack of tools to benchmark Docker Swarm scheduling decision quality, we had to create our own custom workload and extensions to collect metrics. Each evaluation lasted one hour in order to have as much variability as possible and three evaluations were executed for each solution. The resources requested for each workload were saved and used on all attempts for each scheduling algorithm so that they were tested under the same conditions. The following requirements for each workload was generated: CPU; Memory; Request makespan; Workload type; Request rate; Request class.

CPU and memory requirements are generated using an exponential distribution. We decided to use an exponential distribution since it provides a good variability. The number generated by the exponential distribution was mapped to a CPU and memory value. For CPU, the minimum value depends if it was a service or a job. If it was job, the minimum CPU assigned is 204 CPU shares (equal to approximately 20% of a single core utilization). If it was a service, the minimum CPU shares was 2, because services do not require as much CPU as jobs. As for the maximum, it was 1024 CPU shares (equal to 100% of a core utilization). For Memory requirements, the limits are the same for jobs and services, the minimum was 256 MB and maximum was 2GB.

[6] https://hub.docker.com/
[7] https://www.consul.io/
[8] https://golang.org/
[9] https://blog.golang.org/go-slices-usage-and-internals
[10] https://github.com/hyperic/sigar
[11] https://docs.docker.com/engine/reference/commandline/stats/

The request makespan was also generated by the exponential distribution. This makespan was used to control workload's lifetime. Since the evaluations lasted one hour, we needed to limit the duration of the workloads so that new requests could be scheduled. After this makespan elapsed, the task was terminated. The minimum value was 30 seconds and the maximum was 30 minutes.

The workload type was chosen randomly between four types of workloads that we have selected. For each of these application's types, we have selected real and popular Docker applications (with the exception of the non-intensive), in order to be representative of each type. The types and respective application used for that type were the following: **FFMPEG**[12] is the CPU-intensive workload, a video encoding application. We used **Redis**[13] as the memory intensive application which is an in-memory key/value store. For the CPU and memory intensive, we have chosen a **Deep-learning**[14] application, where a neural network is trained to zoom in images. Finally for the non-intensive application, we created a Docker application called **Timeserver**[15] which simply returns the time when requested.

The last thing to be generated was the request class. We give more probability for classes 2 and 3 (30% and 45% chance respectively) because we believe that these would be the most used in a real situation. Class 4 (15% chance) since it has a big depreciation, it would be less used than classes 2 and 3, however in our view, it would still be more used than class 1 requests (10% chance) due to the lack of benefits (in terms of compensation) this class provides.

We compared our solution with our competitors using the following metrics: **scheduling speed**; **failed/successful allocations**; **resource utilization (CPU and Memory)** throughout the experiment;**job makespans**;**services response times**. We also did an individual evaluation to our solution, to see how much it resorts to **cut** and **terminate**, as well as how much CPU and memory was cut. Sending all requests at once is not realistic so we decided to send two requests per second to the Manager. We kept sending requests until a memory or CPU limit was reached. The full memory capacity of the 5 hosts combined is roughly 60GB and the full CPU capacity is 40970 CPU shares. We defined the limit as being 50% (i.e. 200% overbooking) more than the full capacity. So the limit is 90GB for Memory and 61440 CPU shares for CPU. Now that we have seen how the traces are generated, which metrics are used and how the evaluation is executed, the next section presents the results of the evaluations carried out.

5.2 Results

We will see that our solution allows significantly more requests to be allocated, achieving an overall better resources utilization. A natural and unavoidable tradeoff of our solution is a comparatively slower scheduling speed to the other solutions, these differences will be exposed. A possible consequence of overbooking could be that jobs or services, can take longer times to finish or to respond, respectively. Finally, the cuts/terminate ratio is presented and we will see how they are useful, especially the cuts, in order to increase the amount of requests that can be allocated.

[12]https://hub.docker.com/r/jrottenberg/ffmpeg/
[13]https://hub.docker.com/r/redis/
[14]https://hub.docker.com/r/alexjc/neural-enhance/
[15]https://hub.docker.com/r/sergiomendes/timeserver

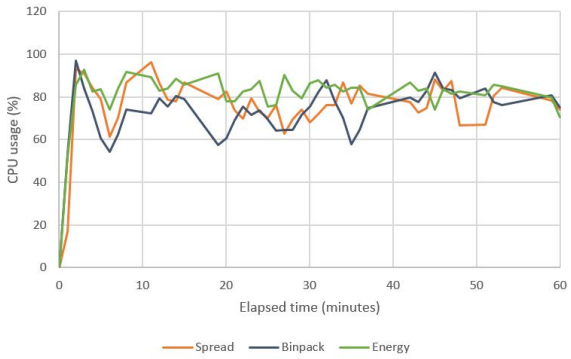

Figure 2: Average hosts CPU utilization of each solution

Successful and failed allocations: The results obtained for the successful and failed allocations are presented at Table 1. We can quickly see that our solution (named Energy) has a significantly higher success rate than the two solutions provided by Docker Swarm. By having such high fail rates, the other solutions would require more machines than our solution does, consequently using more energy. We can also see that our solution deals with less requests than the other two approaches, in an one hour evaluation. This derives from the fact that our algorithm is comparatively slower than the other solutions, due to our solution keeping the resources almost fully utilized for a longer period of time as will be seen afterwards.

This tradeoff is compensated by the high success rate and higher absolute value of successful allocations, since it managed to successfully allocate more requests than both solutions, despite dealing with less requests than those solutions. As will be seen later these values would be lower if more machines were added as can be extrapolated by the data presented on that Section.

Resources utilization: By looking at the graph at Fig.2, which represents the average **CPU utilization** of the worker hosts throughout the evaluation, we can see that our solution (Energy) achieves an overall better CPU utilization. We can see that our solution (green line) is more consistent than the other two, fluctuating most of the time between 75% and 88%. The Binpack solution (blue line) is most of the time bellow 80%. Spread (orange line) is better than Binpack, but worse than our solution, most of the time it is bellow the green line, with some exceptions.

Despite Spread and Binpack having the resources fully allocated, since they are not being used 100% of the time, this resource inefficiency happens. This is even more salient in real life scenarios, where clients after ask for much more resources than they actually need. This clearly indicates that more resources could be allocated to some of the hosts to make them more efficient. This is illustrated

Table 1: Successful and failed allocations

	Success	Failure	Success rate	Failure rate
Spread	1229	904	57.7%	42.3%
Binpack	1256	967	56.5%	43.5%
Energy	1404	274	83.7%	16.3%

Table 2: Average CPU and Memory utilizations

	Avg. CPU utilization	Avg. Memory utilization
Spread	74.9%	39.9%
Binpack	72.3%	36.8%
Energy	80.5%	55.7%

by the results of our solution, where most of the time, the hosts have more than 70% resource utilization.

Memory: Again, our solution presents better results than the existing solutions provided by Docker Swarm. The other solutions never surpass the 60% mark. Our solution achieves it constantly throughout the whole evaluation as can be seen on Figure 3.

Despite being more inconsistent than CPU, our solution provides bigger improvements regarding memory utilization over CPU utilization as can be seen by at Table II. We can see at this table that our solutions provides a 5.6 *p.p* improvement over Spread and 8.2 *p.p* over Binpack, regarding CPU utilization. The memory utilization improvement is much more significant, achieving an improvement of 15.8 *p.p* over Spread and 18.9 *p.p* over Binpack.

Scheduling delays: The significant improvements analyzed previously, unfortunately, do not come without a price. This section presents the results regarding the scheduling delays, i.e. the time to schedule requests.

Table 3 presents a summary about the time it takes to schedule a request on each solution. By looking at the average values, as expected, our solution performs worse than Spread and Binpack. Despite being more complex, our solution is only slightly slower when the system has more resources free, has can be seen by the 50th percentile at 0-5, 25-30 and 50-55. This last note shows that this scheduling delay can be decreased if more machines are added. For the remaining elapsed time, the 50th percentile oscillated between 735.23 and 2190.91 seconds.

Response times: Now we will see that despite allocating more requests as was seen previously, our solution is close to our competitors response times. Table 3 presents the response times obtained for each type of workload used. Redis - 20 indicates that a request rate of 20 to access Redis was used, the same applies for the following columns.

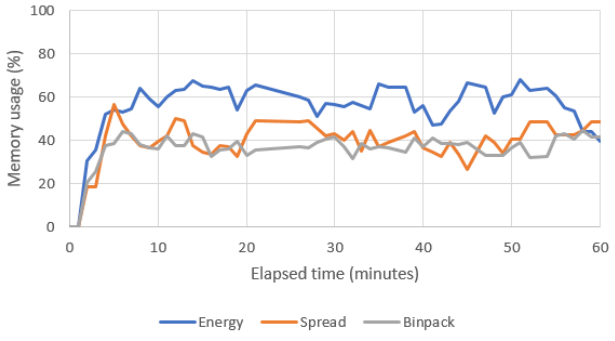

Figure 3: Average hosts memory utilization of each solution

For the CPU-intensive workloads, FFMPEG, we can see that our solution has a better average time than the other two, although it has a higher 50th percentile compared with Binpack. For the CPU/Mem intensive workloads, Deep-learning, we can see that our solution no longer has the best results, but is still better than Spread (better average and 75th percentile results). This decrease in performance compared with Binpack and Spread for this type of workload is unavoidable, because we have significantly more memory utilization rates than the other solutions. Next we have the Redis results, the memory-intensive workload. Redis produced some unstable results as can be seen by the fact that Redis-80, for Binpack, has better results than Redis-40 and Redis-20, which should not be the case, since Redis-80 is twice the request rate of Redis-40, and four times Redis-20. We assume that our solution here would achieve worse times than the other solutions because of what was seen with CPU/Mem-intensive workloads due to the memory impact, potentially worsening as the request rates increased. Finally we have the non-intensive workloads, the Timeserver. Here our solution performs slightly worse than the other solutions at all requests rates.

Cut and Terminate: A total of **636 cuts** were performed throughout the evaluation. This resulted in **112736 CPU shares** and **189.5 GB memory** being cut. These values are the reason why we achieved such a high allocation successful rate. If we resorted only to overbooking such as other approaches in the literature, the successful allocation would be lower because 112736 CPU shares and 189.5 GB memory could not have been allocated. Kills also play and important role, avoiding the hosts from entering extremely high utilization values. Only **202 were terminated** (14,4% of the successfully allocated requests) were executed throughout the experiment. Even if those 202 tasks that were terminated could not be successfully rescheduled and if we considered them as not being allocated, we would still have a higher successful allocation rate than Docker Swarm's solutions.

Table 3: Time to schedule requests

Solution (ms) / Elapsed time (minutes)	**Spread**	**Binpack**	**Energy**
0-5	Average: 2904,87 50th: 9.90 90th: 10002,95 99th: 22974,34	Average: 5278.21 50th: 10.19 90th: 18696.95 99th: 27176.24	Average: 18469.96 50th: 11.94 90th: 72720.68 99th: 128048.71
15-20	Average: 5720.68 50th: 9.64 90th: 16972,52 99th: 68623.37	Average: 5720.68 50th: 9.64 90th: 16972,52 99th: 68623.37	Average: 33644.27 50th: 838.11 90th: 121751.3 99th: 169663.67
25-30	Average: 4281.98 50th: 10.49 90th: 14833.28 99th: 33677.65	Average: 4368.09 50th: 10.09 90th: 18108.94 99th: 28256.79	Average: 33264.4 350th: 15.31 90th: 13997.85 99th: 202526.65
50-55	Average: 4685.35 50th: 10.35 90th: 14950.33 99th: 51035.95	Average: 6079.21 50th: 9.73 90th: 22136.46 99th: 51136.6	Average: 23251.23 50th: 13.7 90th: 19237.15 99th: 223237.94

Table 4: Response times

Workload (ms) / Solution	FFMPEG	Deep-learning	Redis - 20	Redis - 40	Redis - 80	Timeserver - 20	Timeserver - 40	Timeserver - 80
Spread	Average: 333.43 50th: 273 75th: 485	Average: 151.41 50th: 140 75th: 177	Average: 480.53 50th: 115 75th: 587	Average: 560.48 50th: 168 75th: 619.5	Average: 455.08 50th: 322 75th: 880	Average: 1126.04 50th: 800 75th: 944	Average: 2193.75 50th: 1645 75th: 2513.25	Average: 3460.5 50th: 3208 75th: 3477.25
Binpack	Average: 266.51 50th: 189.5 75th: 402.5	Average: 146.76 50th: 137 75th: 163.5	Average: 365.28 50th: 166 75th: 413	Average: 335.91 50th: 197 75th: 220	Average: 239.4 50th: 244 75th: 284	Average: 1475.67 50th: 818 75th: 1126.75	Average: 2380.2 50th: 1669 75th: 2047	Average: 3544.22 50th: 3196 75th: 3477.25
Energy	Average: 250.87 50th: 199 75th: 367	Average: 149.56 50th: 140 75th: 171	Average: 313.2 50th: 247 75th: 393	Average: 393.67 50th: 149 75th: 528	Average: 436.14 50th: 276 75th: 242	Average: 1727 50th: 804 75th: 1315	Average: 2547.48 50th: 1768 75th: 2817	Average: 3570.33 50th: 3332 75th: 3782

6 CONCLUSION

Small clusters with cloud-like services, made of less-resourceful devices, are gaining importance in the *cloud continuum* landscape. These deployments are often supported by containers, running on a given node of the cluster, scheduled by services such as Docker Swarm or Kubernetes. However, few works looked to the possibility of overbooking these clusters, and how to do it in a energy-efficient way. In this paper we extend the base scheduling algorithms of Docker Swarm to allow for overbooking if necessary, while also taking into account the energy state of each node in the cluster. Due to the simplicity of Docker Swarm scheduling algorithms, applying an overbooking strategy would be enough to achieve better results. However, we present new scheduling mechanisms, such as the cut concept, which combines with the overbooking strategy, although some concerns have to be taken into consideration as was seen, to avoid penalties for the clients. The terminate mechanism has the potential to keep the system resources balanced, avoiding global SLA violations. The results obtained revealed that there are many allocated resources wasted due to not being fully utilized, highlighting the opportunity to apply an overbooking strategy to push further the allocated resources, achieving a better energy efficiency, using less machines, which itself allows for more energy savings.

ACKNOWLEDGMENTS

This work was supported by national funds through Fundação para a Ciência e a Tecnologia with reference PTDC/EEI-SCR/6945/2014, and by the ERDF through COMPETE 2020 Programme, within project POCI-01-0145-FEDER-016883. This work was supported by national funds through Fundação para a Ciência e a Tecnologia with reference UID/CEC/50021/2013. This work was partially supported by Instituto Superior de Engenharia de Lisboa and Instituto Politécnico de Lisboa.

REFERENCES

[1] Anton Beloglazov and Rajkumar Buyya. 2012. Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers. *Concurrency Computation Practice and Experience* 24, 13 (2012), 1397–1420. https://doi.org/10.1002/cpe.1867

[2] Alessio Botta, Walter De Donato, Valerio Persico, and Antonio Pescapé. 2016. Integration of Cloud computing and Internet of Things: A survey. *Future Generation Computer Systems* 56 (2016), 684–700. https://doi.org/10.1016/j.future.2015.09.021

[3] Eli Cortez, Anand Bonde, Alexandre Muzio, Mark Russinovich, Marcus Fontoura, and Ricardo Bianchini. 2017. Resource Central: Understanding and Predicting Workloads for Improved Resource Management in Large Cloud Platforms. In *Proceedings of the 26th Symposium on Operating Systems Principles (SOSP '17)*. ACM, New York, NY, USA, 153–167. https://doi.org/10.1145/3132747.3132772

[4] Aurelien Havet, Valerio Schiavoni, Pascal Felber, Maxime Colmant, Romain Rouvoy, and Christof Fetzer. 2017. GENPACK: A generational scheduler for cloud data centers. *Proceedings - 2017 IEEE International Conference on Cloud Engineering, IC2E 2017* (2017), 95–104. https://doi.org/10.1109/IC2E.2017.15

[5] Benjamin Hindman, Andy Konwinski, Matei Zaharia, Ali Ghodsi, Anthony D Joseph, Randy H Katz, Scott Shenker, and Ion Stoica. 2011. Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center. *Proceedings of the 8th USENIX conference on Networked systems design and implementation* (2011), 295–308. https://doi.org/10.1109/TIM.2009.2038002

[6] Wei Huang, Zhen Wang, Mianxiong Dong, and Zhuzhong Qian. 2016. A Two-Tier Energy-Aware Resource Management for Virtualized Cloud Computing System. *Scientific Programming* 2016 (2016).

[7] Md Humayun Kabir, Gholamali C. Shoja, and Sudhakar Ganti. 2014. VM Placement Algorithms for Hierarchical Cloud Infrastructure. *2014 IEEE 6th International Conference on Cloud Computing Technology and Science* (2014), 656–659. https://doi.org/10.1109/CloudCom.2014.53

[8] Sareh Fotuhi Piraghaj, Amir Vahid Dastjerdi, Rodrigo N. Calheiros, and Rajkumar Buyya. 2015. Efficient Virtual Machine Sizing for Hosting Containers as a Service. *Proceedings - 2015 IEEE World Congress on Services, SERVICES 2015* (2015), 31–38. https://doi.org/10.1109/SERVICES.2015.14

[9] Sareh Fotuhi Piraghaj, Amir Vahid Dastjerdi, Rodrigo N. Calheiros, and Rajkumar Buyya. 2016. A Survey and Taxonomy of Energy Efficient Resource Management Techniques in Platform as a Service Cloud. *IGI Global* (2016), 410–454.

[10] Mennan Selimi, Llorenç Cerdà-Alabern, Felix Freitag, Luís Veiga, Arjuna Sathiaseelan, and Jon Crowcroft. 2018. A Lightweight Service Placement Approach for Community Network Micro-Clouds. *Journal of Grid Computing* (28 Feb 2018). https://doi.org/10.1007/s10723-018-9437-3

[11] Leila Sharifi, Llorenç Cerdà-Alabern, Felix Freitag, and Luís Veiga. 2016. Energy Efficient Cloud Service Provisioning: Keeping Data Center Granularity in Perspective. *Journal of Grid Computing* 14, 2 (01 Jun 2016), 299–325. https://doi.org/10.1007/s10723-015-9358-3

[12] Arman Shehabi, Sarah Josephine Smith, Dale A Sartor, Richard E Brown, Magnus Herrlin, Jonathan G Koomey, Eric R Masanet, Nathaniel Horner, Inês Lima Azevedo, and William Lintner. 2016. *United States Data Center Energy Usage Report.* Technical Report June. Ernest Orlando Lawrence Berkeley National Laboratory.

[13] José Simão and Luís Veiga. 2016. Partial Utility-Driven Scheduling for Flexible SLA and Pricing Arbitration in Clouds. *IEEE Transactions on Cloud Computing* 4, 4 (Oct 2016), 467–480. https://doi.org/10.1109/TCC.2014.2372753

[14] Stephen Soltesz, Herbert Pötzl, Marc E Fiuczynski, Andy Bavier, and Larry Peterson. 2007. Container-based operating system virtualization: a scalable, high-performance alternative to hypervisors. *ACM SIGOPS Operating Systems Review* 41, 3 (2007), 275. https://doi.org/10.1145/1272998.1273025

[15] Johan Tordsson, Luis Tom, Luis Tomas, and Johan Tordsson. 2014. An Autonomic Approach to Risk-Aware Data Center Overbooking. *IEEE Transactions on Cloud Computing* 2, 3 (2014), 292–305. https://doi.org/10.1109/TCC.2014.2326166

[16] Ward Van Heddeghem, Sofie Lambert, Bart Lannoo, Didier Colle, Mario Pickavet, and Piet Demeester. 2014. Trends in worldwide ICT electricity consumption from 2007 to 2012. *Computer Communications* 50, 0 (2014), 64–76. https://doi.org/10.1016/j.comcom.2014.02.008

[17] Luis M. Vaquero and Luis Rodero-Merino. 2014. Finding Your Way in the Fog: Towards a Comprehensive Definition of Fog Computing. *SIGCOMM Comput. Commun. Rev.* 44, 5 (Oct. 2014), 27–32. https://doi.org/10.1145/2677046.2677052

[18] Blesson Varghese and Rajkumar Buyya. 2018. Next generation cloud computing: New trends and research directions. *Future Generation Computer Systems* 79 (2018), 849 – 861. https://doi.org/10.1016/j.future.2017.09.020

[19] Minxian Xu and Rajkumar Buyya. 2017. Energy Efficient Scheduling of Application Components via Brownout and Approximate Markov Decision Process. In *Service-Oriented Computing*, Michael Maximilien, Antonio Vallecillo, Jianmin Wang, and Marc Oriol (Eds.). Springer International Publishing, Cham, 206–220.