S-ORCA : A social-based consolidation approach to reduce Cloud infrastructures energy consumption

Simon Lambert
LIP Laboratory
Univ Lyon1, Inria, ENS de Lyon, Ciril GROUP
Lyon, France
simon.lambert@ens-lyon.fr

Eddy Caron
LIP Laboratory
Univ Lyon1, Inria, ENS de Lyon, CNRS
Lyon, France
Eddy.Caron@ens-lyon.fr

Laurent Lefèvre
LIP Laboratory
Univ Lyon1, Inria, ENS de Lyon, CNRS
Lyon, France
laurent.lefevre@inria.fr

Rémi Grivel
Ciril GROUP
Lyon, France
rgrivel@cirilgroup.com

***Abstract*—Information and Communication Technologies (ICT) and Data Centres (DC) have nowadays considerable environmental impacts. The number of applications and services hosted in the cloud is significant, and the associated number of infrastructure is steadily increasing. Cloud Service Providers (CSP) need to respond to this growing demand and size their infrastructures accordingly but users misbehaviour and expectations in terms of service quality lead to oversized infrastructures. Infrastructures are sized to meet peak-demand, resulting in poor resource utilization and additional power consumption. But as their behaviour can increase DC energy consumption and environmental footprint, users can also help reducing them. In this paper, we study how users provided with a simple tool can reduce the power consumption of a virtualization cluster in a cloud company. Using a Virtual Machine (VM) shutdown policy, users can directly contribute to the mitigation of the power consumption of the infrastructure. Part of the paper is dedicated to profile the users and understand their behaviour when it comes to powering off their VMs. To further reduce the energy consumption of the cluster, we combine the VM shutdown policy with a simple consolidation heuristic. Simulations show a 23.95% power consumption reduction, with an additional 8.72% reduction thanks to the users. A production implementation was conducted and results in a 12.58% power consumption reduction over one week.**

***Index Terms*—Cloud, Consolidation, Energy, User involvement**

I. Introduction

Information and Communication Technologies (ICT) are today an integral part of our lives, and the enormous rise in their use has changed our interactions with technology, both as individuals or in companies. The proliferation of services on which society and businesses rely has led to a considerable increase in cloud infrastructures, among which Data Centres (DC) are a cornerstone. The ICT sector and more broadly digital technologies account for 2.1 to 3.9% of global Greenhouse Gas (GHG) emissions [8], with a distribution of emissions between terminals, networks and data centres. For DCs, part of these emissions is linked to electricity consumption, which today accounts for around 2% of global electricity consumption [15], and is rising steadily, particularly with the democratization of generative AI services [4].

But the growth is DCs power consumption did not follow the same trend as the total amount of compute instances and supported workloads [19]. Many factors can explain this limitation in DCs electricity consumption increase. Hardware efficiency, modeled by Koomey's Law, underlines that computations per kilowatt-hour doubled every 1.57 years from 1946 to 2009 [16]. Virtualization [5], through the possibility of sharing computing resources for different operating environments also appears as an enabler in the limitation of DCs growing electricity use. However, these different energy efficiency improvements still led to an increasing use of energy in the IT sector. This paradox, according to which improving the efficiency with which a resource is used ultimately leads to an overall increase in that resource, is called the Jevons paradox and can lead to a rebound effect [6].

Resources management is now a key challenge for DC operators. As cloud infrastructures tend to be under-utilized [7], this represents an operational cost for CSPs and a concern to reduce the environmental impacts of these infrastructures. Global oversizing of cloud infrastructure can be provoked by user misbehaviour, when users themselves oversize their cloud instances or platforms. Involving users to reduce the electricity consumption of such infrastructures [11] therefore appears as a solution to reduce DCs ecological footprint.

In this paper, we study a combination of levers to reduce the power consumption of virtualization infrastructures. By combining a Virtual Machine (VM) shutdown policy with a simple consolidation algorithm, we are able to reduce the power consumption of a virtualization cluster by 12.58%. Our main contributions are :

- A combination of consolidation and shutdown techniques to reduce an infrastructure electricity consumption
- A behavioural study of the infrastructure users to understand their profiles, motivations, obstacles and decision

process

- An implementation of the approach for VMWare, with simulation and experimental results on a production infrastructure
- A qualitative comparison of the approach with a purely technical consolidation methodology: *ORCA*.

The remaining of the paper is organized as follows. Section II discusses related works. Section III presents the methodology of both technical levers and behavioural study. Section IV presents the simulation and experimental results, the users behavioural study as well as the comparison with another consolidation methodology. Section V concludes the paper and outlines the future works.

II. Related Work

A. Virtualization and derivatives for energy efficiency

As for virtualization [5], many techniques were studied to improve the energy efficiency of computing systems and infrastructures [20]. Shutdown techniques for example [21], enable direct energy savings by powering off physical machines (PMs). As for PMs, this process can be translated to VMs in order to free resources of a virtualization infrastructure, therefore facilitating the shutdown of physical machines too. Dynamic Voltage and Frequency Scaling (DVFS) is also used in the cloud to reduce the power consumption of IT equipments by enabling processors to run at different frequencies and voltages [23]. Derived from VM-based virtualization, containerization [2] permits the deployment of applications in containers sharing the same operating-system, enhancing resource management and ultimately reducing power consumption. For both virtualization and containerization, additional techniques can be implemented for further power consumption reduction such as elasticity [14] and consolidation [3].

Consolidation aims to optimize resource usage of DC infrastructures through correct allocation and migration of elements. For VM-based virtualization, it tackles the placement of VM on PMs [13]. Multiple reasons push cloud operators to implement consolidation such as performance, scalability or efficiency. Other parameters can guide consolidation methods and lead to different results: the scale, the placement approach, the objective or the resources on which the method focuses. In [1], Beloglazov et Buyya propose a VM consolidation approach based on a modification of the Best Fit Decreasing (BFD) heuristic : Power Aware BFD (PABFD). With this approach, they aim to reduce energy consumption while limiting Service Level Agreement (SLA) violations. In [17], Lin et al. study a meta-heuristic approach to consolidate containers, using an Ant Colony algorithm to reduce network transmissions overhead and improve cluster services reliability. In [12], Harris et Altiparmak propose variation of the First Fit Decreasing (FFD) heuristic using a Monte Carlo method. Our work differs in the way that we combine a simple consolidation First Fit Decreasing (FFD) approach with a VM shutdown policy to reduce power consumption. This approach is validated both in simulation based on real workload traces and with an experimental implementation on a production infrastructure.

B. User involvement for electricity consumption reduction

In addition to efficiency, digital sufficiency can be used to reduce ICT GHG emissions and environmental impacts. Santarius et al. [22] propose a concept with four dimensions of sufficiency including user sufficiency. In DCs, multiple approaches with users involvement were studied to reduce these infrastructures energy consumption. In [9], Georgiou et al. propose a job scheduling algorithm that priorities jobs based on users' past energy usage, to motivate these users to make more energy-efficient jobs. In [10], Guyon et al. allow users to adapt resource usage and postpone their jobs execution to reduce cloud infrastructure energy consumption. Their simulations give a 5.49% of energy savings at the data centres' scale. In a parallel approach, Madon et al. [18] model 5 users behaviours, including reconfiguration or delay, to meet the same purposes. They also model a renounce behaviour that provides the best results in terms of power consumption reduction, where users do not submit their jobs during the specified windows.

These works rely on user involvement through job submission behaviours to reduce power consumption of DC infrastructures, and validate their results on simulation. In our work, users are involved by the management of a VM shutdown policy that is used for consolidation purposes, which ultimately results in power consumption reduction. To the best of our knowledge, this has not been studied in previous researches nor validated on production infrastructure of a cloud company. Finally, we conducted a user behavioural study to understand our users profiles, motivations or decision processes.

III. Methodology

A. ORCA *consolidation*

This work relies on a consolidation methodology (*ORCA - OuR Consolidation Algorithm*) that benefits from memory overcommitment mechanisms implemented in hypervisors to over allocate VMs on specific PMs. In *ORCA*, VMs are divised into multiple consolidation groups depending on their usage, which is calculated using a simple evaluation algorithm: a VM using less than 1% of its CPU and not performing any network exchange is considered idle over the collection period. Once collected and computed, the usage data is resampled to match a specific consolidation period, using the mean value. This results in a usage matrix containing the average usage of each VM per consolidation period, and over the data retention time. This usage matrix enables the training of prediction models for each VM of a cluster, allowing to perform predictive consolidation. Results are based on XGBoost[1] linear regression models offering fair prediction performances at low training and storage cost in terms of resources.

Group name	G1	G2	G3
Initial usage rate (r)	r<20%	20%<r<80%	r>80%
Associated commitment	3:1	1.5:1	1:1

TABLE I: *ORCA* consolidation groups example

[1] https://xgboost.readthedocs.io/en/stable/#

A group segmentation example of the *ORCA* methodology is defined in TABLE I. In this example, VMs are separated into 3 consolidation groups according to their usage, based on the usage evaluation algorithm given above. Over the consolidation period, which is adjustable, VMs of group G2 will be allocated onto PMs with a memory allocation rate of 1.5:1. This can produce performance degradation on these VMs, which is voluntary here as they are used only between 25 and 50% of the consolidation period time. For each consolidation period, VMs are assigned to their consolidation group and placed on PMs using either a simple FFD heuristic in simulations or using the native hypervisor scheduler in experiments.

Many parameters can affect the results of *ORCA*, such as the consolidation period, the number of groups used, the overcommitment rates or the prediction model training approach. This requires prior knowledge and analysis of the considered infrastructure usage, as for other proactive consolidation approaches. Using appropriate logs and usage traces, it can however be generalized to many virtualization infrastructures.

B. VM Shutdown Policy

The proposed approach is here implemented on an internal virtualization cluster of a cloud company. This cluster of 400 daily-used VMs and 7 PMs hosts multiple applications, software development frameworks, integration and testing VMs, or internal tools. On this cluster, the infrastructure team implemented different rules to prevent misbehaviour and oversights, enabling the infrastructure to be correctly sized. When creating a VM, two fields are mandatory:

1) A VM owner: a person responsible for the VM
2) An expiration date: a date on which the VM is automatically archived

VMs without these 2 fields are automatically shutdown, archived and completely deleted after a period of 15 days. This limits the constant increase in the size of the infrastructure.

Another mechanism was also implemented to increase resource usage and infrastructure performance while limiting the cluster power consumption: an automatic VM shutdown policy. The shutdown policy enables the VM owners to configure a tag on their VMs, which may automatically shutdown their VMs at night and during weekends. Several configurations were proposed to the VM owners:

- *DailySTOP&START*: the VM is shutdown at nights and weekends (9 p.m.), and restarted automatically in the morning (7 a.m.)
- *DailySTOP*: the VM is shutdown at nights and weekends, and restarted manually by the VM owner
- *WeeklySTOP*: the VM is shutdown on Friday evening, and restarted manually by the VM owner
- *NeverSTOPAuto*: the VM is constantly powered-on

Contrary to the VM owner and expiration date tags, the shutdown policy tag is not mandatory on the VMs and no configuration will not cause the VM archiving. A VM with no shutdown policy tag will therefore act as a VM with the *NeverSTOPAuto* tag.

The use of the VM shutdown policy is evaluated here as it is not part of our contributions. After its implementation three years prior to this study, this policy was well received by the infrastructure users and was in line with a dynamic of controlling the computing resources usage in the infrastructure. Fig. 1 shows the amount of VMs configured with the different VM policy tags described above.

The raw distribution depicts a high use of the shutdown policy, with 81.7% of VMs being tagged with one of the described policy tag. Among all the VMs, 61.2% are automatically shutdown daily or weekly. When considering the 400 VMs that are regularly powered-on and used in this cluster, 93% of them are tagged according to the shutdown policy.

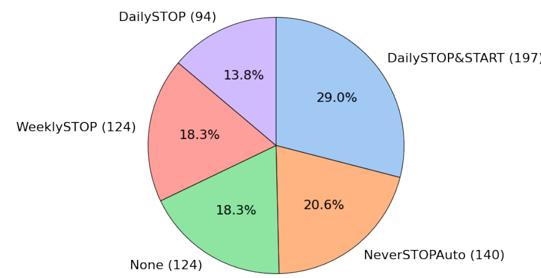


Fig. 1: Distribution of shutdown policy tags per VM

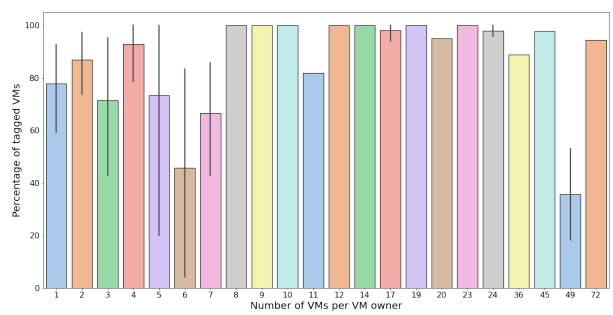


Fig. 2: Percentage of tagged VM per owners' number of VM

Fig. 2 shows the percentage of tagged VMs according to the number of VM owned by the different users with the corresponding variation. The figure underlines different behaviours for the owners responsible of few VMs (up to 7), as well as a good amount of tagged VMs for owners responsible for a more significant number of VMs. We observe one exception for the 2 VM owners responsible of 49 VMs, with respectively 18.3% and 53.06% of tagged VMs.

C. S-ORCA approach

The VM shutdown policy was implemented to mitigate the cluster power consumption by combining it with an embedded feature of the hypervisor: VMWare Distributed Power Management (DPM). This feature enables a vSphere cluster to reduce its power consumption by enabling or disabling PMs according to the cluster resource usage. The activation of DPM did not result in any PM shutdown and therefore, power consumption reduction. The consolidation methodology presented in Section III-A was therefore reworked and adapted to the cluster, considering the VM shutdown policy.

The methodology is here greatly simplified compared with *ORCA*. *S-ORCA* relies on a single consolidation group with no memory overcommitment. It will therefore not cause any performance degradation on the VMs. Instead, the infrastructure is dynamically sized according to the shutdown policy: during periods when the VMs are shutdown by the policy, the consolidation group size will be reduced, and the amount of powered-on PMs will be similarly reduced. Unused PMs are shutdown during these periods of lower activity. Prior to automatic startup of VMs configured with the *DailySTOP&START* tag, the number of required PMs is recalculated and the powered-off PMs are powered-on if needed. This approach eliminates the use of prediction models that create uncertainty in the results of the *ORCA* methodology, both in terms of power consumption reduction and performance for the VMs. The main risk here is service interruption caused by automatic shutdown of the VMs. However, this risk is attributable to the users of the infrastructure and can be rapidly handled.

S-ORCA relies on a simple heuristic to determine the number of required PMs to consolidate all the VMs for a specific iteration t, defined in the Solver of Algorithm 1. The set of PMs P is sorted by decreasing order of PM physical configurations using first memory then number of cores. The total amount of memory of powered-on VMs in the set V for iteration t is then computed as $S_v^m(t)$. The same operations is conducted for total number of vCPUs as $S_v^c(t)$. The number of server required based on both requirements are computed on lines 4 and 5. $CalcPmRam$ and $CalcPmCpu$ are two functions returning the number of required PMs to host an amount of resources (RAM or CPU) from a set P. The total number of server is the max value between each requirement, and the PMs are then selected from the set P.

Algorithm 1 Solving Algorithm - *S-ORCA*

1: **procedure** SOLVER(V, P, t)
2: Sort P in decrease configuration order
3: Get $S_v^m(t)$ and $S_v^c(t)$
4: NumRam ← CALCPMRAM($S_v^m(t)$, P)
5: NumCpu ← CALCPMCPU($S_v^c(t)$, P)
6: numPms←max(NumRam, NumCpu)
7: **return** P[:numPms]
8: **end procedure**

The VM placement problem is not part of this study, as the experimental implementation relies on embedded hypervisor features. VM placement is therefore handled by the hypervisor scheduler. In simulations, VMs are placed on PMs using a First-Fit Decreasing (FFD) approach.

D. User behaviour study

To understand the users behaviour, as well as the levers and obstacles to adopting the VM shutdown policy, we conducted a study among the identified VM owners. Among VM owners, all of them were anonymously questioned to quantify their awareness on DCs environmental impacts as well as their knowledge on the virtualization infrastructure. These two questions were asked to profile the users:

1) From 0 to 5, how would you rate your awareness of the environmental impacts of data centres?
   - 0 : I know nothing about it.
   - 1 : I have a rough idea of the impacts it can have.
   - 2 : I have already seen some information and/or figures on the subject.
   - 3 : I have read up one or more times on the subject.
   - 4 : I have taken one or more specific courses on the subject.
   - 5 : I work daily on the subject, I am an expert.
2) From 0 to 5, how would you rate your knowledge of the virtualization infrastructure and your VMs ?
   - 0 : I don't know what virtualization is.
   - 1 : I know our virtualization infrastructure, but I did not know I was the owner of one or more VMs.
   - 2 : I know I am a VM owner, but I don't know which VMs.
   - 3 : I know some of the VMs I own.
   - 4 : I know all the VMs I own.
   - 5 : I know all the VMs I own, and can describe the services as well as the shutdown policy tag.

In the same time, a representative set of 20 users was selected to respond to short interviews in order to further understand their motivations, obstacles and decision process in the use of the VM shutdown policy. These users were selected based on the number of VMs they are responsible of and the percentage of tagged VMs as presented in Fig. 2. These questions were asked to understand the users behaviour and guide further actions for the infrastructure team:

1) Did you hear about the VM shutdown policy ? If yes, do you use it?
2) What are the profiles of the VMs you are responsible for?
3) What advantages do you identify in the use of the VM shutdown policy?
4) How did you arbitrate the shutdown/non-shutdown of the VMs you are responsible for? Why are some VMs not shutdown?
5) Do you periodically update your VMs configuration?
   a) Have you met any performance degradation on your VMs?
   b) Have you observed any benefits on your VMs?
6) How do you evaluate the investment required to implement the policy on your VMs?
7) Have been informed of the returns and benefits of implementing this policy?

IV. RESULTS

A. S-ORCA consolidation

To evaluate the benefits of *S-ORCA*, both simulations and experiments were conducted on the cluster. In both cases, the resulting power consumption is compared to the initial power consumption using PMs power consumption traces.

1) Simulation: Simulations were conducted to evaluate the potential power consumption reduction enabled by *S-ORCA*. To do so, the consolidation was applied per 1-hour periods over one week, using the cluster inventory for VMs and PMs specifications. For each consolidation period, Algorithm 1 is applied over the inventory to compute the number of required PMs. VM placement is then conducted using a FFD heuristic. VMs affected by the shutdown policy are identified prior to the simulation process and are not considered for periods where these VMs are effectively shutdown. The resulting power consumption is then computed using PM power consumption traces. Two results are compared to this base power consumption: the average power consumption resulting from the sole use of consolidation, and the average power consumption resulting from the combination of consolidation and the shutdown policy.

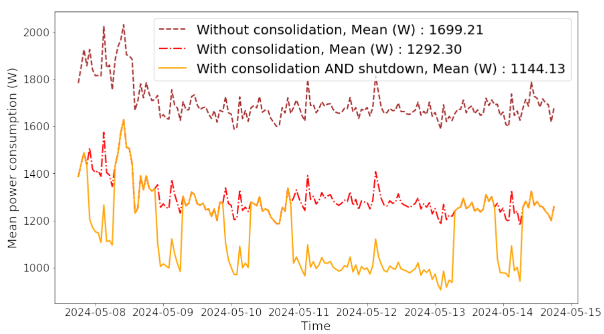


Fig. 3: Base, consolidated and combined power consumption

Fig. 3 illustrates the results of the consolidation. The sole use of consolidation enables a 23.95% power consumption reduction thanks to a better VM placement, the use of an increased memory allocation rate to 1:1, and the absence of fault tolerance in simulations. Thanks to memory management mechanisms such as Transparent Page Sharing[2], there is no risk of memory shortage on the PMs and therefore performance degradation. By combining the consolidation with the shutdown policy, an extra 8.72% reduction is observed, resulting in an overall 32.67% power consumption reduction. This additional consumption reduction is solely enabled by the actions of the VM owners who deliberately shutdown VMs when it is possible.

2) Experimental: To validate the simulation results, a production implementation is conducted over a period of one week. Native features of VMWare vSphere are used to apply the consolidation results on the hypervisor. As one cannot modify the source code, the implementation relies on PowerCli[3], using the *VM Host Affinity rules* feature, which enables to map groups of virtual machines to groups of physical machines. Each rule takes for parameter a VM group, a PM group and an affinity *specification*[4].

Thanks to the use of this feature, the VM placement and migration process is handled by the hypervisor scheduler: vSphere Dynamic Resource Scheduler (DRS). The *Should Run* policy where VMs of a group should, but are not required, to run on PMs of a group, is selected to implement the consolidation results. This enables the system to be reactive to potential PMs breakdown. To imitate PM shutdowns during the experimentation, unused PMs are placed in maintenance mode, meaning that they are not able to host any VM. Complete shutdown of the PMs would require interfacing the implementation with the cluster's monitoring system, which is outside the scope of the study.

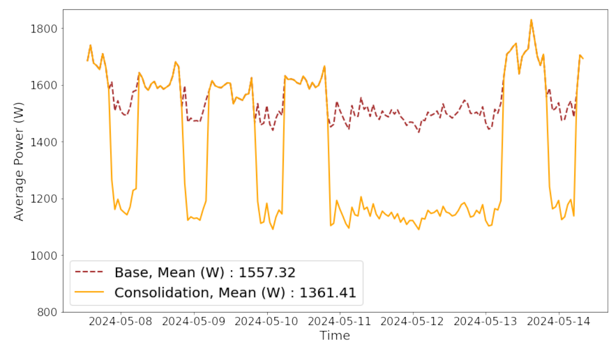


Fig. 4: Experimental results over one week

Fig. 4 presents the experimental results of the production implementation, which show differences and similarities with the simulation. First, the infrastructure base power consumption during the experimentation is lower than the simulation base power consumption. Indeed, the simulation base power consumption is based on the power consumption traces of PMs during a week where the shutdown policy was disabled. This results in a higher power consumption as VM shutdown has a direct impacts on PMs resource usage and power consumption. Second, the production implementation takes into consideration PM fault tolerance. A specific fault tolerance rate is defined by the infrastructure team and set to $2N$, N being the physical configuration of the biggest PM of the infrastructure in terms of CPU and memory. This means that the amount of available resource at all time must be equal or bigger than $2N$, allowing the breakdown of 2 of the biggest PMs without service interruption. Similar to the simulation results however, we can observe a power consumption reduction during nights and weekends. The power consumption reduction obtained from the consolidation approach over this one week period is of 12.58%. This reduction is here solely made possible by the adoption of the automatic shutdown policy by the users, through the VM shutdown operation itself and the consolidation resulting from these VM shutdowns. This power consumption reduction can be separated into 2 main

[2]https://docs.vmware.com/en/VMware-vSphere/7.0/com.vmware.vsphere.resmgmt.doc/GUID-FEAC3A43-C57E-49A2-8303-B06DBC9054C5.html

[3]https://developer.vmware.com/powercli

[4]https://docs.vmware.com/en/VMware-vSphere/7.0/com.vmware.vsphere.resmgmt.doc/GUID-2FB90EF5-7733-4095-8B66-F10D6C57B820.html

components: the first being the reduction in number of VMs to host during VM shutdown policy periods which enables to power-off one PM, the second being the shutdown of VMs themselves, which reduces the base power consumption of the infrastructure. Both components are enabled by the infrastructure users, who therefore act as an important lever.

B. Comparison with other approaches

	Bin Packing (BP)	ORCA	S-ORCA
Memory overcommitment	Yes	Yes	No
VM placement per group	BP Optimal Solver	FFD	FFD
Consolidation groups	Multiple	Multiple	One
Inter group migration	No	Yes	No
Use of prediction models	Yes	Yes	No
User involvement	No	No	Yes
Consumption reduction	19.51%	29.79%	35.79%

TABLE II: Approaches comparison in simulations

To evaluate the relevance of *S-ORCA*, which can be described as a socio-technical approach, we compare it with 2 other approaches defined in TABLE II. We conducted simulations using (1) a Bin Packing Solver with the *ORCA* consolidation groups as in TABLE I, (2) the *ORCA* methodology with the same consolidation groups and (3) *S-ORCA*. (1) separates the VMs into multiple consolidation groups for every consolidation period using the pre-trained prediction models. It then performs VM placement on the PMs using an optimal Bin Packing Solver. (2) uses the same consolidation groups but performs VM placement using a simple FFD heuristic. *ORCA* uses one feature that enables further power consumption reduction: inter-group migrations. After VM placement in the simulations, VMs can be migrated to better QoS (i.e. with a smaller overcommitment rate) groups if this reduces the number of used PMs. This leads to lower power consumption and better performance for VMs. (3) is the defined *S-ORCA* approach that sizes the infrastructure according to the shutdown policy and uses FFD for VM placement. As (1) and (2) are purely technical approaches, they can easily be applied to other virtualization infrastructures to reduce their electricity consumption. *S-ORCA (3)* offers better results in terms of power consumption reduction, but requires actions from infrastructure users which is first not always possible and may require further user behaviour understanding, later discussed in Section IV-C.

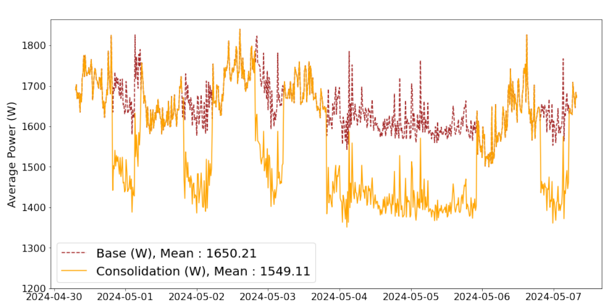
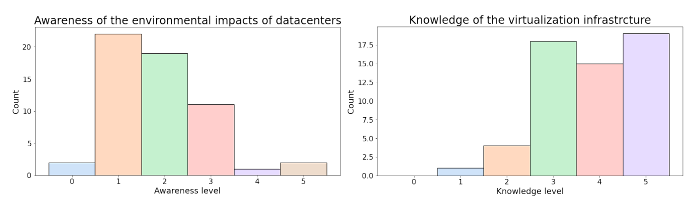


Fig. 5: *ORCA* experimental results on power consumption

We also conducted an experiment to compare the production results of *S-ORCA* with an implementation of *ORCA* using the three consolidation groups of TABLE I. Both experiments use the same fault tolerance rate of $2N$ and are carried out on consecutive weeks to limit the effects of "natural" infrastructure modifications. Fig. 5 shows the base and consolidated power consumption of *ORCA* on the experimental cluster. We can notice that power consumption reduction occurs over similar periods than with *S-ORCA*. This is explained by the fact that prediction models are trained over data where the shutdown policy was already implemented, therefore where a high number of VMs are unused at nights and weekends. The overall power consumption reduction provided by *ORCA* is here of 6.12% over 7 days. The base power consumption of the infrastructure is higher than for *S-ORCA* as the VMs affected by the shutdown policy are not powered-off during the *ORCA* experiment. In terms of power consumption, *S-ORCA* outperforms *ORCA*, with twice the reduction over 7 days. The implementation complexity is also drastically reduced, as *S-ORCA* does not require any prediction models training or updates, and is not dependent on the quality of VMs usage data as well as less exposed to potential performance degradation on VMs. During both experiments, no performance degradation was reported by the infrastructure users. Both *ORCA* and *S-ORCA* offers similar user experience as the prior infrastructure sizing, with a slight performance degradation risk for *ORCA*.

C. User behaviour study

As shown above, the users behaviour and actions enable a significant reduction in the cluster power consumption. To better understand their profiles, as well as their motivations and decision process, a study was conducted among the 83 VM owners of the infrastructure.

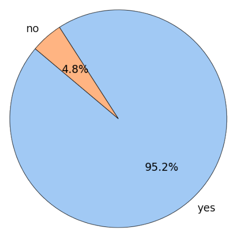


(a) Environmental awareness (b) Virtualization knowledge

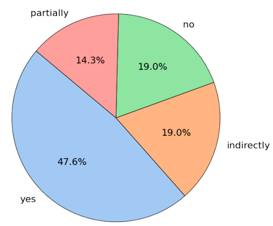
Fig. 6: Answers to the questionnaire

Fig. 6 presents the answers obtained to the questionnaire. Out of 83 VM owners, 52 answers were collected. The results show that the users have low awareness about the DCs environmental impacts, with a mean value of 1.87 on this topic. This corresponds to users having either vague ideas of the impacts of DC, or that saw some numbers or figures about this topic. On the opposite, the users answers underline a good technical knowledge of the virtualization infrastructure with an average answer of 3.82. This corresponds to users being aware of the VMs they are responsible of. The most given answer for this question is answer 5, i.e. a complete knowledge of all the VMs, their hosted applications or services and the associated shutdown policy tags. The questionnaire therefore emphasises a good technical knowledge of the infrastructure with a low awareness level of DCs environmental impacts among the users. This is consistent with the profiles of the users of this infrastructure, that contains software developers, integrators or system administrators for example. Reducing the environmental impact of DCs does not seem to depend on prior knowledge of these impacts here.

Among the 20 selected VM owners for the interviews, 19 were able to participate. Only one respondent was not aware of the existence of the shutdown policy, as shown in Fig. 7a.



(a) Policy knowledge



(b) Policy use

Fig. 7: (a) Did you hear about the VM shutdown policy ? (b) Do you use the VM shutdown policy ?

The answers given in Fig. 7b are sensible to the semantic understanding of the word “use”. For some users, it refers to the assignation of a policy tag on their VMs, whether or not this will result in a shutdown of a VM. For other users, this refers to the assignation of a tag that will shutdown their VM at one point. Some users replied that they did not use the policy, even though 100% of their VMs were actually tagged.

Questioning the users on the profiles of their VMs was another way to assess their technical knowledge of the infrastructure, and the interviewees were able to respond in the vast majority of the cases. Among the VMs, multiple tools and profiles were identified : VMs for software testing, software integration, infrastructure VMs, logs and data processing VMs, VMs for demos or for employees training program. These VM profiles rapidly enables to identify a corresponding shutdown policy tag. Regarding whether or not to shutdown VMs, all the users do so on the basis of the VM utility or usage.

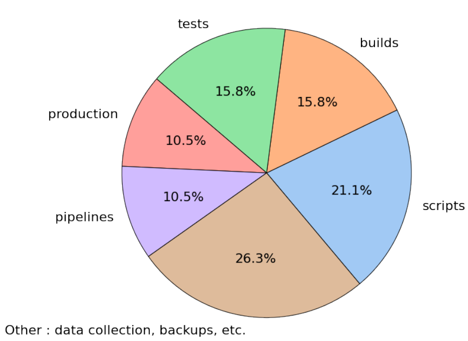


Fig. 8: Why are some of your VMs not powered off ?

Fig. 8 illustrates reasons to keep the VMs powered-on. The answers converge on keeping multiple parts of the infrastructure operational, which can be interconnected, from data collection to automatic tests and scripts executions for example. Powered-off VMs are those used exclusively during working hours, with a distinction between automatically and manually restarted VMs. Manually restarted VMs are generally used on rarer occasions, for specific manual tests for example.

As the shutdown policy requires additional actions from users to act as a lever for a power consumption reduction, the interviewees were asked about the time investment to initially implement the policy on their VMs as well as the need to periodically update the policy. The answers for the initial investment on the policy implementation usually highlights very little effort: *“very low”, “almost null”, “near zero”, “super easy”, “few clicks”, “15 minutes per year”*. For policy updates, only 15% of the respondent performs periodical checkups on their VMs to update the shutdown policy. This is in line with the responses obtained concerning potential errors or benefits after the shutdown policy implementation. Some users experienced errors or undesirable behaviours after the policy implementation, but were always able to correct them by adjusting the tags. Some users were even able to troubleshoot and fix bugs caused by the VM restart operations, identifying these as benefits obtained from the policy. Last, one user observed direct performance benefits on its VMs, especially on software compilation times.

Users were finally asked about their motivation on the usage of the VM shutdown policy, through the identification of the benefits of such mechanisms.fig Fig. 9 depicts these results.

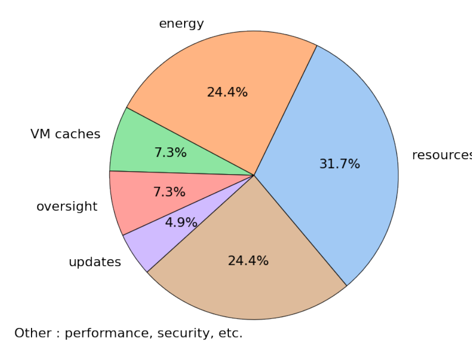


Fig. 9: Why do you power off VMs during the night ?

The first driver for our users is the resource usage and their potential wastage. As they have a technical profile, these users know that computation resources are not free, and that freeing some of them over periods of time will benefit to other VMs, thus enhancing performances. Linked to the resource usage, the electricity use appear as a motivation here, though there might be a bias as the interview was conducted by someone working directly on the DC electricity consumption. Other technical aspects such as VM caches, updates or security concerns appear to a lesser extent, validating again the technical profile of the users however.

This behavioural study was highly beneficial to identify a user profile enabling power consumption reduction in a virtualization infrastructure. Here, the low awareness on DC environmental impacts is counterbalanced by a high technical knowledge of the users, which do not fear service interruptions on their VMs. This is further reinforced by the fact that users are not directly rewarded for their actions: users are here employees of the company who act solely to do their best. One can imagine that customers would expect retribution for acting this way, or that their technical knowledge would be limited. It is therefore difficult to know whether the approach can be extended to a large number of infrastructures. To tackle this issue, the results were compared to the *ORCA* methodology described in Section III-A.

V. Conclusion and perspectives

Cloud infrastructures tend to be oversized nowadays partly due to user misbehaviour and the need for infrastructure operators to be able to meet peak demand. To reduce cloud infrastructures GHG emissions as well as other environmental impacts, CSPs need to deal with resources wastage that leads for example to increased electricity consumption. In this paper, we show that under the right circumstances, infrastructure users with no prior knowledge of DCs environmental impacts can positively affect their environment by acting at their scale. With an automatic VM shutdown policy, they are directly able to limit an infrastructure resource and power consumption. The combination of this tool with a simple consolidation algorithm reduces energy consumption by 32.67% in simulations, of which 8.72% is directly attributable to our users via the shutdown policy. *S-ORCA* also results in a 12.58% reduction in production without any performance degradation. This socio-technical approach that combines a consolidation heuristic and user involvement offers better power consumption reduction than purely technical methodologies.

In future works, we will study how both approaches can be combined at the scale of a company operating multiple infrastructures. Furthermore, we will evaluate how the implementation of a consolidation methodology can affect the infrastructure team's perception of the environmental impacts of data centres, as well as how this can help reducing other environmental impacts derived from LCA such as Abiotic Depletion (ADP) or Primary Energy use (PE) in the long-term.

References

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

[2] Ouafa Bentaleb, Adam S. Z. Belloum, Abderrazak Sebaa, and Aouaouche El-Maouhab. Containerization technologies: taxonomies, applications and challenges. *The Journal of Supercomputing*, 78(1):1144–1181, January 2022.

[3] Belen Bermejo, Carlos Juiz, and Carlos Guerrero. Virtualization and consolidation: a systematic review of the past 10 years of research on energy and performance. *The Journal of Supercomputing*, 75(2):808–836, February 2019.

[4] Adrien Berthelot, Eddy Caron, Mathilde Jay, and Laurent Lefèvre. Estimating the environmental impact of Generative-AI services using an LCA-based methodology. In *CIRP LCE 2024 - 31st Conference on Life Cycle Engineering*, pages 1–10, Turin, Italy, June 2024.

[5] Susanta Nanda Tzi-cker Chiueh. A Survey on Virtualization Technologies. *RPE Report*, 142, 2005.

[6] Vlad C. Coroamă and Friedemann Mattern. Digital Rebound - Why Digitalization Will not Redeem us our Environmental Sins. In *Proceedings 6th international conference on ICT for sustainability.*, Finland, 2019.

[7] Christina Delimitrou and Christos Kozyrakis. Quasar: resource-efficient and QoS-aware cluster management. In *Proceedings of the 19th international conference on Architectural support for programming languages and operating systems*, ASPLOS '14, pages 127–144, New York, NY, USA, February 2014. Association for Computing Machinery.

[8] Charlotte Freitag, Mike Berners-Lee, Kelly Widdicks, Bran Knowles, Gordon S. Blair, and Adrian Friday. The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns*, 3(8), August 2022. Publisher: Elsevier.

[9] Yiannis Georgiou, David Glesser, Krzysztof Rzadca, and Denis Trystram. A Scheduler-Level Incentive Mechanism for Energy Efficiency in HPC. In *2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, pages 617–626, May 2015.

[10] David Guyon, Anne-Cécile Orgerie, and Christine Morin. Energy - Efficient IaaS-PaaS Co-Design for Flexible Cloud Deployment of Scientific Applications. In *2018 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, pages 69–76, September 2018. ISSN: 1550-6533.

[11] David Guyon, Anne-Cécile Orgerie, Christine Morin, and Deb Agarwal. How Much Energy can Green HPC Cloud Users Save? In *PDP 2017 - 25th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*, Saint Petersbourg, Russia, March 2017.

[12] Bryan Harris and Nihat Altiparmak. Monte Carlo Based Server Consolidation for Energy Efficient Cloud Data Centers. In *2019 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 263–270, December 2019. ISSN: 2330-2186.

[13] Leila Helali and Mohamed Nazih Omri. A survey of data center consolidation in cloud computing systems. *Computer Science Review*, 39:100366, February 2021.

[14] Nikolas Herbst, Samuel Kounev, and Ralf Reussner. Elasticity in Cloud Computing: What it is, and What it is Not. In *10th International Conference on Autonomic Computing (ICAC 2013)*, pages 23–27, San Jose, CA, June 2013.

[15] IEA. Electricity 2024 – Analysis. https://www.iea.org/reports/electricity-2024, January 2024.

[16] Jonathan Koomey, Stephen Berard, Marla Sanchez, and Henry Wong. Implications of Historical Trends in the Electrical Efficiency of Computing. *IEEE Annals of the History of Computing*, 33(3):46–54, March 2011. Conference Name: IEEE Annals of the History of Computing.

[17] Miao Lin, Jianqing Xi, Weihua Bai, and Jiayin Wu. Ant Colony Algorithm for Multi-Objective Optimization of Container-Based Microservice Scheduling in Cloud. *IEEE Access*, 7:83088–83100, 2019.

[18] Maël Madon, Georges Da Costa, and Jean-Marc Pierson. Characterization of Different User Behaviors for Demand Response in Data Centers. In José Cano and Phil Trinder, editors, *Euro-Par 2022: Parallel Processing*, pages 53–68, Cham, 2022. Springer International Publishing.

[19] Eric Masanet, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey. Recalibrating global data center energy-use estimates. *Science*, 367(6481):984–986, February 2020.

[20] Anne-Cécile Orgerie, Marcos Dias de Assuncao, and Laurent Lefèvre. A Survey on Techniques for Improving the Energy Efficiency of Large-Scale Distributed Systems. *ACM Computing Surveys*, 46(4):31, April 2014.

[21] Issam Raïs, Anne-Cécile Orgerie, Martin Quinson, and Laurent Lefèvre. Quantifying the impact of shutdown techniques for energy-efficient data centers. *Concurrency and Computation: Practice and Experience*, 30(17):e4471, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.4471.

[22] Tilman Santarius, Jan C. T. Bieser, Vivian Frick, Mattias Höjer, Maike Gossen, Lorenz M. Hilty, Eva Kern, Johanna Pohl, Friederike Rohde, and Steffen Lange. Digital sufficiency: conceptual considerations for ICTs on a finite planet. *Annals of Telecommunications*, 78(5):277–295, June 2023.

[23] Chia-Ming Wu, Ruay-Shiung Chang, and Hsin-Yu Chan. A green energy-efficient scheduling algorithm using the DVFS technique for cloud datacenters. *Future Generation Computer Systems*, 37:141–147, July 2014.