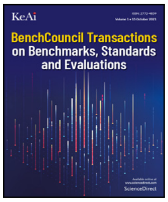
BenchCouncil Transactions on Benchmarks, Standards and Evaluations 2 (2022) 100060

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: https://www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/



Review article

Performance and energy consumption tradeoff in server consolidation

Belen Bermejo [*], Carlos Juiz

Computer Science Department, Universitat de les Illes Balears, Spain

ARTICLE INFO

Keywords:
Metrics
Consolidation
Performance
Energy
Servers
Benchmarking

ABSTRACT

Server consolidation is one of the techniques used to increase energy efficiency in datacentres. Nevertheless, the server consolidation has an inherent trade-off between performance degradation and energy consumption which has to be quantified to be managed. In this paper, the CiS^2 index is proposed to quantify the mentioned trade-off. We validated de use of the CiS^2 index through real experimentation. Also, these observations lead us to propose the second contribution, which focuses on the consolidation overhead. We proposed a general method to quantify this overhead and be able to manage its effect on performance degradation. To sum up, this paper improved the management of energy efficiency in datacentres' servers through the CiS^2 index and the server consolidation determination method.

1. Introduction

In the last years, organizations started to be concerned about the impact of information technology (IT) on energy consumption. For this reason, the Green IT initiatives appeared to make companies more environment-friendly [1,2].

The datacentres consume a huge amount of power and emit greenhouse gases in the form of CO_2. In a current datacentre, 30% of servers either are even not used or their utilization is very low, around 5%–10% [3,4]. Also, servers are the most power-demand device of a datacentre [5].

During last years, Green IT was used as an umbrella covering overlapping concepts like server consolidation and power management, among many others. Then, the aspiration of Green IT is achieving higher energy efficiency in the use of the IT devices and to increase the utilization of already installed devices in datacentres using the virtualization technology, specifically the server consolidation technique [1].

The server consolidation technique is based on the reallocation of virtual servers (could be virtual machines) among different physical servers using machine migration (see Fig. 1). As a consequence, the utilization of physical servers increases and the number of switched-on physical servers can be reduced.

Therefore, server consolidation increases the utilization of physical servers. However, due to the possibility of switch-off some physical servers, the power consumption is reduced. Nevertheless, as [6] states, the energy consumption depends on the overhead inherent to virtualization. The virtualization overhead is the extra workload that the physical server has to perform due to being virtualized, that is, tasks of managing virtual machines and coordinating the access to physical resources. As a consequence, the larger the number of consolidated virtual machines is, the higher the overhead is because of the coordination of simultaneously demanding resources access [5].

In certain cases, the energy-saving is not compensated with the performance degradation, which will be very high. It could also be the opposite case, that is, high performance of the datacentre may not be able to compensate servers to reduce it [5]. The current challenge in server consolidation is how to determine if a consolidated server is efficient or not in terms of energy consumption and performance degradation.

Therefore, the research question we attempted to solve in this work is: **could the performance-energy trade-off of physical servers when consolidating virtual machines be quantified?**

2. State of the art

In this work, we are interested in the server consolidation point of view tracking the management of these issues proposing several metrics to quantify the performance and the efficiency of a datacentre and servers.

The main developed work [7] explores the diverse metrics that are currently available to measure numerous datacentre infrastructure components behaviour. Also, they proposed a taxonomy of metrics based on datacentre dimensions. In addition, authors argue for the design of new metrics considering factors such as locations and resource co-locations, to assist in the strategic datacentre design and operations processes. One of the challenges authors announced is that it is hard to know the energy consumption due to datacentre sub-components, as operating systems and virtual machines. Due to that, in this work, we

---

[*] Corresponding author.
E-mail address: belen.bermejo@uib.es (B. Bermejo).

https://doi.org/10.1016/j.tbench.2022.100060
Received 26 April 2022; Received in revised form 10 May 2022; Accepted 10 May 2022

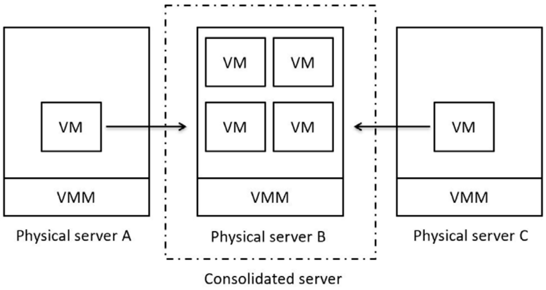


Fig. 1. Virtual machine consolidation, from [9].

proposed a metric for performance-energy trade-off, which took into account the number of allocated virtual servers.

On the one hand, the performance metrics attempt to quantify the suitability of the amount of work accomplished by a server or a datacentre. The values can be directly monitored from the system or inferred [8]. In the same manner, the energy and power metrics quantify the consumption of power or energy of a datacentre and/or a physical server. To take into account both previous aspects simultaneously, it is necessary to measure the relationship between performance and energy. These metrics relate to the performance of servers or datacentres with the power or energy consumption.

Power management metrics and techniques at different levels in datacentres are shown in [7]. System's administrators may measure information from software and hardware optimization. In this work, we are focused on metrics regarding software-oriented optimizations, specifically virtual machine consolidation, and, hardware-oriented optimizations, focused on the power and energy reduction in physical servers.

As we can observe from the previous works, all the used metrics are focused on the performance degradation and energy consumption, but, they are not considering the fact of having virtual servers consolidated. Then, with the current metrics it is not possible to know the efficiency of a consolidated server, or which number of virtual servers is more efficient in a specific scenario.

As a result, to the best of our knowledge, this is the first attempt to define metrics to quantify the performance and energy trade-off in server consolidation.

3. The CiS^2 index

In this section, we presented a new metric: Consolidated index for CPU- Server Saturation (CiS^2) which attempts to quantify the performance and energy trade-off taking into account the number of consolidated virtual machines (or containers) per server [10].

The CiS^2 index is defined as the product of the speed-up of the performance and the ratio of the consumed energy (see Eq. (1)). The speed-up of the performance is calculated as the ratio between the mean response time of the consolidation scenario and the physical server execution ($SP_p = R^c/R^p$). In the same manner, the ratio of the consumed energy is the division between the energy consumed in the consolidated scenario and the physical one ($SP_e = E^c/E^p$) [10].

Therefore, it is a simple quadratic efficiency and also the name of CiS^2. Also, in Fig. 2 the desirable index in function of the possible number of consolidated virtual machines (or containers) is represented. In the vertical axis, the combined performance and energy efficiency values are represented. In the horizontal axis the number of machines to be consolidated, either they are available in the datacentre or they are only considered for future capacity planning and forecasting bottlenecks.

$$CiS^2 = SP_p \cdot SP_e \tag{1}$$

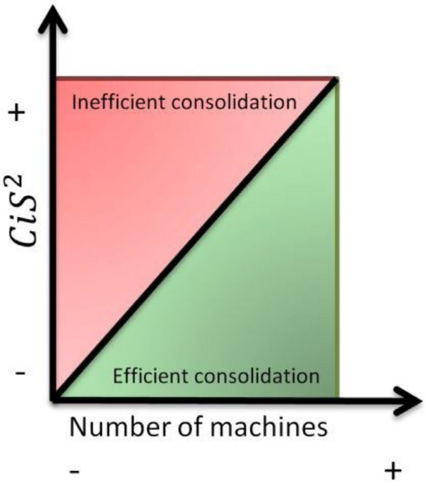


Fig. 2. CiS^2 index values and reference diagonal.

3.1. Graphical representation and interpretation

The representation of the CiS^2 index values, concerning the incremental number of consolidated machines, would be a square where the reference diagonal separated the scenario configurations of high performance and low energy from those that degrade and/or consume excessive energy about the following rules of thumb:

- N virtual machines consolidated in one physical machine should be N times slower than N physical machines (linear performance degradation).
- N virtual machines in one physical machine should consume as energy as N physical machines (energy conservation).

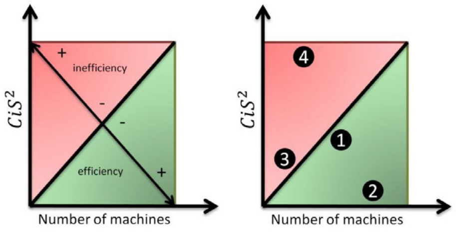
One the hand, the more consolidated virtual servers hosted in physical servers, the more performance degradation and consequently, the energy consumption would increase due to the increment of the mean response time. On the other hand, the more physical machines used, the more power is consumed, and consequently, the energy is also increased. Thus, we argue that it is possible to measure the balance between both situations. Due to that, the CiS^2 index compares different configurations in the performance-energy trade-off between different server consolidation scenarios.

From the energy efficiency point of view, the balanced efficiency metric shown through CiS^2 should be the one in which the average energy of a number of consolidated physical machines in a number of a virtual servers is exactly the same of using corresponding physical machines, i.e. the energy ratio is equal to one ($SP_e = 1$). However, from the performance speed-up point of view, the balanced efficiency shown through CiS^2 should be the one in which performance degradation is linear, that is, the slowdown is the same of the number of consolidated virtual servers per physical machine, i.e. N is the number of machines ($SP_p = N$).

3.2. Desirable values

Being CiS^2 values the result of the product of performance and energy speed-up, CiS^2 index acts as a qualifier of the consolidation of several virtual machines in comparison with this balanced (and pessimistic) reference diagonal described by the application of the rules of thumb.

Therefore, we also defined the CiS^2 reference diagonal as the imaginary border separating the" inefficient" CiS2 values (above the line) from the "efficient" CiS^2 values (below the lines) as we depicted in Fig. 3. This reference diagonal represents the linearity of the



(a)

(b)

Fig. 3. Desirable values o CiS^2 index.

Table 1
Physical servers of the experimentation.

Server	Number of CPUs	RAM size (GB)
Fujitsu RX600S5–1	48	1024
Dell PowerEdge T330	16	16
Dell PowerEdge T430	16	8
Dell PowerEdge R310	4	4
Dell PowerEdge T3400	2	8

consolidation in terms of performance and energy so that increasing consolidation would mean lowering proportionally the performance and also it means exchanging power by energy [11].

Having two different areas to distinguish among different consolidations one server or to compare different servers' consolidation spending on the area position of values, another interesting feature of the CiS^2 index for system administrators. Any consolidation configuration is more efficient or more inefficient, depending on the Euclidean distance of the index to the reference diagonal, above or below the reference diagonal, respectively as it is shown in Fig. 3a.

For example, the point 2 represented in 3b, which is on the green area, is more efficient than the point 1 because it is far from the diagonal. On the contrary, the point 3 represented in 3b, which is on the red area, is more efficient than the point 4 because it is closer to the diagonal than the point 4.

3.3. CiS^2 index evaluation

In the real experimentation, several factors should be considered, such as the hypervisor type, the benchmark or workload kind and the server hardware features. To simplify these factors, we represented the system using a black-box model. The workload is submitted in the system (consolidated servers) and we monitor the system behaviour (mean response time and power consumption) until the workload is completed [12,13].

The experimental set-up is composed of a set of different physical servers, which the number of CPUs and the RAM size are described in Table 1. Besides, the power consumption was measured by the Chroma 66200 power meter, the used hypervisors to deploy the consolidation are KVM, Virtual Box and Docker. In addition, it is important to note that the physical CPU executes the workload under the saturation condition, that is, the % of utilization is around 100%. The selected workloads are the Sysbench-CPU and the Stress-ng, which are intensive-based CPU workloads [14].

3.4. CiS^2 evaluation's results

In Fig. 4 the CiS^2 values for each physical server in the function of the number of consolidated virtual machines can be shown. The first that can observe is that the CiS^2 values have the same shape, that is, it starts increment and then it goes down when the physical machine has a certain number of allocated virtual machines. Therefore, there is an inflexion point that determines the number of minimum consolidated virtual machines the server needs to have a good CiS^2 value. The inflexion point depends on the physical server, and, as a consequence, it depends on the physical resources the server has.

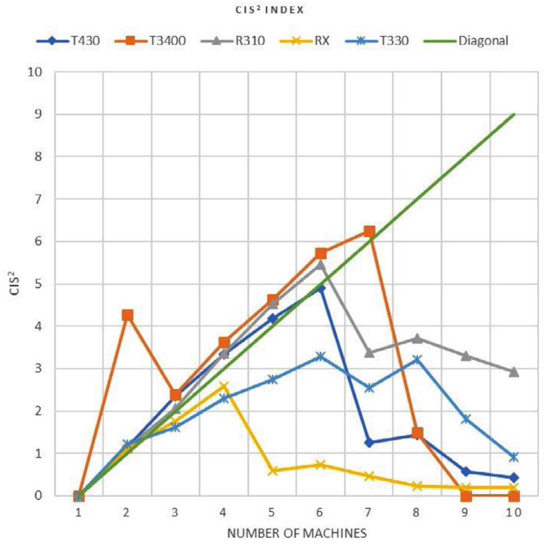


Fig. 4. CiS^2 index values.

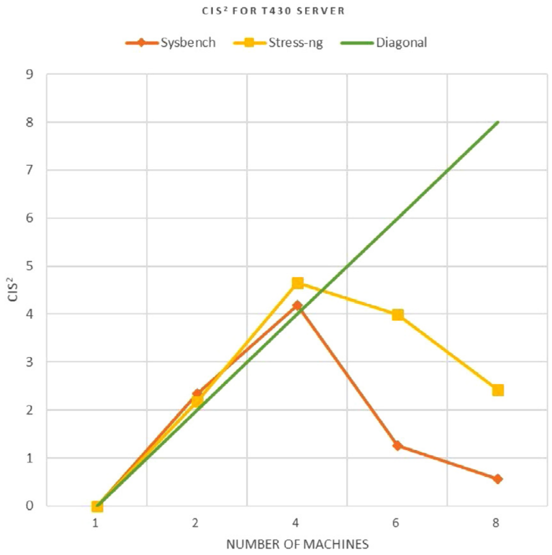
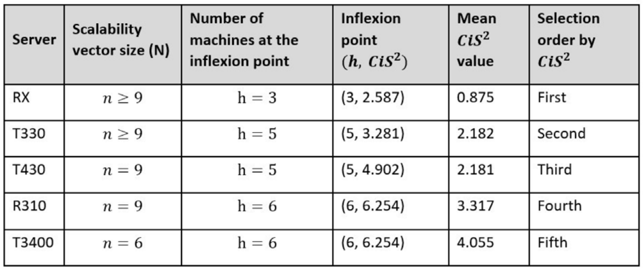


Fig. 5. CiS^2 index values for different workload.

Besides, the graphical representation of the CiS^2 index allows us to distinguish between the efficient and non-efficient consolidation configurations. For example, taking the T430 server from Fig. 4, it can be observed that $N = 6$ is more efficient than $N = 3$ because it is under the diagonal. In addition, for the HPI server, $N = 4$ is more efficient than $N = 3$ because it is far away from the diagonal.

Moreover, in Fig. 5 the CiS^2 index for the Sysbench workload in comparison with the Stress-ng workload for the T430 server is shown. It can be observed that for different nature of CPU workload, the behaviour of the CiS^2 index is the same. It starts growing, and it goes down after the inflexion point. Also, the inflexion point is the same for both workloads.

In previous sections, we stated that the CiS^2 index can be used for server's benchmarking and comparison. In Fig. 6 the physical servers by their consolidation efficiency considering the CiS^2 value is shown. It can be observed that the RX server is the most efficient because its CiS^2 value at the inflexion point is the lowest one.



Server	Scalability vector size (N)	Number of machines at the inflexion point	Inflexion point $(h,\ CiS^2)$	Mean CiS^2 value	Selection order by CiS^2
RX	$n \geq 9$	h = 3	(3, 2.587)	0.875	First
T330	$n \geq 9$	h = 5	(5, 3.281)	2.182	Second
T430	$n = 9$	h = 5	(5, 4.902)	2.181	Third
R310	$n = 9$	h = 6	(6, 6.254)	3.317	Fourth
T3400	$n = 6$	h = 6	(6, 6.254)	4.055	Fifth

Fig. 6. Server selection depending on the values of parameters obtained from benchmarking (sorted by server efficiency).

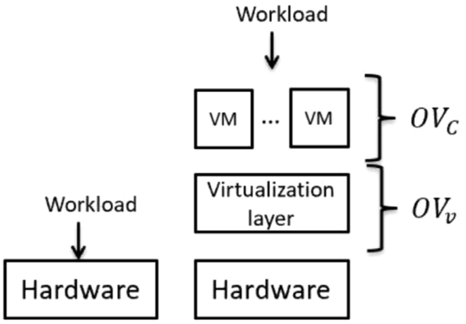


Fig. 7. Server consolidation overhead sources.

4. Consolidation overhead quantification method

As we observed in previous results, the value of the CiS^2 index depends on the hardware features, the number of allocated virtual machines and the workload nature, together with the performance degradation inherent to the server consolidation. Therefore, the second contribution of this work regards the server consolidation overhead [9].

The server consolidation overhead is defined as the extra workload that the system has to perform to manage the consolidation. This extra workload comes from the fact of having a hypervisor and the current access to physical resources from several consolidated virtual machines (or containers). Therefore, there are two sources of overhead (see Fig. 7) [15]:

- OV_v: overhead of virtualization.
- OV_c: overhead of consolidation.

Regarding the server consolidation overhead, the aim is to provide a general method for quantifying OV_v and OV_c. Let us define R^C and the mean response of the consolidated server, R^V as the mean response time of the physical server with a single consolidated virtual machine, and R^{PM} as the mean response time of the physical server. The OV_v can be defined as the difference between R^V and R^{PM} (see Eq. (2)). In the same manner, OV_c can be defined as the difference between R^C and R^V (see Eq. (3)).

$$OV_v = R^V - R^{PM} \tag{2}$$

$$OV_c = R^C - R^V \tag{3}$$

The main advantage of the proposed method is that it can be applied to any consolidation scenario considering any physical server, hypervisor and workload type.

The evaluation of the proposed method was performed using the previous experimental set-up, monitoring the mean response time of the required scenarios. We represented for each consolidation configuration the value of OV_v, OV_c and the useful work in percentage. The useful work represents the portion of the time that the system is executing just the workload, in this case, the CPU operations.

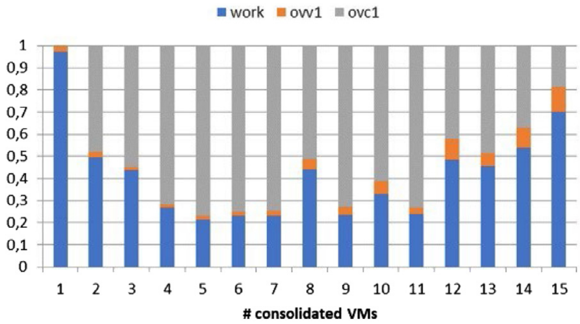


Fig. 8. Consolidation overhead representation for T430 server and KVM hypervisor.

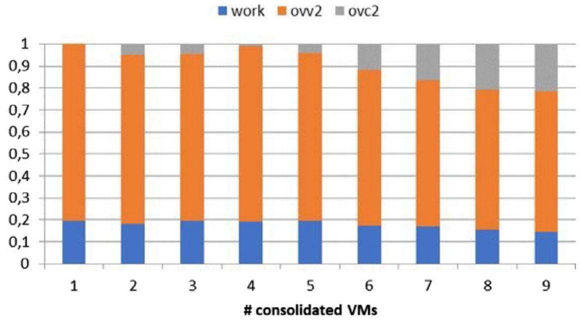


Fig. 9. Consolidation overhead representation for T430 server and Virtual-Box hypervisor.

In Figs. 8 and 9 the values of OV_v (blue), OV_c (orange) and %work (grey) for the T430 are represented, in function of the number of consolidated virtual machines. It can be observed, that for the KVM hypervisor the values of OV_v and OV_c are smaller than the values for the Virtual-Box hypervisor. Also, it can be seen that these values depend on the number of consolidated machines and the hypervisor. However, in any case, the consolidation is not for free, being more than 50% for Virtual-Box hypervisor configurations.

5. Conclusion and future work

This paper aims to measure the performance-energy tradeoff in server consolidation. Since there are no metrics to capture how server consolidation is managed considering the relationship between performance and energy, the CiS^2 index is proposed to achieve this aim. As the results show, this index can be applied to any type of server, under any virtualization platform and any level of use of its resources, in this case, the CPU. In addition, it enables the datacentre administrator to make better consolidation decisions thanks to the proposed graphical representation.

Also, the proposed index reflects a set of behaviours inherent to consolidated servers. The second contribution of this paper consists of the classification and quantification of the factors that affect the behaviour of server consolidation, in this case, two types of overhead (OV_v and OV_c). By the application of a simple method, these overheads can be quantified through the proposed method, which is also independent of the type of server, the executed workload, the virtualization and the percentage of CPU utilization.

Therefore, through this work, a step has been made towards a more efficient management of virtualized servers, and the datacentres. Now, the performance and energy balance of servers can be measured through the CiS^2 index and graphically analysed with a general method. Besides, system's administrators dispose of a method to go in-depth the overhead caused by the consolidation of servers and thus be able to make better decisions regarding the improvement of these systems.

As future work, the CiS^2 index can be extended to multiple devices. Also, it can be extended for scales workload and considering different workload distributions. Moreover, system properties could be described by the CiS^2 index. Regarding the overhead quantification method, it could be extended considering the power and energy consumption.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Uddin, A.A. Rahman, Energy efficiency and low carbon enabler green IT framework for data centers considering green metrics, Renew. Sustain. Energy Rev. 16 (6) (2012) 4078–4094.

[2] C.-J. Tang, M.-R. Dai, H.-C. He, C.-C. Chuang, Evaluating energy efficiency of data centers with generating cost and service demand, Bull. Netw. Comput. Syst. Softw. 1 (1) (2012) pp–16.

[3] L.A. Barroso, U. Hölzle, The case for energy-proportional computing, Computer (12) (2007) 33–37.

[4] L. Minas, B. Ellison, Energy Efficiency for Information Technology: How to Reduce Power Consumption in Servers and Data Centers, Intel Press, 2009.

[5] F. Abaunza, A.-P. Hameri, T. Niemi, EEUI: A new measure to monitor and manage energy efficiency in data centers, Int. J. Prod. Perform. Manag. 67 (1) (2018) 111–127.

[6] G. Lovász, F. Niedermeier, H. De Meer, Performance tradeoffs of energy-aware virtual machine consolidation, Cluster Comput. 16 (3) (2013) 481–496.

[7] A.M. Ferreira, B. Pernici, Managing the complex data center environment: An integrated energy-aware framework, Computing 98 (7) (2016) 709–749.

[8] X. Molero, C. Juiz, M. Rodeño, Evaluación Y Modelado Del Rendimiento De Los Sistemas Informáticos, Pearson Educación London, 2004.

[9] B. Bermejo, C. Juiz, C. Guerrero, Virtualization and consolidation: A systematic review of the past 10 years of research on energy and performance, J. Supercomput. (2018) 1–29.

[10] C. Juiz, B. Bermejo, The CiS2: A new metric for performance and energy trade-off in consolidated servers, Cluster Comput. 23 (4) (2020) 2769–2788.

[11] B. Bermejo, C. Juiz, C. Guerrero, On the linearity of performance and energy at VMC: the 2 index for CPU workload in server saturation, in: IEEE High Performance Computing and Communications, HPCC-2018, 2018.

[12] B. Bermejo, C. Juiz, N. Thomas, On the virtualization overhead and energy consumption in consolidated servers, in: UK- Performance Engineering Workshop, UKPEW, 2018.

[13] S.K. Panda, P.K. Jana, An energy-efficient task scheduling algorithm for heterogeneous cloud computing systems, Cluster Comput. (2018) 1–19.

[14] E. Casalicchio, A study on performance measures for auto-scaling CPU-intensive containerized applications, Cluster Comput. (2019) 1–12.

[15] B. Bermejo, C. Juiz, On the classification and quantification of server consolidation overheads, J. Supercomput. 77 (1) (2021).