Proceedings Paper:

Yang, R, Moreno, IS, Xu, J orcid.org/0000-0002-4598-167X et al. (1 more author) (2014) An Analysis of Performance Interference Effects on Energy-Efficiency of Virtualized Cloud Environments. In: Proceedings of the 2013 IEEE 5th International Conference on Cloud Computing Technology and Science. 2013 IEEE 5th International Conference on Cloud Computing Technology and Science (CloudCom), 02-05 Dec 2013, Bristol, UK. IEEE , pp. 112-119. ISBN 978-0-7695-5095-4

https://doi.org/10.1109/cloudcom.2013.22

An Analysis of Performance Interference Effects on Energy-Efficiency of Virtualized Cloud Environments

Renyu Yang[2], Ismael Solis Moreno[1], Jie Xu[1, 2], Tianyu Wo[2]

School of Computing[1]
University of Leeds
Leeds, UK
{scism, J.Xu}@leeds.ac.uk

School of Computer Science and Engineering[2]
Beihang University
Beijing, China
{yangry, woty}@act.buaa.edu.cn

Abstract **—Co-allocated workloads in a virtualized computing environment often have to compete for resources, thereby suffering from performance interference. While this phenomenon has a direct impact on the Quality of Service provided to customers, it also changes the patterns of resource utilization and reduces the amount of work per Watt consumed. Unfortunately, there has been only limited research into how performance interference affects energy-efficiency of servers in such environments. In reality, there is a highly dynamic and complicated correlation among resource utilization, performance interference and energy-efficiency. This paper presents a comprehensive analysis that quantifies the negative impact of performance interference on the energy-efficiency of virtualized servers. Our analysis methodology takes into account the heterogeneous workload characteristics identified from a real Cloud environment. In particular, we investigate the impact due to different workload type combinations and develop a method for approximating the levels of performance interference and energy-efficiency degradation. The proposed method is based on profiles of pair combinations of existing workload types and the patterns derived from the analysis. Our experimental results reveal a non-linear relationship between the increase in interference and the reduction in energy-efficiency as well as an average precision within +/-5% of error margin for the estimation of both parameters. These findings provide vital information for research into dynamic trade-offs between resource utilization, performance, and energy-efficiency of a datacenter.**

Keywords — Cloud computing, energy-efficiency, performance interference, performance estimation, workload analysis

I. INTRODUCTION

Cloud computing is experiencing rapid growth as it promises to reduce maintenance and management costs in comparison with in-house infrastructure [1,2]. Despite its commercial advantage of reduced energy consumption on the client side, Cloud providers still need to address a number of key challenges, such as striking a balance between optimal energy-efficiency and satisfying the increasing demand and high performance expectations of users. By using virtualization, the first generation of energy-efficient Cloud computing approaches have introduced mechanisms to dynamically resize the pool of servers based on actual demand [3, 4]. Additionally, others such as [5, 6] have proposed to extend these mechanisms with enhanced live-migration and server activation policies to reduce Service Level Agreement (SLA) violations. However, potential inefficiencies at a fine-grained level such as the overhead produced by the high competition for resources in virtualized environments [7] are usually ignored by these approaches. Consequently, their claimed energy-efficiency and performance improvements may be significantly diminished under real conditions.

Defined by [16], workload is "*the amount of work assigned to, or done by, a client, workgroup, server, or system in a given time period*". In the context of Cloud computing, workloads are the different tasks submitted by all the customers and executed at the Cloud providers' datacenters. Workloads by themselves have properties or attributes that describe their behavior. These attributes are normally expressed by the type and amount of resources consumed, geographical location requirement, or specific hardware constraints such as those described in [17]. As discussed in [18], as more and more customers adopt Cloud platforms to fulfill their IT requirements, Cloud providers need to be prepared for handling highly heterogeneous workloads to maximize the datacenter utilization. Hence, analyzing the impact that specific workload types have on others is critical to improve the Cloud datacenter's management.

In multi-tenant Cloud environments, workloads are generally encapsulated into Virtual Machines (VMs) and co-allocated into the same servers sharing the underlying physical infrastructure. Despite the environmental and fault isolation offered by virtualization, the high-competition for resources among running workloads will lead to a negative impact on the expected performance specified in SLAs. This phenomenon is known as *Performance Interference* and its effect on the Quality of Service (QoS) of workloads has been previously analyzed in [8-12]. However, current approaches have not yet considered the impact of such interference on a datacenter´s energy-efficiency. Understanding the relationship between performance interference and its impact on the energy-efficiency is critical if we are to design energy-efficient mechanisms that maintain performance under realistic environmental conditions.

Additionally, the levels of performance interference tend to vary significantly depending on the number and types of co-allocated workloads. In particular, for m different types of n workloads hosted in the same server there are C_{m+n-1}^{m} combinations. In production environments where the workload density per physical server tends to be high, a large n will lead to a considerable variety of combinations. Because Cloud datacenters are highly dynamic and transient environments, it is impractical for providers to characterize the levels of interference and the impact to the energy-efficiency produced by each possible mixture of workloads. Therefore, it is very important to rely on models to estimate both parameters in an effective way.

In this paper, we present an analysis of the performance interference impact on energy-efficiency by conducting experiments from three different perspectives: the decrement of work computed in a fixed period of time, the impact produced on different server configurations and the increment of elapsed time on a fixed amount of work. Furthermore, we propose an approach to quantify the levels of performance interference and energy-efficiency reduction when multiple workloads are deployed in the same virtualized node. The core idea is to exploit the real measurements taken from profiling pair-combinations of the existing workload types and the growth correlation patterns derived from the performed analysis. Then regression analysis is applied to determine the estimation models. In order to conduct this study, we emulate different workload types derived from the Google Cloud tracelog[14], and execute them on the iVIC Virtual Computing Infrastructure [15] to measure their interference and energy consumption. In addition, we evaluate the performance and energy-efficiency for different combinations of 3 to 9 workloads in order to analyze the phenomenon when the number of aggregated workloads grows. Our experimentation shows that while performance interference increases linearly, the impact on the energy-efficiency stop growing with the increment of co-allocated workloads, creating an exponential relationship. Moreover, the obtained results demonstrate that using these patterns and the profiled pair-combinations, it is possible to accurately estimate the levels of performance interference and energy-efficiency decrement when multiple workloads are co-allocated. The contributions of the work in this paper are:

- A comprehensive analysis conducted to determine the impact of performance interference on energy-efficiency in virtualized Cloud environments.
- An approach to estimate the interference levels and energy-efficiency decrement in virtualized nodes by conducting pair-combination profiling and exploiting the growth patterns outlined from the analysis.

The remaining sections are structured as follows: Section 2 introduces the problem of energy-efficiency decrement when virtualization interference occurs; Section 3 describes the performed analysis methodology and the proposed approach; Section 4 describes the experimental results; Section 5 presents the estimation for the interference level and the energy-efficiency degradation; Section 6 discusses related work; Section 7 presents the conclusions and discusses future work.

II. Decrement of Energy-Efficiency Due to Performance Interfence

The impact of performance interference in virtualized environments has been typically measured in terms of QoS such as throughput, latency or response time. However, performance interference induced by workload combinations also affects other critical factors that include the energy-efficiency of the overall datacenter. Specifically, when performance interference occurs, co-allocated workloads essentially fight for common resources while creating overhead that increases the energy consumption of individual servers.

This can be demonstrated with the example in which we have co-allocated 3 KVM-based VMs repeatedly running CPU-bounded workloads for 10 hours in the same virtualized server. The utilized server has the following characteristics: Intel Core i7 860@2.80GHz CPU (8-cores) and 16G RAM with Linux Debian 2.6.32. Each workload computes the 50th Fibonacci number using naive recursion. While the performance is measured in terms of execution time which is recorded when a workload is completed, the power is measured in 5 second intervals using a Voltech PM1000+ power analyzer.

Each workload requires on average 91.5 seconds to be completed when running in isolation, but when running all together the performance for some of the workloads is reduced in some periods of time during which the interference occurs. In this preliminary experiment, we observe that one workload primarily keeps the control of the resources considerably affecting the performance of the other two. This increases their execution time to around 178 seconds. We also observe that during the period of time when one single VM dominates the physical resources, the power consumption steadily remains about 115 Watts on average. On the other hand, during the periods of time when the three VMs have a fair access to the physical resources, the average execution time of each workload turns to 94 seconds indicating the reduction of the mutual interference. The corresponding power consumption in this case increases up to 135 Watts on average. Despite that the increase in power is close to 16%, it is still small in comparison to the performance improvement close to 50% for each affected workload. The experiment suggests that when the interference decreases, the energy-efficiency is improved due to the increase of operations computed per Watt consumed.

Therefore, it is important for providers to understand the characteristics of the co-existing workloads and the levels of interference that they produce. Moreover, approximating the interference and consequential decrement of energy-efficiency can facilitate mechanisms for efficient workload allocation and mitigate the negative effects of performance interference within Cloud environments.

III. Methodology of Analysis

A. Workload Characterization

The first step in order to analyze the impact of performance interference on energy-efficiency is to determine the characteristics of workloads from a realistic scenario. To this end, we analyze a tracelog made available by Google in [14] and derived a task clusterization based on their resource utilization patterns. The tracelog contains information about 930 different users submitting 25 million running-tasks on a Cloud Computing cluster composed of 12,000 servers for a period of a month. Furthermore, the tracelog offers information about the utilization ratios of the principal

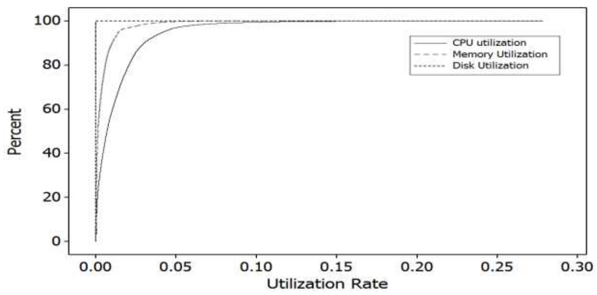


Figure 1. CDFs of the three analyzed resources.

resources: CPU, memory and disk for each running tasks. To determine the dimensions of our clusterization schema, we analyze the distributions of these three resources as well as the length of tasks from a representative sample. From this analysis we observe that while CPU and memory utilization vary significantly among all the analyzed tasks, the utilization of disk is almost uniform with 98% of the tasks consuming very similar proportions as observed in Fig. 1. Therefore disk usage is discarded and task's length, CPU and memory consumption are selected as the cluster dimensions.

Three different types of workloads can be outlined from the cluster centroid analysis which is obtained by applying *k*-means method [19] and the *k*-selection algorithm presented in [20]. These have been labeled as "*Small*", "*Medium*", and "*Large*" due to the proportions *P* of their 3 dimensions as presented in Table I. For example, *Medium* tasks are on average 5 times larger, using 5 times more CPU and 6 times more memory respectively than *Small* tasks. During the clusterization, the values for task's length, CPU, and memory are normalized based on the maximum and minimum values from the tracelog to avoid skewed results due to the use of different metric units.

TABLE I. CLUSTER CENTROIDS AND PROPORTIONS.

	Length	*P*	CPU	*P*	Memory	*P*
S	0.0007	1	0.0149	1	0.0089	1
M	0.0038	5	0.0810	5	0.0585	6
L	0.0107	15	0.2206	14	0.2556	28

B. Impact Analysis Based on Workloads Emulation

In order to analyze the impact of performance interference on energy-efficiency in a real environment, we emulate the 3 task types derived from the analyzed tracelog. Sysbench [21] "*memory-test*" is used to stress CPU and memory based on the task cluster centroids and their proportions *P* shown in Table I. Sysbench is a modular, cross-platform and multi-threaded benchmark tool for evaluating system parameters under intensive loads. In our emulation, each workload is synthesized by one or more Sysbench commands which execute a number of write operations on pre-established memory blocks creating required CPU and memory usage patterns. The CPU utilization of each type can be indicated by the number of Sysbench commands while the length of emulated type is determined by the total number of operations to be executed by the set of commands running on individual VMs. The emulation configuration for each workload type is presented in Table II.

TABLE II. WORKLOAD TYPES CONFIGURATION.

Type	Length (Number of Operations)	Sysbench Commands	Memory Allocation (MB)
Small	1707	1	60
Medium	8535	5	360
Large	23898	15	1680

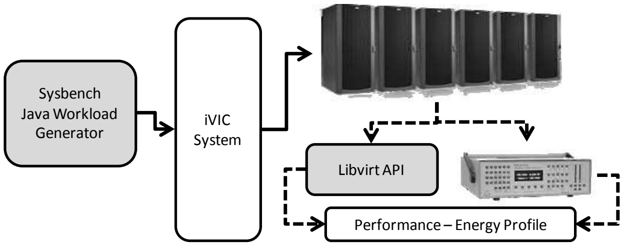

Sysbench Java Workload Generator
iVIC System
Libvirt API
Performance – Energy Profile

Figure 2. Overview of the assessment environment.

In order to assess the impact of interference on energy-efficiency, we setup a virtualized environment as illustrated in Fig. 2. We implement a workload generator that continuously submits and co-allocates instances of the emulated workload types in a virtualized cluster of 32 physical nodes managed by iVIC system [15, 22]. iVIC is a KVM-based Virtual Computing Infrastructure which provides flexible on-demand access to virtual computing environment on top of shared resources. It allows users to dynamically create, customize, migrate and scale VMs over clustered physical servers. The characteristics of the utilized servers are listed in Table III. The resource utilization of each workload is recorded using the libvirt API whilst the performance is calculated based on the number of operations completed per workload type and corresponding completion time. The transient power and total energy consumption is monitored through a Voltech PM1000+ power analyzer unit. In this environment, we conduct the experimentation and analysis in 3 different scenarios:

i) Over a fixed period of 12 hours we continuously submit different combinations from 2 to 9 workloads on servers T1500. The objective is to evaluate the impact of different workload combinations on the produced amount of interference and the reduction of energy-efficiency. Additionally, we analyze the performance and energy patterns when the number of co-allocated workloads increases by randomly selecting combinations from 3 to 9 workloads.

ii) Over a fixed period of 12 hours we continuously submit different pair-combinations of workloads on servers T3400. The objective is to compare the levels of

TABLE III. USED SERVER CONFIGURATIONS.

Server Family	Description
Dell Precision T1500	Intel Core i7 860, 2.80GHz CPU(8cores),16G RAM, Linux Debian 2.6.32
Dell Precision T3400	Intel Core 2 Duo, 2.33Ghz, 8GB RAM, Linux Debian 2.6.32

interference introduced by different server configurations processing the same workload types.

iii) We continuously submit pair-combinations of workloads until a fixed amount of operations are completed. The objective is to analyze changes on workloads' completion time which can significantly increase the overall energy consumption.

C. Interference and Energy-Efficiency Decrement Metrics

The effects of performance interference in each workload combination are measured by extending the Combined Score (CS) proposed in [11] to calculate the "*Combined Interference Score*" (CIS) as described in (1).

$$CIS(s) = \sum_{i=1}^{n} \frac{P_i - B_i}{B_i} \tag{1}$$

Where n is the total number of co-allocated workloads in the server s, P_i is the performance of the i-workload when combined with others, and B_i is the performance of the i-workload when running in isolation. Regarding to the decrement of energy-efficiency, it is calculated as described in (2) where E is the expected energy-efficiency and A is the actual energy-efficiency obtained for each combination. In both cases energy-efficiency is defined as the ratio of work (performed or expected) by the total amount of energy consumed. While the expected work is the aggregated operations computed by individual workloads when running in isolation, the completed work is the total operations achieved by all combined workloads.

$$\Delta EE(s) = \frac{E - A}{E} \tag{2}$$

In order to determine the expected amount of work, we initially benchmark the performance of each workload type when running in isolation for 12 hours on servers from the described families. Considering the number of completed executions during the fixed period of time and the amount of operations per execution as described in Table II, we obtain the total completed operations for each type as presented in Table IV.

IV. Experimental Results

A. Impact on Energy-Efficiency Considering Fixed Time

When workloads are combined and performance interference is produced, there is a significant impact on the amount of completed operations in comparison to the expected number. Therefore, the energy-efficiency is negatively impacted since the number of operations per Watt consumed is drastically reduced. As observed in Table V, the ΔEE increases along with the CIS for each evaluated combination. However, while the performance affectation linearly grows, the impact on energy-efficiency decreases in relation to the number of co-allocated workloads. For example, the average increment from combining 2 to 3 workloads is 1.045 and 0.2635 whilst from 3 to 4 workloads is 0.97 and 0.114 for CIS and ΔEE respectively. That is, while CIS increment remains close to 1.0, the ΔEE is proportionally reduced by 56%. These trends can be observed in Fig. 3 where all the evaluated combinations are plotted.

TABLE IV. BENCHMARK OF WORKLOADS IN ISOLATION

Workload type	Dell T1500		Dell T3400	
	Executions	Total Operations	Executions	Total Operations
Small(S)	2046	3492522	1085	1852095
Medium(M)	427	3644445	217	1852095
Large(L)	149	3815145	78	1997190

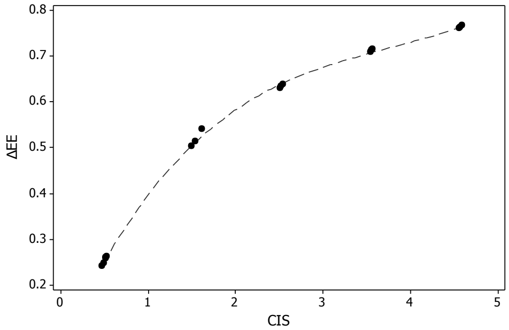
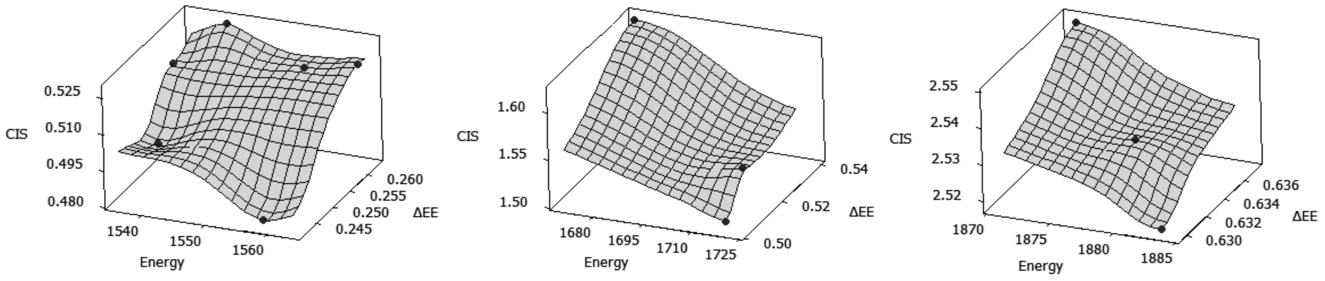


Figure 3. Relationship between performance interference and energy-efficiency degradation of the analyzed workloads combinations.

Another important observation from the results in Table V is that apart from the number of co-allocated workloads, different combinations produce different impact on the performance and energy-efficiency of virtualized servers. For example, in the case of pair-combinations, *LL* produces less interference than *MM* but more than *SS*. This creates the opportunity for developing workload-aware scheduling mechanisms to reduce the performance degradation while the energy-efficiency is maintained at a high number of co-



(a) (b) (c)

Figure 4. Impact on performance interference, energy consumption and energy-efficiency when combining (a) two, (b) three and (c)four workload types.

TABLE V. PERFORMANCE AND ENERGY PROFILES OF WORKLOAD COMBINATIONS RUNNING IN SERVERS T1500.

Combination	Work (Operations)		Energy (Whr)	CIS	ΔEE
	Completed	Expected			
SS	5298528	6985044	1558.92	0.482	0.241
SM	5283165	7136967	1555.80	0.519	0.259
SL	5424846	7307667	1535.88	0.516	0.257
MM	5385585	7288890	1563.60	0.522	0.261
ML	5496540	7459590	1541.64	0.527	0.263
LL	5735520	7630290	1538.40	0.496	0.248
MMM	5317305	10933335	1717.80	1.540	0.513
LMS	5448744	10952112	1718.94	1.507	0.502
SMM	4958835	10781412	1669.60	1.619	0.540
SSLM	5300235	14444634	1879.00	2.532	0.633
SSMM	5175624	14273934	1871.10	2.549	0.637
SSSS	5173917	13970088	1883.60	2.518	0.629
SMMLL	5305356	18411702	2022.30	3.559	0.711
MMMLL	5291700	18563625	2021.00	3.575	0.714
SSSML	5214885	17937156	1998.40	3.546	0.709
SMMMMM	5074911	21714747	2027.60	4.597	0.766
SSSMMM	5069790	21410901	2033.70	4.578	0.763
MLLLLL	5428260	22720170	2044.32	4.566	0.761

allocated workloads. The impact on performance, energy, and energy-efficiency decrement produced by different combinations is observed in Fig. 4 where co-allocations of 2, 3 and 4 workloads are illustrated. An important observation from this graph is that although different workload combinations produce a different interference level, this difference is significantly reduced when the number of co-allocated VMs grows. This suggests that at higher VM density the interference is mainly driven by the number of co-allocated VMs and slightly influenced by the different combinations mixture.

B. Impact Comparison Between Server Configurations

Running the same workload types on servers with different characteristics produces diverse performance interference and energy-efficiency decrement patterns. Because of the capacity of T3400 servers is almost 50% lower, the level of performance interference is as expected higher in comparison to T1500 servers. As observed in Fig. 5, the average CIS for pair-combinations in T1500 and T3400 are 0.5107 and 0.7732 respectively. This represents a nearly 51% increment compared with the CIS in T1500 that matches the proportions of the servers' capacities. However, what is more important to observe is the different impact level that the analyzed combinations create on different server configurations. For example, the *SM* and *MM* combinations running on T1500 have proportionally higher impact than those running on T3400. Therefore, in order to mitigate the performance interference while efficiently exploiting the resources in virtualized environments, it is important to understand not only the relationship between performance interference and energy-efficiency but also the impact that workload combinations have on the different server configurations.

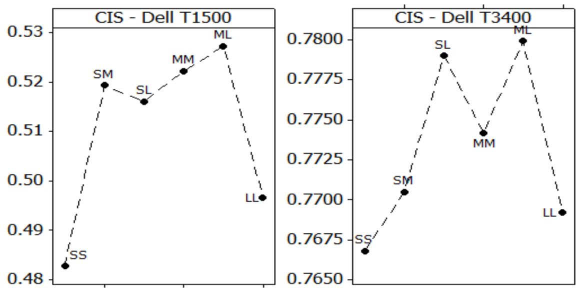


Figure 5. Performance interference comparison of workload pair-combinations with different server configurations.

C. Impact on Energy-Efficiency Considering Fixed Amount of Work

Besides the impact on the amount of work computed per Watt consumed, the performance interference can also increase the completion time of co-allocated workloads as well as the energy consumption at the datacenter. Namely, workloads running for longer time require more resources for them to be completed. This can be especially the case of those long-term computing-intensive applications running on Cloud environments. To evaluate time delays and their consequent increase in energy consumption, we continuously submit pair-combinations of workloads until the expected number of operations for 12 hrs are completed. For example, in the case of the combination *SM*, the expected amount of work for 12 hrs running on servers T3400 is 3,704,190 operations according to the values in Table IV. However, when interference occurs the required time to complete the same amount of work is extended by 7.52 hrs. This produces an increase in the energy consumption close to 64% due to an average execution delay on each workload equivalent to 62.38 seconds. Although the delay per execution is short, the aggregated time produces a high impact on the overall energy consumption in a long term. Table VI lists the time delays measured for all the pair-combinations as well as the increment on energy consumption introduced by performance interference for each case.

TABLE VI. DELAYS INTRODUCED BY PERFORMANCE INTERFERENCE.

Workload Combination	Average Delay per Execution (Seconds)	Total Workload Delay (Hrs)	Energy Increment (Watts)
SS	12.34	7.44	1297.50
SM	62.38	7.52	1310.67
SL	176.77	7.66	1324.00
MM	20.71	7.49	1316.50
ML	23.43	7.57	1314.00
LL	92.87	7.61	1322.00

V. ESTIMATING INTERFERENCE AND ENERGY-EFFICIENCY DECREMENT

In this section we describe a simple but effective approach to estimate the CIS and the ΔEE when multiple workloads are co-allocated. It is based on the exposed correlation patterns between CIS and ΔEE as illustrated in Fig. 3 and on the profiling of workload pair-combinations as presented in Table V.

In order to obtain an effective approximate estimation for *CIS* and ΔEE, we calculate the pair-based Combined Interference Score (*pbCIS*) of *n*-workloads by adding the resulting C_n^2 pair-combinations from server *s*.

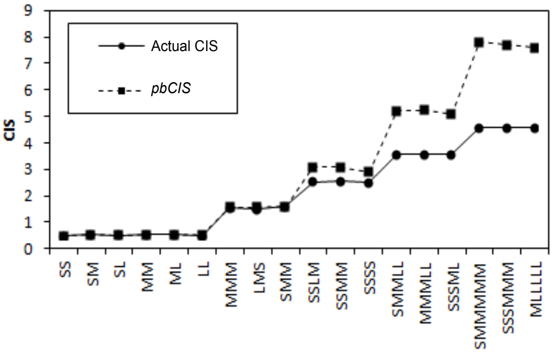


Figure 6. CIS estimation based on pair-combinations compared against the actual CIS.

$$pbCIS(s) = \sum_{i=1}^{{}^{n}C_{2}} CIS(p_i) \qquad (3)$$

Where p_i is an element of the set $P = \{SS, SM, SL, MM, ML, LL\}$ and $CIS(p_i)$ is the measured CIS of each pair-combination. For example, to estimate the *pbCIS* of the combination *LMS,* the measured values of *LM*, *LS* and *MS* are added together.

If we estimate the CIS based only on the effect of pair combinations, it is possible to observe that while the actual CIS grows linearly, the estimation of *pbCIS* in Eq. (3) grows exponentially producing a substantial margin of error particularly when *n* increases as illustrated in Fig. 6. The reason for this is because although the actual *CIS* varies according to the workload combinations as illustrated in Fig. 4, the variation is also significantly influenced by the increments of *n.* Consequently, the estimation of *CIS* for a given *n*-workload combination depends on two variables: the number of co-allocated workloads *n* and their resulting *pbCIS.* Therefore, the estimation model need to find a fitting *f* and *g* as in Eq. (4) and (5).

$$EstimatedCIS(s) \;=\; f\,(pbCIS(s),\, n) \qquad (4)$$

$$Estimated\ \Delta\mathrm{EE} = g\ (EstimatedCIS(s)) \qquad (5)$$

The fitting function can be outlined and determined by using regression analysis based on the data and patterns derived from the mixing workloads experimental results in the previous section. Analyzing the relationship of these variables, it is observable from Fig. 7 that while *CIS* linearly grows along with *n*, the impact on *CIS* produced by *pbCIS* gets reduced when the latter increases. Therefore, we can formalize the observed relationships leveraging the linear and quadratic regression analysis based on the data obtained from combining 2 to 6 workloads. These are described in Eq. (6) and Eq. (7) for *CIS(n)* and *CIS(pbCIS)* respectively. In addition, we combine these equations to approximate *CIS* based on *n* and *pbCIS* for servers T1500 as presented in Eq. (8). The estimation of CIS using these equations is contrasted to the actual measurements described in Section IV. As observed in Fig. 8, this produces a better fitting in comparison to only using the *pbCIS.*

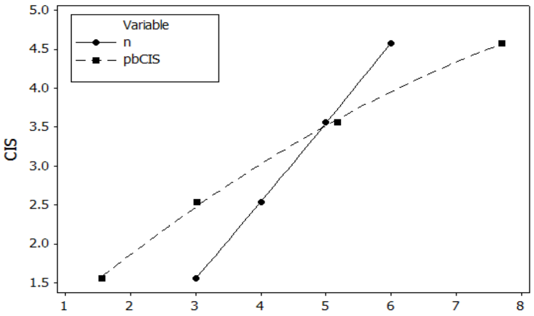


Figure 7. Relationship of CIS with n and pbCIS.

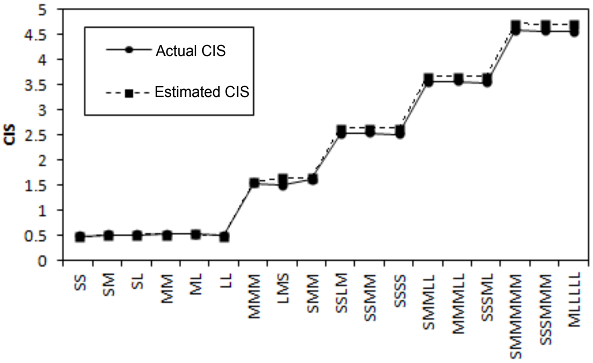


Figure 8. Estimated CIS compared against actual CIS.

$$CIS(n) = 1.011n - 1.492 \qquad (6)$$

$$CIS(pbCIS) = 0.5212pbCIS - 0.0084pbCIS^2 + 0.9538 \qquad (7)$$

$$CIS(n, pbCIS) = 0.505n + 0.260pbCIS - 0.004pbCIS^2 - 0.269 \; (8)$$

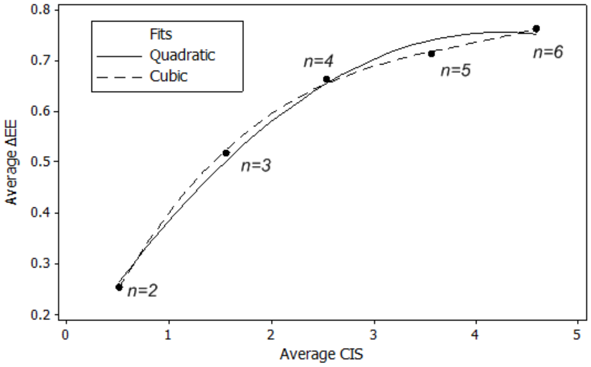


Figure 9. Regression fit of CIS and ΔEE centroids.

In order to approximate the reduction in energy-efficiency produced by *CIS*, we cluster the points in Fig. 3 based on the number of workloads. Driven by the data distribution on the graph, we fit all the cluster centroids to quadratic and cubic models as illustrated in Fig. 9. Applying regression analysis, it is determined that quadratic and cubic models fit the centroids distribution in 98.10% and 99.7% respectively. Finally, substituting the values of estimated *CIS* in the generalized cubic model and the parameters derived from the regression analysis, we can approximate the *ΔEE* for T1500 servers as described in Eq. (9). The evalution results can be observed in Fig. 10 where the estimated ΔEE is compared against the actual ΔEE measured during the experimentation.

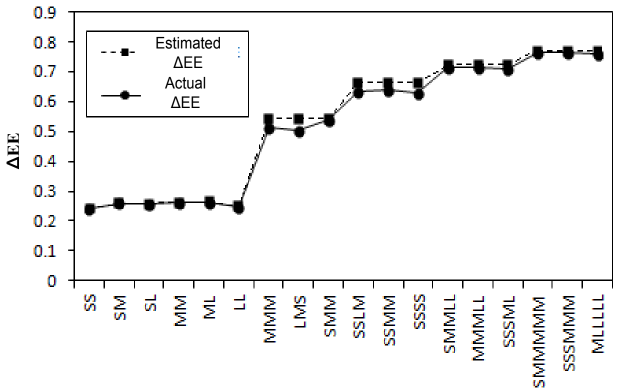


Figure 10. Estimated ΔEE compared against actual ΔEE

$$\Delta EE(CIS) = 0.310CIS - 0.048CIS^2 + 0.002CIS^3 + 0.123 \qquad (9)$$

The models are derived using the data obtained from combining 2 to 6 workloads. However, we also assess the accuracy against the data obtained from combining 7, 8 and 9 workloads. The derived estimation models closely match the real measurements for both CIS and ΔEE. As observed in Table VII, the average error in both cases is contained within a margin of 5%. The largest discrepancies can be found at very low number of co-allocated VMs where the workload types have a stronger influence. However as *n* increases the percentage of error is noticeably reduced. It is also important to remark that the coefficients of the models change for each different server configuration even considering the same workload types because as discussed previously the levels of *CIS* and *ΔEE* are affected by the capacity and characteristics of the physical server. Nevertheless, the proposed estimation approach can be generally applied to different architectures and workload types to determine the performance interference and energy-efficiency models of specific environment. In fact, characterizing few combinations to derive the parameters of the estimation model following the proposed approach is more feasible than evaluating every possible workload combination.

TABLE VII. ESTIMATED CIS AND ΔEE RESULTS

Workload Combination	**Est. CIS**	**Real CIS**	**% Error**	**Est. ΔEE**	**Real ΔEE**	**% Error**
MMM	1.656	1.540	**7.53**	0.542	0.513	**5.63**
LMS	1.656	1.507	**9.89**	0.542	0.502	**7.97**
SMM	1.655	1.619	**2.22**	0.542	0.540	**0.45**
SSLM	2.637	2.532	**4.15**	0.663	0.633	**4.79**
SSMM	2.637	2.549	**3.45**	0.663	0.637	**4.07**
SSSS	2.624	2.518	**4.21**	0.662	0.629	**5.18**
SMMLL	3.662	3.559	**2.89**	0.721	0.711	**1.42**
MMMLL	3.664	3.575	**2.49**	0.722	0.714	**0.99**
SSSML	3.654	3.546	**3.05**	0.721	0.709	**1.74**
SMMMMM	4.723	4.597	**2.74**	0.769	0.766	**0.43**
SSSMMM	4.714	4.578	**2.97**	0.769	0.763	**0.77**
MLLLLL	4.707	4.566	**3.09**	0.768	0.761	**1.00**
LLMMMMS	5.620	5.600	**0.36**	0.803	0.800	**0.38**
LLLMMSS	5.605	5.573	**0.57**	0.802	0.796	**0.85**
LMSSSSS	5.548	5.575	**0.48**	0.802	0.796	**0.70**
LLLMMSSS	6.654	6.621	**0.50**	0.815	0.827	**1.39**
LLLLMMMS	6.669	6.608	**0.92**	0.815	0.826	**1.23**
LMMSSSSS	6.625	6.615	**0.15**	0.815	0.826	**1.38**
LLMMSSSSS	7.643	7.608	**0.46**	0.834	0.845	**1.21**
LLLLLMMSS	7.661	7.600	**0.80**	0.835	0.871	**4.18**
LLLMMMMSS	7.682	7.628	**0.71**	0.835	0.846	**1.28**

VI. RELATED WORK

The negative effect of performance interference in virtualized environments has been previously analyzed. This section describes and discusses the most relevant related work approaching the problem. Younggyun et al. [12], present a study that evaluates the performance impact of co-allocating pairs of different applications in virtualized servers by analyzing system-level characteristics including CPU, memory, and disk utilization. In this paper the authors proposed a model to predict the performance of a new incoming application based on previous observations. Gupta et al. [13], discuss the sources of interference at Xen's Virtual Machine Monitor (VMM) for I/O intensive workloads. They propose a set of primitives implemented at hypervisor-level to improve the resource sharing mechanisms and mitigate the performance impact caused by co-allocated VMs. Pu et al. [11], present a complete analysis of performance interference in Xen hypervisor. In this analysis they demonstrate that co-allocating different types of workloads reduces the performance interference in virtualized environments. Moreover, they present a set of performance metrics to outline points of conflict among the studied workloads. Govindan et al. [10], also analyze the phenomenon of performance interference at Low-level Cache (LLC). They propose a technique to predict the performance interference due to shared processor cache using synthetic cache loader benchmarks to profile the performance of mixed applications.

It is observable that the main related approaches have focused on QoS aspects but completely neglected the impact on energy-efficiency produced by this phenomenon. If this is not considered, it can drastically diminish the claimed energy-efficiency improvements by energy-aware mechanisms when applied under real conditions. Furthermore, most of previous analyses with the exception of [10] have been limited to the study of pair-combinations. However, in real virtualized multi-tenant environments where multiple VMs can be allocated in the same server the combined workload is significantly more complex. Finally, previous studies are largely based on unrealistic workload characteristics that can lead to misleading results in real operational environments. This is mainly caused by the lack of tracelogs and their consequent workload analysis from real and large-scale Cloud computing environments.

VII. CONCLUSIONS AND FUTURE WORK

In this paper we have characterized workload heterogeneity derived from a real Cloud environment, and presented a comprehensive analysis to assess the impact of performance interference on a virtualized datacenter's energy-efficiency. Moreover, we have presented an approach to estimate both performance interference and energy-efficiency decrement based on workload pair-combination profiles and the correlation patterns derived from the presented analysis. Experimental results demonstrate an exponential relationship between the increase in interference and the reduction in energy-efficiency. Additionally, they also demonstrate that using the outlined correlation patterns as well as the measurements taken from pair-combinations it is possible to accurately estimate both parameters when multiple workloads are co-allocated. From our presented study, the following conclusions can be drawn:

- *Performance interference significantly affects the energy-efficiency. However, the intensity of that impact is reduced according to the increment of co-allocated workloads.* As indicated in our experimental results, the impact on energy-efficiency is reduced when the number of workloads increases despite the considerable rise of interference levels and its associated energy-efficiency decay. Therefore, the interference not only affects the QoS of individual workloads, but can also produce a significant impact on the energy-efficiency of virtualized servers and the overall datacenter if not properly handled. It makes critical the need for mechanisms to find a tradeoff between the QoS guarantees and the levels of the energy-efficiency.

- *Although creating a general interference estimation models is complex, it is still feasible to estimate the performance interference and its impact to the datacenter for concrete scenarios*. Our results show that interference is affected not only by workload types but also by server characteristics. However, by exploiting the exposed patterns and profiles of pair-combination, it is possible to approximate the interference levels and energy-efficiency impact by deriving platform-specific models.

- *Understanding the levels of interference and the impact to the energy-efficiency produced by the combination of diverse workloads can lead to an improved resources allocation in virtualized environments.* This allows the development of scheduling mechanisms to select the hosting servers based on the amount of interference produced by the current co-allocated workloads during specific instants of time.

- *Relying on real data is critical to understanding the real challenges in Cloud Computing and formulating assumptions under realistic operational circumstances.* This is especially true in very dynamic environments such as Cloud datacenters, where precise behavioural modeling is required to improve environmental energy-efficiency.

As future work, we are planning to perform more experimentation to determine what other factors affect performance and energy-efficiency in Clouds, such as the effect of using different hypervisors. Furthermore, a deeper study about the exposed interference impact on energy-efficiency needs to be conducted in order to formulate holistic models considering hardware, software, and workload patterns. It is also necessary to develop a framework of practical tools to effectively conduct the described analysis under different environmental characteristics. Finally, we are interested in evaluating the impact of performance interference on energy-efficiency when resources in the Cloud datacenter are over-allocated, in order to improve server availability whilst reducing interference effects.

ACKNOWLEDGMENTS

The work in this paper has been supported in part by the National Basic Research Program of China (973) (No. 2011CB302602), the Mexican Council of Science and Technology CONACyT (No. 213247), China 863 program (No .2011AA01A202), the UK EPSRC WRG platform project (No. EP/F057644/1) and National Nature Science Foundation of China (No.91118008, 61272165, 61170294).

REFERENCES

[1] D. Amrhein, *et al.*, "Cloud Computing Use Case," Cloud Computing Use Case Discussion Group, White paper, 2010.

[2] B. Gain. (2010, January 1) Cloud Computing & SaaS In 2010 *Processor Mag.* 12.

[3] R. Nathuji, *et al.*, "Exploiting Platform Heterogeneity for Power Efficient Data Centers," in *Proc. of the IEEE International Conference on Autonomic Computing* Washington, DC, USA 2007, pp. 5-15.

[4] K. Ley, *et al.*, "Cost- and Energy-Aware Load Distribution Across Data Centers," presented at the 22nd ACM Symposium on Operating Systems Principles, Montana, USA, 2009.

[5] R. Buyya, *et al.*, "Energy-Efficient Management of Data Center Resources for Cloud Computing: A Vision, Architectural Elements, and Open Challenges," in *Proc. of the 2010 International Conference on Parallel and Distributed Processing Techniques and Applications*, Las Vegas, NV, USA, 2010, pp. 1-12.

[6] J. L. Berral, *et al.*, "Towards energy-aware scheduling in data centers using machine learning," in *Proc. of the 1st International Conference on Energy-Efficient Computing and Networking*, Passau, Germany, 2010, pp. 215-224.

[7] M. Hauck, *et al.*, "Towards Performance Prediction for Cloud Computing Environments based on Goal-oriented Measurements," in *in Proc. CLOSER*, 2011, pp. 616-622.

[8] R. Nathuji, *et al.*, "Q-clouds: managing performance interference effects for QoS-aware clouds," presented at the Proceedings of the 5th European conference on Computer systems, Paris, France, 2010.

[9] G. Casale, *et al.*, "A Model of Storage I/O Performance Interference in Virtualized Systems," presented at the Proceedings of the 2011 31st International Conference on Distributed Computing Systems Workshops, 2011.

[10] S. Govindan, *et al.*, "Cuanta: quantifying effects of shared on-chip resource interference for consolidated virtual machines," presented at the Proceedings of the 2nd ACM Symposium on Cloud Computing, Cascais, Portugal, 2011.

[11] X. Pu, *et al.*, "Who is Your Neighbor: Net I/O Performance Interference in Virtualized Clouds," *Services Computing, IEEE Transactions on,* vol. PP, pp. 1-1, 2012.

[12] K. Younggyun, *et al.*, "An Analysis of Performance Interference Effects in Virtual Environments," in *Performance Analysis of Systems & Software, 2007. ISPASS 2007. IEEE International Symposium on*, 2007, pp. 200-209.

[13] D. Gupta, *et al.*, "Enforcing performance isolation across virtual machines in Xen," presented at the Proceedings of the ACM/IFIP/USENIX 2006 International Conference on Middleware, Melbourne, Australia, 2006.

[14] Google. *Google Cluster Data V2*. Available: http://code.google.com/p/googleclusterdata/wiki/ClusterData2011_1

[15] J. Huai, *et al.*, "CIVIC: a Hypervisor based Virtual Computing Environment," presented at the International Conference on Parallel Processing Workshops, Xi'an, China, 2007.

[16] M. A. El-Refaey and M. A. Rizkaa, "Virtual Systems Workload Characterization: An Overview," in *Proc of the18th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises*, 2009, pp. 72-77.

[17] B. Sharma, *et al.*, "Modeling and synthesizing task placement constraints in Google compute clusters," presented at the Proceedings of the 2nd ACM Symposium on Cloud Computing, Cascais, Portugal, 2011.

[18] J. Zhan, *et al.*, "PhoenixCloud: Provisioning Resources for Heterogeneous Workloads in Cloud Computing," *arXiv preprint arXiv:1006.1401,* 2010.

[19] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters,* vol. 31, pp. 651-666, 2010.

[20] D. T. Pham, *et al.*, "Selection of K in K-means clustering," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science,* vol. 219, pp. 103-119, January 1, 2005.

[21] A. Kopytov. (2012, July). *Sysbench Manual*. Available: http://sysbench.sourceforge.net/docs/

[22] W. Tianyu, *et al.*, "NeTrOS: A Virtual Computing Environment towards Instant Service of Network Software," in *Semantics, Knowledge and Grids (SKG), 2012 Eighth International Conference on*, 2012, pp. 24-31.