IEEE Xplore: https://ieeexplore.ieee.org/document/6821020

DOI: https://doi.org/10.1109/CLOUDCOM-ASIA.2013.52

Current Challenges and Approaches for Resource Demand Estimation in the Cloud

Markus Ullrich, Jörg Lässig
University of Applied Sciences Zittau/Görlitz
Department of Computer Science
Görlitz, Germany
Email: {mullrich,jlaessig}@hszg.de

Abstract—The increasing popularity of Cloud computing, especially for high performance computing (HPC) applications offers a huge potential to optimize the consumption of compute resources. Since hybrid Cloud platforms in particular offer the best balance between data security, performance, business agility and mobile support, they are getting used more and more frequently. In this work, we are highlighting the most important challenges that arise for resource demand estimation systems, especially in public and hybrid Cloud environments. We present existing approaches, separated in load-balancing, or single resource type systems, and Cloud or virtual machine (VM) type selection, or multiple resource type systems. The approaches are analyzed including their potential to overcome the presented challenges and their applicability in different Cloud environments. Our research reveals that not all of the issues have been solved yet, but the means to achieve that are available. We conclude our work with useful suggestions that can help to overcome the remaining challenges.

Index Terms—cloud computing, resource demand estimation, survey

1 Introduction

Cloud computing nowadays has a huge impact on many IT solutions [1] for mostly big, but also small and medium sized enterprises [2], [3]. Many popular Cloud definitions, e.g. from Vaquero et al. [4] and the NIST definition of Cloud computing [5] highlight the most important benefits of Cloud computing. Those benefits are the flexibility and scalability, the cost benefits of the pay-as-you-go principle and the possibility to access it from almost everywhere worldwide with different devices. On the other hand, there still exist many challenges in Cloud computing that seem to be not or only partially solved [6], [7], [8]. The biggest challenges found in the literature are data security and privacy but also vendor lock-in and uncertainty about the actual cost and time benefit play quite a key role. We highlight the most important expectations and concerns that enterprises raise, based on those benefits and challenges, whilst considering to move the whole or parts of their IT infrastructure into the Cloud. This work further elaborates on the question if they are justified or not. Finally, we particularly emphasize the importance of resource demand estimation in this particular area.

1.1 Enterprises and Cloud Computing

Due to the above characteristics of Cloud computing, many enterprises aim to reduce their capital expenditure (CAPEX) and the overall cost, and to increase their flexibility. A survey by Narasimhan et al. revealed that companies which already use the Cloud are more interested in gaining flexibility and improved mobile access for their services than cost reduction [9]. Gupta et al. also discovered that the number one reason for especially small and medium sized enterprises (SMEs) to use the Cloud is surprisingly not cost reduction but

business agility [3]. For the most part, cost reduction is still an important factor which leads to the conclusion, that the objective for most companies is to provide business agility and mobile access while keeping the cost as low as possible. One of the biggest concerns about Cloud computing since the beginning are data security and privacy [10], [11]. That is still the case for many Cloud adopters that are already using the Cloud but other factors are almost as important – factors like reliability, customizability, user adoption and ease of integration. In fact, 28 percent of the cloud adaptors, which already are using the cloud, state that the security concerns are one of the biggest misconceptions about Cloud computing. More than half of the respondents believe that Cloud solutions perform better in terms of security compared to on-premise applications which is also the case for many other important properties. Those are shown in Figure 1. The survey also revealed that almost 70 percent of the Cloud adopters are planning to move at least 50 percent of their IT infrastructure to the Cloud in the next three years, as shown in Figure 2.

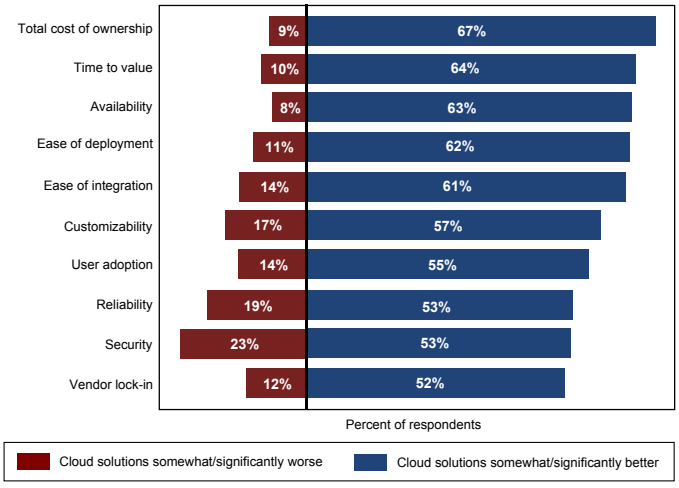
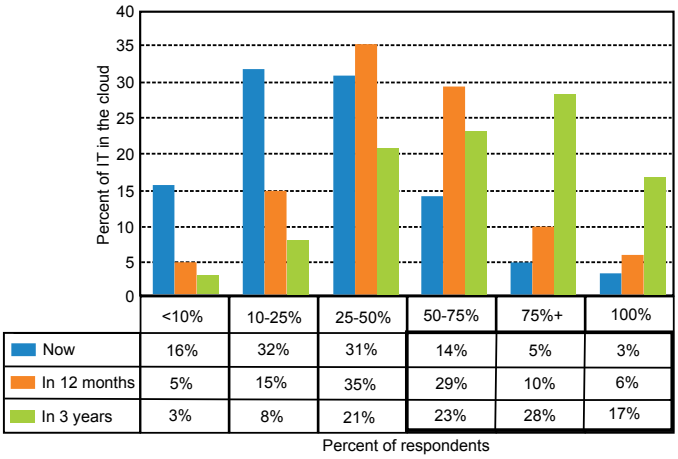


Fig. 1. Cloud adopters' perception of cloud applications versus on-premises applications. Percentages do not add up to 100 because 'about the same' responses are excluded [9].

Since business agility and cost are still important concerns, it is useful to be able to predict the amount of necessary resources in advance so that a sufficient amount of them can be set



	<10%	10-25%	25-50%	50-75%	75%+	100%
Now	16%	32%	31%	14%	5%	3%
In 12 months	5%	15%	35%	29%	10%	6%
In 3 years	3%	8%	21%	23%	28%	17%

Fig. 2. Percent of IT in the cloud over time [9].

up in advance. That is not only useful because it usually takes some time for a virtual compute instance to be up and running from the time it has been started. It also may be used to the extend where those instances can be reserved, which is usually cheaper than requesting them on the fly [12].

1.2 Computational Expensive Tasks

The execution of scientific applications or analytical processes in the Cloud is also rapidly gaining popularity. In many previous work it has been shown that outsourcing of computational expensive tasks in the Cloud can be beneficial. In [13] Deelman et al. simulated the cost performance of different execution and resource provisioning plans for a real-life astronomy application and found out that cost can be significantly reduced with no significant impact on application performance by provisioning the right amount of storage and compute resources. We discovered that for parallel applications the correct setup of instances can even reduce the runtime and cost of an application due to the fact that virtual compute instances are usually provided as packed resources, consisting of CPU, hard disk and memory [14]. The application we used in our study consumes mostly compute power. Thus, it is more efficient to select smaller compute instances with less memory but a better processing unit. On the other hand, it would be fatal to choose an instance with less memory

than the application actually requires since that would result in swapping. He et al. measured already in 2010 that virtualization technology adds only a little performance overhead whilst executing HPC applications in the Cloud [15]. However, they also pointed out that due to the slow networking performance of virtual compute nodes in public clouds, the application of a private cloud is more beneficial. Nonetheless, they promoted that public Cloud platforms can also be utilized by scientists despite those deficiencies. Although private Clouds seem to be the best solution concerning runtime, especially network performance, and security they are not applicable in a scenario where data or services have to be provided publicly. Also many public Cloud providers offer different types of virtual compute instances which offer different benefits for one application. Therefore, it is helpful to predict in advance, which platform offers the instances with the best performance for a specific task or if it is even useful to utilize multiple platforms at once.

The remainder of this paper is organized as follows. In Section 2 we highlight different challenges in resource demand estimation. After general difficulties, we focus on specific challenges for public Clouds and hybrid Clouds in particular. Next, in Section 3 we present different existing solutions for resource demand estimation which represent the current state in this area of research. We focus on their applicability in solving the challenges we mentioned earlier and especially in hybrid Cloud scenarios. Finally, we conclude this work in Section 4 and present possible solutions to the problems which are still remaining.

2 Challenges

Many researchers concerned themselves with resource demand estimation and the associated challenges in this research area. Already in 1998, Dilleya et al. worked out different factors affecting the performance of a single web server [16]. Those are for example the web server pool size, the underlying network topology and server system configuration properties like the HTTP object cache size. An obvious difficulty is to consider many different types of hardware or arbitrary applications with different properties. Especially for distributed computing, resource demand estimation is most effective, if only one type of resource package has to be considered. In the optimal case, a certain number of those resources has to be allocated to compute jobs that all have the same properties. Unfortunately, that is generally not the case. That the distribution of jobs in Cloud computing is important has already been demonstrated by researchers. Stantchev confirmed that replication configurations in Cloud computing can have a positive effect on non-functional properties using benchmarking [17]. He measured an increase of transactions per second and a decrease of response time with his experimental setup. Chieu et al. discovered that an effective use of the dynamic scaling property in Cloud computing is possible for Web applications [18].

Another important challenge besides network and Cloud related issues is the capability of an estimation system to detect necessary resource adaptions in real-time if not even in advance, since it usually takes up to one minute for a compute instance in the Cloud to be available after requesting it. To be cost efficient, it is also important to not reserve more resources than actually required. Too many unused resources can be a major problem in a data center as well, especially in private Clouds where those resources can not be utilized by other applications. Thus, a tradeoff has to be made between meeting quality of service (QoS) requirements and high resource utilization [19]. Stefano et al. discovered many factors affecting the design of load balancing algorithms in distributed systems [20]. The most important ones include the current workload of a host and information about the structure of a process. Thus, it is required to monitor the application state for applications with a variable load for efficient load-balancing. Based on the amount of information that is available from monitoring, three different categories for load-balancing algorithms exist. Black-box approaches, where no information about the actual workload is available and only the execution time can be measured, grey-box approaches, that allow at least access to operat-

ing system specific values like memory consumption or CPU utilization, and white-box approaches, which allow the complete monitoring of the whole system including application specific variables. For the most part, it is only possible to use a grey-box if not only a black-box approach. If only the latter is possible, another level of complexity is added to the resource demand estimation system. A load-balancing approach called Sandpiper that relies on a grey-box and a black-box approach to monitor and detect hotspots in virtual environments and to eliminate them [21] has been presented by Wood et al.. They confirm, that a grey-box approach can indeed improve the responsiveness of their system.

2.1 Public Clouds

To reduce the effort of setting up and configuring a data center, to increase business agility, reduce the CAPEX and allow better access for mobile devices, public Clouds offer a convenient solution. However, new challenges arise with the application of public Clouds like frequently changing APIs. This alone is nowadays not problematic, but since vendor lock-in is still a major concern about public Clouds, it almost became mandatory to provide support for multiple platforms. Unfortunately, every platform has its own characteristics and, due to the lack of standardization, its own interfaces and APIs for accessing them [22]. Managing that is not an easy task, so a good resource demand estimation application should make use of a good Cloud integration strategy or utilize existing middleware for that task as well. Another reason, why the utilization of different public Clouds is preferable are the huge performance differences between them [23]. That means, the performance of an application and thus its resource allocation strongly depends on the Cloud provider, but it may also differ in a single Cloud. For example, it can depend on points in time, the physical locations of the machines and the choice of the virtual system type [24]. Dejun et al. also concluded that different instances of the same type can have a different performance whilst the performance of a single instance is relatively stable [25].

Further problems for resource demand estimation are the throughput instability and delay variations in virtual networks discovered by Wang and Ng [26]. They observed, that certain types of virtual compute instances only receive a 40% to 50% share of the processor which can cause very unstable TCP/UDP throughput among those instances. They further observed abnormally large packet delay variations among all instance types and concluded, that the unstable network performance can indeed dramatically influence the results of network performance techniques. This problem however, seems to be only related to public Clouds. He et al. pointed out that due to the slow networking performance of virtual compute nodes in public clouds, the application of a private cloud is more beneficial [15].

2.2 Hybrid Clouds

Although a public Cloud has the benefit of reduced CAPEX and better deployment speed, private Clouds are even more popular amongst enterprises according to a survey by IDG in June 2013[1]. The survey revealed that companies tend to optimize existing infrastructure with the implementation of a private Cloud which results in a lower total cost of ownership (TCO). Despite that fact, 59 percent of the respondents have a portion of their IT environment in the public Cloud. Therefore, they are in fact working in a hybrid Cloud environment. Considering the above statement about slow virtual networks in public Clouds, that is a good decision, but it leads to another level of complexity for resource demand estimation algorithms. After determining if a certain task can be executed in the public Cloud or not, based on security policies, it must be decided if it is faster or maybe cheaper to execute this task in the public or in the private Cloud or if both should be utilized for complex applications. Factors like the resource capacity of a private Cloud, which is usually much less then of a public Cloud, play an important role and, as mentioned before, the fact that some parts of an application may need to be publicly available.

1. http://www.eweek.com/cloud/enterprises-prefer-private-clouds-survey/ - 2013-09-18

3 Existing Approaches

Numerous approaches to predict or estimate resource demand in the Cloud exist already. Here, we classify them as either single or multiple resource type systems. The former perform best if only one type of resource package or virtual machine (VM) is available and predict the amount of VMs allocated to one specific application or a number of different applications. Usually, load-balancing systems belong to this category. The latter aim to select the best VM type for one application out of different sized VMs. This selection process can also involve multiple Clouds. Conditions for the selection process are usually based on a service level agreement (SLA), QoS requirements or simply minimizing runtime or cost. An optimal solution should be able to select the most appropriate Cloud provider, including private Clouds, and the necessary number of compute instances of all available VM types for arbitrary applications.

3.1 Single Resource Type Solutions

Although most load-balancing system cannot be used to estimate the resources for an application that is not already running, they become vital for managing resources for running applications with a variable load. Some very good approaches do already exist, e.g. the Sandpiper system by Wood et al. [21] which we mentioned above. An approach for dynamically estimating CPU demands of applications using CPU measurements from previous executions has been presented by Pacifici et al. in 2008 [27]. They formulated a multivariate linear regression problem and used a linear model to solve it. They further addressed practical issues like insignificant flows, collinear flows, space and temporal variations and background noise. Problems with their approach have been the long response times of the system, capturing background noises and missing scaling and standardization techniques. Gong et al. introduced PRESS, a PRedictive Elastic ReSource Scaling scheme for Cloud systems, which makes a prediction of future demand based on patterns extracted from previous executions [28]. It uses the average value of the samples in each prediction window or alternatively, a state-based prediction approach using a discrete-time Markov chain. PRESS showed high accuracy resulting in low over provisioning and almost none under provisioning. However, it is not applicable in a hybrid Cloud environment and not able to consider different types of virtual machines at the moment. Isci et al. implemented a completely non-utility based approach for CPU demand estimation which performed well in comparison to other utility based solutions [29]. They considered an increase in the service time of a virtual machine as indicator for performance degradation. Thus, their technique is independent from the structure of an application as well as the underlying operating system which makes it more flexible. The downside is, that this black-box approach cannot really predict a resource demand in advance, but on the other hand, it is quick enough to respond to demand changes almost immediately. Kousiouris et al. presented a two-level generic black-box approach for behavioral-based management across different Cloud layers [30]. They identified patterns in high-level information and translated them to low-level resource attributes. This approach showed notably good prediction accuracy and seems to be even generic enough to work with different Cloud environments including private Clouds. It is further fast enough to predict changes in real time. Unfortunately, it may not be applicable in a hybrid Cloud environment, which would be preferred to optimize the runtime and cost across multiple Clouds.

A method to measure the elasticity in Cloud computing has been developed by Islam et al. [31]. Their approach is QoS related and also of particular interest for resource demand estimation in order to make assumptions on how quick a certain Cloud platform can scale up or down. To measure the elasticity they use a penalty model that penalizes imperfections in elasticity for a given workload in monetary units. We found, that this model can also be used for an optimization function to support resource demand estimation approaches that rely on data mining techniques. The developed algorithms for calculating the over and under provisioning penalties are given in Algorithm 1

and Algorithm 2 respectively. It is important to mention that the automation of this approach is also possible if the QoS measurements of the customer are known. The downside of this method is that it requires extensive benchmarking.

Algorithm 1 Calculation of the over provisioning penalty P_o according to Islam et al. [31] - t_s and t_e represent the start- and end time, c_i the cost for a resource i and $R_i(t)$, $M_i(t)$ and $D_i(t)$ the available, chargable and actual supply at time t respectively.

$$P_o(t_s, t_e) = \sum_i P_{o,i}(t_s, t_e)$$

$$P_{o,i}(t_s, t_e) = \int_{t_s}^{t_e} c_i \cdot d_i(t) dt$$

$$d_i(t) = \begin{cases} M_i(t) - D_i(t) & \text{if } R_i(t) > D_i(t), \\ M_i(t) - R_i(t) & \text{if } M_i(t) > R_i(t) \\ & \text{and } D_i(t) >= R_i(t), \\ 0 & \text{otherwise} \end{cases}$$

Algorithm 2 Calculation of the under provisioning penalty P_u according to Islam et al. [31] - Q is a set of QoS properties and $p_q(t)$ the amount of unsatisfactory behaviour at time t. This amount is mapped to the financial impact in f_q. The limit of accaptable unsatisfactory behaviour at time t is donated by $p_q^{\text{opt}}(t)$.

$$P_u(t_s, t_e) = \sum_{q \in Q} P_{u,q}(t_s, t_e)$$

$$P_{u,q}(t_s, t_e) = \int_{t_s}^{t_e} (f_q(p_q(t)) - f_q(p_q^{\text{opt}}(t))) dt$$

In [32] Islam et al. developed a method to predict resource provisioning using machine learning algorithms. They verified their approach with experimental results, using neural networks and linear regression as learning algorithms. The prediction accuracy is quite notable which allows to achieve on-demand resource allocation in the Cloud even with several minutes delay in the hardware resource allocation. However, the experiments were only performed using the TPC-W benchmark [33] yet and for just a single VM type. Accordingly, it is not certain that this method works for arbitrary applications or in hybrid Cloud environments.

Many of the previously examined approaches have in common, that they were only designed and tested for scalable web applications, which are a suitable target to apply dynamic scaling or load-balancing algorithms to. But for HPC or even arbitrary applications, those approaches are either not applicable or have not been thoroughly tested yet.

3.2 Multiple Resource Type Solutions

Most resource demand estimation solutions that allow the selection of a specific resource type focus on selecting one out of multiple Cloud providers for a certain application. This approach is popular because of the huge performance differences between Cloud providers which we explained already. However, it is almost as important to select the correct types and number of VMs that should be used. Amazon for example offers 17 different compute instance types with different specifications each[2].

Li et al. performed several benchmarks to compare public Cloud providers [23] and derived a performance estimation from them to select the best provider for a certain job. Although their approach, which is called CloudCmp, may not include the accurate prediction of runtime or cost for the selected Cloud provider, it delivers a pretty good estimation, which provider suits best for a specific application. An alternate approach has been developed by Kaisler et al. [12]. Their decision framework is designed to assist managers to decide which Cloud alternative is the right for their application case based on specific requirements like business objectives, QoS attributes and architectural decisions. They also pointed out, that different pricing strategies offered to the user by a single Cloud provider should be taken into account, i.e. many providers offer pre-paid instances which are cheaper than usual. They state in their conclusion, that especially SMEs or Cloud beginners should start off with a small private Cloud and then eventually

2. http://aws.amazon.com/ec2/instance-types/ - 2013-09-24

move critical parts of their IT to a public Cloud. However, although this approach may help selecting a Cloud provider, it does not offer any real resource demand estimation.

The most promising approach, to our knowledge, has been presented by Li et al. With their approach called CloudProphet they are able to predict application performance in the Cloud for arbitrary applications [34]. In their work they use a trace and replay method which means the application is executed locally and the same workload is emulated in the Cloud. The performance of the agent that emulated the workload is used to predict the performance after migration. The accuracy of this method is notably high and in combination with CloudCmp, this approach offers very good support for selecting the right Cloud provider for an application without executing it in the Cloud. However, their approach has been tested on a single compute instance type only so it does not consider different instance types yet. An obvious difficulty for this goal is the high effort and expenses for executing an agent on all different instance types. This effort needs to be reduced. Although that goal could be achieved with little effort using regression methods, it would reduce the prediction accuracy by a notable amount which is not desirable. Hybrid Clouds were also not considered in the literature yet, but as this approach works with multiple public Cloud providers already, this seems to be only a minor problem. Another problem is that it is not certain if this approach works for scientific or high performance applications as well, since it has not been tested with those kind of applications either. Supposedly, this is due to the usually huge workload of those applications so the trace and replay method would be very slow and expensive in such a case.

4 Conclusion

In this work, we presented current challenges for resource demand estimation in Cloud computing and analyzed existing approaches for solving those challenges. We discovered, that many of these problems can already be solved by most of the solutions but some important research challenges remain. For most of the approaches it is still unclear whether the presented approaches are applicable in a hybrid Cloud environment. This problem could be solved by utilizing Cloud integration techniques or middleware to support the estimation algorithm. Another benefit of this approach is the resulting simplification of the development process. Further, the increasing number of HPC applications in the Cloud contain a huge potential for saving resources and thus energy and cost if resource demand estimation algorithms can be properly applied. Current solutions can not yet meet that demand since usually extensive benchmarking is required to be able to accurately predict the resource demand for future jobs. For HPC applications, this effort is not feasible. To overcome this issue, a method to predict the demand of a large scale experiment based on previously executed benchmarks of small scale experiments needs to be developed, possibly by utilizing state of the art regression techniques. However, this method may only be applicable to grey-box, if not only white-box approaches, and the prediction accuracy will possibly still be reduced by a notable amount. More research in that direction should help to discover exactly how the prediction accuracy is influenced and to derive possible solutions for that. Finally, taking different VM types of each Cloud provider into account is still not supported by most of the approaches which is due to frequent changes and lack of standardization in this area. Data mining techniques like regression could be applied as well to partially solve this problem but, as discussed already, the prediction accuracy may suffer from this additional complexity as well. To fully overcome this last challenge, more standardized resource descriptions and reliable provisioning of resources in Cloud computing are necessary.

Acknowledgments

This research project was promoted and funded by the European Union and the Free State of Saxony. The authors take the responsibility for the content of this publication.

REFERENCES

[1] S. Fremdt, R. Beck, and S. Weber, "Does cloud computing matter? an analysis of the cloud model software-as-a-service and its impact on operational agility," in *46th Hawaii International Conference on System Sciences (HICSS), 2013*, Jan 2013, pp. 1025–1034.

[2] Y. Alshamaila, S. Papagiannidis, and F. Li, "Cloud computing adoption by smes in the north east of england: A multi-perspective framework," *Journal of Enterprise Information Management*, vol. 26, no. 3, pp. 250 – 275, 2013.

[3] P. Gupta, A. Seetharaman, and J. R. Raj, "The usage and adoption of cloud computing by small and medium businesses," *International Journal of Information Management*, vol. 33, no. 5, pp. 861–874, Oct 2013.

[4] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A break in the clouds: towards a cloud definition," *ACM SIGCOMM Computer Communication Review*, vol. 39, pp. 50–55, Jan 2008.

[5] P. Mell and T. Grance, *The NIST Definition of Cloud Computing*, NIST Std., Jan 2011.

[6] T. Dillon, C. Wu, and C. E., "Cloud computing: Issues and challenges," in *IEEE AINA*, vol. 24, 2010, pp. 27–33.

[7] R. Moreno-Vozmediano, R. Montero, and I. Llorente, "Key challenges in cloud computing to enable the future internet of services," *IEEE Internet Computing*, vol. PP, p. 1, 2012.

[8] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, Apr 2010.

[9] B. Narasimhan and R. Nichols, "State of cloud applications and platforms: The cloud adoptors' view," *Computer*, vol. 44, no. 3, pp. 24–28, Mar 2011.

[10] D. Chen and H. Zhao, "Data security and privacy protection issues in cloud computing," in *ICCSEE*, vol. 1. Shenyang, China: Northeastern University, Mar 2012, pp. 647–651.

[11] K. Ren, C. Wang, and Q. Wang, "Security challenges for the public cloud," *IEEE Internet Computing*, vol. 16, no. 1, pp. 69–73, Jan 2012.

[12] S. Kaisler, W. Money, and S. Cohen, "A decision framework for cloud computing," in *System Science (HICSS), 2012 45th Hawaii International Conference on*, Jan 2012, pp. 1553–1562.

[13] E. Deelman, G. Singh, M. Livny, B. Berriman, and J. Good, "The cost of doing science on the cloud: the montage example," in *ACM/IEEE conference on Supercomputing*, 2008.

[14] M. Ullrich, K. ten Hagen, and J. Lässig, "Public cloud extension for desktop applications–case study of a data mining solution," in *Network Cloud Computing and Applications (NCCA), 2012 Second Symposium on*. IEEE, 2012, pp. 53–64.

[15] Q. He, S. Zhou, B. Kobler, D. Duffy, and T. McGlynn, "Case study for running hpc applications in public clouds," in *ACM International Symposium on High Performance Distributed Computing*, vol. 19, 2010, pp. 395–401.

[16] J. Dilleya, R. Friedrich, T. Jina, and J. Roliab, "Web server performance measurement and modeling techniques," *Performance Evaluation*, vol. 33, no. 1, pp. 5–26, Jun 1998.

[17] V. Stantchev, "Performance evaluation of cloud computing offerings," in *IEEE ADVCOMP*, vol. 3, Dec 2009, pp. 187–192.

[18] T. Chieu, A. Mohindra, A. Karve, and A. Segal, "Dynamic scaling of web applications in a virtualized cloud computing environment," in *IEEE International Conference on e-Business Engineering, 2009. ICEBE '09.*, Oct 2009, pp. 281–286.

[19] P. Padala, K. G. Shin, X. Zhu, M. Uysal, Z. Wang, S. Singhal, and A. M. K. Salem, "Adaptive control of virtualized resources in utility computing environments," in *Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems*, 2007, pp. 289–302.

[20] A. D. Stefano, L. L. Bello, and E. Tramontana, "Factors affecting the design of load balancing algorithms in distributed systems," *Journal of System and Software*, vol. 48, no. 2, pp. 105–117, Oct 1999.

[21] T. Wood, P. Shenoy, A. Venkataramani, and M. Yousif, "Sandpiper: Black-box and gray-box resource management for virtual machines," *Computer Networks*, vol. 53, no. 17, pp. 2923 – 2938, 2009, ¡ce:title¿Virtualized Data Centers¡/ce:title¿.

[22] J. Peng, X. Zhang, Z. Lei, B. Zhang, W. Zhang, and Q. Li, "Comparison of several cloud computing platforms," in *IEEE ISISE*, vol. 2, Dec 2009, pp. 23–27.

[23] A. Li, X. Yang, S. Kandula, and M. Zhang, "Cloudcmp: comparing public cloud providers," in *ACM SIGCOMM*, vol. 10, 2010, pp. 1–14.

[24] J. Schad, J. Dittrich, and J.-A. Quian-Ruiz, "Runtime measurements in the cloud: observing, analyzing, and reducing variance," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 460–471, Sep 2010.

[25] J. Dejun, G. Pierre, and C.-H. Chi, "Ec2 performance analysis for resource provisioning of service-oriented applications," in *Service-Oriented Computing. ICSOC/ServiceWave 2009 Workshops*, vol. 6275, 2010, pp. 197–207.

[26] G. Wang and T. Ng, "The impact of virtualization on network performance of amazon ec2 data center," in *IEEE INFOCOM*, Mar 2012, pp. 1–9.

[27] G. Pacifici, W. Segmuller, M. Spreitzer, and A. Tantawi, "{CPU} demand for web serving: Measurement analysis and dynamic estimation," *Performance Evaluation*, vol. 65, no. 67, pp. 531 – 553, 2008, ¡ce:title¿Innovative Performance Evaluation Methodologies and Tools: Selected Papers from ValueTools 2006¡/ce:title¿.

[28] Z. Gong, X. Gu, and J. Wilkes, "Press: Predictive elastic resource scaling for cloud systems," in *Network and Service Management (CNSM), 2010 International Conference on*, 2010, pp. 9–16.

[29] J. W. I. S. M. K. J. Isci, C.; Hanson, "Runtime demand estimation for effective dynamic resource management," in *Network Operations and Management Symposium (NOMS), 2010 IEEE*, Apr 2010, pp. 381,388.

[30] G. Kousiouris, A. Menychtas, D. Kyriazis, S. Gogouvitis, and T. Varvarigou, "Dynamic, behavioral-based estimation of resource provisioning based on high-level application terms in cloud platforms," *Future Generation Computer Systems*, no. 0, pp. –, 2012.

[31] S. Islam, K. Lee, A. Fekete, and A. Liu, "How a consumer can measure elasticity for cloud platforms," in *ICPE*, vol. 3, 2012, pp. 85–96.

[32] S. Islam, J. Keung, K. Lee, and A. Liu, "Empirical prediction models for adaptive resource provisioning in the cloud," *Future Generation Computer Systems*, vol. 28, no. 1, pp. 155–162, Jan 2012.

[33] D. Garcia and J. Garcia, "Tpc-w e-commerce benchmark evaluation," *Computer*, vol. 36, no. 2, pp. 42–48, Feb 2003.

[34] A. Li, X. Zong, S. Kandula, X. Yang, and M. Zhang, "Cloudprophet: towards application performance prediction in cloud," in *ACM SIGCOMM*, vol. 41, Aug 2011, pp. 426–427.