Exploring Inter-Cloud Load Balancing by Utilizing Historical Service Submission Records

Stelios Sotiriadis, University of Derby, UK

Nik Bessis, University of Derby, UK

Nick Antonopoulos, University of Derby, UK

ABSTRACT

Cloud computing offers significant advantages to Internet users by deploying hosted services via bespoke service-provisioning environments. In advance, the emergence of Inter-Cloud increases the competences and opportunities of clients for a wider resource provision selection. This extends current capabilities by decoupling users from cloud providers while at the same time cloud providers offer an augmented service delivery mean. In practice, cloud users make use of their brokering component for selecting the best available resource, in terms of computational power and software licensing of a datacenter based on service level agreements for service execution. However, from the cloud perspective, the overall choice for balancing the different workloads within the Inter-Cloud is a complex decision. This article explores the performance of an Inter-Cloud to measure the utilization levels among their sub-clouds for various job submissions. With this in mind, the solution is modeled for achieving load balancing based on historical records from past service execution experiences. The record files are composed in the form of log files that keep related information about the size of the Inter-Cloud, basic specifications, and job submission parameters. Finally, the solution is integrated in a simulated setting for exploring the performance of the approach for various heavy workload submissions.

Keywords: Cloud Computing, Cloud Load Balancing, Cloud Scheduling, Historical Records on Job Delegations, Inter-Cloud

INTRODUCTION

A cloud-computing environment includes the delivery of hosted services that are located in a remote location via the public Internet to everyday users. Although various opinions include a narrow view of clouds – mainly as an enterprise server-based datacenter – in this article we have taken an inclusive perspective of clouds as to encompass various kinds of resources. At a first glance, cloud computing share similar fundamental elements with other large scale and/or distributed computing paradigms e.g., clusters and grids. In addition, a broad definition of cloud computing includes a service-oriented virtualization environment. This generic vision – from the perspective of the service – change the focus on how to or-

DOI: 10.4018/jdst.2012070106

chestrate the cloud service distribution, rather than aim to the management and deployment of the underlying infrastructure.

In such environments, massive computing capacity resides at a remote space and could be delivered in the form of software and/or hardware (Carolan & Gaede, 2011). These offered services are identical to job submissions that have been encapsulated in application execution requests that have been posed by the end-users. Although cloud computing is still in its infancy due to the facility-orientation (physical space), it needs to be evolved to a more distributed infrastructure with a broader propagation of services. This could be achieved by utilizing available resources (clusters, high performance computing and grids) relying at the lower level (infrastructure). By transforming the cloud infrastructure to go beyond its premises we could facilitate a wider set of deployed services and applications.

The study aims to the InterCloud load balancing which represents an interconnected global cloud of clouds (Sotiriadis et al., 2012a). The generic idea as presented by Bessis et al. (in press-b) and Buyya et al. (2010) is to decouple resource consumers from providers and allowing providers to offer resources on demand and on an ad hoc basis. For achieving this model, a new structure should be established to contain the required conceptions of resource decoupling. These are protocols to control trust standards, discovering, systems for naming, scheduling of services, portability and workload exchange. We focus majorly on the service management concept, and eventually the load balancing mechanism during the service submissions. The target is to effectively achieve an enhanced quality of service by methodologically assign services in the form of job tasks (processes) to resources. These services (jobs) encapsulate various capabilities of the cloud environment (e.g., provisioning of software and/or hardware). Thus, the challenge is to identify the rationality behind the decisions of the cloud provider to manage the Inter-Cloud service execution for efficient load balancing.

Specifically, the users submit their requests in a broker with the latter communicating and monitoring the whole service exchanging procedure. This component is responsible for autonomous decisions by selecting a datacenter for forwarding the request. Then, each request is sandboxed in a virtual machine (VM) that satisfies these requirements. Various criteria are implemented in this level that includes the user-defined quality of service levels e.g., pricing, homogeneity in terms of hardware and software, and generic specification of the cloud VM. These are enclosed in service level agreements (SLAs) that formally define the level of agreed terms between provider and client. Usually, this is related with the required computational power (performance) and time constraints.

To this extend, we have presented a meta-brokering solution in Sotiriadis et al. (2012b) as a novel component that is placed on the top of each broker. The aim was to achieve the decentralization of the setting in which meta-brokers collaborate with each other for SLAs trading. By using cloud meta-brokers an InterCloud is formed into an autonomously management setting of interconnected sub-clouds. Current efforts in this direction organize (meta-) centralized topologies of brokers, so various drawbacks derived from this narrow view. Herein, the work is inspired from the meta-computing concept and based upon the model of a decentralized meta-broker. Specifically, we measure the utilization level of the Inter-Cloud and we save results in log files named as historical records. Then, we explore the performance of the Inter-Cloud for various sub-cloud numbers and job variations.

With this in mind, the following section presents the motivation of our work, which is related, with the area of large-scale systems. The rest of the paper is organized as follows; we present the InterCloud load balancing solution that is followed by the presentation of the experimental analysis and results of the Inter-Cloud utilization levels and the load-balancing mechanism. At last we conclude our work by illustrating the concluding remarks and the future work section.

MOTIVATION

The Inter-Cloud as a term has been emphasised by the leading vendors in cloud services area such as HP, Intel, Yahoo, etc. (Buyya et al., 2010). It is noticeable that their state-of-the-art efforts have led to the establishment of a federation of collaborated clouds with joint initiatives. However, this vendor-oriented endeavour of Inter-Clouds has a specific control plane rather than a setting that it is based on future standards and open interfaces which are available to be shared in the academic community. In addition, knowledge sharing, experimentation and testing within their systems have been limited to the wide range of researchers.

In a different direction, the work of GICTF (2010) suggests a blueprint of Inter-Clouds, including network protocol and format. The authors propose the concept of a cloud operated by one service provider to inter-operate with clouds operated by another, thus forming a federation of clouds. They suggest that interoperable network protocols at the lower level must be presented inside and in-between of clouds in order to achieve dynamic workload migration. However, despite the fact that there aren't any experimental evidences, the lower level communication of Inter-Clouds is out of the focus of our present study.

In contrast to aforementioned works, the vision of Inter-Clouds as an inter-cooperative infrastructure including inter-enterprises has been introduced by Bessis et al. (in press-a, in press-b) yet from a federated perspective. They suggest a utility-oriented federation of various cloud computing environments and conclude to a business model of system architecture including the most important elements (requirements) of Inter-Clouds in terms of services. These functional specific requirements are presented and target to highlight the most important features for the strategic architectural plan. However, it is essential to introduce the most important characteristics of clouds first and map characteristics to inter-enterprises, by focusing in Inter-Clouds generic requirements extracted from Sotiriadis et al. (2011).

Having said that, our primary interest here is to design a load balancing mechanism for Inter-Cloud environments. On this basis, we have presented a work that aims to model the meta-brokering solution for Inter-Cloud. Existing efforts organize brokers in a (meta-) centralized topology, therefore various drawbacks derived from this perspective, like central point of failure and bottleneck in concurrent requests. For addressing these, we model a total decentralized component that is positioned on top of each traditional broker for achieving interoperation among clouds namely as Inter-Cloud meta-broker. The purpose is to distribute the whole setting and allow meta-brokers to collaborate with its other for trading SLAs and resources. By using this model the Inter-Cloud is transformed into an autonomously manage setting of interconnected sub-clouds. This moves the complexity of inter-coordination from data centres to meta-brokers; thus achieving a decoupling of users from data centres. We further design the Inter-Cloud meta-broker to be total decentralized, decoupled and dynamic by enhancing the decision making process for resource allocation and execution during cloudlet. Thus, in the next section we model the Inter-Cloud load balancing solution that encompasses its most important characteristics regarding the effectiveness of distributing workloads among clouds datacenters.

Modeling the Inter-Cloud Load Balancing Solution

The proposed cloud load balancer allows services submitted by the users to be redirected among each other for offering an improved service provision. This is to achieve a better distribution of submissions with the aim of keeping resource availability in high levels. During the cloud life cycle the cloudlets submitted by the users through a broker to the virtual machines (VMs) for execution. It should be mentioned that these jobs are identical to a real cloud application abstraction, thus we assume that the cloudlets could represent either a single job or the smallest manageable chunk of a large parallel job submission.

Specifically, a user interacts with the broker for demanding service executions. The last one acts on behalf of the user and requests from the environment specific resources in the form of resource capacity (Xhafa & Abraham, 2010). In the cloud setting, the general management platform offers the operational and business functionalities for responding to the user request. Specifically, various processes take place within both components e.g., operational management involves security control, fault tolerance management, and eventually scheduling coordination. The operational management is also responsible for the core middleware functionalities including VM orchestration via the hypervisor. The last one is software for controlling the VM deployment. The business functionalities, on the other hand, include the SLA communication process involving payments and debts, which are decided prior to the service submission and scheduling.

During the exchange phase, cloudlets are submitted to the VMs through the broker for job executions. The users do not have any knowledge on that, as the process is organized by the broker who is responsible for binding cloudlets to VMs. Together various other components initialize communication within the datacenter. It should be mentioned that each a cloud registry monitors the whole procedure and the communication among users, broker and datacenters.

For performing load balancing we explore the performance of an Inter-Cloud system for various combinations of service submissions and number of datacenters. Then, we keep a log file of these submissions and we utilize records for enhancing further scheduling decisions. Then, during a service submission a component decides the job distribution based on the historical records.

EXPERIMENTAL ANALYSIS AND RESULTS

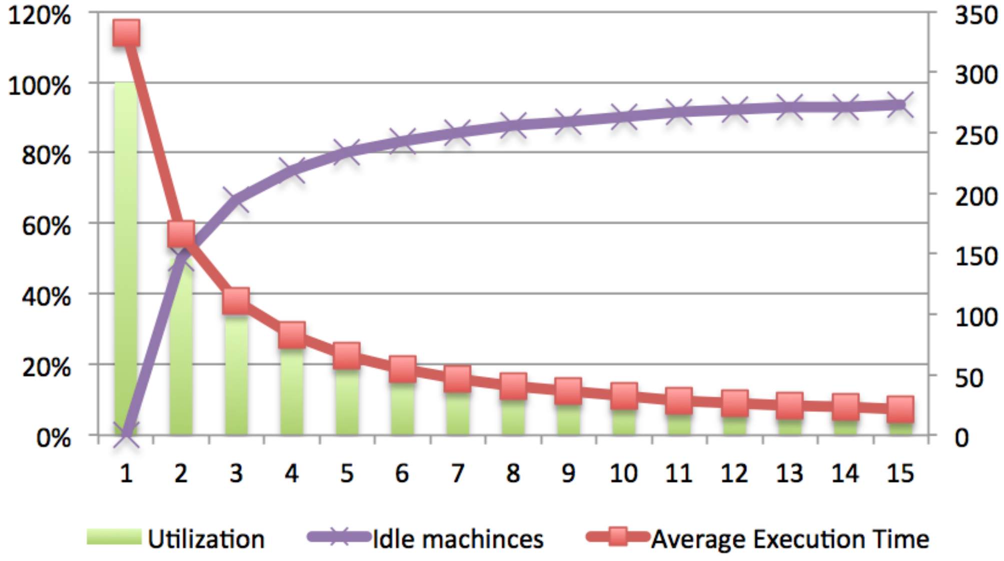
The simulation environment is developed using the Cloudsim version 3.0 (Calheiros et al., 2011; Buyya et al., 2010), a framework for modeling and simulating clouds and their services. Specifically, Cloudsim allows control of a) large scale clouds, b) datacenters, brokers and scheduling policies in a self-contained fashion, c) adaptability of the virtualization technology for creating multiple virtualization services, and d) flexibility of the processing cores to switch between time and space shared allocation policies (Melbourne Clouds Lab, n.d.). Based on that we configure our experiment to contain the characteristics of Table 1.

Our first experiment includes the training of the testbed to extract values for the average execution time and the utilization percentage when running 1000 services in a setting of 100 VMs. The system encompasses an Inter-Cloud

Table 1. The experiment configuration

Host		**VM**		**Cloudlet**	
PEs:	1	PEs:	1	PEs	1
RAM:	2048	RAM:	1024	Length:	1000
Storage:	1000000	Size:	1000	File size:	100
Mips:	1000	Mips:	150	Output size:	100
Bandwidth:	10000	VM name:	Xen	Utilization model:	Full Utilization
Allocation Policy:	Simple	Allocation Policy:	Space shared		

Figure 1. The utilization percentage, the idle machines number and the average execution times



of 1 to 15 sub-cloud datacenters. It is apparent from the figure that the higher the number of datacenters the lower the utilization level and the higher the number of idle machines. Figure 1 demonstrates the experimental results.

Figure 2 illustrates the remaining utilization level of a typical Inter-Cloud system that includes the same configuration with the previous experiment. Specifically, when the number of datacenters is seven (7) or higher, the idle utilization level is high, thus more service requests could be forwarded or new recourse availability could be offered to users.

Figure 3 demonstrates a service submission for 1000 to 10000 jobs in an Inter-Cloud of 1 to 15 datacenters. The diagram shows the behavior of the testbed for the selected job variation

Based on that, Figure 4 illustrates the utilization levels of 1 to 15 datacenters along with the linear trend line. It is apparent that when the number of datacenters is nine (9) or higher the utilization level remains the same, thus, the system keeps almost identical number of idle cloud datacenters.

To test the Inter-Cloud load balancer we concurrently add 1000 to 10000 jobs during submission to the same environment as discussed. This denotes that the system make use of the previous job delegations for deciding the number of the datacenters to be utilized by achieving an average execution time in acceptable levels while keeping most datacenters in idle status.

Figure 5 demonstrates the system decision to utilize three to five (3-5) datacenters bases on the previous submissions.

Specifically, these datacenters achieve an average execution time that keeps the value almost identical with higher datacenters collaboration. Figure 6 illustrates the utilization levels that varying from under 35% to 20%. In addition, the worst-case scenario implies that ten (10) datacenters remain in idle situation. This denotes that two sets of five (5) datacenters can concurrently run two user submissions without affecting the whole performance by keeping identical execution times and low utilization levels.

Finally, in Figure 7 we demonstrate an experiment that encompasses three concurrent

Figure 2. The cloud utilization and the remaining utilization levels

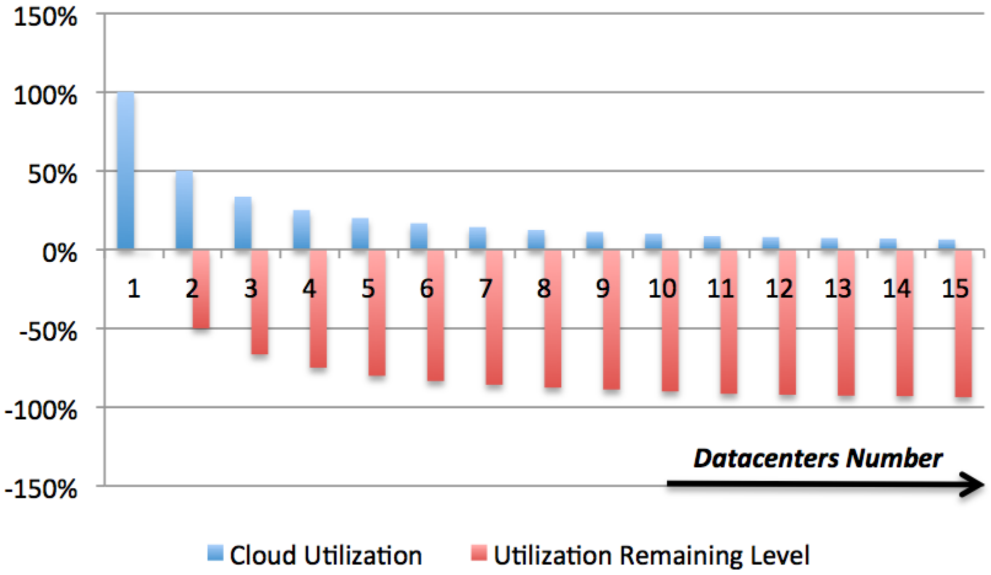


Figure 3. Service submission for 1000 to 10000 jobs in an Inter-Cloud of 1 to 15 datacenters

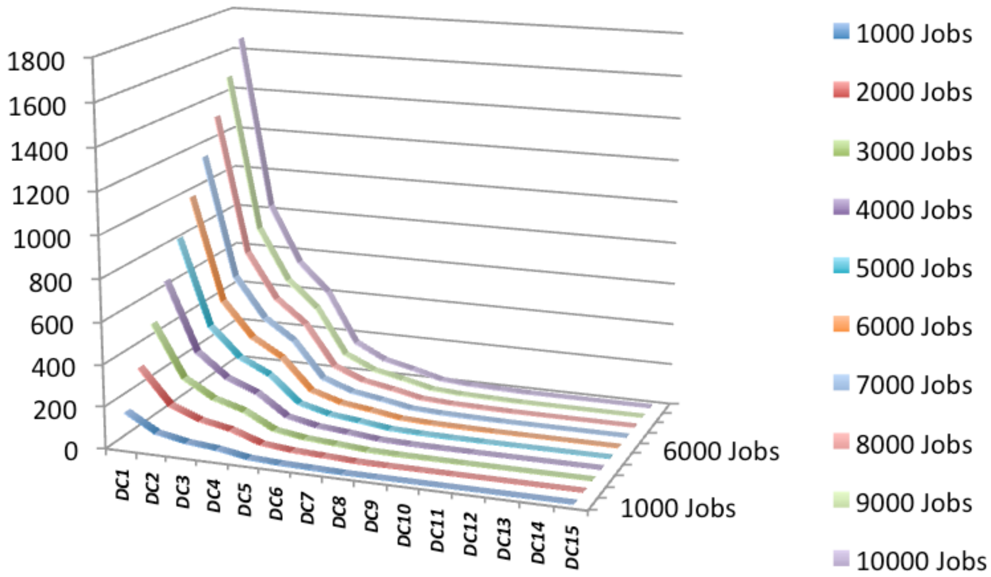


Figure 4. Utilization levels (percentages) for 1 to 15 datacenters along with the linear trend line

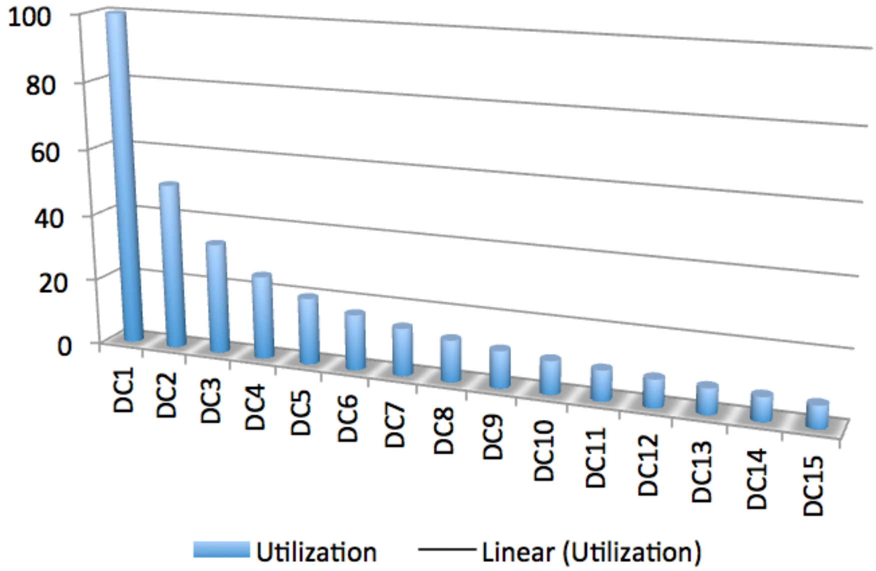


Figure 5. Average job execution times for three to five datacenters

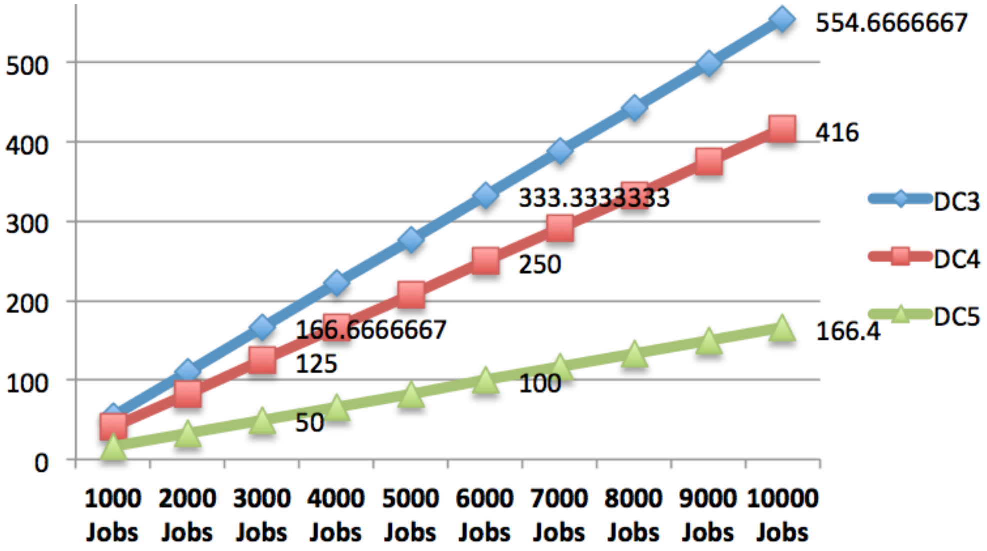


Figure 6. Utilization levels for job submissions within a three to five datacenters setting

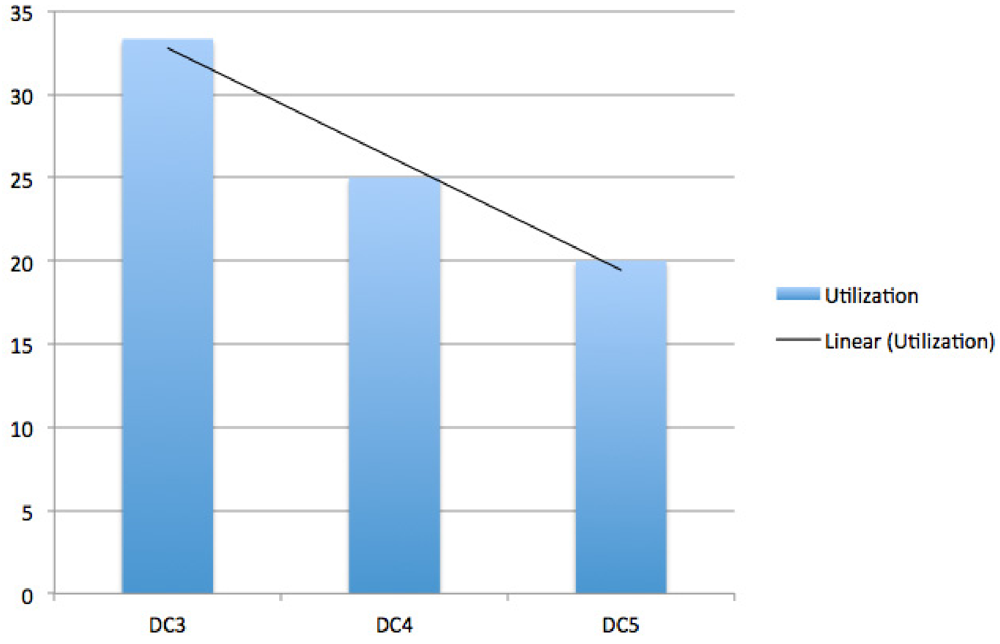
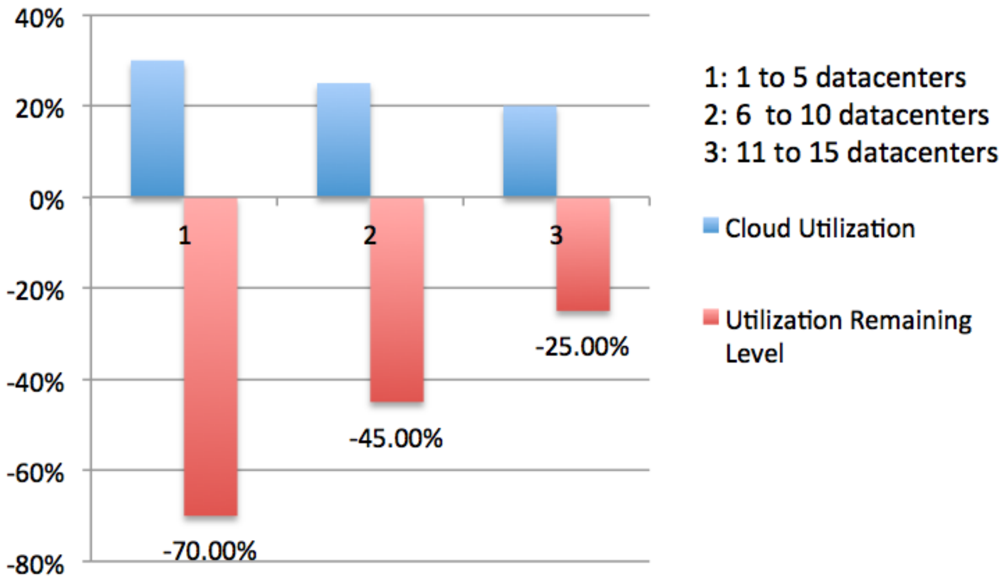


Figure 7. Three concurrent user service submissions of 1000 to 10000 jobs along with the consumed and available utilization levels



user service submissions of 1000 to 10000 jobs. In this simulation, the second user submits the jobs after few seconds from the first user submission. The same interval is utilized for the third user respectively. The load balancer reorganizes the job distribution in a maximum five (5) datacenters. This keeps the utilization levels of datacenters under 25%.

CONCLUSION AND FUTURE WORK

In this study we have explored the performance of an Inter-Cloud to measure the utilization levels among their sub-clouds for various job submissions. By modeling our solution we aimed to achieve load balancing based on historical records from past service execution experiences. The experimental analysis shows that based on basic historical records, an Inter-Cloud can decide the number of datacenters to be utilized based on the number of jobs submitted to the system. This experiment is mostly designed to explore the Inter-Cloud capabilities in this direction, thus a finer modeling is required.

However, further research challenges should aim in extending the functionality of the allocation policies and the utilization model in order to achieve an additional improvement of the simulation time. A future direction is to incorporate cloud datacenters and allow tasks to be migrated between different hosts belonging to various datacenters.

REFERENCES

Bessis, N., & Sotiriadis, S. Cristea, V., & Pop, F. (in press-a, March 26-29). An architectural strategy for meta-scheduling in InterCloud. In *Proceedings of 1st International Workshop on InterCloud and Collective Intelligence in conjunction with the 26th IEEE International Conference on Advanced Information Networking and Applications*, Fukuoka, Japan.

Bessis, N., Sotiriadis, S., Cristea V., & Pop, F. (in press-b). Meta-scheduling issues in interoperable HPCs, grids and clouds. *International Journal of Web and Grid Services, 8*(2).

Buyya, R., Ranjan, R., & Calheiros, R. N. (2010). InterCloud: Utility-oriented federation of cloud computing environments for scaling of application services. In C.-H. Hsu, L. T. Yang, J. H. Park, & S.-S. Yeo (Eds.), *Proceedings of the 10th International Conference on Algorithms and Architectures for Parallel Processing* (LNCS 6081, pp. 13-31).

Calheiros, R. N., Ranjan, R., Beloglazov, A., De Rose, C. A. F., & Buyya, R. (2011). CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software, Practice & Experience*, *41*(1), 23–50. doi:10.1002/spe.995

Carolan, J., & Gaede, S. (2011). *Introduction to cloud computing architecture* (1st ed.). Santa Clara, CA: Sun Microsystems.

Global Inter-Cloud Technology Forum (GICTF). (2010). *Use cases and functional requirements for inter-cloud computing*. Retrieved February 4, 2012, from http://www.gictf.jp/doc/GICTF Whitepaper20100809.pdf

Melbourne Clouds Lab. (n.d.). *Cloudsim: A framework for modeling and simulation of cloud computing infrastructures and services*. Retrieved February 4, 2012, from http://www.cloudbus.org/cloudsim/

Sotiriadis, S., Bessis, N., & Antonopoulos, N. (2011). Towards InterCloud schedulers: A survey of meta-scheduling approaches. In *Proceedings of the 6th International Conference on P2P, Parallel, Grid, cloud and Internet Computing* (pp. 59-66).

Sotiriadis, S., Bessis, N., & Antonopoulos, N. (2012). *Decentralized meta-brokers for inter-cloud: Modeling brokering coordinators for interoperable resource management.* Manuscript under review, Proceedings of the 8th International Conference on Natural Computation.

Sotiriadis, S., Bessis, N., & Antonopoulos, N. (in press). From meta-computing to interoperable infrastructures: A literature review of schedulers and simulators. In *Proceedings of the Advanced Information Networking and Applications*, Fukuoka, Japan.

Xhafa, F., & Abraham, A. (2010). Computational models and heuristic methods for grid scheduling problems. *Future Generation Computer Systems*, *26*(4), 608–621. doi:10.1016/j.future.2009.11.005

Stelios Sotiriadis is currently a Postgraduate Teaching Assistant at the School of Computer and Mathematics at University of Derby (UK) and a member of the Distributed and Intelligent Systems (DISYS) research group. At the same time he is doing his PhD in the area of scheduling in inter-Clouds at the same university. He obtained a BSc in Computer Science, and an MSc in Computer and Internet Application from the University of Bedfordshire. His research interests are in the area of algorithmic design for resource discovery and scheduling in distributed computing including Grids, Clouds and Inter-Clouds. He is a regular reviewer in and has published over 20 publications in refereed international journals and conferences.

Nik Bessis is currently a Head of Distributed and Intelligent Systems (DISYS) research group, a Professor and a Chair of Computer Science in the School of Computing and Mathematics at University of Derby, UK. He is also an academic member in the Department of Computer Science and Technology at University of Bedfordshire (UK). He obtained a BA (1991) from the TEI of Athens, Greece and completed his MA (1995) and PhD (2002) at De Montfort University (Leicester, UK). His research interest is the analysis, research, and delivery of user-led developments with regard to trust, data integration, annotation, and data push methods and services in distributed environments. These have a particular focus on the study and use of next generation and grid technologies methods for the benefit of various virtual organizational settings. He is involved in and leading a number of funded research and commercial projects in these areas. Prof. Bessis has published over 100 papers, won 2 best paper awards and is the editor of several books and the Editor-in-Chief of the International Journal of Distributed Systems and Technologies *(IJDST). In addition, Prof. Bessis is a regular reviewer and has served several times as a keynote speaker, conferences/workshops/track chair, associate editor, session chair and scientific program committee member.*

Nick Antonopoulos is currently a Head of School of Computing and Mathematics at the University of Derby, UK. His main research interests are Peer-to-Peer and Cloud Computing and he has published over 100 publications in these areas. Nich is also an editor of books and an associate editor in several international journals. In addition, Prof. Antonopoulos is a regular reviewer and has served several times as a keynote speaker and as a chair in conferences/workshops/tracks.