INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

# A SURVEY ON RESOURCE PROVISIONING IN CLOUD COMPUTING

**Hamid Reza Qavami[1], Shahram Jamali[2]**
*[1]University of Mohaghegh Ardabili, qavami@gmail.com*
*[2] University of Mohaghegh Ardabili, jamali@iust.ac.ir*
*Author Correspondence: Computer Engineering and Information Technology Department of University of Mohaghegh Ardabili, Ardabil, Iran, Tel.: +98 4515 513917; fax: +98 4515510141, qavami@gmail.com*

## Abstract

The growing tendency of information technology users to use cloud computing based services encourages service providers to offering services in a way that is different in functional and non-functional features. Providing resource on demand from a resource pool is one of the best offers from cloud computing based systems. This type of service offering is not limited to resources rather includes spread range of services. Services are offered from infrastructure layer to software layer in this way. Based on supply on demand rules and because of daily growth of the services that is offered, plus the importance of affordable access to reliable high-performance hardware and software resources, reducing maintenance and users' costs, energy consumption and environmental issues, etc the need of resource management techniques rises every day by day. Managing the applications more efficiently in cloud computing motivates the challenge of provisioning and allocating resource on demand in response to dynamically changing workloads. Dynamic resource provisioning in cloud computing addresses the methods that provide resources on demand based on workloads or users requirements. There are several investigations in this area in different levels of cloud basic architecture. In this paper authors tried to review several proposed approaches in dynamic resources provisioning from qualified publishers and also tried to cover all levels of provisioning in cloud computing.

**Keywords***:* Cloud computing, Resource provisioning, Dynamic provisioning, Efficiency.

## 1. Introduction

With recent progressions in Information Technology the need for computations when ever and where ever on the one hand and also the need of individuals and organizations for cost effective heavy duty computation powers on the other hand, have increased the desire for computation as a utility paradigm. Cloud computing is the latest answer to these tendencies where IT resources are offered as services. Providing different computer services every where every time on users' demands is one of the most important ideas of Cloud computing. Cloud computing also offers the user an infinite resource pool (e.g. processing capacity, Memory, Storage etc.); an intrinsic feature of cloud computing that severs it from traditional hosting services.

Everything in computer world need infrastructures and so cloud computing needs an infrastructure too which usually is a Data Center. With daily growth in the amount and the size of the datacenters and warehouse scale computer farms the overall amount of energy consumption would rise considerably. This is not just because of the energy consumption of the devices themselves, also because of the cooling system energy consumption. Between 2000 and 2006 the amount of energy consumed by data centers around the world has doubled and in 2008, the average data center consumption were measured something as much as 25,000 households (J. Kaplan 2008). Computation needs are growing day by day and the big data idea also encourages the cloud owners for growing up their data centers. These growths are expensive and also harmful to the environment, much energy consumption means much GHGs (Greenhouse Gases) .But by utilizing the current resources in a much more efficient manner some part of this desire can be satisfied. That means by optimizing resource usage cloud providers can save money and even save the environment. The environmental issues is interested in green computing topics and the resources optimization The section 2 of this paper explains about the cloud computing from definition to service model. Section 3 studies the proposed approaches in resource provisioning while several subsections are included to categorize the studies. Finally, the section 4, conclusion concludes the main text while references and author biography complete the article.

## 2. Cloud Computing

### *2.1 Cloud Computing Definition*

There are multiple definitions out there for cloud, but some efforts have been made to standardize the definition of it. Beyond them the definition of the NIST (National Institute of Standard and Technology of the U.S.A) (Mell and Grance 2011) seems to be more accepted by the people. Based on the official NIST definition, “cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction”.

### *2.2 Cloud Computing Essential Characteristics*

**On-demand self-service:** A consumer can order several services to each cloud provider, wherever and whenever, without any human interaction requirement.

**Broad network access:** Resource Pools and services which are located in geographically distributed cloud providers can be reachable over the network and accessed through standard mechanisms that boosts use by inharmonious thin or thick client platforms (e.g. PCs, workstations, laptops, tablets and mobile phones).

**Resource pooling:** Providing non-united either physical or virtual resource (e.g. Processing, memory, storage etc.) for the consumers in a way that it appears like a united infinite resource pool. This means that the users have no vision about the certain location of resource because of abstraction techniques. With a higher level of abstraction users may be able to define location of their requested resource (e.g. country, state, or datacenter).

**Rapid elasticity:** Cloud users can request resources as much as they need and scale rapidly up or down them whenever they want. To the consumer, this capability means resource scalability on demand which can be done automatically in some cases (Qavami, Jamali et al. 2013).

**Measured service:** Cloud systems automatically monitor, control, and report resource usage by different level of abstraction based on resource or capability type (e.g. Processing, memory, storage, balancing services etc.). It would allow making transparent billing services to cloud users and also transparent user accounting for cloud providers.

### *2.3 Cloud Basic Service Models*

**Software as a Service (SaaS):** Enables the customers to use applications which are provided on cloud infrastructure. The users can use the applications whenever they need them and they will pay the price just for the time that the applications were used. Usually the processes of the application would be done on the cloud side and customers can access them using different device types even thin clients throw simple program interfaces or web based interfaces (e.g. Gmail, Google Docs). Customers do not have any concentration to keep the application up to date, or to extend application licenses or even to manage the servers, operating system, etc. They will just lean on their seats and enjoy using application.

**Platform as a Service (PaaS):** Platform is usually a less equipped system comparing to the application service layer. It usually includes a server on cloud infrastructure with an operating system with some specific development tools based on platform requirements (e.g. Apache Hadoop) in which applications can be deployed and performed. Customer does not have any administration control on underlying infrastructures, operating system, etc; in some cases user has control over the deployed applications and configuration options for the hosting environment of applications.

**Infrastructure as a Service (IaaS):** Infrastructure refers to underlying physical components that are fundamental parts of a computing system. IaaS provisions processing units, storages, network communication, and other basic computing resources to the consumers. This gives cloud users the power of full managing provisioned resources for deploying preferred operating systems, applications etc. Even with this level of

Administration controls on operating systems, storage, and deployed applications and may be some network components, the consumers are not supposed to control the underlying cloud infrastructure.

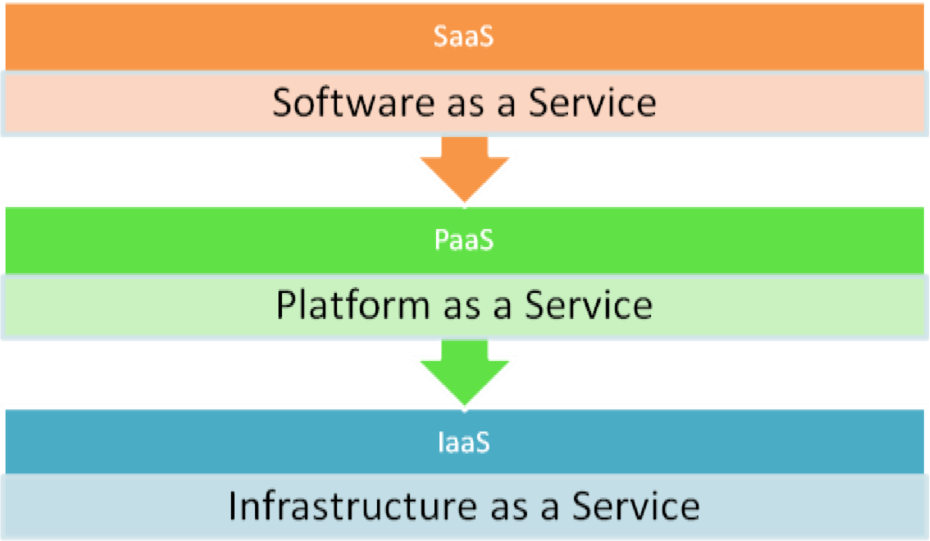

**Figure 1:** Cloud computing basic service models

## 3. Cloud computing Resource Provisioning (CCRP)

### *3.1 CCRP definition*

Generally the term Resource Provisioning in Cloud Computing are used for the taking in, deployment and managing an application on Cloud infrastructure, which includes controlling resources in all layers of the cloud. There are entirely two generic way of resource provisioning, first is *Static Resource Provisioning* which usually provides the peak time needed resource all the time for the application. In this kind of provisioning must of the time the resources are wasted because the workload is not peaked, but resource providers provide the maximum (peak time) required resource to preventing SLA[1] violation. This type of provisioning has mutual disadvantage for both the provider and user, provider must implement more infrastructure to expand its business and serve more user while there are Non-loaded resource which are assigned to the current users, and on the other side the user must pay for the resources which are not used in the most of the time. The second way which is used in recently advanced cloud data centers is *Dynamic Resource Provisioning*. The basic fundamental idea in the latter way is providing the resources based on the application needs, this helps the provisioner to assign the Non-loaded resources (which become free to used now) to the new users. This method reduces a fraction of providers' development costs by utilizing current available resources and beside that the user can happily just pay for the amount of the resources which were really used; this type of billing

Which seems fairer, are usually called “Pay As you Go”. The term *Adaptive Resource Provisioning* sometimes are used instead of *Dynamic Resource Provisioning*, because in this type of provisioning the amount of resources that belongs to a specific job are adapting continuously during the run time. The focus of the investigations in this area is on the methods which adapt the resource.

In this paper writers tried to review a few of the current dynamic resource provisioning investigations to demonstrate different ideas and to make it easier for other researchers to compare them.

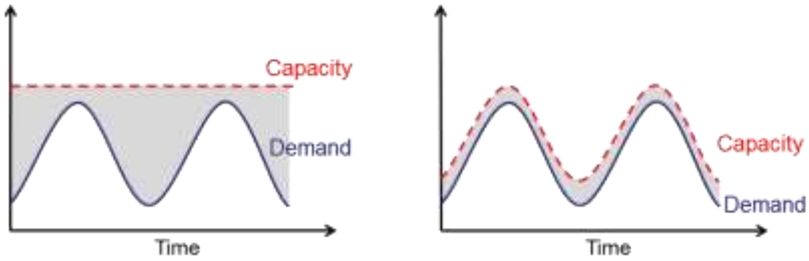

**Figure 2:** Static Resource Provisioning vs. Dynamic Resource Provisioning

### *3.2 CCRP challenges*

With all advantages of dynamic resource providing, we should not forget its implementation and development challenges and difficulties. For provisioning planning we should take appropriate provisioning times. Provisioning resources too soon will wastes our resources and therefore our money, on the other side provisioning resources too late will cause potentially service level agreement (SLA) violations and makes the users angry. Of course we must know the amount of resources that are needed to be provisioned, under or over provisioning are not acceptable with the same prior reasons.

With all efforts and considerations it is clear that a completely exact prediction is not possible and always there would be at least minor errors. Therefore as a feasible method we use estimation and we accept that it is not really and completely an exact way. Estimation methods try to estimate future needs as correct as possible while they must face some difficulties such as high varying environments and changes with different intensities either in loads and resources themselves.

Some really important issue that must be considered in this area is to keep the users satisfied of quality of services (QoS). Dynamic resource provisioning is important because it reduce our costs, energy consumption, save the environment etc. but it must work in a way without harming users demands. It must balance the efficiency in a side and the users’ demands such as service level agreement (SLA) contents and QoS parameters on the other side.

### *3.3 Proposed approaches for CCRP*

Several approaches for resource provisioning proposed in this literature and were categorized based on approaches in this section. Different investigations in this area are studied from section 3.3.1 to 3.3.15 from the year 2001 to later 2013.

### *3.3.1 Host consolidation:*

One of the first resource provisioning investigations in data center level was (Pinheiro, Bianchini et al. 2001) . The main idea was about addressing power conservation for clusters of workstations or personal computers. They presented an algorithm to consolidate workload to as few nodes as possible in a cluster of computers and to turnoff remaining idle nodes for power saving and would only turn them on if necessary for upcoming loads. They have developed their idea and published it latter as (Pinheiro, Bianchini et al. 2003).

### *3.3.2 Capital Market framework:*

One of the other headmost investigations about power management was carried out by (Chase, Anderson et al. 2001). Considering the energy optimization problem in internet hosting servers, they modeled the problem using a capital market framework and solved the problem with auction-bid model. The virtualization technology provides easier and more efficient ways for cloud providers to provide their services to cloud users. Now providers can divide their giant servers to smaller ones and offer them to users by a new unit called virtual machine with different configurations. Some cloud providers use a mechanism in which there is a fixed cost for selling virtual machines and after that remaining resources will enter in an auction mechanism. (Zaman and Grosu 2012) believed that current auction based virtual machine provisioning systems contain a shortcoming which is they use offline systems. This being offline means they gather the information periodically and do their actions based on them, so they are not really online system. They proposed a new bid based (capital market model) approach for responding to the users' requests which is really online and response to users immediately. User sends its request including virtual machine type and required renting time and system tries to answer the maximum possible responds to these requests. The system does accounting base on this request for accepted requests.

### *3.3.3 A Novel called Self-Resource Scaling:*

In (Nathuji and Schwan 2007) beside scaling and consolidation of virtual machines authors developed a novel approach called Self-Resource Scaling; using that by the virtual machine manager the size of virtual machines would change to an appropriate level base on requirements.

### *3.3.4 Combinatorial policies:*

In a web environment and assuming single application running host considering a specific response time we can change the CPU frequency on demand. (Elnozahy, Kistler et al. 2003) combined Dynamic Voltage Frequency Scaling (DVFS) with dynamically turning on/off method called VOVO (vary-on/vary off) to reduce power consumption during reduction period of workload. (Raghavendra, Ranganathan et al. 2008) also described and analyzed five distinct policies for power management in server cluster with web workloads.

### *3.3.5 Limited Look ahead control:*

(Kusic, Kephart et al. 2008) described the problem of power management in a heterogeneous virtualized environments as a sequential optimization and demonstrated it using Limited Look ahead Control (LLC). The goal was to maximize the resource provider's profit by minimizing both power consumption and SLA violation. Kalman filter was used to predict the number of next coming requests to predict the future state of the system and perform necessary reallocations, but the approach seems to produce some sensible overheads due to its high level of complexity.

### *3.3.6 Min-Max algorithm:*

Power optimized virtual machine allocation was also done using Mean-Max algorithm. In a heterogeneous environment this can demonstrate minimum and maximum ratio of allocating CPU resource to virtual machines which are using a common resource (Cardosa, Korupolu et al. 2009).

### *3.3.7 Bin Packing problem:*

(Verma, Ahuja et al. 2008) solved the problem of power-aware dynamic placement of applications in virtualized and heterogeneous environment considering power consumption, by using Bin Packing problem. In their study, the total available power is considered as distributable amounts that must be distributed in an efficient way between resources.

### *3.3.8 Constraint satisfaction problem:*

(Van, Tran et al. 2010) developed an optimization method using a utility function. They emphasized private cloud which in them the applications are so heterogeneous from batch processing work to online request with powerful service requests. Their works contains three main portions: a utility-based dynamic Virtual Machine (VM) provisioning manager capable of balancing application SLA compliance with energy consumption, a dynamic virtual machine placement manager which consolidates virtual machines on the minimum number of physical hosts through virtual machine live migration so that idle hosts can be turned off to save energy, a two-level resource management middleware framework with a clear separation between application-specific management and a generic resource management substrate. They initiated a few fixed classes of virtual instances (like public clouds) and a matrix which showed how many of each virtual machine classes are assigned to each application. The main object in resource provisioning division is to determine this matrix entry. For modeling both provisioning and allocating problem they used Constraint Satisfaction Problem (CSP).

### *3.3.9 Dynamic Round Robin:*

(Lin, Liu et al. 2011) purposed a new Round Robin algorithm called Dynamic Round Robin (DRR) to allocation and migration of Virtual Machines between hosts. The DRR consolidates virtual machine in physical host to minimize the number of hosting machines and as a result, reducing energy consumption and increasing in use hosts' utilization. They defined two main rules for their system .The first rule is that if a virtual machine has finished and there are still other virtual machines hosting on the same physical machine, this physical machine (called retiring machine) will not accept any new virtual machines, and this avoids adding extra virtual machines to a retiring physical machine so it could be shut down. The second rule is that if a physical machine is in the retirement state for a sufficiently long period of time, it will be forced to migrate remaining virtual machines to other physical machines which will speed up the consolidation process.

### *3.3.10 Threshold based virtual machine varying:*

(Lin, Wang et al. 2011) introduced a dynamic Virtual Machine-Varying Based resource allocation using a threshold. Using this threshold their algorithm decides that the current counts of virtual machines which are assigned to an application are sufficient or not, it is the same for over provisioning. They have defined two other parameters in threshold formulation; one is a rate called normal rate which demonstrates the average amount of workload that one individual virtual instance can tolerate without any over utilization and the other is a parameter that would be defined by system admin based on the work load; those two made the approach very parametric which seems to be a weakness. The basic differences and advantages of our study as compared to the latter are that first, our work does not need any human admin interferences and is able to estimate next workload instead of a reactive action.

### *3.3.11 Queuing model and analytical performance:*

Several investigations in scheduling or even in resource provisioning have used QoS parameter as a criterion for majoring quality and capability of investigated approaches. Considering the QoS parameters is very important to achieve service level agreement (SLA) service quality goals. The provisioning process is complicated in cloud environment because the provisioner must obtain an exact calculation for configuring hardware and software in the best way according to user requirements. This is very important for providing best possible QoS to the users. Some uncertain behaviors in the cloud environment like high variability in workloads, network elements or even virtual resources it selves make this process more difficult; also the approximation is not exact anyway. (Calheiros, Ranjan et al. 2011) addressed workload prediction and resource adaption using a queuing model and analytical performance. The QoS parameters which were considered for this analytical performance method are response time and job rejection rate. The disadvantage of this approach like previous work is dependency on human control parameter that here is the *k* parameter in their M/M/1/k model.

### *3.3.12 Heuristic bottleneck detection approach:*

With spread web services and their popularity in the world, hosting web servers in the clouds can be considered as a serious challenge for cloud providers. Beside that response time for web users is a very sensible and important parameter; this is one of the reasons that it appears frequently in web SLAs. With such difficulties it can be concluded that giving a guarantee for response time in web sphere is not an easy task. Various application workload traffics and the multi tier nature of the web structure make it difficult or may be impossible for a human administrator to detect bottlenecks and of course SLA violations after that. For resource provisioning in this level, (Iqbal, Dailey et al. 2011) aimed a bottleneck detection system for multi tier web application using a heuristic approach. This mechanism is able to detect bottlenecks in every tier of system with consideration of response time and provision extra virtual servers to them, for over provisioned areas the system acts vice versa.

### *3.3.13 PSO based approach:*

The fundamental part of a cloud infrastructure is datacenters. Assuming a centralized data center (Jeyarani, Nagaveni et al. 2012) developed a data center manager system with energy consumption centrality; a meta scheduler which its main goal is to map virtual machines to physical servers efficiently. In addition inside an environment by resources and loads with dynamic changes, the system must provide enough resources for user loads. Authors developed a dispatcher using a new PSO (Particle Swarm Optimization) method Called SAPSO (Self-Adaptive PSO) to dispatch virtual machine instances among physical servers efficiently.

### *3.3.14 Neural Network and Linear Regression:*

Islam et al. (Islam, Keung et al. 2012) advanced a new machine learning technique by developing a Neural Network system called ECNN (Error Correction Neural Network) and using it side by side with a Linear Regression. These two are able to predict utilization of resources in the future and automatically scale the host system. The advantage of our work comparing to this, is more simplicity which means less overhead for whole system.

### *3.3.15 Markov based approach*

An efficient enough resource provisioning method by utilizing current resources can reduce costs for both consumer and cloud provider. This matter can be implemented in all several layers of a cloud system. Thank to virtualization and self service panels cloud computing is now the most flexible beyond similar computation service. In (Qavami, Jamali et al. 2013) a heuristic markovian approach was proposed for changing the number of virtual machines which are running a specific application based on the workload demands. They developed a state output machine based on a quasi markovian idea. Each transaction that is between two states contains a probability to happen. These probabilities would be updated using a heuristic approach and after that transaction from current state with greatest probability would be chosen. This method must adapt virtual machine number dynamically to reduce the cost for the application owner, while it must consider the application demands to avoid QoS parameter violation from the users of application point of view.

## 4. Conclusion

The investigations which were studied above, are trying to optimize and utilize the resources for power and financial cost reduction besides satisfying the user’s desires in QoS and SLA. Several method were mentioned here which used different parameters as a goal for resource provisioning such as response time, rejection rate, SAL violation rate, cost etc. The common difficulty between all of them seems to be the balancing between efficiency goals and QoS parameters. It means it is difficult to reduce inefficiency (e.g. in cost or energy) and stay in a reasonable level of QoS for the user. Several ideas were reviewed and it can be concluded from them that managing such big resources for human administrators is not possible anymore and administrators are going to be replaced with managing systems. These systems must use techniques that are able to estimate and allocated resource in the most efficient way while avoiding SLA violations.
With respect to the findings in this paper for the future work, preparing a suitable test bed for implementing and comparing similar approaches is considered.

## REFERENCES

Calheiros, R. N., R. Ranjan, et al. (2011). Virtual Machine Provisioning Based on Analytical Performance and QoS in Cloud Computing Environments. Parallel Processing (ICPP), 2011 International Conference on.

Cardosa, M., M. R. Korupolu, et al. (2009). Shares and utilities based power consolidation in virtualized server environments. Integrated Network Management, 2009. IM '09. IFIP/IEEE International Symposium on.

Chase, J. S., D. C. Anderson, et al. (2001). "Managing energy and server resources in hosting centers." SIGOPS Oper. Syst. Rev. **35**(5): 103-116.

Elnozahy, E. N., M. Kistler, et al. (2003). Energy-efficient server clusters. Proceedings of the 2nd international conference on Power-aware computer systems. Cambridge, MA, USA, Springer-Verlag**:** 179-197.

Iqbal, W., M. N. Dailey, et al. (2011). "Adaptive resource provisioning for read intensive multi-tier applications in the cloud." Future Generation Computer Systems **27**(6): 871-879.

Islam, S., J. Keung, et al. (2012). "Empirical prediction models for adaptive resource provisioning in the cloud." Future Generation Computer Systems **28**(1): 155-162.

J. Kaplan, W. F., N. Kindler (2008). Revolutionizing Data Center Energy Efficiency, McKinsey & Company.

Jeyarani, R., N. Nagaveni, et al. (2012). "Design and implementation of adaptive power-aware virtual machine provisioner (APA-VMP) using swarm intelligence." Future Generation Computer Systems **28**(5): 811-821.

Kusic, D., J. O. Kephart, et al. (2008). Power and Performance Management of Virtualized Computing Environments Via Lookahead Control. Proceedings of the 2008 International Conference on Autonomic Computing, IEEE Computer Society**:** 3-12.

Lin, C.-C., P. Liu, et al. (2011). Energy-Aware Virtual Machine Dynamic Provision and Scheduling for Cloud Computing. Cloud Computing (CLOUD), 2011 IEEE International Conference on.

Lin, W., J. Z. Wang, et al. (2011). "A Threshold-based Dynamic Resource Allocation Scheme for Cloud Computing." Procedia Engineering **23**(0): 695-703.

Mell, P. and T. Grance (2011). The NIST Definition of Cloud Computing. C. S. D. Department of Commerce, Information Technology Laboratory. Gaithersburg, U.S.A, NIST Special Publication 800-145.

Nathuji, R. and K. Schwan (2007). "VirtualPower: coordinated power management in virtualized enterprise systems." SIGOPS Oper. Syst. Rev. **41**(6): 265-278.

Pinheiro, E., R. Bianchini, et al. (2001). Load balancing and unbalancing for power and performance in cluster-based systems.

Pinheiro, E., R. Bianchini, et al. (2003). Dynamic cluster reconfiguration for power and performance. Compilers and operating systems for low power, Kluwer Academic Publishers**:** 75-93.

Qavami, H. R., S. Jamali, et al. (2013). Dynamic Resource Provisioning in Cloud Computing: A Heuristic Markovian Approach. 4th International Conference on Cloud Computing. Wuhan, People's Republic of China, EAI.

Raghavendra, R., P. Ranganathan, et al. (2008). "No "power" struggles: coordinated multi-level power management for the data center." SIGOPS Oper. Syst. Rev. **42**(2): 48-59.

Van, H. N., F. D. Tran, et al. (2010). Performance and Power Management for Cloud Infrastructures. Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on.

Verma, A., P. Ahuja, et al. (2008). pMapper: power and migration cost aware application placement in virtualized systems. Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware. Leuven, Belgium, Springer-Verlag New York, Inc.**:** 243-264.

Zaman, S. and D. Grosu (2012). An Online Mechanism for Dynamic VM Provisioning and Allocation in Clouds. Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on.

## A Brief Author Biography

***Hamid Reza Qavami*** – Attended M.Sc. student in University of Mohaghegh Ardabili, Iran. He received his B.Sc. degree from the Dept. of Computer Engineering, Islamic Azad University of Kashan, Iran in 2011. He studied multiple issues in computer science and engineering. His major fields of study include cloud computing, computer networks and robotics. Mr. Qavami is member of the Cloud Research Center in Amirkabir University of Technology, Iran.

***Shahram Jamali*** – Associate Professor of University of Mohaghegh Ardabili, Iran. He received his M.Sc. and Ph.D. degree from the Dept. of Computer Engineering, University of Science and Technology, Iran in 2001 and 2007, respectively. He has published several papers in different fields of computer science. His study interests include computer networks congestion control, network security, natural optimization algorithms etc. Dr. Jamali is the head of Computer Engineering and Information Technology Department and also the head of Computer and Network Research Lab in University of Mohaghegh Ardabili, Iran.