



RESOURCE MANAGEMENT IN CLOUDS: OUTLOOK AND REFLECTIONS

Aneena Ann Alexander¹ and S. Durga²

¹Department of Computer Science and Engineering, Karunya University, Coimbatore, India

²Department of Information Technology, Karunya University, Coimbatore, India

E-Mail: aneenaalex@gmail.com

ABSTRACT

Cloud computing is a common buzz word in today's computing environment where processing, storage, network and software are provided as an on demand service to their customers. The resources required by different users depend on their respective personalized applications. Advances in technologies on the other hand, lead to the migration from traditional desktop devices to smart mobile devices. Resource management in cloud is a complex problem due to its inherent nature such as heterogeneity in the resource types and their interdependencies, unpredictable load and scalable nature of the datacenters. This paper reviews various provisioning and load balancing strategies which outlines a conceptual framework of the resource management in various clouds. Additionally, an outlook of common research challenges such as meeting customer demands, application requirements, achieving global manageability of the cloud computing resources, developing energy efficient strategies are discussed for an efficient cloud resource management.

Keywords: cloud computing, resource provisioning, load balancing, mobile cloud computing, resource management.

INTRODUCTION

Cloud computing is a new emerging trend in IT environment with huge requirements of infrastructure as well as resources. The basic principle of cloud computing is that the data is not stored locally in the machine, but will be available in the data centers. The users can access the stored data by using an application programming interface in the terminal equipment, provided it should be connected to the internet. The processing units in cloud are called virtual machines, and in order to reduce the execution time VM should run in parallel.

The cloud computing framework is made up of certain components. There is a back end and a front end which needs to connect and communicate. This communication can be accomplished by using internet. The back end as the name suggests is the "not seen" end of the network, which is the group of machines that form the network. The front end consists of the client hardware and software that helps us to connect to the cloud network. The services provided by the cloud can be classified as Infrastructure as a service (IaaS), Platform as a service (PaaS) and Software as service (SaaS). Cloud computing encompasses both provisioning the resources to the third parties on a leased, usage-based basis and the private infrastructure maintained and utilized by private individuals. The former one is called public clouds and the later one is referred to as private clouds. Cloud bursting is the term which is used when an enterprise tries to extend the capacity of its private cloud by leasing resources from public cloud.

A. Entities of cloud

The following are the main entities of cloud and their roles are discussed below [31].

End user

They refer to the consumers of cloud. It is required that all users must adhere to the SLA specified by

the cloud provider. They use the services on an on-demand basis and they are required to pay only for what they are using. Before signing SLA (Service Level Agreement), the users of cloud must verify that SLA contains certain Quality of service (QoS) parameters which are the prerequisites of the consumer before using the cloud services. Therefore, on the perspective of an end user cloud computing is a scenario where users have access to any kind of services like IaaS, PaaS, and SaaS on a pay as you use manner.

Cloud provider

It manages a set of data center hardware or software system resources. They can offer private, public or hybrid clouds.

Private clouds [21] are owned by the organisations or enterprises for their internal use. They may use it to store and manage the Big-Data of their organization or to provide enough resources on an on demand basis to its team of consumers. Some of the examples of private cloud include Open Stack [23], VMware [24] and Cloud Stack [25].

Public clouds [21] can be used by either individuals or an organization depending upon their requirements. They offer higher level of efficiency in shared resources. Confidentiality is the major security issue in using public cloud. Some of the public clouds include Amazon web services [26], Google Compute Engine [27], Microsoft Azure [28], HP cloud [29] etc.

A hybrid cloud [21] is a combination of public and private clouds. It allows businesses to manage some resources internally within organization and some externally. The main disadvantage is that the complexity of overall management of resources increases along with security concerns.



Cloud developer

This component lies between the end user and cloud provider. Cloud developer has the responsibility of taking into consideration both the view of end user and cloud provider. The developer of cloud must adhere to all the technical details of the cloud which are essential to meet the requirements of both, the end user as well as the cloud provider. In addition to the objective of satisfying customer SLAs, the Cloud Provider may also have objectives relating to the management of its data center infrastructure which include: balanced load, whereby resources are allocated in such a manner that utilization is balanced across all resources of a particular type; fault tolerance, whereby resources are allocated in a manner such that the impact of a failure on system performance is minimized; or energy use minimization, whereby data center resources are allocated in a manner that the amount of energy required to execute a given workload is minimized.

This paper covers resource management in cloud environment. The rest of the paper is organized as follows: Section II discusses Provisioning and load balancing techniques in cloud computing environment. Section III discusses some related work to provisioning and load balancing algorithms. Section IV discusses the challenges and future directions. The study is concluded in Section V.

RESOURCE PROVISIONING AND LOAD BALANCING

A. Resource provisioning

In cloud computing, resource provisioning means the process of selection, deployment and run time management of the software (load balancers) and hardware resources (CPU, storage). Resource allocation is

the process of assigning the available resources to the needed cloud applications. Resource allocation starves services if the allocation is not managed precisely. Resource provisioning solves this problem by allowing the service providers to manage the resources for each individual module. This resource provisioning takes Service Level Agreement to consideration, which is the initial agreement between cloud users and cloud service providers which ensures QoS parameters like performance, availability, reliability etc. Resources should be provisioned in such a way that there should be reduced SLA-violations. In order to achieve this goal the cloud user has to request the cloud service provider to provision the resources either statically or dynamically based on the application needs.

The provisioning techniques can be classified as follows based on the needs of the application. By provisioning the cloud service provider will know how many instances and what all resources are needed for a particular application. The classifications of the provisioning techniques are as follows.

Static / advance / reserved provisioning techniques-These provisioning technique is used for those applications which have predictable or constant workloads. In this technique the customer will enter into a contract with the provider and prepares the appropriate resources in the start of the service.

Dynamic Provisioning techniques-In those applications which are having changing workloads the dynamic provisioning techniques can be used, where virtual machines may be migrated to the new nodes in the cloud. In this technique the provider allocates more resources as they are needed and removes when not in use. Common Parameters of provisioning are shown in Table-1.

Table-1. Parameters of resource provisioning.

Parameter	Definition	Focus
Response time	Amount of time taken by an algorithm to respond.	Low
Fault tolerance	The ability of the algorithm to continue providing the services in case of a node failure	High
SLA(service level agreements)	Contains certain QoS parameters which are the pre- requisites of consumers before using cloud services.	Minimal SLA violations
Power consumption	The amount of power consumed for provisioning.	Low
Cost	Two perspectives: Cloud users point of view and Cloud providers point of view	Low (in terms of user), High(Profit- in terms of provider)

Resource Allocation (RA) is the process of assigning available resources to the needed cloud applications. Resource provisioning solves this problem by allowing the service providers to manage the resources for each individual module.

A good resource allocation algorithm should avoid the following criteria

a) Resource contention-This situation arises when two applications try to access the same resource at the same time.



b) Scarcity of resources -This situation occurs when there are no sufficient resources for the fulfillment of the application needs.

c) Over-provisioning-The situation that arises when the application gets surplus resources than demanded.

d) Under-provisioning -This occurs when the applications are assigned fewer number of resources than demanded.

B. Load balancing

Load balancing is one of the main issues of cloud computing. These loads can be a capacity of CPU, memory, network or delay load. To improve the resource utilization and for better performance of the system, it is

always necessary to share work load among the various nodes of the distributed system. By balancing the load we can avoid the situation where nodes are either heavily loaded or under loaded in the network. Load balancing is the process of assuring the evenly distribution of work load on the processor or pool of system node so that without disturbing, the running task is completed. The goals of load balancing [8], [9] are to:

- 1) Increase the performance.
- 2) Maintenance of system stability
- 3) Achieving fault tolerant system by creating backup.

C. Classification of load balancing algorithms

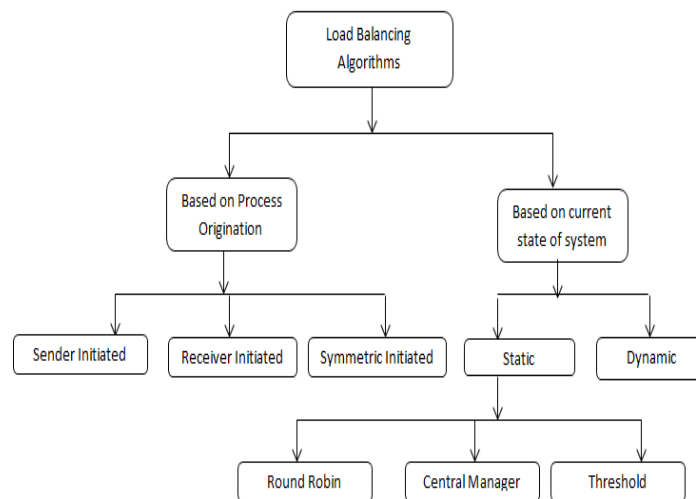


Figure-1. Classification of load balancing algorithms.

Based on process origination, load balancing algorithms can be classified as [1], [2], [3].

Sender initiated: In this type of load balancing algorithm the client keeps on sending the request until a receiver is assigned to receive his workload which means here the sender initiates the process.

Receiver initiated: This type of load balancing works in such a way that the receiver sends a request to acknowledge a sender who is ready to share the workload which means the receiver initiates the process.

Symmetric: These classes are a combination of both sender and receiver initiated type of load balancing algorithm.

Based on the current state of the system there are two other types of load balancing algorithms [1], [2], [10].

Static algorithm: In static algorithm the load is distributed equally among all the available servers. This

algorithm requires only a little understanding of system resources, and therefore the decision of transferring the load on the system does not depend on the current state of the system. This scheme of load balancing is highly suited in the scenario where the fluctuations in load are comparatively low.

The different types of Static load balancing algorithm include Round Robin algorithm, Central Manager Algorithm, Threshold algorithm and randomized algorithm.

Dynamic algorithm: In dynamic load balancing algorithms, the lightly loaded servers will be considered for balancing the load. For finding the lightly loaded server, a communication with the network is needed which can increase the traffic in the whole network. Unlike in static algorithms, current state of the system is used to make conclusions regarding the management of loads. The common dynamic load balancing algorithms involve central queue and local queue algorithms. Table-2 shows the parameters of load balancing.

**Table-2.** Parameters of load balancing.

Parameter	Definition	Focus
Throughput	Number of jobs completed per unit time.	High
Associated overhead	It includes inter process communication and movement of tasks	Minimum
Migration time	Amount of time taken to transfer the load from one node to another.	Minimum
Resource utilization	The extent to which resources are utilized completely.	High
Fault-tolerance	It is the ability of the algorithm to perform correctly and uniformly even in conditions of failure at any arbitrary node in the system	High
Efficiency	Represents the effectiveness of the system. If all the above parameters are satisfied optimally then it will highly improve the performance of the system	High

RELATED WORKS**A. Resource provisioning in cloud computing**

There are various models suggested to solve the problem of resource provisioning in cloud computing. The main aspect of resource provisioning is dynamic allocation of resources in the cloud the size and the complexity of

datacenters are growing day by day to meet the increased demand for resources. Cloud computing service providers must allocate enough resources in order to meet the specified SLA. Most resource provisioning algorithms are designed so that it must guarantee both minimal response time and minimal resource usage cost. A comparison of various provisioning algorithms is shown in Table-3.

Table-3. Comparison of provisioning algorithms in cloud.

Provisioning technique	Cost in users point of view	SLA violation	Power consumption	Throughput	Fault tolerance
Failure aware policy[22]	Low	Low (32%improvement in QoS)	-	High	High
QoS based provisioning [5], [14]	Low	Low	-	High	-
Elastic application container [29]	Low	-	Low	-	-
Priority based approach [4]	-	-	-	High	-
APA-VMP [30]	Low	Low	Low	High	-
Energy aware [7]	Low	-	Low (30% reduction in power)	High	-
Inter cloud resource provisioning	-	Low	-	High	-
Genetic algorithm	-	Low	-	-	-
Workload model [35]	Low	Low(20% reduction in deadline violation rate)	-	-	-
Server consolidation	Low	Low	Low (32% reduction in power)	High	High

Chandrashekar *et al* proposed a priority based dynamic resource allocation algorithm [4] which dynamically responds to the fluctuating workload by preempting the current executing task having low priority with high priority task and if pre-emption is not possible due to same priority then by creating the new VM from globally available resources. Hariharasudhan *et al* proposed an energy aware resource allocation engine [7] where the arbitrator is aided by an energy aware allocation

engine, which will distribute the workload tasks optimally among the service providers that will ensure that the data providers do not drain valuable energy. AmitKumar *et al* proposed an intelligent approach for virtual machine and QoS provisioning [5], where the target QoS has been met by controlling the admittance of requests so that the system does not get overloaded. The analyzer first checks if the available resources are sufficient for servicing newly requested job. If it finds that the job can be served within



the time promised in SLA, it lets the job to enter the computing premise, and it is placed in a queue. If no queue is found, new queue is formed and it is placed in the new queue. Xiaoming *et al* [22] proposed an optimal resource allocation for multimedia cloud in priority service scheme where a queuing model is proposed, which consist of three queuing systems, namely schedule queue, computation queue and transmission queue. The request in the scheduling queue is sent to the computing server at a scheduling rate S by the master server, which always schedules the highest priority request first. In transmission queue results are stored and are then delivered to the customers. Khalid. *et al* described a cloud assisted mobile architecture [6], focusing on offloading. The mobile provider decides on the best executional plan and where offloading will be beneficial. Here cloud offers dynamic on demand provisioning for mobile users. Natasha *et al* proposed a method to handle same priority requests [13], by managing resources using priority based approach. Here the requests with attached priorities are received by the resource allocator, which is the initiator. Here all priorities will be extracted sorted and same priority request will be grouped together. Then the load needed by the request is calculated, and it is sorted. Then the total available load of the server is calculated and threshold is found. Zhang *et al* [15] proposed a dynamic heterogeneity aware resource provisioning in cloud which is capable of performing DCP in heterogeneous data centers here task classification followed by resource prediction is done, and after that DCP is done. A software platform infrastructure

model is proposed by George *et al* [14], where all the QoS requirements are met. Here two level generic black box approaches is used based on behavioral management to access cloud. It can predict the user behavior.

B. Resource provisioning in mobile cloud computing

Deboosere *et al.* [11] proposed an algorithm for assigning resources for mobile users in grid environment. They faced two problems viz. protocols used in LAN are not applicable to WAN for grid environment and selection appropriate server for mobile so that it can provide quality of service to MC. They focused on network latency rather than bandwidth of network because due to mobility nature of MC, network latency can increase significantly. The problem of network latency is resolved by creating three managers viz. service manager, resource manager and server selection manager at the provider central site. These managers are responsible for computing environment. They provided a thematic classification of all distributed application processing frameworks into six main categories depending on area of use and theme of framework. They discussed issues and challenges in current frameworks and suggested future areas for optimum distributed application processing frameworks development. In 2014, Shiraz *et al.* [12] investigated the runtime overload on mobile device while offloading mobile applications over the cloud. Before offloading mobile application on the cloud, it is profiled and partitioned for locating computational extensive components.

Table-4. Comparison of load balancing algorithms in cloud.

Load balancing technique	Makespan	Resource utilization	Response time	Efficiency	Fault tolerance
Max-min algorithm[16]	Low	High (79.5%)	Low	-	-
OLSRA[20]	Low	High	Low	-	-
Ga-max-min [17]	Low	High	-	-	-
Ant colony ptimization [18]	Low	High	Low	-	High
OLB and LBMM [20]	Low	High	Low	High (80%)	-
Min-min [19]	Low	High (76.5%)	Low	-	-
Dynamic LB algorithm [10]	Low	High	Low	High	-
HTV Dynamic LB[1]	-	Optimal	Low	-	-
ADLB for parallel file system[9]	Low	High	Low	-	-
Hybrid ACO[18]	Low	Efficient	Low	-	-

Profiling and partitioning require additional computation resources from mobile device. Results show that CPU utilization of mobile device increases when partitioning is done for mobile application.

C. Load balancing in cloud

Various max-min algorithms are described in [16] and [17]. In [16], a unique modification of Max-min

algorithm is proposed. The algorithm is built based on comprehensive study of the impact of RASA algorithm in scheduling tasks and the atom concept of Max-min strategy. It is proposed to outperform scheduling map at least similar to RASA map in total complete time for submitted jobs. Improved Max-min is based on the expected execution time instead of complete time as a selection basis. In turn scheduling tasks within cloud



computing using Improved Max-min demonstrates achieving schedules with comparable lower makespan rather than RASA and original Max-min. The concept of load balancer aimed to distribute the tasks to different Web Servers to reduce response times was introduced in [17]. The Web Services have gained considerable attention over the last few years. This paper presents an efficient heuristic called Ga-max-min for distributing the load among servers. Heuristics like min-min and max-min are also applied to heterogeneous server farms and the result is compared with the proposed heuristic for VOD Servers. Ga-max-min was found to provide lower makespan and higher resource utilization than the genetic algorithm. In paper [19] a Load Balanced Min-Min (LBMM) algorithm is proposed that reduces the makespan and increases the resource utilization. When the number of the small tasks is more than the number of the large tasks in a meta-task, the Min-Min algorithm cannot schedule tasks, appropriately, and the makespan of the system gets relatively large. Furthermore it does not provide a load balanced schedule. To overcome the limitations of Min-Min algorithm, a new task scheduling algorithm, are proposed. It has two-phases. In the first phase the traditional Min-Min algorithm is executed and in the second phase the tasks are rescheduled so as to use unutilized resources effectively. A comparison of various load balancing algorithms is shown in Table-4.

CHALLENGES AND FUTURE DIRECTIONS

Based on the analysis various areas for future investigation are identified. Some of them are really hard to solve theoretically and require special practical considerations to implement efficiently in a cloud system. Here various challenges that arise from the evolution of the cloud computing paradigm are also identified.

A. Issues in mobile cloud computing

Compared to desktop computers, mobile devices are having lots of constraints both in terms of energy and bandwidth. The basic challenge of mobile cloud computing is how to overcome or mitigate these constraints through effective practical engineering practices. With the inherent limitations of these devices it is necessary to find a provisioning mechanism that can provision resources to such users before their battery goes down or before their connectivity fluctuates.

The major issues in mobile cloud computing involves

- Low bandwidth
- Computational offloading issues
- Quality of service.

B. Developing energy efficient strategies and acquiring global manageability of datacentres

Global Manageability is the process of controlling a system in such a way to achieve a certain set of objectives [31]. A specific challenge in global management is the management of datacentres, to achieve a desired level of availability of services and robustness,

or to meet a specified performance target. Nowadays green computing is also gaining popularity. So management of datacentres in such a way to reduce the overall energy consumption is also required. Virtual machine clustering and turning off the idle machines can be implemented to reduce the overall power

Consumption of data centres. A related challenge is the management of a federated cloud infrastructure, either by managing a hybrid cloud or co-operative management of a shared infrastructure.

C. Achieving predictability in performance

The predictability in performance can be achieved by following different methods. Specific architectures are devised for this purpose including multi-tiered architecture models, Feedback loops are used which will enable to dynamically resize the services in response to increasing workloads. Finally demand forecasting or demand profiling is also one of the most important factors which affect the accuracy of prediction.

CONCLUSIONS

Resource management is an important concern in Cloud computing environment in order to achieve maximum utilization of resources. In this paper, various resource provisioning and balancing schemes are reviewed each having its own pros and cons. Here, the need for resource management is also discussed which clearly show a set of fundamental research challenges that must be tackled. The issues related to cloud mobile applications are also outlined. It is believed that the issues that are outlined will lead to future research proposals.

ACKNOWLEDGEMENT

The authors are grateful to the faculties and colleagues of Karunya University for providing helpful comments and suggestions regarding this work

REFERENCES

- [1] Ali M Alakeel, "A Guide To Dynamic Load Balancing In Distributed Computer Systems", International Journal of Computer Science and Network Security, Vol. 10 No. 6, June 2010.
- [2] Ram Prasad Padhey, P. Goutam Prasad Rao, "Load Balancing in Cloud Computing Systems", Department of Computer Science and Engineering, National Institute of Technology, May 2011
- [3] Abhijit A Rajguru, S.S. Apte, "A Comparative Performance Analysis of Load Balancing Algorithms In Distributed Systems Using Qualitative Parameters", International Journal of Recent Technology and Engineering, Vol. 1, Issue 3, August 2012.
- [4] Chandrashekhar S. Pawar and Ranjnikant B Wagh, "Priority Based Dynamic resource allocation in Cloud Computing", IEEE 2013.



- [5] Amit Kumar Das, Choong Seon Hong. "An Intelligent Approach for Virtual Machine and QoS Provisioning in Cloud Computing", IEEE 2013
- [6] Khalid Elgazzar, Patrick Martin Hossam S Hassanein, "Empowering Mobile Service Provisioning Through Cloud Assistance", IEEE 2013, DOI 10.1109/UCC.2013.20.
- [7] Hariharasudhan Viswanathan, Eun Kyung Lee, Ivan Rodeero and Dario Pompil, "An Autonomic Resource Provisioning Framework for Mobile Computing Grids", ACM International Conference (2012).
- [8] David Escalante and Andrew J. Korte, "Cloud Services: Policy and Assessment", EDUCAUSE Review, Vol. 46, July/August 2011.
- [9] Parin. V. Patel, Hitesh. D. Patel, Pinal. J. Patel, "A Survey on Load Balancing in Cloud Computing" IJERT, Vol. 1, Issue 9, November 2012
- [10] R. X. T and X. F. Z, "A Load Balancing Strategy Based on the Combination of Static and Dynamic, in Database Technology and Applications", 2nd International Workshop, 2010.
- [11] L. Deboosere, P. Simoens, J.D. Wachter, B. Vankeirsbilck, F.D. Turck, B. Dhoedt, P. Demeester, Grid design for mobile thin client computing, Future Gener. Computer Syst. 27 (6) (2011) 681–693.
- [12] J. Park, H. Kim, Y.S. Jeong, E. Lee, Two-phase grouping-based resource management for big data processing in mobile cloud, Int. J. Commun. Syst. (2013), <http://dx.doi.org/10.1002/dac.2627>
- [13] Natasha, Nivit Gill, "Enhanced Priority Based Resource Allocation in Cloud Computing", IEEE 2013
- [14] George Kousiouris, Andreas Mencychtas, Dimosthenis Kyriazis, Spyridon Gogouvis, "Dynamic, behavioral-based estimation of resource provisioning based on high-level application terms in cloud platforms" Elsevier Journal of Future Generation Computer Systems, 2012, DOI: /10.1016/j.future.2012.05.009
- [15] Qi Zang, Mohamed Fatn Zhani, Raouf Boutaba and Joseph L Hellerstein, "Dynamic Heterogeneity Aware Resource Provisioning in the Cloud", IEEE 2013, DOI 10.1109/ICDS.2013.28.
- [16] O. M. Elzeki, M. Z. Reshad, M. A. Elsoud "Improved Max-Min Algorithm in Cloud Computing" International Journal of Computer Applications (0975 - 8887) Volume 50 - No.12, July 2012.
- [17] Alok Kumar Prusty, Bibhudatta Sahoo "Heuristics Load Balancing Algorithms for Video on Demand Servers" Department of Computer Science and Engineering, National Institute of Technology, Rourkela, ODISHA, INDIA
- [18] Ratan Mishra and Anant Jaiswal "Ant Colony Optimization: A Solution of Load balancing in Cloud" International Journal of Web and Semantic Technology. Vol. 3, No. 2. pp. 33-50. April 2012
- [19] T. Kokilavani, Dr. D.I. George Amalarethnam "Load Balanced Min-Min Algorithm for Static Meta-Task Scheduling in Grid Computing" International Journal of Computer Applications (0975 – 8887) Volume 20– No.2, April 2011
- [20] Minxian Xu, Wenhong Tian "An online load balancing scheduling algorithm for cloud data centers considering real-time multi-dimensional resource", Proceedings of IEEE CCIS2012, Hangzhou, China, Oct 30 2012–November 1 2012.
- [21] Zhang, Q., Cheng, L. and Boutaba, R. (2010). Cloud computing: State-of-the-art and research challenges. Journal of Internet Services and Applications, 1(1), 7-18. DOI 10.1007/s13174010-0007-6.
- [22] Bahman Javadi, Jemal Abawajy, Rajkumar Buyya "Failure-aware resource provisioning for hybrid Cloud infrastructure", Elsevier Journal of Parallel and Distributed Computing, Volume. 72, Issue 10, pp 1318-1331, October 2012, DOI: /10.1016/j.jpdc.2012.06.012.
- [23] Open Stack: An Overview. Retrieved from www.openstack.org/downloads/openstack-overview-datasheet. Pdf.
- [24] White Paper (2013). Measuring the Business Value of VMware Horizon View.
- [25] Huang, A. Software Architect, Citrix Systems Apache Cloud-Stack Architecture
- [26] Amazon Elastic Compute Cloud (EC2). Retrieved from <http://www.amazon.com/ec2/> (accessed on April 18, 2010).
- [27] Google App Engine, <http://appengine.google.com> (April 18, 2010)
- [28] Chappell, D. (2008). Introducing the Azure services platform. White paper.
- [29] Sijin He, Li Guo, Yike Guo, Chao Wu, Moustafa Ghanem, Rui Han, "Elastic Application Container: A Lightweight Approach for Cloud Resource Provisioning", IEEE International Conference on



Advanced Information Networking and Applications,
DOI:/ 10.1109/AINA.2012.74.

- [30] R. Jeyarani, N. Nagaveni, R. Vasanth Ramc, "Design and implementation of adaptive power-aware virtual machine provisioner (APA- VMP) using swarm intelligence", Elsevier Journal of Future Generation Computer Systems, Volume 28, Issue 5, May 2012, pp. 811-821, DOI:/ 10.1016/j.future.2011.06.002.
- [31] B. Jennings and R. Stadler. Resource management in clouds: Survey and research challenges. Journal of Network and Systems Management, pages 1–53, 2014.