International Journal of Advance Research and Innovation
Vol. 1(3), Jul-Sept 2013, pp. 20-28
Doi: 10.51976/ijari.131303
www.gla.ac.in/journals/ijari

Article Info
Received: 01 Nov 2013 | Revised Submission: 20 Nov 2013 | Accepted: 30 Nov 2013 | Available Online: 15 Dec 2013

Study on Cloud Computing Resource Allocation Strategies

Mahendra Singh Sagar, Babita Singh** and Waseem Ahmad****

ABSTRACT

Cloud computing is offering utility-oriented IT services to users worldwide. Based on pay-as-you-go model, it enables hosting of pervasive applications from consumer, scientific, and business domains. It is a revolution of traditional data centre's and offers subscription-based access to infrastructure, platforms, and applications that are popularly referred to as Infrastructure, Platform and Software as a Service. Numerous IT vendors are promising to offer computation, storage, and application hosting services and to provide coverage in several continents. These vendors required a huge amount of energy for contributing to high dynamic cost along with a drawback of environment pollution. Therefore, current scenario needs Green computing to save energy and reduce dynamic costs too. Because of increasing demand of high speed computation and data storages, distributed computing system has beckon a lot of contemplation. Resource allocation plays an indispensable role in distributed system where clients have service level agreements. Due to these IT Vendors total profit depends on these Service level agreements.

Keywords: *Cloud Computing; Virtual Machine; Resource Allocation Strategy etc*.

1.0 Introduction

Cloud Computing Resource allocation is constituent, unfolds part of many data centre management dilemma i.e.; virtual machine induction in data centres, network virtualization, and multi-path network routing. The processing, data storage, and communication resources are considered as three dimensions in which optimizations are performed. [1]

Now a day's Cloud computing has become a prevailing exemplary of infrastructural services rehabilitation long-established data centres and providing distributed applications with the resources that they need to serve Internet-scale workloads. The Cloud computing model is centred on the use of virtualization technologies to take the advantage of statistical multiplexing on a shared infrastructure.

Resource allocation protocol in public Clouds are today broadly materialistic to engrossment that distributed applications have from their underlying infrastructure.

As a result, assumptions about data-centre topology that are built-into distributed data-intensive applications are often violated, impacting performance and availability goals. [2][3]

1.1 Resource allocation and its significance

In cloud computing, Resource Allocation (RA) is the process of assigning available resources to the needed cloud applications over the internet. Resource allocation starves services if the allocation is not managed precisely. Resource provisioning solves that problem by allowing the service providers to manage the resources for each individual module.[4] [5]

Resource Allocation Strategy is all about integrating cloud provider activities for utilizing and allocating scarce resources within the limit of cloud environment so as to meet the needs of the cloud application. It requires the type and amount of resources needed by each application in order to complete a user job. The order and time of allocation of resources are also an input for an optimal RAS [6]. An optimal Resource Allocation Strategy should avoid the following criteria as follows.

First criteria are Resource Contention in which demand exceeds supply for a shared resource, such as

**Corresponding Author: Department of Computer Science & Engineering, NIT Hamirpur, India (E-mail: Mailbabita17@gmail.com)*
***Department of Computer Science & Engineering, Alfalah School of Engineering & Technology, Dhauj Faridabad, Haryana, India*
****Department of Computer Science & Engineering, Alfalah School of Engineering & Technology, Dhauj Faridabad, Haryana, India*

memory, CPU, network or storage. In modern IT, where cost cuts are the norm, addressing resource contention is a top priority. The main concern with resource contention is the performance degradation that occurs as a result. [7][8]

Second Criteria is Scarcity of Resource which happens when there are limited resources and the demand for resources is high. In such situation user can not avail facility of resource.

Third criteria are Resource Fragmentation –In these criteria resources are isolated. There would be enough resources but cannot allocate it to the needed application due to fragmentation into small entities. If fragmentation is done into big entities then we can use it optimum.

Forth criteria is Over Provisioning - Over provisioning arises when the application gets surplus resources than the demanded one. Due to this investment high and revenue almost low

Fifth criteria is Under Provisioning , which occurs when the application is assigned with fewer numbers of resources than it demanded.

Cloud users estimate of resource demands to complete a job before the estimated time may lead to an over-provisioning of resources. Resource providers' allocation of resources may lead to an under-provisioning of resources. To overcome the above mentioned discrepancies, inputs needed from both cloud providers and users for a RAS. From the cloud user's angle, the application requirement and Service Level Agreement (SLA)are major inputs to RAS. The offerings, resource status and available resources are the inputs required from the other side to manage and allocate resources to host applications by RAS.

The outcome of any optimal RAS must satisfy the parameters such as throughput, latency and response time. Even though cloud provides reliable resources, it also poses a crucial problem in allocating and managing resources dynamically across the applications. [4][9]

The request for virtualized resources is described through a set of parameters detailing the processing, memory and disk needs. Provisioning satisfies the request by mapping virtualized resources to physical ones.

The hardware and software resources are allocated to the cloud applications on-demand basis. For scalable computing, Virtual Machines are rented. [10][11]

2.0 Related Work

Very few literature is available on this survey paper in cloud computing paradigm. Shikharesh et al in paper describes the resource allocation challenges in clouds from the fundamental point of resource management. The paper has not addressed any specific resource allocation strategy. [12]

Patricia et al., investigates the uncertainties that increase difficulty in scheduling and matchmaking by considering some examples of recent research.

It is evident that the paper which analyzes various resource allocation strategies is not available so far. The proposed literature focuses on resource allocation strategies and its impacts on cloud users and cloud providers. It is believed that this survey would greatly benefit the cloud users and researchers. [9]

3.0 Resource Allocation Strategies at a Glance

The input parameters to RAS and the way of resource allocation vary based on the services, infrastructure and the nature of applications which demand resources. 0RESOURCE ALLOCATION STRATEGIES employed in cloud are as follows-

3.1 Electrocution time

Various of resource allocation mechanisms are proposed in cloud. Jiani at.al proposed , actual task execution time and preempt able scheduling is considered for resource allocation. It overcomes the problem of resource contention and increases resource utilization by using different modes of renting computing capacities.

But estimating the execution time for a job is a hard task for a user and errors are made very often. But the VM model considered in is heterogeneous and proposed for IaaS. Using the above-mentioned strategy, a resource allocation strategy for distributed environment is proposed by Jose et al. [13], [14] [15]

3.2 Policy

Since centralized user and resource management lacks in scalable management of users, resources and organization level security policy Decentralized user concept proposed by Dongwan et al. And virtualized resource management for IaaS by adding a new layer called domain in between the user and the virtualized resources.

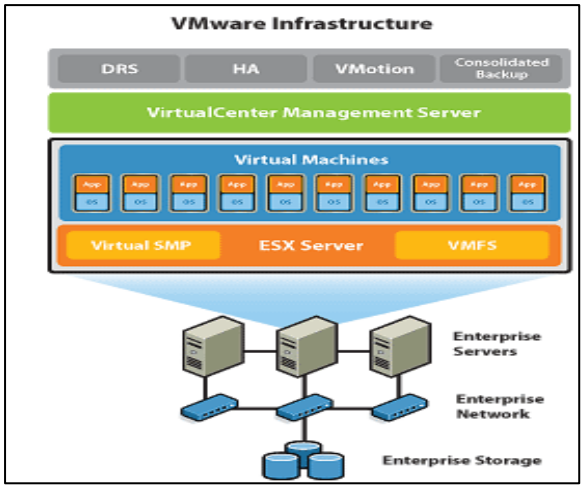
One another challenges for resource allocation is resource fragmentation in multi-cluster environment ,which is controlled by the work done by Kuo-Chan et al., which proposed to use the most fit processor policy for resource allocation. The mostfit processor policy assign a job to the cluster, which produces a left over processor distribution, leading to the most number of immediate subsequent job allocations. Observations shows that the most-fit processor policy has higher time complexities but the time overheads are negligible compared to the system long time operation. This policy is practical to use in a real system. [16] [17]

3.3 Virtual machine technology

A system knows as Virtual machine which can automatically scale its infrastructure resources. The system composed of a virtual network of virtual machines capable of live migration across multi-domain physical infrastructure. [18]

Virtualization is one of the main characters of cloud computing. The technology started being used on mainframes and has been recently adopted for other classes of servers. With system virtualization, a thin hypervisor layer, sometimes called virtual machine monitor (VMM), sits between the physical hardware resources and the operating systems. A system knows as Virtual machine which can automatically scale its infrastructure resources. The system composed of a virtual network of virtual machines capable of live migration across multi-domain physical infrastructure. [19] [20]

Virtualization has the power to divide a single physical computer into multiple logical computers, or virtual servers, each running its own guest operating system (GOS), as described in Fig.



The VMM role is to multiplex and arbitrate access to resources of the host platform so that they can be shared among multiple VMs.

Zhen Kong et al. have proposed a mechanism design to allocate virtualized resources among selfish VMs in a no cooperative cloud environment . By non-cooperative means, VMs care essentially about their own benefits without any consideration for others.

They have utilized stochastic approximation approach to model and analyze QoS performance under various virtual resource allocations. The proposed stochastic resource allocation and management approaches enforced the VMs to report their types truthfully and the virtual resources can be allocated efficiently.

3.4 Buzz

Cloud environment differs in terms of clusters, servers, nodes, their locality reference and capacity. The problem of resource management for a large-scale cloud environment is addressed in and general buzz protocol is proposed for fair allocation of CPU resources to clients.[21]

A buzz-based protocol for resource allocation in largescale cloud environments is proposed in [9]. It performs a key function within distributed middleware architecture for large clouds. Each node has a specific CPU capacity and memory capacity. The protocol implements a distributed scheme that allocates cloud resources to a set of applications that have time dependent memory demands and it dynamically maximizes a global cloud utility function.[22]

In the work by Paul et al. cloud resources are being allocated by obtaining resources from remote nodes when there is a change in user demand and has addressed three different policies to avoid over-provisioning and under provisioning of resources. Recent research on sky computing focuses on bridging multiple cloud providers using the resources as a single entity which would allow elastic site for leveraging resources from multiple cloud providers [23][24]

Related work is proposed in but it is considered only forpreemptable tasks. Yang et al. have proposed a profile based approach for scaling the applications automatically by capturing the experts' knowledge of scaling application servers as a profile. This approach greatly improves the system performance and resource utilization. Utility based RAS is also proposed for PaaS in. [24][25][26]

In Gossip based co-operative VM management with VM allocation and cost management is introduced. By this method, the organizations can cooperate to share the available resources to reduce the cost. Here the cloud environments of public and private clouds are considered. They have formulated an optimization model to obtain the optimal virtual machine allocation. Network game approach is adopted for the cooperative formation of organizations so that none of the organizations wants to deviate. This system does not consider the dynamic co-operative formation of organizations. Related work is discussed in [2] that use desktop cloud for better usage of computing resources due to the increase in average system utilization. The implication for a desktop cloud is that individual resource reallocation decisions using desktop consolidation and decision based on aggregate behavior of the system.[27][28]

3.5 Utility function

There are many proposals that dynamically manage VMs in IaaS by optimizing some objective function such as minimizing cost function, cost performance function and meeting QoS objectives. The objective function is defined as Utility property which is selected based on measures of response time, number of QoS, targets met and profit etc. There are few works that dynamically allocate CPU resources to meet QoS objectives by first allocating requests to high priority applications. The authors of the papers do not try to maximize the objectives. Hence the authors' Dorian et al. proposed Utility (profit) based resource allocation for VMs which use live VM migration (one physical machine to other) as a resource allocation mechanism . This controls the cost-performance trade-off by changing VM utilities or node costs. This work mainly focuses on scaling CPU resources in IaaS.

A few works ,that use live migration as a resource provisioning mechanism but all of them use policy based heuristic algorithm to live migrate VM which is difficult in the presence of conflicting goals.[29][30][31]

For multitier cloud computing systems (heterogeneous servers), resource allocation based on response time as a measure of utility function is proposed by considering CPU, memory and communication resources in . HadiGoudarzietal. characterized the servers based on their capacity of processing powers, memory usage and communication bandwidth.

For each tier, requests of the application are distributed among some of the available servers. Each available server is assigned to exactly one of these applications tiers i.e. Server can only serve the requests on that specified server. Each client request is dispatched to the server using queuing theory and this system meets the requirement of SLA such as response time and utility function based on its response time . [32]

It follows the heuristics called force-directed resource management for resource consolidation. But this system is acceptable only as long as the client behaviors remain stationary.

But the work proposed in considers the utility function as a measure of application satisfaction for specific resource allocation (CPU, RAM). The system of data center with single cluster is considered in that support heterogeneous applications and workloads including both enterprise online applications and CPU-intensive applications. [33]

The utility goal is computed by Local Decision Module(LDM) by taking current work load of the system. The LDMs interact with Global Decision Module (GDM) and that is the decision making entity within the autonomic control loop.

This system relies on a two-tier architecture and resource arbitration process that can be controlled through each application's weight and other factors. F. Hardware Resource Dependency In paper to improve the hardware utilization, Multiple Job Optimization (MJO) scheduler is proposed. Jobs could be classified by hardware-resource dependency such as Cupboard, Network I/O-bound, Disk I/O bound and memory bound. MJO scheduler can detect the type of jobs and parallel jobs of different categories. Based on the categories, resources are allocated. This system focuses only on CPU and I/O resource. [34]

Eucalyptus, Open Nebula and Nimbus are typical open source frame works for resource virtualization management. The common feature of these frameworks is to allocate virtual resources based on the available physical resources, expecting to form a virtualization resource pool decoupled with physical infrastructure.

Because of the complexity of virtualization technology, all these frameworks cannot support all the application modes.

The system called Vega Ling Cloud proposed in paper supports both virtual and physical resources leasing from a single point to support heterogeneous application modes on shared infrastructure. [35]

Cloud infrastructure refers to the physical and organizational structure needed for the operation of cloud. Many recent researches address the resource allocation strategies for different cloud environment. Xiaoping Wang et al. have discussed adaptive resource co-allocation approach based on CPU consumption amount . The stepwise resource co-allocation is done in three phases. The first phase determines the co-allocation scheme by considering the CPU consumption amount for each physical machine (PM).

The second phase determines whether to put applications on PM or not by using simulated annealing algorithm which tries to perturb the configuration solution by randomly changing one element. During phase 3, the exact CPU share that each VM Occupies is determined and it is optimized by the gradient climbing approach. This system mainly focuses on CPU and memory resources for co-allocation and does not considered the dynamic nature of resource request.[9]

HadiGoudarzi et al in paper proposed a RAS by categorizing the cluster in the system based on the number and type of computing, data storage and communication resources that they control. All of these resources are allocated within each server. The disk resource is allocated based on the constant need of the clients and other kind of resources in the servers and clusters are allocated using Generalized Processor Sharing (GPS).

This system performs distributed decision making to reduce the decision time by parallelizing the solution and used greedy algorithm to find the best initial solution.

The solution could be improved by changing resource allocation. But this system cannot handle large changes in the parameters which are used for finding the solution. [36]

3.6 Auction

Cloud resource allocation by auction mechanism is addressed by Wei-Yu Lin et al. in. The proposed mechanism is based on sealed-bid auction. The cloud service provider collects all the users' bids and determines the price. [36]

The resource is distributed to the first the highest bidders under the price of the (k+1) the highest bid. This system simplifies the cloud service provider decision rule and the clear cut allocation rule by reducing the resource problem into ordering problem. But this mechanism does not ensure profit maximization due to its truth telling property under constraints.

The aim of resource allocation strategy is to maximize the profits of both the customer agent and the resource agent in a large data center by balancing the demand and supply in the market. It is achieved by using market based resource allocation strategy in which equilibrium theory is introduced (RSA-M) [41]. RSA-M determines the number of fractions used by one VM and can be adjusted dynamically according to the varied resource requirement of the workloads. One type of resource is delegated to publish the resource's price by resource agent and the resource delegated by the customer agent participates in the market system to obtain the maximum benefit for the consumer. Market Economy Mechanism is responsible for balancing the resource supply and demand in the market system.

3.7 Application

Resource Allocation strategies are proposed based on the nature of the applications in .In the work by Tram et al., Virtual infrastructure allocation strategies are designed for workflow based applications where resources are allocated based on the workflow representation of the application. For work flow based applications, the application logic can be interpreted and exploited to produce an execution schedule estimate. This helps the user to estimate the exact amount of resources that will be consumed for each run of the application. Four strategies such as Naive, FIFO, Optimized and services group optimization are designed to allocate resources and schedule computing tasks. [37][38] Real time application which collects and analyzes real time data from external service or applications has a deadline for completing the task. This kind of application has a light weight web interface and resource intensive back end. To enable dynamic allocation of cloud resources for back-end mashups, a prototype system is implemented and evaluated for both static and adaptive allocation with a test bed cloud to allocate resources to the application. The system also accommodates new

requests despite a-priori undefined resource utilization requirements. This prototype works by monitoring the CPU usage of each virtual machine and adaptively invoking additional virtual machines as required by the system.[38]

David Irwin et al. have suggested the integration of high bandwidth radar sensor networks with computational and storage resources in the cloud to design end-to-end data intensive cloud systems. Their work provides a platform that supports a research on broad range of heterogeneous resources and overcomes the challenges of coordinated provisioning between sensors networks, network providers and cloud computing providers.

In that work, the resource allocation module divides the resource (CPU, Memory and DB replicas) allocation problem in two levels. The first level optimally splits the resources among the clients whereas the database replicas are expandable (dynamic) in the second level, based on the learned predictive model. It achieves optimal resource allocation in a dynamic and intelligent fashion. [39][40]

Popovivi et al. havemainly considered QoS parameters on the resource provider's side such as price and offered load. [41]

Moreover Lee at.al have addressed the problem ofprofit driven service request scheduling in cloud computing by considering the objectives of both parties such as service providers and consumers. [42] But the author Linlin Wu et al. have contributed to RAS by focusing on SLA driven user based QoS parameters to maximize the profit for SaaS providers.

The mappings of customer requests in to infrastructure level parameters and policies that minimize the cost by optimizing the resource allocation within a VM are also proposed in. Managing the computing resources for SaaS processes is challenging for SaaS providers.

Therefore a framework for resource management for SaaS providers to efficiently control the service levels of their users is contributed by Richard et al.

It can also scale SaaS provider application under various dynamic user arrivals/departures. All the above mentioned mainly focus on SaaS providers' benefits and significantly reduce resource waste and SLO violations. [43][44]

4.0 Advantages and Limitations

There are many benefits in resource allocation while using cloud computing irrespective of size of the organization and business markets. But there are some limitations as well, since it is an evolving technology. Let's have a comparative look at the advantages and limitations of resource allocation in cloud.

4.1 Advantages

The biggest benefit of resource allocation is that user neither has to install software nor hardware to access the applications, to develop the application and to host the application over the internet. The next major benefit is that there is no limitation of place and medium.

We can reach our applications and data anywhere in the world, on any system. The user does not need to expend on hardware and software systems. Cloud providers can share their resources over the internet during resource scarcity.

4.2 Limitations

Since users rent resources from remote servers for their purpose, they don't have control over their resources. Migration problem occurs, when the users wants to switch to some other provider for the better storage of their data. It's not easy to transfer huge data from one provider to the other. In public cloud, the clients' data can be susceptible to hacking or phishing attacks. Since the servers on cloud are interconnected, it is easy for malware to spread.

Peripheral devices like printers or scanners might not work with cloud. Many of them require software to be installed locally. Networked peripherals have lesser problems. More and deeper knowledge is required for allocating and managing resources in cloud, since all knowledge about the working of the cloud mainly depends upon the cloud service provider.

5.0 Conclusion

Cloud computing technology is increasingly being used in enterprises and business markets. In cloud paradigm, an effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers.

Some of the strategies are discussed above mainly focus mainly on CPU, memory resources but are lacking in some factors. The response time based on the different allocation of resources for different servers and the clusters is modeled and used in the profit optimization problem. This paper is based on a cloud computing architecture that was continuously monitored to check for proper resource utilization. From the above learning and experiments it can be concluded that cloud service is not only data storage and software provision, but it is about intelligent utilization of the available computing resources also. At the same time the service is on demand. Customers can use computation capacities on the cloud when needed, and release those upon completion of their tasks to prevent wastage. There are a number of ways from which the performance analysis can be improved. The resource utilization analysis is based on memory usage. The metrics that identify CPU utilization can be used to enhance the research on the basis of minimizing resource wastage. The research only concentrates on IaaS but PaaS and SaaS should also be considered for further researches. Research needs to be done on how two cloud architectures can be made interoperable.

References

[1] Maximizing Profit in Cloud Computing System via Resource Allocation Hadi Goudarzi and Massoud Pedram ,University of Southern California, Los Angeles, CA 90089 .

[2] Anshul Rai, Ranjita Bhagwan Saikat Guha, "Generalized Resource Allocation for the Cloud", Microsoft Research India

[3] Ioannis Kitsos,Antonis Papaioannou, Nikos Tsikoudis and Kostas Magoutis, " Adapting Data-Intensive Workloads to Generic Allocation Policies in Cloud Infrastructures", Institute of Computer Science (ICS) Foundation for Research and Technology Hellas (FORTH) Heraklion GR-70013, Greece

[4] V. Vinothina, Dr. R. Sridaran, Dr. Padmavathi Ganapathi, " A Survey on Resource Allocation Strategies in Cloud Computing", (IJACSA) International Journal of Advanced Computer Science and Applications, 3(6) 2012

[5] Ronak Patel, Sanjay Patel", Survey on Resource Allocation Strategies in Cloud Computing", International Journal of Engineering Research & Technology (IJERT) 2(2), February- 2013, ISSN: 2278-0181.

[6] V. Vinothina, Dr. R. Shridaran, and Dr. Padmavathi Ganpathi, "A survey on resource allocation strategies in cloud computing", International Journal of Advanced Computer Science and Applications, 3(6):97--104, 2012.

[7] http://apmdigest.com/best-practices-to-resolve-resource-contention-in-the-cloud

[8] http://www.intel.com/support/netport/sb/cs-015182.htm3

[9] Patricia Takako Endo et al. "Resource allocation for distributed cloud", Concept and Research challenges (IEEE, 2011), pp.42-46.

[10] http://www.systemsarchitecture.co.uk/server/

[11] A. Singh, M. Korupolu and D. Mohapatra, "Server-storage virtualization: Integration and Load balancing in data centers". In Proc.2008 ACM/IEEE conference on supercomputing (SC'08) pages 1-12, IEEE Press 2008.

[12] Shikharesh Majumdar: "Resource Management on cloud: Handling uncertainties in Parameters and Policies" (CSI communications, 2011, edn) pp.16-19.

[13] Jiyani et al.: Adaptive resource allocation for preemptable jobs in cloud systems (IEEE, 2010), pp.31-36.

[14] Jose Orlando Melendez & Shikharesh Majumdar, "Matchmaking with Limited knowledge of Resources on Clouds and Grids".

[15] Shikharesh Majumdar, "Resource Management on cloud: Handling uncertainties in Parameters and Policies" (CSI communications, 2011, edn) pp.16-19.

[16] Dongwan Shin and HakanAkkan, "Domain-based virtualized resource management in cloud computing"

[17] Kuo-Chan Huang & Kuan-Po Lai, "Processor Allocation policies for Reducing Resource fragmentation in Multi cluster Grid and Cloud Environments",(IEEE, 2010), pp.971-976.

[18] P. Ruth, J. Rhee, D. Xu, R. Kennell and S. Goasguen, "Autonomic Adaptation of virtual computational environments in a multi-domain infrastructure", IEEE International conference on Autonomic Computing, 2006, pp.5-14.

[19] M. Altino, Sampaio & G. Jorge Barbosa, "Dynamic Power- and Failure-Aware Cloud Resources Allocation for Sets of Independent Tasks", IEEE International Conference on Cloud Engineering, 2013.

[20] O. Niehorster, A. Brinkmann, G. Fels, J. Kruger, and J. Simon, "Enforcing SLAs in Scientific Clouds," Proc. 2010 IEEE International Conference on Cluster Computing (CLUSTER), pp. 178-187, doi: 10.1109/CLUSTER.2010.42

[21] Rerngvit Yanggratoke, Fetahi Wuhib and Rolf Stadler, "Gossip-based resource allocation for green computing in Large Cloud" 7th International conference on network and service management, Paris, France, 24-28 October, 2011.

[22] Fetahi Wuhib and Rolf Stadler, "Distributed monitoring and resource management for Large cloud environments", (IEEE, 2011), pp.970-975

[23] Paul Marshall, Kate Keahey & Tim Freeman: Elastic Site (IEEE, 2010), pp.43-52.

[24] P. Ruth, J. Rhee, D. Xu, R. Kennell and S. Goasguen, "Autonomic Adaptation of virtual computational environments in a multi-domain infrastructure", IEEE International conference on Autonomic Computing, 2006,pp.5-14

[25] Yang wt.al, "A profile based approach to Just in time scalability for cloud applications", IEEE international conference on cloud computing, 2009, pp 9-16.

[26] Hien et al.," Automatic virtual resource management for service hosting platforms", cloud'09, pp 1-8.

[27] Dusit Niyato, Zhu Kun and Ping Wang, "Cooperative Virtual Machine Management for Multi-Organization Cloud Computing Environment

[28] Andrzej Kochut et al., "Desktop Workload Study with Implications for Desktop Cloud Resource Optimization", 978-1-4244-6534-7/10 2010 IEEE.

[29] D. Gmach, J.Rolia and L.cherkasova, "Satisfying service level objectives in a self-managing resource pool, In Proc", Third IEEE international conference on self-adaptive and self organizing system.(SASO'09) pages 243-253.IEEE Press 2009

[30] X. Zhu et al. "Integrated capacity and workload management for the next generation data center", in proc.5thinternational conference on Automatic computing (ICAC'08), pages 172-181, IEEE Press 2008

[31] T. Wood et al., "Black Box and gray box strategies for virtual machine migration", In Proc 4th USENIX Symposium on Networked Systems Design and Implementation (NSDI 07),pages 229-242.

[32] Hadi Goudaezi and Massoud Pedram, Multidimensional SLA-based Resource Allocation for Multi-tier Cloud Computing Systems IEEE 4thInternational conference on Cloud computing 2011, pp.324-331

[33] Hien Nguyen et al., "SLA-aware Virtual Resource Management for Cloud Infrastructures", IEEE Ninth International Conference on Computer and Information Technology 2009, pp.357-362.

[34] Weisong Hu et al., "Multiple Job Optimization in Map Reduce for Heterogeneous Workloads", IEEE Sixth International Conference on Semantics, Knowledge and Grids 2010,pp.135-140.

[35] Xiaoyi Lu, Jian Lin, Li Zha and Zhiwei Xu, "Vega Ling Cloud: A Resource Single Leasing Point System to Support Heterogeneous Application Modes on Shared Infrastructure", (IEEE, 2011), pp.99-106.

[36] Wei-Yu Lin et al., "Dynamic Auction Mechanism for Cloud Resource Allocation", IEEE/ACM 10th International Conference on Cluster, Cloud and Grid Computing, pp.591-592, 2010

[37] Tram Truong Huu & John Montagnat, "Virtual Resource Allocations distribution on a cloud infrastructure" (IEEE, 2010), pp.612-617.

[38] Waheed Iqbal, Matthew N. Dailey, "Imran Ali and Paul Janecek & David Carrera, "Adaptive Resource Allocation for Back-end Mash up Applications on a heterogeneous private cloud".

[39] David Irwin, Prashant Shenoy, Emmanuel Cecchet and Michael Zink, "Resource Management in Data-Intensive Clouds, Opportunities and Challenges", This work is supported in part by NSF under grant number CNS-0834243.

[40] Pencheng Xiong, Yun Chi, Shenghuo Zhu, Hyun Jin Moon, Calton Pu & Hakan Hacigumus, "Intelligent Management Of Virtualized Resources for Database Systems in Cloud Environment", (IEEE,2011),pp.87-98.

[41] I. Popovici et al,"Profitable services in an uncertain world", in proceedings of the conference on supercomputing CSC2005

[42] Y. C Lee et.al,"Project driven service request scheduling in clouds", In proceedings of the international symposium on cluster & Grid Computing.(CC Grid 2010), Melbourne, Australia.

[43] Linlin Wu, Saurabh Kumar Garg and Raj kumar Buyya, "SLA –based Resource Allocation for SaaS Provides in Cloud Computing Environments" (IEEE, 2011), pp.195-204 .

[44] Richard T.B. Ma, Dah Ming Chiu and John C.S.Lui, Vishal Misra and Dan Rubenstein, "On Resource Management for Cloud users :a Generalized Kelly Mechanism Approach'