Survey on Dynamic resource allocation techniques for Overload avoidance and green cloud computing

Saima Israil[1], Dr. Rajeev Pandey[2], Uday Chaurasia[3]
[1](Student, M. Tech. (ME) Computer Science and Engineering, UIT, RGPV, Bhopal, India)
[2, 3](Assistant Professor, Department of Computer Science and Engineering, UIT, RGPV, Bhopal, India)

ABSTRACT—*Cloud Computing is a flourishing technology nowadays because of its scalability, flexibility, availability of resources and other features. Resource multiplexing is done through the virtualization technologyin cloud computing. Virtualization technology acts as a backbone for provisioning requirements of the cloud based solutions. At present, load balancing is one of the challenging issues in cloud computing environment. This issue arises due to massive consumer demands variety of services as per their dynamically changing requirements. So it becomes liabilityof cloud service provide to facilitate all the demanded services to the cloud consumers. However, due to the availability of finite resources, it is very challenging for cloud service providers to facilitate all the demanded services efficiently. From the cloud service provider's perspective, cloud resources must be allocated in a fair manner. This paper mainly addresses the existing techniques for resource allocation in cloud computing environment. It also focuses on the key issues, challenges of various resource allocation techniques.*

Keywords *– Cloud computing, Dynamic resource allocation, overload avoidance, green computing.*

I. INTRODUCTION

Cloud computing has become more and more popular with the widely deployment of several cloud infrastructures [1]. The underlying principle of cloud computing is to deliver the required services from shared hardware through virtualization technology. The goal of this computing model is to make a better use of distributed resources, put them together to make higher throughput and to handle large-scale computation problem efficiently and economically. Cloud computing can been broadly categorizedinto three levels of use model or cloud computing services.

Infrastructure-as-a-service (IaaS): Cloud computing replaces mainly computer hardware. Users of IaaS can manage to support operating systems andapplications, but don't desire to buy server, storage and networking hardware and a data centre to house the hardware. Examples of those providers are companies such as Amazon, ENKI, GoGrid[2].

Platform-as-a-service (PaaS): Cloud computing replaces an execution environment for a computer language by providing a system ready to execute the user's software. The user of PaaS is the programmer. Examples of those providers are companies such as Engine Yard or Google [3].

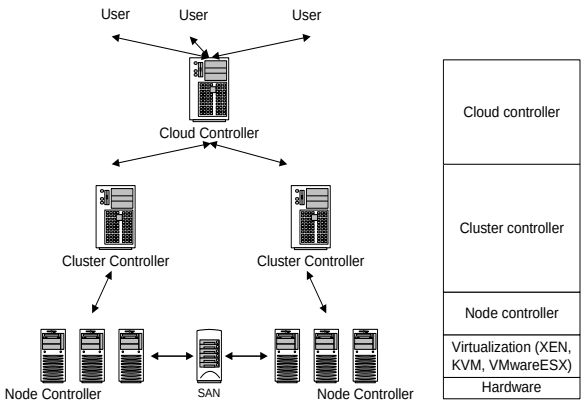

Figure 1.Classic IaaS cloud architecture

Software-as-a-Service (SaaS): The cloud user interacts directly with the Cloud software provided by CSP and often pays for usages only in place of computer time. Examples of those providers are NetSuite, SalesForce.com, Google Apps[4].
Typical architecture of an IaaS cloud is presented in Figure 1. Scope of this paper mainly focuses on the IaaS cloud. The IaaS cloud has various computing nodes grouped together to form clusters. For every node, there is an associated special purpose operating system called virtualization component. Its main function is to creates and maintains the VMs and further serves their requests for accessing to the required hardware resources.
The Node Controller (NC) executes on every node which hosts VM instances. NC further makes queries to discover the node's physical resources which include information about the number of cores, memory size,and available disk space. It also

gathers information about the state of VM instances on the node. The vital information congregated is further propagated up to the Cluster Controller.Cluster front-end machine generally executes the Cluster Controller (CC). It has three principal functions which include issue running instances to specific NCs;controlling instances of virtual network overlay and gathering information about a set of node controllers.

Cloud Controller is the interface point between cloud used and cloud service providers. The cloud controller queries node managers for information about the resources. It usually makes resource allocation decisionsbased on gathered information and implements them by making requests to cluster controllers.

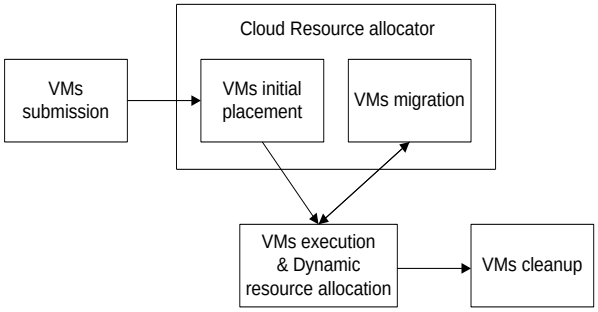

Figure 2.Resource allocation in VM life cycle

Resource allocation module is very important component of the IaaS cloud software stack. It assigns resources to virtual machines. Figure 2depicts the functionality of resource allocation in virtual machines life cycle. Once user submits request to the IaaS cloud system, the cloud resource allocation module will find the acceptable VMs and decides the initial places to run those virtual machines. While the VMs are in execution process, the cloud system may decide to migrate VMs from initial place to other computing nodes. The cloud resource allocation module identifies which nodes to migrate. While the node is executing virtual machines, the OS of the node may perform coarse-grained dynamic resource allocation to VMs [5].

II. OVERVIEW OF THE CLOUD RESOURCE ALLOCATION PROBLEM

In cloud computing environments, efficient resource provisioning and management is a challenging task for cloud service provider. Dynamically changing needs of the cloud users and the need to satisfy heterogeneous resource requirements further exacerbate the resource allocation management concern. In such a dynamic environments where cloud users can connect or disconnect at any time, there shall be provision by the cloud service provider to be able to makeaccurate decisions for scaling up or down its data-centers resources.

The resource allocation technique responds to find the resource allocation solution that satisfies a specific goal of the cloud service provider. While managing the resource numerous utility criteria shall be considered like delay of virtual resources setup, migration of existing processes, the resource utilization, energy consumption, overload avoidance, minimise resource wastage, ensuring service level agreement etc. So it become essential to allocate the resources appropriately but the static allocations have some constraints. Dynamic resource allocation can overcome these constraints [6]. Using the virtualization techniques, virtual machines can migrate to physical machines effectively [7].In order to resolve the resource allocation issue and to fulfil the could service provider and the end-users requirements, an efficient and dynamic resource allocation strategy becomes mandatory.
In view of essential characteristics of cloud and specific requirements for dynamic resource management, in this paper, we provide an overview of the recent research and practice advancement in cloud computing dynamic resource management.

III. RELATED WORK

According to our knowledge, we have not noticed any comprehensive journal article on IaaS cloud Dynamic resource allocation approaches. However, a number of related research papers and reviews that referred to IaaS cloud resource allocation have been published. In this section, we describe relative mechanisms and the methods which are implemented earlier and also the advantages and disadvantages of each method is described briefly.

Ying Song*et al.* [8] has proposed A two-tiered on-demand resource allocation mechanism, including the local and global resource allocation, based on a two-level control model. A well designed on demand resource allocation algorithm may minimize the waste of resources as well as guarantee the quality of the hosted applications. The local on-demand resource allocation on each server optimizes the resource allocation to VMs within a server taking the allocation threshold into account, while the global on-demand resource allocation optimizes the resource allocation among applications at the macro level by adjusting the allocation threshold of each local resource allocation.

A novel two-tiered on-demand resource allocation mechanism with feedback to optimize the resource allocation for VM-based data centres. In order to guide the design of the on-demand resource allocation algorithm, model the resource allocation using optimization theory. Base on the two-tiered on-demand resource allocation mechanism and model, local and global resource allocation algorithms to optimize the dynamic resource provision for VMs.

Christopher Clark*et al.* [9]Live OS migration is an extremely powerful tool for cluster administrators, allowing separation of hardware and software considerations, and consolidating clustered hardware into a single coherent management domain. If a physical machine needs to be removed from service an administrator may migrate OS instances including the applications that they are running to alternative machine(s), freeing the original machine for maintenance. Similarly, OS instances may be rearranged across machines in a cluster to relieve load on congested hosts. In the situations, the combination of virtualization and migration significantly improves manageability. Live migration refers to the process of making running virtual machines or applications between different physical machines without disconnecting the client or application. Memory, storage and network connectivity of the virtual machines are transferred from the original host machine to the destination. Migration processes have certain steps to perform.

Stage 0: Pre-Migration Begin with an active VM on physical host A. To speed any future migration, a target host may be preselected where the resources required to receive migration will be guaranteed.

Stage 1: Reservation A request is issued to migrate an from host A to host B. Initially confirm that the necessary resources are available on B and reserve a VM container of that size. Failure to secure resources here means that the VM simply continues to run on Aunaffected.

Stage 2: Iterative Pre -Copy During the first iteration, all pages are transferred from A to B. Subsequent iterations copy only those pages dirtied during the previous transfer phase.

Stage 3: Stop-and-Copy Suspend the running OS instance at A and redirect its network traffic to B. CPU state and any remaining inconsistent memory pages are then transferred. At the end of this stage there is a consistent suspended copy ofthe VM at both A and B. The copy at A is still considered to be primary and is resumed in case of failure.

Stage 4: Commitment Host B indicates to A that it has successfully received a consistent OS image. Host A acknowledges this message as commitment of the migration transaction: host A may now discard the original VM, and host B becomes the primary host.

Stage 5: Activation The migrated VM on B is now activated. Post-migration code runs to reattach device drivers to the new machine and advertise moved IP addresses.

Marvin McNett*et al.* [10] in his paper reveal that Usher provides a simple abstraction of a logical cluster of virtual machines, or virtual cluster. Usher users can create any number of virtual clusters of arbitrary size, whileUsher multiplexes individual virtual machines on available physical machine hardware.The Usher core implements basic virtual cluster and machine management mechanisms, such as creating, destroying, and migrating VMs. Usher clients use this core to manipulate virtual clusters. These clients serve as interfaces to the system for users as well as for use by higher-level cluster software. For example, an Usher client called ush provides an interactive command shell for users to interact with the system. And also implemented an adapter for a high-level execution management system, which operates as an Usher client that creates and manipulates virtual clusters on its own behalf. Two modules are there, first modules enable Usher to interact with broader site infrastructure, such as authentication, storage, and host address and naming services. Second, pluggable modules enable system administrators to express site-specific policies for the placement, scheduling, and use of VMs. As a result, Usherallows administrators to decide how to configure their virtual machine environments and determine the appropriate management policies.

On the other hand, Usher provides a framework that allows system administrators to express site-specific policies depending upon their needs and goals. By default, the Usher core provides, in essence, a general-purpose, best-effort computing environment. It imposes no restrictions on the number and kind of virtual clusters and machines, and performs simple load balancing across physical machines. Here believe this usage model is important because it is widely applicable and natural to use. Requiring users to explicitly specify their resource requirements for their needs, for example, can be awkward and challenging since users often do not know when or for how long they will need resources. Further, allocating and reserving resources can limit resource utilization; guaranteed resources that go idle cannot be used for other purposes. However, sites can specify more elaborate policies in Usher for controlling the placement, scheduling, and migration of VMs if desired. Such policies can range from batch schedulers to allocation of dedicated physical resources. Usher maintains a clean separation

between policy and mechanism. The Usher core provides a minimal set of mechanisms essential for virtual machine management. Usher provides a set of hooks to integrate with existing infrastructure. A Plugin API to enhance Usher functionality.

Xiaoyun Zhu*et al*. [11] AutoControl a resource control system that automatically adapts to dynamic changes in a shared virtualized infrastructure to achieve application SLOs. AutoControl is a combination of an online model estimator and a novel multi-input, multi-output resource controller. The model estimator captures the complex relationship between application performance and resource allocation, while the MIMO controller allocates the right amount of resources to achieve application SLOs. Virtualization is causing a disruptive change in enterprise data centres and giving rise to a new paradigm: shared virtualized infrastructure. In this new paradigm, multiple enterprise applications share dynamically allocated resources. These applications are also consolidated to reduce infrastructure and operating costs while simultaneously increasing resource utilization. As a result, data centre administrators are faced with growing challenges to meet service level objectives in the presence of dynamic resource sharing and unpredictable interactions across many applications. These challenges include:

Gong Chen*et al*. [12] Load skewing algorithms that allow significant amount of energy saving without sacrificing user experiences, i.e. maintaining very small number of SIDs. Understanding how power is consumed by connection servers provides insights on energy saving strategies. Connection servers are CPU, network, and memory intensive servers. There is almost no disk IO in normal operation, except occasional log writing. Since memory is typically preallocated to prevent run-time performance hit, the main contributor to the power consumption variations of a server is the CPU utilization .if pack connections and login requests to a portion of servers, and keep the rest of servers hibernating, here it can achieve significant power savings. However the consolidation of login requests results in high utilization of those servers, which may downgrade performance and user experiences. Hence, it is important to understand the user experience model before address the power saving schemes for large-scale Internet service.

T. R. Gopalkrishnan Nair*et al*. [13]presented a model, named as Ruled Based ResourceAllocation (RBRAM) which deals with the efficient resource utilization in M-P-S (Memory-Processor-Storage) Matrix Model. Authors say that resource allocation rate should be greater than resource request rate. Major components of the system are: cloud priority manager, cloud resource allocation,virtualization system manager and end result collection. To analyse the performance of the cloud system authors considered the Cloud Efficiency Factor. However, authors also identified other parameters of Cloud System for future work.

Justin Y. Shi*et al*. [14] explored a simple quantitative Timing Model method for cloud resource planning. For the same they considered the estimated resource usage times in steady state. Authors had calculated Speed up for Parallel Resource Planning based on Parallel Matrix Multiplication. To investigate multiple important dimensions of a program's scalability, authors proposed quantitative application dependent instrumentation method instead of qualitative performance models. Authors had mainly focused on application inter dependencies for cost effective processing.

Chu-Fu Wang*et al*. [15] in *"A PredictionBased Energy Conserving ResourcesAllocation Scheme for Cloud Computing*", hasdevelop an Energy Conserving Resource Allocation Scheme with Prediction (ECRASP) for cloud computing systems. The prediction mechanism can predict the trend of arriving jobs (dense or sparse) in the near future and their related features, so as with help the system to make adequate decisions. Simulation results show that our proposed ECRASP method performs well compared to conventional resource allocation algorithms in the energy conserving comparisons. The proposed method canarrange each arrival job to appropriate PMs and can make adequate decisions on when to shut down a PM or to start up a new PM to conserve power consumption.

Stefan Spitz*et al*. [16]authored paper in which he present approaches which improve current trust models according to the problems mentioned. Thereby, the degree of automation in the trust evaluation process increased. Finally, in combination with an adjusted trust level workflow, the presented approach allows an optimal resource allocation for grid or cloud computing service providers in combination with a trust model, a service provider can evaluate a resources' performance based on a set of trust and QoS requirements. As a result, trust relations can be established between a service provider and the associated resources. This allows the service provider to assign resources which are not only capable of processing a given task but also will most likely perform well. Future research in this field includes the identification ofadditional trust

aspects to further refine and strengthen the trust evaluation.

IV. COMPARISONS OF THE PREVAILING STRATEGIES

ExistingSystem	Methods	Advantages	Disadvantages
A Two Tired On-Demand Resource Allocation Mechanism for VMBased Data Centers.	Two Tiered allocation Mechanism 1)Local resource Scheduler RC 2)Global RC	It addresses the problems of availability and scalability. If global resource allocation failure occurs then the local resource allocation will work, vice versa.So no failure of resource allocation is occurred.	Application workload scheduling is not considered. Mismatch between the on demand resource and workload dispatch.
Qualitative timing Model for Resource Planning	Timing Model with Amazon EC	Use capacity measures to capture the quantitative dependencies between a computer programme and its processing environments It explores the multiple dimension of programmes quantitatively to gain non trivial insights	Mainly Focus on cost dependency parallel processing.
Usher:An Extensive Framework for Managing Clusters of Virtual Machines.	Usher Framework: Plugin API is for adding modules	Provide a best effort computingenvironment. Performs load balancing acrossphysical machines. Usher can be used for controllingthe placement scheduling, andmigration of VM's if desired.	For using the usher in another siteneeds to modify the existing plugin/ rewrite it. No plugin's for managing clusters of physicalmachines is written.
Rule Based Resource Allocation	RBRAM	Efficient resource utilization in M-P-S (Memory-Processor-Storage) Matrix Model.	Resource allocation rate should be greater than resource request rate. Performance of the cloud system is based on Cloud Efficiency Factor.
Automated Control of Multiple Virtualized Resource.	Auto Control: An automatic control system	Performance assurance: All Applications can be meets theirperformance. Without human intervention allocation decision should be made automatically. Various workloads can beadopted. Scalability can be achieved.	Auto Control only does not deal the bottleneck problems. It does not control any memory control.
Trust Based Resource Allocation and Evaluation of Workflows	Trust Level Workflow TLWF	Service provider can evaluate a resources performance based on a set of trust and QoS requirements. Service provider to assign resources which can provide the service qualitatively	It doesn't include user opinion in trust establishment process.
Prediction Based Energy Conserving Resources Allocation Scheme	ECRASP: Energy Conserving Resource Allocation Scheme with Prediction	Exponential smoothing prediction to predict the status of the forthcoming jobs Better energy conservation as compared to conventional resource allocation.	Extent of energy conservation is dependent on accuracy of prediction. Frequent switching of servers.
Live Migration of Virtual machines.	Live migration: Migrating application into another system	It is extremely powerful tool for clusters administrators. It will spare the original machine for maintenance. Relieve the load on the congest hosts.	Sending of the VM's memory will consume the entire bandwidth. It only considers the live migration among the well-connected data centre.
Energy-Aware Server Provisioning and Load Dispatching for Connection-Intensive Internet Services.	Skewness algorithm	Load prediction will help to reduce the frequently turning on and off servers. Load prediction will reduce the power consumption. Load balancing is considered.	Sometimes load prediction may cause the failure. Migration of the load is not considered. Power is only considered as the parameter.

Table 1:Comparison of Resource Allocation Techniques

CONCLUSION

Cloud Computing is revolutionising the computing paradigm for delivering computing services. The success and beauty behind cloud computing is due to the shared resource through virtualization. However, due to the availability of finite resources, it is very important for cloud service providers to manage and assign all the required resources in time to cloud consumers as their requirements are changing dynamically. So in this paper, diverse techniques for ensuring optimized resource allocation in cloud computing environments have been surveyed and investigated. Many authors have proposed methods for dynamic resource allocation in cloud computing. Few of them have been compared with its merits and limitations. In brief, an efficient Resource Allocation Technique should adhere to achieve Quality of Service aware utilization of resources, cost reduction and energy consumption minimization. Interest of many authors now a days is oriented toward efficient dynamic resource allocation to achieve green cloud computing. The eventual objective of resource allocation in cloud computing is to optimize the profit for cloud service providers and minimize the cost for cloud consumers.

REFERENCES

[1] Rimal, B.P., Choi, E., Lumb, I., 2009, A Taxonomy and Survey of Cloud Computing Systems, Proceeding of the Fifth International Joint Conference on INC, IMS and IDC, pp. 44 – 51.

[2] http://aws.amazon.com/ec2/ ; http://www.enki.co/ ; http://www.gogrid.com/

[3] http://www.engineyard.com/products/cloud ; http://www.google.com/apps/intl/en/business/cloud.html

[4] http://www.netsuite.com/portal/home.shtml;http://www.salesforce.com; https://developers.google.com/

[5] Waldspurger, C. A., 2002, Memory Resource Management in VMware ESX Server, ACM SIGOPS Operating Systems Review - OSDI '02: Proceedings of the 5th symposium on Operating systems design and implementation, pp. 181-194.

[6] Zhen Xiao, Senior member, IEEE, weijia song and Qi chen "*Dynamic Resource allocation using Virtual Machines For Cloud Computing Environment* ," IEEE Transaction on parallel and distributed systems, vol.24, No.6 june 2013.

[7] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield, *Xen and the Art of Virtualization*," Proc. ACM Sy mp. Operating Systems Princip les Oct. 2003.

[8] Ying Song, Yuzhong Sun, Member, IEEE, and Weisong Shi, Senior Member, IEEE "*A Two-TieredOn-Demand Resource Allocation Mechanism for VMBased Data Centers",* IEEE t ransactions on services computing, vol. 6, no. 1, january-march 2013.

[9] Christopher Clark, Keir Fraser, Steven Hand, Jacob Gorm Hansen, Eric Jul, Christian Limpach, Ian Pratt, Andrew Warfield *"Live Migration of Virtual Machines",*University of Cambridge Computer Laboratory 15 JJ Thomson Avenue, Cambridge, UK.

[10] Marvin McNett, Diwaker Gupta, Amin Vahdat, and Geoffrey M. Voelker "*Usher: An Extensible Framework For Managing Clusters Of Virtual Machines",* Proceedings of Large Installation System Administration Conference 2007 pp. 167-181.

[11] PradeepPadala, Kai-Yuan Hou Kang G. Shin, Xiaoyun Zhu, Mustafa Uysal, Zhikui Wang, SharadSinghal, Arif Me rchant "*Automated Control of Multiple Virtualized Resources",* The University of Michigan, Hewlett Packard Laboratories.

[12] Gong Chen, Wenbo He, Jie Liu, SumanNath, Leonidas Rigas, Lin Xiao, Feng Zhao "*Energy-Aware Server Provisioning and Load Dispatching for Connection-Intensive Internet Services",*Dept. of Computer Science, University of Illinois, Urbana-Champaign, IL 61801.

[13] T.R. Gopalkrishnan Nair, Vaidehi M, "*Efficient Resource Arbitation And Allocation Stratargies In Cloud Computing Through Virtualization*" in Proceedings of IEEE CCIS2011, 978-1-61284-204-2/11.

[14] Justin Y. Shi, Moussa Taifi and Abdallah Khreishah, "*Resource Planning for Parallel Processing in the Cloud*" in IEEE International Conference on High Performance Computing and Communications, 978-0-7659-4538-7/11, Nov. 2011.

[15] Wang Chu-Fu, Wen-Yi, Hung, and Yang Chen-Shun, *"A Prediction Based Energy Conserving ResourcesAllocation Scheme for Cloud Computing*"2014 IEEE International Conference on Granular Computing (GrC), 978-1-4799-5464-3/14 ©2014 IEEE, pp.321-324.

[16] Stefan S, Patrick B , York T*"Trust-based Resource Allocation and Evaluation of Workflowsin Distributed Computing Environments",*2010 2nd International Conference on Software Technology and Engineering(ICSTE) 978-1-4244-8666-3/10, 2010 IEEE, pp.IV-372-76.