Succinct Observation on Resource Management In Cloud Environment

[1,]Akilandeshwari.K

*[1,]PG Scholar Dept.Of Computer Science Adhiyamaan College Of Engineering,
Hosur, Tamilnadu. India*

---**ABSTRACT** ---
Cloud computing is ubiquitous and promise a cost-effective realization of the utility computing principle, allowing users and providers easy access to resources in a self-service, pay-as-you-go fashion, thus decreasing cost for system administration and improving resource utilization and accounting. Hence there is no need for getting license for individual products .The dominance of this technology is variety of users can be simultaneously access the variety of cloud services, so it is necessary to provide the sufficient resources to the user in a well organized manner. However one of the major perils in cloud is related to optimizing the resource being allocated and other challenge of resource allocation are meeting customer demands and application requirements. Here in this paper, a survey is meted out on the realm of resource management strategies; this would be beneficial for both cloud users and researchers.

KEYWORDS: *Cloud Computing, Resource Allocation, Load balancing.*

Date of Submission: 2 December 2013   Date of Acceptance: 25 December 13

I.   INTRODUCTION

CLOUDs are dynamic environment that guarantee availability, reliability and related quality aspects through automated, elastic management of the hosted services-that is hardware and software services such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS) or Software as a Service(SaaS) [10].The automated management thereby aims at optimizing the overall resource utilization whilst maintaining the quality constraints. Distributed processing, parallel processing and grid computing together emerged as cloud computing. The basic principle of cloud computing is that user data is not stored locally but is stored in the data center of internet. So the user can access the data anywhere at any time by using API which is provided by the cloud. The most popular definition is probably the one provided by the National Institute of Standards and Technology (NIST) "Cloud computing is a model for enabling ubiquitous, convenient, on demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [8]. Resource allocation is a process of assigning the resource based on the application demands by using resource provisioning over the internet. Fig.1. shows three different perspectives on cloud, from the perspective of cloud user is obviously the customer or end-user that makes actual use of the resources / services offered. As clouds are economy driven, the main incentives for their uptake (and thus provisioning) are cost aspects, expressed through a wide range of factors, such as outsourcing of resources and management, higher availability and thus better service provisioning.
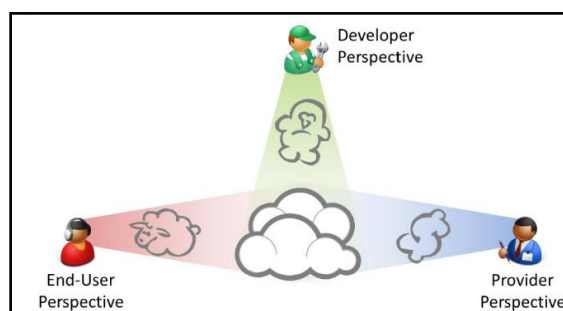

Fig. 1 Schematic on three different perspectives on cloud [8].

From the provider perspective it hosts the resources that the cloud offering (P/S/IaaS) is executed on. It's responsible for managing and providing the resources in a way that economic incentives are met. From the developer perspective it has to realize the capacities for the customer and the provider, the developers (and researchers) have to incorporate according features in their services, platforms and frameworks. Due to the limited resources and dynamic nature of resource, effective resource allocation is required. Cloud resources contain physical and virtual resources. Through the use of virtualization and provisioning [9] physical resources are mapped to the virtual machines. Fig.2. provides the mapping between virtual and physical resources. Virtualization is used to separate the OS and application from the hardware gives its ideal properties to best deliver these on-demand services. For scalability virtual machines are rented [11].
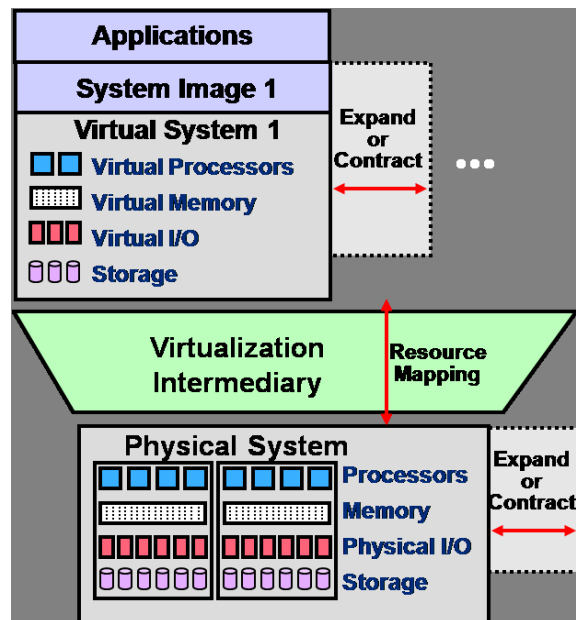

Fig. 2 Resource mapping between virtual and physical machines

Now a day's demands of resources are increasing in order to handle the growing amount of resources, users, but also scope of devices, use cases etc. some form of automated resource management is required, It reduces the overhead for the administrator to cater for this diversity and scale. This relates not only to the machines (respectively processing units) themselves, but also to the communication network and to handling the data. The amount of data consumed, produced and distributed increases constantly – way faster than the underlying network, thus leading to increasing challenges with respect to replicating and distributing data, localizing it, routing to it, compensating for bandwidth and latency limitations etc. Resource management must thereby respect multiple positions, i.e. not only does the own infrastructure grow and become more complex in management, but also the connection to the user, the user devices, the usage context etc. all vary increasingly and have to be handled by the providers.

We can therefore identify the following main sub-concerns without going into technical details:
- Efficient handling of an increasing amount of more and more heterogeneous resources.
- Efficient handling of an increasing amount of information and communication.
- Efficient handling & processing of large quantities of data.
- Interoperability and portability between the resources employed.

  To achieve this, it is necessary that resources, their usage etc. can be monitored and (dynamically) reconfigured – ideally autonomously.

  The rest of the paper is organized as follows: Various resource allocation strategies and their impacts in cloud environments are discussed in section II. Some of the advantages and restraints of resource allocation in cloud is specified in section III. In section IV it talks about the merits and demerits of the prevailing strategies. Finally the conclusion of the paper is given as section V.

II. RESOURCE ALLOCATION STRATEGIES & ALGORITHMS

Resource Allocation Strategy is all about integrating cloud provider activities for utilizing and allocating scarce resources in the cloud environment. In this section we are aiming to analyze some of the

resource allocation methods and load balancing techniques that are already present in the cloud environment in briefly.

A. Black-Box and Gray-Box Strategies (BG Algorithm)
Sandpiper [4] technique it assumes a large cluster of possibly heterogeneous servers. The hardware configuration of each server-its CPU, network interface, disk and memory characteristics, sandpiper system that automates the task of monitoring and detecting hotspots, determining a new mapping of physical to virtual resources and initiating the necessary migrations. Sandpiper implements a black-box approach that is fully OS and application-agnostic and a gray-box approach that exploits OS- and application-level statistics. Black-box techniques can make decisions by simply observing each virtual machine from the outside and without any knowledge of the application resident within each VM which is shown in Fig.3. It presents a Gray-box approach that assumes access to a small amount of OS-level statistics in addition to external observations to better inform the migration algorithm. Since a black-box approach is more general by virtue of being OS and application-agnostic, an important aspect is to understand if a black-box approach alone is sufficient and effective for hotspot detection and mitigation. But Sandpiper system supports black-box, gray-box, or combined techniques.
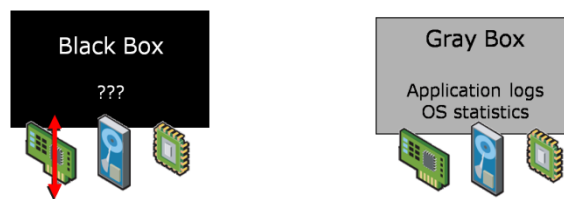

Fig. 3 Black-box and Gray-box approach

Sandpiper runs a nucleus on each physical server; the nucleus runs inside a special virtual server and is responsible for gathering resource usage statistics on that server (see Fig 4). It employs a monitoring engine that gathers processor, network interface and memory swap statistics for each virtual server. The nuclei periodically relay these statistics to the Sandpiper control plane. The control plane runs on a distinguished node and implements much of the intelligence in Sandpiper. It comprises three components: a profiling engine, a hotspot detector and a migration manager. The profiling engine uses the statistics from the nuclei to construct resource usage profiles for each virtual server and aggregate profiles for each physical server. The hotspot detector continuously monitors these usage profiles to detect hotspots, if it detect hotspot means it instruct migration manager to migrate the hotspot to balance load.
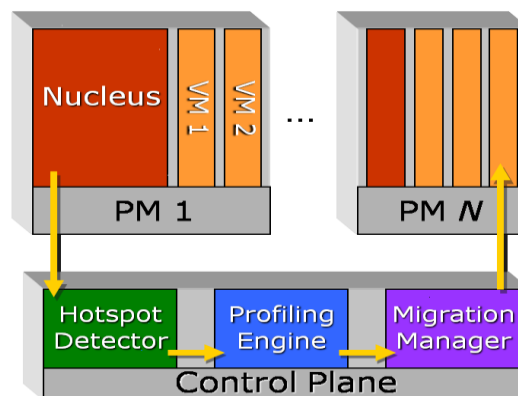

Fig. 4 Design of Sandpiper Architecture [4]

Black-box is possible to track network and disk I/O activity of various VMs by observing the driver activity in domain0. But tracking of memory is challenging which is limited to reactive. In the gray box it is feasible to gather OS and application level statistics. Sandpiper supports gray-box because it uses some light weight daemon installed inside the VM to gather the resource statistics, so it enables the proactive detection and mitigation of memory hotspots. Black-box has some limitation. But gray-box allows proactive decision this is the main advantage. Drawback of this paper is BG algorithm requires more number of migrations to resolve the hotspot.

B. Integrated Server-Storage Virtualization

According to the design in [6], Data centers form a key building block for new cloud computing. Agile data center is designed with integrated server and storage virtualization using load balancing algorithm called *Vector Dot* that handles the hierarchical and multi-dimensional resource constraints. HARMONY and *Vector Dot* through experiments on both a real data center setup as well as simulated large scale data center environments. HARMONY extracts an end-to-end view of the SAN including performance and usage characteristics. HARMONY consists of a *Configuration and Performance Manager*, which is responsible to collect the configuration and performance information from individual server and storage management, server virtualization management and storage virtualization management components. Once a load is identified in the collected data, HARMONY initiates the optimization planning component – *Vector Dot*.
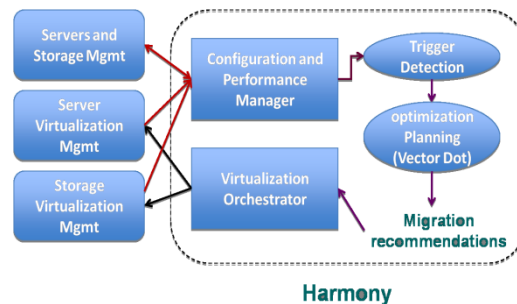

Fig. 5 Design Architecture of Harmony [6]

The latter generates recommendations for migrating one or more VMs or Vdisks to alleviate the hotspot. These are then processed by the v*irtualization Orchestrator* by using appropriate server and storage virtualization managers. The merit of this paper is *Vector Dot* scheme well address the complexity introduced by the data center topology and the multidimensional nature of the loads on resources. The drawback of this paper is uneven distribution of residual resource makes it hard to be fully utilized in the future and prediction is not convincing.

C. Automated Control of Multiple Virtualized Resources

AutoControl [5], a resource allocation system that automatically adapts to dynamic workload 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 (MIMO) resource controller. Main contributions of this autocontrol: First, online model estimator to dynamically determine and capture the relationship between application level performance and the allocation of individual resource shares. Second, two layered multi-input, multi-output (MIMO) controller to *automatically* allocate multiple types of resources to multiple enterprise applications to achieve their SLOs. The first layer consists of a set of application controllers that automatically determines the amount of resources necessary to achieve individual application SLOs, using the estimated models and a feedback approach. The second layer is comprised of a set of node controllers that detect resource bottlenecks on the shared nodes and properly allocate resources of multiple types to individual applications. In overload cases, the node controllers can provide service differentiation by prioritizing allocations among different applications. *AutoControl* can detect CPU and disk bottlenecks across multiple nodes and can adjust resource allocation to achieve end-to-end application level SLOs. Advantage of this strategy is it can enforce performance targets for different application-level metrics, including throughput and response time. It is mainly affected by time taken to collect statistics from clients.

D. Dynamic Resource Allocation using virtual machines

This system uses virtualization technology to allocate data center resources dynamically based on application demands. "Skewness" [7] concept is used to measure the unevenness in the multidimensional resource utilization of a server. By minimizing Skewness, we can combine different types of workloads nicely and improve the overall utilization of server resources. It support green computing [12] it is environmental responsible use of servers and also it reduces the resource consumption and saves energy.

[1]   Load Prediction: Load prediction algorithm that can capture the future resource usages of applications accurately without looking inside the VMs.
[2]   Usher [3]: Local node manager (LNM) on domain zero of Xen [1] which collects the usage statistics of resources for each VM on that node is forwarded to the Usher central controller (Usher CTRL) where our VM scheduler runs.
[3]   VM Scheduler: The scheduler has several components shown in Fig.6. The predictor predicts the future resource demands of VMs and the future load of PMs based on past statistics. The hot spot solver in our

VM Scheduler detects if the resource utilization of any PM is above the hot threshold (i.e., a hot spot). If so, some VMs running on them will be migrated [2] away to reduce their load. The cold spot solver checks if the average utilization of actively used PMs (APMs) is below the green computing threshold. If so, some of those PMs could potentially be turned off to save energy. It identifies the set of PMs whose utilization is below the cold threshold (i.e., cold spots) and then attempts to migrate away all their VMs. It then compiles a migration list of VMs and passes it to the Usher CTRL for execution.
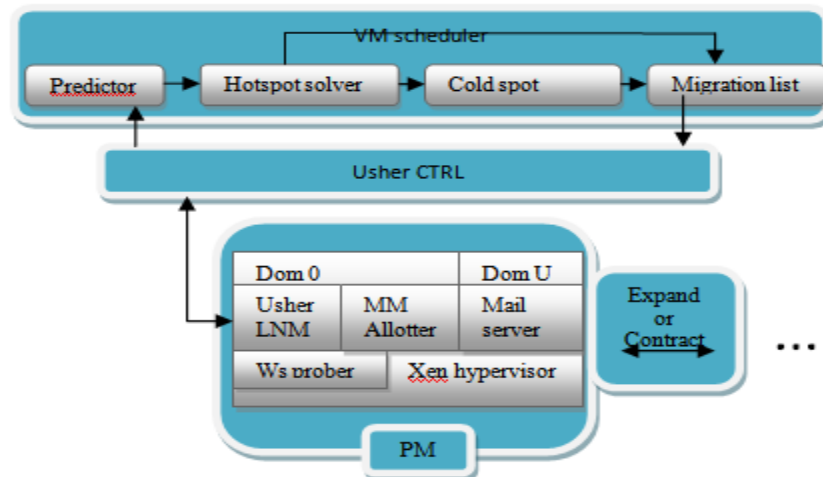

Fig. 6 Load balancing technique based on prediction and Skewness

Actually many of the loads balancing techniques failed because of resource contention when the available resource is not sufficient. But in Skewness it achieves overload avoidance because resources are monitored and dynamically reconfigured- ideally autonomously and this system multiplexes virtual to physical resources adaptively based on the changing demand. Compare to BG and VectorDot algorithm Skewness achieve overload avoidance. In cloud environment Skewness strategy and prediction it helps to overcome many limitations which we have discussed earlier section.

III ADVANTAGES AND RESTRAINTS
We have a comparative look at the advantages and restraints of resource allocation in cloud.

A. Advantages
- Cost reduction. Ability to increase or decrease the consumption of hardware or software resources immediately and in some cases automatically.
- Lower investment, reduced risk. Immediate access to the improvements in the resource proposed (hardware and software) and debugging.
- Greater security and accessibility. Access to resources from any geographical point and the ability to test and evaluate resources at no cost.
- Scalability. Users can set their own on how much resources are needed for example how much data space that he would need. Users can set how much data to be processed.
- Virtualization. Infrastructure devices can support each other and share among them. This increases the utilization of the device. The software can be moved from one server to another server.

B. Restraints
- *Dependence.* Migration of data from one application service providers to other service providers may be difficult; Resource usage information and other user activity rely heavily on service providers.
- *Very little control over assets.* Data audit and asset inspection is difficult because users don't have the actual information where the system located and how it operates.
- *Data security* becomes very complicated. Security systems are generally better but because the amount of data is very much the complexity of the problem is greatly increased data security. Users have little right to view the security log, especially if there is something odd.
- *Integration* to the cloud computing system using a standalone application from other service provider or third party tends to be very difficult. Several cloud providers offer their API but this is practically useful for big companies only.

IV MERITS AND DEMERITS OF THE PREVAILING STRATEGIES

Table 1. Comparison Between The Resource Allocation Strategies

Author	Method	Hypervisor	Controller	Merits	Demerits
T.Wood [4]	Black-box and Gray-box Strategy (BG Algorithm)	Xen	Nucleus and Monitoring engine	Gray-box enables proactive decision making.	1.Black-box limited to reactive decision making. 2. BG requires more number of migrations
A.Singh[6]	Integrated Server Storage Virtualization. (VectorDot Algorithm)	Xen	Configuration and Performance Manager	Less Complexity	1.Prediction is not convincing. 2. Uneven distribution of residual resource makes it hard to be fully utilized in the future.
P.Padala [5]	Automated Control of Multiple Virtualized Resources	Xen	AutoControl	Achieves end-to-end application level SLOs.	Time taken to collect statistics from clients.
Zhen Xiao [7]	Dynamic Resource Allocation(Skewness and Prediction Algorithm)	Xen	Usher	1. No overheads 2. High performance. 3. Requires less number of migrations. 4. Residual resource is friendly to virtual machines. 5. Improves Scheduling effectiveness.	Not cost effective.

V CONCLUSION

Cloud computing technology is increasingly being used in enterprises and business markets. In cloud archetype, an effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. This paper epitomizes the various resource allocation strategies and its impacts in cloud system. It tries to analyze the resource allocation strategies based on various strategies and it points out that some of the strategies are efficient than others in some aspects.

REFERENCES

[1] 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 Symp. Operating Systems Principles (SOSP '03), Oct. 2003.

[2] C. Clark, K. Fraser, S. Hand, J.G. Hansen, E. Jul, C. Limpach, I.Pratt, and A. Warfield, "Live Migration of Virtual Machines,"Proc. Symp. Networked Systems Design and Implementation (NSDI '05), May 2005.

[3] M. McNett, D. Gupta, A. Vahdat, and G.M. Voelker, "Usher: An Extensible Framework for Managing Clusters of Virtual Machines," Proc. Large Installation System Administration Conf. (LISA '07), Nov. 2007.

[4] T. Wood, P. Shenoy, A. Venkataramani, and M. Yousif, "Black-Box and Gray-Box Strategies for Virtual Machine Migration," Proc.Symp. Networked Systems Design and Implementation (NSDI '07), Apr. 2007.

[5] P. Padala, K.-Y. Hou, K.G. Shin, X. Zhu, M. Uysal, Z. Wang, S. Singhal, and A. Merchant, "Automated Control of Multiple Virtualized Resources," Proc. ACM European conf. Computer Systems (EuroSys '09), 2009.

[6]   A. Singh, M. Korupolu, and D. Mohapatra, "Server-Storage
         Virtualization: Integration and Load Balancing in Data Centers,"Proc. ACM/IEEE Conf. Supercomputing, 2008.
[7]   Zhen Xiao, Weijia Song and Qi Chen "Dynamic Resource Allocation Using Virtual Machines for Cloud Computing
         Environment", IEEE Transactions on Parallel and Distributed Systems, vol.24, June 2013.
[8]   Lutz schubert, Keith Jeffery "Advances in CLOUDS" the expert group report public version 1.0, European commission, 2012.
[9]   M.SuhailRehman, MajdF.Sakr, "Initial Findings for provisioning Variation in Cloud Computing" (IEEE, 2010), pp.473-479.
[10]   ZhixiongChen, JongP.Yoon,"International Conference on P2P, Parallel, Grid, Cloud and Internet Computing", 2010 IEEE: pp
         250-257.
[11]   Dorian Minarolli and Bernd Freisleben, "Uitlity–based Resource Allocations for virtual machines in cloud computing", IEEE,
         2011.
[12]   Saurabh Kumar Garg, Rajkumar Buyya"Green cloud computing and environmental sustainability" cloud computing and
         distributed laboratory, Oct 2010.