

# Dynamic Resource Allocation For Cloud Data Center Using Live Migration Of Virtual Machines

Sankar.B, Arunprakash.T

**Abstract**— Cloud computing permits business customers to scale up and down their asset usage based on desires. Numerous of the touted gains in the cloud model arrive from resource multiplexing through virtualization expertise. In this paper, we present a system that values virtualization expertise to allocate facts and figures center assets dynamically founded on application claims and support greencomputing by optimizing the number of servers in use. We insert the concept of “skewness” to measure the unevenness in the multidimensional asset utilization of a server. By minimizing skewness, we can combine distinct types of workloads nicely and improve the general utilization of server resources. We evolve a set of heuristics that prevent overload in the scheme effectively while saving power utilized. Trace propelled replication and trial outcomes demonstrate that our algorithm achieves good performance.

**Index Terms**— APM ( Active Physical Machine),DCLC (Delay Constrained Least Cost), IaaS (Infrastructure as a Service), PaaS (Platform as a Service), PM (Physical Machine).SaaS (Software as a Service), VM (Virtual Machine).

## 1 INTRODUCTION

THE elasticity and the lack of upfront capital investment offered by cloud computing is appealing to many businesses. There is a lot of discussion on the benefits and costs of the cloud model and on how to move legacy applications onto the cloud platform. Over provisioning for the peak demand .The cloud model is expected to make such practice unnecessary by offering automatic scale up and down in response to load variation. Besides reducing the hardware cost, it also saves on electricity which contributes to a significant portion of the operational expenses in large data centers. Virtual machine monitors (VMMs) like Xen provide a mechanism for mapping virtual machines (VMs) to physical resources .

This mapping is largely hidden from the cloud users. Users with the Amazon EC2 service[4], for example, do not know where their VM instances run. It is up to the cloud provider to make sure the underlying physical machines (PMs) have sufficient resources to meet their needs. VM live migration technology makes it possible to change the mapping between VMs and PMs while applications are running [5], [6].

However, a policy issue remains as how to decide the mapping adaptively so that the resource demands of VMs are met while the number of PMs used is minimized. This is challenging when the resource needs of VMs are heterogeneous due to the diverse set of applications they run and vary with time as the workloads grow and shrink.

The capacity of PMs can also be heterogeneous because multiple generations of hardware coexist in a data center. We aim to achieve two goals in our algorithm. The capacity of a PM transmission should be sufficient to satisfy the resource needs of all VMs running on it. Otherwise, the PM is overloaded and can lead to degraded performance of its VMs.

Green computing: The number of PMs used should be minimized as long as they can still satisfy the needs of all VMs. Idle PMs can be turned off to save energy. There is an inherent tradeoff between the two goals in the face of changing resource needs of VMs. For overload avoidance, we should keep the utilization of PMs low to reduce the possibility of overload in case the resource needs of VMs increase later. For green computing, we should keep the utilization of PMs reasonably high to make efficient use of their energy.

In this paper, we present the design and implementation of an automated resource management system that achieves a good balance between the two goals We introduce the concept of “skewness” to measure the uneven utilization of a server. By minimizing skewness, we can improve the overall utilization of servers in the face of multidimensional resource constraints.. We design a load prediction algorithm that can capture the future resource usages of applications accurately without looking inside the VMs. The algorithm can capture the rising trend of resource usage patterns and help reduce the placement churn significantly.

## 2 SYSTEM IMPLEMENTATION

### 2.1 Dynamic Resource Allocation

Resource Allocation (RA) is the process of assigning available

• Sankar.B is currently pursuing master degree program in computer science and engineering in Gnanamani college of technology,Anna university,India..

• Arunprakash.T ME,Assistant professor in Gnanamani college of technology, Anna university,India.

resources to the needed cloud applications over the internet. Resource allocation starves services if the allocation is not managed accurately. Resource provisioning solves that problem by allowing the service providers to manage the resources for each individual module.

## 2.2 Cloud Service Provider

The cloud service provider is responsible for maintaining an agreed-on level of service and provisions resources accordingly. A CSP, who has significant resources and expertise in building and managing distributed cloud storage servers, owns and operates live Cloud Computing systems. It is the central entity of cloud. 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 resource allocation.

## 2.3 Cloud Consumer

Cloud consumer represents a person or organization that maintains a business relationship with, and uses the service from, a cloud provider. Users, who stores data in the cloud and rely on the cloud for data computation, Cloud consists of both individual consumers and organizations. Cloud consumers use Service-Level Agreements (SLAs) for specifying the technical performance requirements to be fulfilled by a cloud provider.

## 2.4 Virtual Machine Environment

Virtualization provides an efficient solution to the objectives of the cloud computing paradigm by facilitating creation of Virtual Machines (VMs) over the underlying physical servers, leading to improved resource utilization. Virtualization refers to creating a virtual version of a device or a resource such as a server, a storage device, network or even operating system where the mechanism divides the resource into one or more execution environments.

- When a physical server is considered to be overloaded requiring live migration of one or more VMs from the physical server under consideration.
- Selection of VMs that should be migrated from an overloaded physical server. VM selection policy (algorithm) has to be applied to carry out the selection process.
- Finding a new placement of the VMs selected for migration from the overload and physical servers and finding the best physical.

## 2.5 Resource Manager

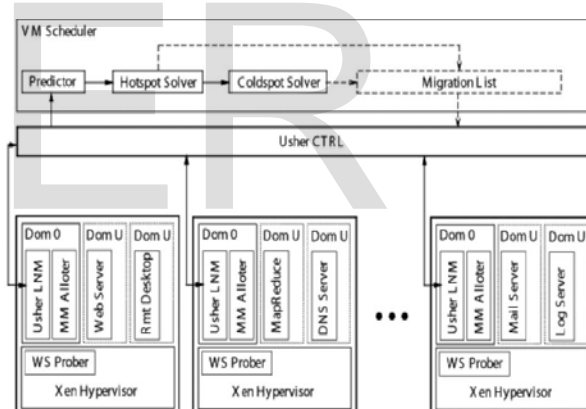
Service management in this context covers all the data center operations activities. This broad discipline considers the necessary techniques and tools for managing services by both cloud providers and the internal data center managers across these physical, IT and virtual environments. The availability of Service computing clouds gives researchers access to a large set of new resources for running complex scientific applications. However, exploiting cloud resources for large numbers of jobs requires significant effort and expertise.

## 2.6 Performance Evaluation

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 discussed above mainly focus on CPU, memory resources .secured optimal resource allocation algorithms and framework to strengthen the cloud computing paradigm.

## 3 SYSTEM DESIGN

The rest of the paper is organized as follows. Section 2 provides an overview of our system and Section 3 describes our algorithm to predict resource usage.



The architecture of the system is presented .Each PM runs the Xen hypervisor (VMM) which supports a privileged domain 0 and one or more domain U [3]. Each VM in domain U encapsulates one or more applications such as Web server, remote desktop, DNS, Mail, Map/Reduce, etc. We assume all PMs share a backend storage. The multiplexing of VMs to PMs is managed using the Usher framework [7]. The main logic of our system is implemented as a set of plug-ins to Usher. Each node runs an Usher local node manager (LNM) on domain 0 which collects the usage statistics of resources for each VM on that node.

The CPU and network usage can be calculated by monitoring the scheduling events in Xen. The memory usage within a VM, however, is not visible to the hypervisor. One approach is to infer memory shortage of a VM by observing its swap activities . Unfortunately, the guest OS is required to install a separate swap partition. Furthermore, it may be too late to adjust

the memory allocation by the time swapping occurs. Instead we implemented a working set prober (WS Prober) on each hypervisor to estimate the working set sizes of VMs running on it. We use the random page sampling technique as in the VMware ESX Server .

The statistics collected at each PM are forwarded to the Usher central controller (Usher CTRL) where our VMs scheduler runs. The VM Scheduler is invoked periodically and receives from the LNM the resource demand history of VMs, the capacity and the load history of PMs, and the current layout of VMs on PMs. The scheduler has several components. The predictor predicts the future resource demands of VMs and the future load of PMs based on past statistics. We compute the load of a PM by aggregating the resource usage of its VMs. The details of the load prediction algorithm will be described in the next section. The LNM at each node first attempts to satisfy the new demands locally by adjusting the resource allocation of VMs sharing the same VMM.

### 3.1 Predicting Future Resource Needs

We need to predict the future resource needs of VMs. As said earlier, our focus is on Internet applications. One solution is to look inside a VM for application level statistics, e.g., by parsing logs of pending requests. Doing so requires modification of the VM which may not always be possible. Instead, we make our prediction based on the past external behaviors of VMs.

Although seemingly satisfactory, this formula does not capture the rising trends of resource usage. For example, when we see a sequence of 10; 20; 30, and 40, it is reasonable to predict the next value to be 50. Fortunately, when  $\alpha$  is between 0 and 1, the predicted value is always between the historic value and the observed one. To reflect the "acceleration," we take an innovative approach by setting  $\alpha$  to a negative value. When  $\alpha < 0$ , the above formula can be transformed into the following: On the other hand, when the observed resource usage is going down, we want to be conservative in reducing estimation.

Hence, we use two parameters,  $\alpha$  and  $\beta$ , to control how quickly adapts to changes when  $\alpha$  is increasing or decreasing, respectively. We call this the Fast Up and Slow Down (FUSD) algorithm. Effectiveness of the FUSD algorithm for experience with traces collected for several Internet applications.) Now the predicted values are higher than the observed ones most of the time: 77 percent according. The median error is increased to 9.4 percent because we trade accuracy for safety. It is still quite acceptable nevertheless. So far we take  $\alpha$  as the last observed value.

Most applications have their SLOs specified in terms of a certain percentiles of requests meeting a specific performance level. More generally, we keep a window of  $W$  recently observed values and take  $\alpha$  as a high percentile of them.  $\alpha$  shows the result when  $W = \frac{1}{4} \cdot 8$  and we take the 90% the percentile of the

peak resource demand. The figure shows that the prediction gets substantially better. We have also investigated other prediction algorithms. Linear auto regression (AR) models, for example, are broadly adopted in load prediction by other. It models a predictive value as linear function of its past observations. Model parameters are determined by training with historical values. AR predictors are capable of incorporating the seasonal pattern of load change. For instance, the SPAR(4,2) [10] estimate the future logging rate of MSN clients from six past observations, two of which are the latest observations and the other four at the same time in the last four weeks.

### 3.2 The Skewness Algorithm

We introduce the concept of skewness to quantify the unevenness in the utilization of multiple resources on a server. Let  $n$  be the number of resources we consider and  $r$  be the utilization of the  $i$ 'th resource. We define the resource skewness of a server  $p$  as  $S_p$  where  $r$  is the average utilization of all resources for server  $p$ . In practice, not all types of resources are performance critical and hence we only need to consider bottleneck resources in the above calculation. By minimizing the skewness, we can combine different types of workloads nicely and improve the overall utilization of server resources. In the following, we describe the details of our algorithm.

**Hot and Cold Spots:** Our algorithm executes periodically to evaluate the resource allocation status based on the predicted future resource demands of VMs. We define a server as a hot spot if the utilization of any of its resources is above a hot threshold. This indicates that the server is overloaded and hence some VMs running on it should be migrated away. We define the temperature of a hot spot  $p$  as the square sum of its resource utilization beyond the hot threshold where  $R$  is the set of overloaded resources in server  $p$  and  $r$  is the hot threshold for resource  $r$ .

The temperature of a hot spot reflects its degree of overload. If a server is not a hot spot, its temperature is zero. We define a server as a cold spot if the utilizations of all its resources are below a cold threshold. This indicates that the server is mostly idle and a potential candidate to turn off to save energy. However, we do only when the average resource utilization of all actively used servers (i.e., APMs) in the system is below a green computing threshold. A server is actively used if it has at least one VM running. Otherwise, it is inactive. Finally, we define the warm threshold to be a level of resource utilization that is sufficiently high to justify having the server running but not so high as to risk becoming a hot spot in the face of temporary fluctuation of application resource demands. Different types of resources have different thresholds. For example, we can define the hot thresholds for CPU and memory resources to be 90 and 80 percent, respectively. Thus a server is a hot spot if either its CPU usage is above 90 percent or its memory usage is above 80 percent.

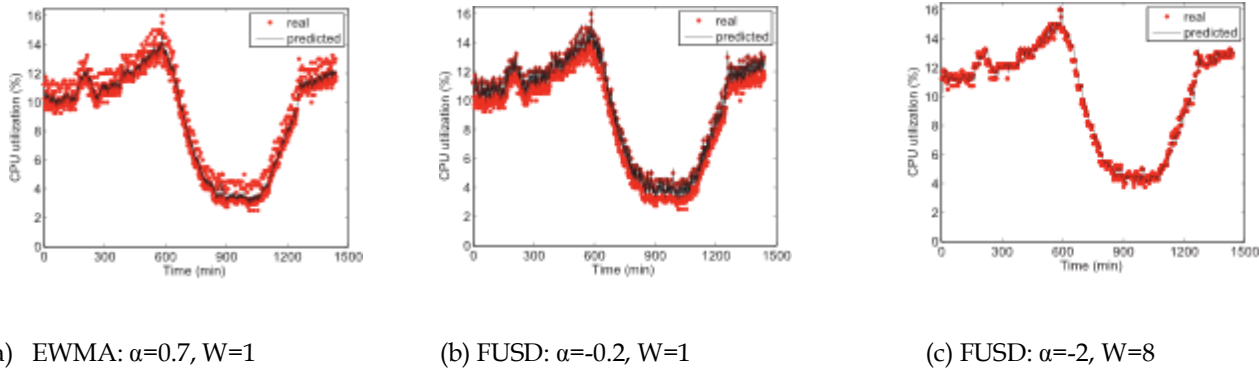


Fig: CPU load prediction for the DNS server. W is measurement window.

**Hot Spot Mitigation:** We sort the list of hot spots in the system in descending temperature (i.e., we handle the hottest one first). Our goal is to eliminate all hot spots if possible. Otherwise, keep their temperature as low as possible. For each server  $p$ , we first decide which of its VMs should be migrated away. We sort its list of VMs based on the resulting temperature of the server if that VM is migrated away. We aim to migrate away the VM that can reduce the server's temperature the most. In case of ties, we select the VM whose removal can reduce the skewness of the server the most. For each VM in the list, we see if we can find a destination server to accommodate it. The server must not become a hot spot after accepting this VM. Among all such servers, we select one whose skewness can be reduced the most by accepting this VM. Note that this reduction can be negative which means we select the server whose skewness increases the least.

### 3.3 Green Computing

When the resource utilization of active servers is too low some of them can be turned off to save energy. This is handled in our green computing algorithm. The challenge here is to reduce the number of active servers during low load without sacrificing performance either now or in the future. We need to avoid oscillation in the system. Our green computing algorithm is invoked when the average utilizations of all resources on active servers are below the green computing threshold sort the list of cold spots in the system based on the ascending order of their memory size. Since we need to migrate away all its VMs before we can shut down an underutilized server, we define the memory size of a cold spot as the aggregate memory size of all VMs running on it. Recall that our model assumes all VMs connect to a shared back-end storage. Hence, the cost of a VM live migration is determined mostly by its memory footprint explains why the memory is a good measure in depth. We try to eliminate the cold spot with the lowest cost first The above consolidation adds extra load onto the related servers. This is not as serious a problem as in the hot spot mitigation case because green computing is initiated only when the load in the system is low. Nevertheless, we wan to bound

the extra load due to server consolidation. Were strict the number of cold spots that can be eliminated in each run of the algorithm to be no more than a certain percentage of active servers in the system. This is called the consolidation limit.

**Consolidated Movements:** The movements generated in each step above are not executed until all steps have finished. The list of movements are then consolidated so that each VM is moved at most once to its final destination. For example, hot spot mitigate on may dictate a VM to move from PM A to PM B, while greencomputing dictates it to move from PM B to PM C. In the actual execution, the VM is moved from A to C directly.

### 3.4 Simulations

We evaluate the performance of our algorithm using tracedriven simulation. Note that our simulation uses the same code base for the algorithm as the real implementation in the experiments. This ensures the fidelity of our simulation results. Traces are per-minute server resource utilization, such as CPU rate, memory usage, and network traffic statistics, collected using tools like "perfmon" (Windows), the "/proc" file system (Linux), pmstat/vmstat/netstat" commands (Solaris), etc.. The raw traces are pre-processed into "Usher" format so that the simulator can read them. We collected the traces from a variety of sources.

**Web Info Mail:** The largest online Web archive in China (i.e., the counterpart of Internet Archive in the US) with more than three billion archived Web pages. Real Course. The largest online distance learning system in China with servers distributed across 13major cities. Amazing Store. The largest P2P storage system in China.

**Effect of Thresholds on APMS:** We first evaluate the effect of the various thresholds used in our algorithm. We simulate a system with 100 PMs and 1,000 VMs (selected randomly from the trace). We use random VM to PM mapping in the initial

layout. The scheduler is invoked once per minute. The bottom part show the daily load variation in the system. The x axis is the time of the day starting at 8 am. The y-axis is overloaded with two meanings: the percentage of the loader the percentage of APMs (i.e., Active PMs) in the system. Recall that a PM is active (i.e., an APM) if it has at least one VM running. As can be seen from the figure, the CPU load demonstrates diurnal patterns which decreases substantially after midnight. The memory consumption is fairly is table over the time.

### 3.5 Scalability Of The Algorithm

The algorithm rises with the system dimensions. The hasten of increase is between linear and quadratic. We break down the decision time into two components: warm spot mitigation (marked as "warm") and green computing (marked as "cold"). We find that warm spot mitigation contributes more to the conclusion time. We furthermore find that the conclusion time for the synthetic workload is higher than that for the genuine find due to the large variety in the synthetic workload. With 140 PMs and 1,400 VMs, the decision time is about 1.3 seconds for the synthetic workload and 0.2 second for the real find. Fig. 5b displays the average number of migrations in the whole scheme throughout each decision.

The number of migrations is little and rises approximately linearly with the scheme dimensions. We find that hot location contributes more to the number of migrations. We also find that the number of migrations in the synthetic workload is higher than that in the genuine find. With 140 PMs and 1,400 VMs, on average each run of our algorithm acquires about three migrations in the entire scheme for the synthetic workload and only 1.3 migrations for the genuine find. This is furthermore verified by Fig. 5c which computes the mean number of migrations per VM in each conclusion. The figure indicates that each VM experiences a tiny, roughly constant number of migrations throughout a decision run, unaligned of the system size.

This number is about 0.0022 for the synthetic workload and 0.0009 for the genuine find. This converts into roughly one migration per 456 or 1,174 decision gaps, respectively. The steadiness of our algorithm is very good. We also conduct simulations by varying the VM to PM ratio. With a higher VM to PM ratio, the burden is distributed more equally amidst the PMs. The outcomes are offered in part 4 of the supplementary document, which is accessible online.

## 4 IMPLEMENTATION

### 4.1 Effect Of Load Prediction

We contrast the execution of our algorithm with and without load proposition in Fig. 6. When burden proposition is hand-capped, the algorithm easily values the last observed burden in its conclusion making. Fig. 6a displays that burden proposition considerably decreases the mean number of hot spots in

the scheme throughout a decision run. especially, prediction stops over 46 percent hot spots in the replication with 1,400 VMs.

This illustrates its high effectiveness in preventing server overload proactively. Without prediction, the algorithm devours to consolidate a PM as shortly as its burden lets slip underneath the threshold. With proposition, the algorithm rightly foresees that the load of the PM will boost overhead the threshold soon and therefore takes no action. This departs the PM in the "cold location" state for a while. However, it furthermore reduces placement churns by bypassing pointless migrations due to temporary load fluctuation.

Consequently, the number of migrations in the system with burden prediction is lesser than that without prediction as shown in Fig. 6c. We can adjust the conservativeness of load proposition by tweaking its parameters, but the current configuration mostly serves our reason (i.e., mistake on the side of caution). The only downside of having more freezing spots in the scheme is that it may increase the number of APMs. This is investigated in Fig. 6b which displays that the mean figures of APMs remain vitally the same with or without burden prediction (the difference is less than 1 percent).

This is appealing because significant overload protection can be accomplished without forfeiting resources effectiveness. Fig. 6c compares the mean number of migrations per VM in each decision with and without burden proposition. It shows that each VM familiarity 17 per hundred less migrations with load prediction.

### 4.2 Resource Allocation At The Application Level

Automatic climbing of Web submissions was before studied in [14] and [15] for facts and figures center environments. In MUSE [14], each server has replicas of all web submissions running in the scheme. The dispatch algorithm in a frontend L7-switch makes certain demands are reasonably served while minimizing the number of underutilized servers. Work [15] values mesh flow algorithms to allocate the load of an submission amidst its running examples. For connection oriented Internet services like Windows Live Messenger, work [10] presents an integrated approach for burden dispatching and server provisioning. All works above do not use virtual machines and need the applications be organized in a multitier architecture with burden balancing supplied through an front-end dispatcher. In compare, our work goals Amazon EC2-style natural environment where it locations no restriction on what and how submissions are assembled inside the VMs.

A VM is treated like a blackbox. Resource management is done only at the granularity of entire VMs. Map Reduce [16] is another kind of popular Cloud service where data locality is the key to its presentation. Quincy [17] adopts min-cost flow form in task arranging to maximize facts and figures locality

while holding fairness amidst different jobs. The "Delay Scheduling" algorithm [18] trades execution time for facts and figures locality. Work [19] assign dynamic main concerns to jobs and users to facilitate asset allocation.

### 4.3 Resource Allocation By Live VM Migration

VM reside migration is a broadly utilized method for dynamic resource share in a virtualized environment [8], [12],[20]. Our work also pertains to this class. Sandpiper combines multidimensional burden information into a long Volume metric [8]. It kinds the list of PMs based on their volumes and the VMs in each PM in their volume-to-size ratio (VSR). This regrettably abstracts away critical data needed when making the migration decision. It then considers the PMs and the VMs in the presorted alignment. We give a solid example in Section 1 of the supplementary document, which is accessible online, where their algorithm chooses the incorrect VM to migrate away during overload and fails to mitigate the warm spot. We also contrast our algorithm and theirs in genuine trial.

The outcomes are analyzed in part 5 of the supplementary file, which is accessible online, to display how they behave distinctly. In supplement, their work has no support for green computing and differs from ours in many other facets such as load proposition. The HARMONY system concerns virtualization expertise over multiple resource layers [20]. It benefits VM and facts and figures migration to mitigate warm spots not just on the servers, but also on mesh apparatus and the storage nodes as well. It inserts the expanded Vector Product (EVP) as an sign of imbalance in asset utilization. Their load balancing algorithm is a variant of the Toyoda procedure [21] for multidimensional knapsack difficulty. different our system, their system does not support green computing and burden proposition is left as future work we investigate the phenomenon that Vector Dot behaves differently contrasted with our work and issue out the cause why our algorithm can utilize residual resources better.

Dynamic placement of virtual servers to minimize SLA violations is studied in [12]. They model it as a receptacle cramming difficulty and use the well-known first-fit approximation algorithm to calculate the VM to PM layout periodically. That algorithm, although, is designed mostly for offline use. It is expected to acquire a large number of migrations when directed in online natural environment where the resource needs of VMs change dynamically.

## 5 CONCLUSION

Virtualization, in computing is the creation of a virtual i.e., rather than actual version of a storage device or network resources. Using some interfaces we can access the data in cloud. This paper gives about the cloud data management interface by using storage virtualization mechanism. The open cloud

computing interface is an emerging standard for interoperable interface management in the cloud. Cloud computing can solve complex set of tasks in shorter time by proper resource utilization. The usage of cloud effectively with best resource allocation strategies have to be employed. Utilization of resources is one of the most significant task in cloud computing environment where user's jobs are scheduled to different machines. The various strategies have been studied and classified. The different features of the algorithms have been studied.

The development of the better allocation algorithm which is in heterogeneous and works in dynamic environment using virtual machines. In addition, for cloud computing to be used in a wide scale and really deliver on its promised benefits of elasticity, scalability, flexibility, and economies of scale, the focus of security needs to shift towards devising techniques to enable federation of security functions that are used today. For example, federation of audit, identity management, authentication, authorization, and incident response must all be explored in greater detail.

The focus of federation should be to enable a breadth of computing capabilities provided by multiple providers with different qualities of service to be consumed by customers with varying computing needs in a cohesive and secure fashion. Further, the federation should allow the cloud consumers to commission and decommission services from various CSPs with flexibility and agility. To further improve efficiency and application quality of service, we manage workloads by integrating the workload placement approach with a workload migration controller.

Finally, interest research problems will arise when we consider cloud computing security together with classical quality-of-serve issues and distributed computing issues in a network-wide scope where cloud (storage) systems are implemented in a distributed manner.

## 6 ACKNOWLEDGEMENTS

At the outset, we express our heartfelt gratitude to PAR-ENTS and GOD, who has been our strength to bring this project to light. I expand sincere thanks to our Head of the Department, Mr.S.K.MURUGARAJA and my Project Guide Mr.T.ARUNPRAKASH, GNANAMANI COLLEGE OF TECHNOLOGY, PACHAL for his kind guidance and encouragement during this project.

I would also like to express our thanks to all the Staff members of our department, friends who helped us directly and indirectly in all aspects of the project work to get completed successfully.

## REFERENCES

1. N. Bobroff, A. Kochut, and K. Beaty, "Dynamic placement of

virtual machines for managing SLA violations," in Proc. of the 10th IFIP/IEEE Intl. Symp. on Integrated Network Management (IM), 2007, pp. 119-128.

2. "C Strack - 2012 "Performance and Power Management for Cloud Infrastructures

3." X Meng - 2010" Efficient Resource Provisioning in Compute Clouds via VM Multiplexing"

4." A Iosup-2010" on the performance variability of production cloud services

5." C Tang-2007" A Scalable Application Placement Controller for Enterprise Data Centers

6. D. Gmach, J. Rolia, L. Cherkasova, G. Belrose, T. Turicchi, and A. Kemper, "An integrated approach to resource pool management: Policies, efficiency and quality metrics," in Proc. of the 38<sup>th</sup> IEEE Intl. Conf. on Dependable Systems and Networks (DSN), 2008, pp. 326-335.

7. D. Gmach, J. Rolia, L. Cherkasova, and A. Kemper, "Resource pool management: Reactive versus proactive or lets be friends," Computer Networks, vol. 53, no. 17, pp. 2905-2922, 2009.

8. S. Kumar, V. Talwar, V. Kumar, P. Ranganathan, and K. Schwan, "vManage: Loosely coupled platform and virtualization management in data centers," in Proc. of the 6th Intl. Conf. on Autonomic Computing (ICAC), 2009, pp. 127-136.

9. A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," Concurrency and Computation: Practice and Experience

(CCPE), 2012, DOI: 10.1002/cpe.1867, (in press).

10. J. Koomey, Growth in data center electricity use 2005 to 2010. Oakland, CA: Analytics Press, 2011.