Scientific Journal of Impact Factor (SJIF): 3.134

E-ISSN (O): 2348-4470
P-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 12,December -2015

# Survey on Dynamic Resource Allocation Strategies (RAS) & Contention Management in Cloud Computing Environment

Khushbu Maurya[1], Richa Sinha[2], Midhi Purohit[3]

[1]Gujarat Technological University, Computer Department,A.I.T. Ahmedabad
[2]Gujarat Technological University, Computer Department,K.I.R.C. Kalol,
[3]Indus University, Computer Department,Indus institute of technology & engineering, Ahmedabad,

**Abstract:** *Cloud computing, an amalgam of existing technologies ranging from distributed computing to cluster computing, to grid computing, to virtualization (which forms the foundation of these technologies) has changed the way organizations use Information and Communication Technology (ICT). Instead of acquiring resources for on-premise ICT departments, these resources are provisioned as service. It usually involves a pool of resources that multiple users can tap into and make use of (in parallel) whenever there is need to. These resources are also provisioned dynamically and are scaled up/down depending on demand. In addition, like any other utility, payment is done on a pay-per-use model thus reducing the huge initial cost of acquiring on-premise IT infrastructure. Since inception, there has been a steady increase in the number of users migrating to the clouds. Based on this increase, there is need to optimally allocate cloud resources so as to ensure that users perceived satisfaction is guaranteed. This work is an exposé on the challenges of resource allocation in cloud computing and works done in order to surmount these challenges. The work further goes on to juxtapose the various resource allocation strategies in order to identify their strengths and weakness based on how well they avoid situations such as resource under provisioning, over provisioning, contention, fragmentation and scarcity.*

**Key Words:** *Cloud Computing, Resource Allocation, Service Oriented Architecture, Quality of Service, Service Level Agreement, Contention Management.*

## I. INTRODUCTION

Cloud computing which evolved as an amalgamation of various existing technologies, virtualization being the key technology, enables users to have access to a pool of resources as a service. Other technologies that power this relatively new paradigm of computing include automated provisioning (servers have software installed automatically) and internet connectivity technologies to deliver the services [5][4][9].

This paradigm is gradually relinquishing control from in-house Information Technology infrastructures (necessary hardware and software installed within the premises) to service providers (that provision these hardware and software resources), enabling business organizations to focus more on core business activities and strategic development for their organizations. Furthermore, the cost of setting up a standard IT department is drastically reduced. Only the resources needed at any given time by Cloud computing users are received as services from the provider on a pay-as-you-go basis. This attribute makes Cloud computing elastic. According to a technical report published by the University of California, Berkley and [2] there is no precise definition for cloud computing. It is defined according to changes in services rendered by different organizations that provide the cloud solutions. The main reason for this indefinite description according [3] is as a result of statistical multiplexing of datacenters made possible with the efforts of researchers in areas such as Distributed Computing, Grid Computing, Web Technologies, Service Computing and Virtualization. Based on this, [28] concluded that the existence of different perceptions of Cloud Computing is that it is not a new technology, but rather a new model that brings together a set of existing technologies, identified in [3] to develop and run applications in a different way.

However, [4] defined the Cloud as “a large pool of easily usable and accessible virtualized resources (such as hardware, development platforms and/or services). These resources can be dynamically reconfigured to optimum resource utilization. This pool of resources is typically exploited by a pay-per-user model in which guarantees are offered by the Infrastructure Provider by means of customized SLA's”.

Despite the numerous advantages that the Cloud offers, business organizations hesitated and expressed a level of doubtfulness about adopting the services. This was because the thought of their data residing on someone else’ server heighted their fear concerning security of these data. Also, because resources in Cloud computing are obtained as a service, questions about Quality of Service arose. However recent studies [4] [5] [6] show that there has been a steady increase in Cloud adoption since 2009. According to a current research carried out by Right Scale, Infrastructure as a Service (IaaS) grew 45 percent from 2012 to 2013. Currently it is estimated to be a $15 billion business and it expected to grow to about $31 billion by 2015 while in 2012, Software as a Service (SaaS) was a $13 billion market, but is now predicted to grow to over $30 billion by 2016. In 2008, revenue from worldwide cloud services was $46.4 billion; in 2013, it is expected to reach $150 billion, a jump of just over 225 percent [8]

In addition, a survey carried out by Avanade, businesses have increased investments in resources to secure, manage, and support Cloud Computing.

## II. CHARACTERISTICS OF CLOUD COMPUTING

This definition implies that there are some key characteristics that are peculiar to Cloud Computing. The following are the characteristics:

- **Resource Pooling:** by using virtualized software model to share physical services, storage, and networking capabilities, service providers are able to create a pool of resources that multiple users can dynamically acquire for use and release when there is no longer need for it [9]. The user has no explicit knowledge of the physical location of the resources being used, except when the consumer requests to limit the physical location of his data to meet legal requirement.

- **Rapid Elasticity:** Cloud users are provisioned with the resources needed per time. Should there be a need for more resources at anytime, more resources are rapidly matched to cater for the rise in demand. Likewise, when there is reduction in the demand for resources, the excess resources are quickly released. This automated process decreases the procurement time for new computing capabilities when the need is there, while preventing an abundance of unused computing power when the need has [10].

- **On-demand self-service:** provisioning of resources to Cloud computing users is done without human interaction. This automated process reduces the personal overhead of the Cloud provider, cutting costs and lowering the price at which the services are offered [29].

- **Measured Service:** unlike traditional IT infrastructure that have inflexible computing capacities that are left unutilized when work load is not much and maxed out when work load is high, Cloud Computing automatically allows resources to be harnessed according the volume of work that is being done. This automated process decreases the procurement time for new computing capabilities when the need is there, while preventing an abundance of unused computing power when the need has subsided.

- **Broad network access:** services offered by Cloud Computing providers are done on-demand over the Internet with a plethora of devices that have thin client platforms.

## III. SERVICE DELIVERY METHODS IN CLOUD COMPUTING
## JJJ.

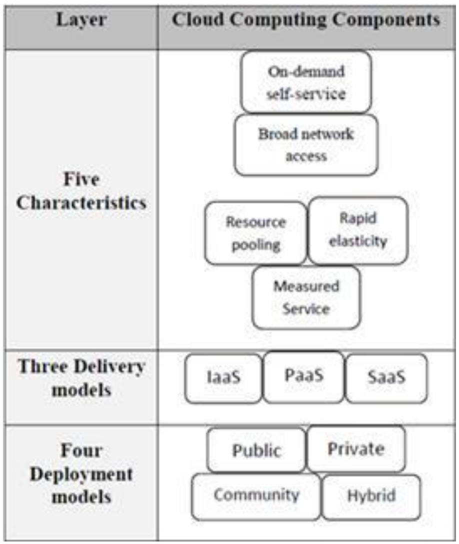
According to a general consensus [11] [4] [9] [3], the services offered in Cloud computing can be classified into the following depending on the actors involved and the services they provide:

- **Infrastructure as a Service (IaaS)** which provides virtual platforms, storage for servers, storage systems and data centers.
- **Platform as a Service (PaaS**) which provides a computing platform for design, development and testing applications.
- **Software as a Service (SaaS)** which provides application software on a pay-as-you-go basis over the Internet using a thin client.

## IV. SERVICE DEPLOYMENT MODELS IN CLOUD COMPUTING

Cloud computing service may be deployed in one of the following ways:

- **Private cloud:** this is the deployment of Cloud computing service for exclusive use by a single organization consisting of multiple users. The private Cloud may be owned and operated by the organization using it or a third party. It also could an amalgam of both.
- **Public cloud:** this refers to the deployment of cloud services on a fine-grained, self-service basis over the Internet, via web applications/web services, from an off-site third-party provider who shares resources and bills on a fine-grained utility computing basis for the consumption of the general public [2] [10].
- **Hybrid cloud:** this is a concatenation of two or more distinct cloud service deployment models.
- **Community cloud:** this is deployed to serve several organizations that have a common theme and share a common concern. It may be owned, organized, and functioned by one or more of the organizations in the community, a third party, or some blend of them, and it may exist on or off premises.



| Layer | Cloud Computing Components |
|---|---|
| Five Characteristics | On-demand self-service<br>Broad network access<br>Resource pooling<br>Rapid elasticity<br>Measured Service |
| Three Delivery models | IaaS<br>PaaS<br>SaaS |
| Four Deployment models | Public<br>Private<br>Community<br>Hybrid |

*Figure 1: Cloud Environment Architecture [12]*

## V. CLOUD COMPUTING SERVICE-ORIENTED ARCHITECTURE

Cloud resources can be seen as any physical or virtual resource that users may request from the Cloud. These include network requirements, storage, computational needs such as CPU time, or even software applications [14]. These resources are usually placed in multi-tenant data center that are able to match the resources with the volume of work being done at any point in time such that an expansion in business activities leads to more resources being provisioned and a contraction leads to less resources being provisioned. Cloud is defined as both the applications delivered as services over the Internet and the hardware and systems software in the data centers that provide those services [3]. According to this, delivery of application as services (SaaS - Software as a Service) over the Internet and hardware services (IaaS - Infrastructure as a Service) is both parts of cloud computing phenomena. From hardware service (utility computing) point of view, there are few new aspects in cloud [3] [14], the most prominent being the illusion of infinite computing resources and the ability to pay for use of computing resources on a short-term basis as needed.

Because all the resources in cloud computing are delivered as a service, it is a referred to as SOA [30]. As businesses and other organizations migrate from the traditional in-house IT infrastructure to SAO, the issues of quality of service and reliability become a challenge. This is because with in-house IT infrastructure, the resources can be easily monitored by in-house technical staff to ensure that the resources are reliable and that the quality of service is met. However, due to the dynamic nature of demand from users, it is not certain that service providers may be able to fully satisfy these demands. The service-oriented architecture or service-based systems pose a new challenge in the areas of quality, availability, usability, and reliability of services provided [31].

## VI. SERVICE LEVEL AGREEMENT (SLA)

Due to the complex nature of service demand from cloud users and the inability for service providers to totally satisfy the needs and fulfill their expectations. It is necessary for both Cloud service provider and consumers to come to a consensus of what the user expects and what the provider can offer. Thus the SLA provides a facility to agree upon QoS between an End-User and Provider and define End-User resource requirements and Provider guarantees, thus assuring an End-User that they are receiving the services they have paid for. Quality of Service in Distributed Systems is referred to as the resource reservation control mechanisms in place to guarantee a certain level of performance and availability of a service. QoS provides a level of assurance that the resource requirements of various web contents are strictly supported.

## VII. RESOURCE ALLOCATION

Resource allocation is the process of assigning available resources to complete cloud services optimally in an economic way. It could also be seen as any mechanism that aims to guarantee that the applications’ requirements as stated in the SLA are attended to correctly by the provider’s infrastructure. Resource allocation is defined as the process of integrating cloud provider activities for utilizing and allocation scarce resources, which may seem unlimited to users, within the limit of cloud environment so as to meet the needs of the cloud application in an elastic and transparent manner [13] [14] [15]. Resource allocation strategies help the two major players (users and service providers) in cloud computing to achieve their goals. Because of the Service-oriented nature of Cloud computing, users are concerned with quality and reliability, hence users may wish to estimate the resource demands to complete a job before the estimated time. This however could lead to the situation described as over-provisioning. On the other hand, providers wish to maximize their profit by using fewer resources per user in order to accommodate more users and make more profit. This will lead to under provisioning. However, it is difficult to allocate resources in a mutually optimal way due to the lack of information sharing between them. Moreover, ever-increasing heterogeneity, variability of the environment and uncertainty of resources in the node which cannot be satisfied with traditional resource allocation pose harder challenges for both parties [16]

Inputs from users and providers are put together in order to optimally allocate resources to prevent the problems of under/over provisioning of resources. From the users, application requirement and SLA are required while from the providers, offerings, available resources, status of the resources and SLA are required [17]

Requirements from both parties are put together in order to optimally allocate resources so as to satisfy various user requirements, a resource allocation strategy must circumvent the following scenarios as opined by [17] [16] [15] [34].

1. **Under provisioning of resources:** a situation where an application is assigned fewer resources that required satisfying the QoS requirements.
2. **Over provisioning of resources:** a situation where an application receives more resources than are needed to satisfy the QoS requirements.
3. **Resource contention:** a situation where multiple application try to access the same resource at the same time
4. **Resource starvation/scarcity of resources:** a situation where the available resources are limited while the demands on these resources are high.
5. **Resource fragmentation:** this occurs when resources are superfluous but cannot be used by applications that need them because they are not contagious.

## VIII. RESOURCES ALLOCATION STRATEGIES

Multiple services are often hosted by the same server in Service-Oriented Architecture and the services in the same server compete for limited available resources of the server which include CPU time and memory and network resources such as bandwidth. Different resource allocations will result in different QoS in runtime. It is important to note that the user may see those limited resources as unlimited and the tool that makes that possible is the Resource Allocation Strategy [14]. A variety of resource allocation strategies are examined.

### 1. Execution Time - Based RAS

To overcome the challenge of resource contention while allocating resources to tasks using the principle of parallel processing in the Cloud, [19] proposed the use of actual execution time to preemptively schedule how these resources are provisioned for effectiveness. This is done by adjusting the resource allocation adaptively based on the update of the actual task executions. However, there is a challenge in estimating the executing time for a job. This is usually because execution time allocations for jobs are over-estimated. A job with overestimated executing time may take less time than estimated or may get terminated before the preempted time expires thus leading to gross waste of resources. This may eventually lead to a degradation of the system performance because jobs that could have been attended to by idle resources are eventually turned away.

### 2. Just-In-Time (JiT) RAS

By incorporating Toyota's Just-in-Time philosophy, [20] were able to address the problems that arise from capacity planning in Cloud Computing. By incorporation this philosophy for the provision of cloud data centers, computational infrastructure of a cloud computing provider are assemble based on the costs that have already been absorbed by the core businesses that use them. Just in Time Clouds represents a new service category in which the provider allocates resources only when demanded and until there is use for them. Built upon the amortized resources from a supply chain, JiT Clouds may represent an attractive alternative for many types of clients and applications both in price and in scalability. Amortized resources are gotten as a result of a federation of low scale resources already existing.

Just in Time Provider is a public cloud computing provider that instead of assembling and maintaining a structure of data centers for supporting its own service makes use of a federation of low scale amortized resources already existing into private contexts. Unlike proxies of conventional providers of cloud computing, a Just in Time Provider does not represent any public cloud provider, but acts as a legitimate and fully autonomous provider that takes advantage of resources that would be irretrievably wasted without its intervention.

### 3. Linear Scheduling RAS

Scheduling of resources to tasks on an individual basis is usually associated with high waiting/response time. This challenge led to the formulation of linear scheduling resource allocation strategy [21] which focuses on distribution of resources among jobs that are capable of maximizing cost function as they arrive. This advantage of this strategy is that it improves throughput since jobs that can make effective use of the available resources are attended to. However this strategy is not suitable for real-time systems because jobs are not attended to on a "first come first serve" or "first in first out" basis.

**4. Policy Based RAS**

In order to resolve resource fragmentation in multi cluster grid and cloud computing, [22] proposed a resource allocation strategy based on the most-fit processor. This works by allocating a job to the cluster which produces a leftover processor distribution, leading to the most number of immediate subsequent job allocations. In order for this resource allocation strategy to function properly, a complex search process, involving simulated annealing activities is required to determine the target cluster. Under this resource allocation strategy the following are the basic assumption: the clusters are assumed to be homogenous and the clusters are assumed to be geographically distributed.

**5. VM-Based RAS**

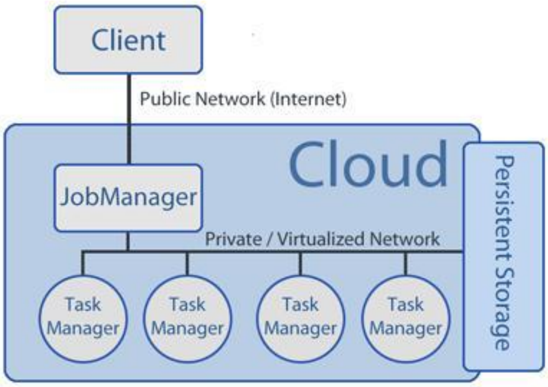


Figure 2: Nephele Architecture [23]

[23] Proposed a framework designed for efficient parallel data processing in the cloud. This framework is shown in figure 2.This allocation framework dynamically allocate/de-allocated different compute resources from a cloud in its scheduling and during job execution. Particular tasks of processing job can be assigned to different types of virtual machines which are automatically instantiated and terminated during the job execution.
Architecture of the Nephele framework

1 User starts a virtual machine (VM) in the Cloud which starts up the Job Manager (JM).
2 Users submits a Nephele compute job .
3 The JM accepts jobs from the client, schedules them, and coordinated their execution.
4 The JM is capable of communication with the cloud controller (the interface the cloud operator provides to control the instantiation of VMs.
5 The JM is able to allocate/de-allocate VMs according to the current job execution.
6 The actual task execution is done by the Task Manager (TM) that is instantiated by the Cloud controller.

A TM receives one or more tasks from the JM at a time, executes and after that informs the JM about the completion.

**6. Gossip Based RAS & Topology Aware Resource Allocation (TARA)**

A gossip-based protocol for resource allocation in large scale cloud environments is proposed in [9]. It performs a key function within distributed middleware architecture for large clouds. In the thesis, the system is modeled as a dynamic set of nodes that represents the machines of cloud environment. 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. The simulation results show that the protocol produces optimal allocation when memory demand is smaller than the available memory in the cloud and the quality of the allocation does not change with the number of applications and the number of machines. But this work requires additional functionalities to make resource allocation scheme is robust to machine failure which spans several clusters and datacenters.

**Topology Aware Resource Allocation (TARA)**

[24] proposed an architecture for optimized resource allocation in Infrastructure-as-a-Service-Based cloud systems that gathers information about hosted application requirements without the explicit user input. This information is used to forecast the performance of a particular resource allocation. This architecture is referred to as TARA and it is made up of a prediction engine that uses a lightweight simulator to estimate the performance

of a given resource allocation and search engine that makes use of genetic algorithm to the find an optimal solution in a large search space. Figure 3 gives the architecture of the TARA.

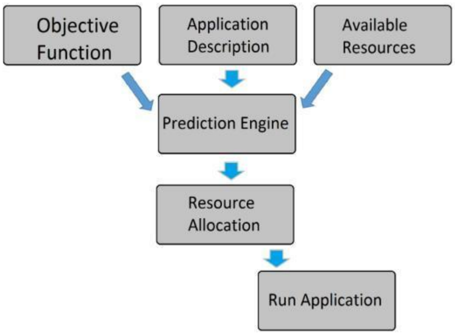


Figure 3 Architecture of TARA [24] Prediction Engine

The prediction engine combines objective functions, application description, and the available resources to determine how best to allocate resources. It maps resource allocation candidates to scores that measure their appropriateness with respect to a given objective function.

**Objective function**

This defines the metric that needs to be optimized. In TARA, the objective function uses MapReduce job completion time as the optimization metric for the fact that it indirectly maps the monetary cost of executing a job on an IaaS system. Application Description

This part of the TARA can be broken down into three sub sections: the framework type, the workload specific parameter, and a request for resources including Virtual Machines, storage, and bandwidth amongst many others.

### 7. Reservation Based RAS

Two provisioning plans for computing resources are majorly in use. Reservation plan and On-demand plan. [25] proposed the Robust Cloud Resource Provisioning (RCRP) algorithm to achieve the best advance reservation. The RCRP came as an improvement on the existing work of [26] that used demand and price to find an optimal solution for resource provisioning and VM placement. The RCRP considers four uncertainties (demand, profit, resource utilization and cost uncertainty) to get a more robust solution.

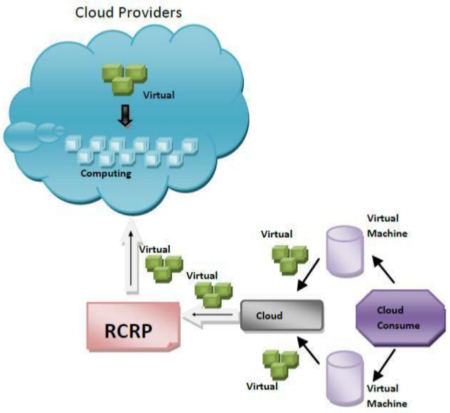


Figure 4: RCRP System Model for provisioning Cloud resources [25]

The model consists of four components: Cloud Providers, Cloud consumer, Virtual Machine repository, Cloud Agent. Job executions are requested from the service providers by the Cloud consumer. The Cloud agent traverses the network and gathers requirements from consumers and then allocates computing resources to jobs that have been requested by these consumers using the RCRP algorithm. The RCRP uses optimization technique to find the appropriate cloud provider. Optimization is done by calculating four uncertainty parameters viz., wait-time, idle-time, cost and profit.

- **Wait-time**- The time at which the users have to wait for before getting the requested resources allocated.
- **Idle-time-** The time at which the cloud consumer has to wait after allocating the requested VM to a particular loud provider.
- **Cost**- The amount of requested resources that cloud consumer has to pay.
- **Profit**- The benefit that a provider gets when allocating resources to cloud consumer.

## 8 Heuristic Based RAS

Based on the concept of virtualization, Cloud computing paradigm enables administrators and various organizations to freely migrate computing services between physical servers in cloud computing data centers. This is done by creating Virtual Machines over the underlying physical servers and it leads to better resource utilization and abstraction. VMs offer great benefits such as load balancing, server consolidation, online maintenance, and proactive tolerance. [27] however opined that the cost of VM migration requires thorough consideration. They further posited that each VM migration may result in SLA violation. Therefore it is pertinent to minimize the number of migrations to the barest minimum. In order to address that challenge, an algorithm was developed to heuristically place/replace VMs. The heuristic based VM migration scenario is partitioned as follows:

1. Determining when a physical server is considered to be overloaded requiring live migration of one or more VMs from the physical server under consideration
2. Determining when a physical server is considered as being under loaded hence it becomes a good candidate for hosting VMs that are being migrated from overloaded physical servers.
3. Selection of VMs that should be migrated from an over loaded physical server. VM selection policy (algorithm) has to be applied to carry out the selection process.
4. Finding a new placement of the VMs selected for migration from the overload and physical servers and finding the best physical server.

## 9 Priority Based Resource Allocation Model for Cloud Computing:

Cloud computing is a model which enables on demand network access to a shared pool computing resources. A cloud environment consists of multiple customers requesting for resources in a dynamic environment with possible constraints. In existing system cloud computing, allocating the resource efficiently is a challenging job. The cloud does not show the Qos, SLA.

This paper proposed allocates resource with minimum wastage and provides maximum profit. The developed resource allocation algorithm is based on different parameters like time, cost, No of processor request etc.

- Priority algorithm:
  Priority algorithm that mainly decides priority among different user request based on many parameters like cost of resource, time needed to access, task type, number of processors needed to run the job or task.
- Resource Allocation Model:
  In this model client send the request to the cloud server. The cloud service provider runs the task submitted by the client. The cloud admin decides the priority among the different users request.
  Each request consists of different task and it have the different parameters such as ,
  - **Time**- computation time needed to complete the particular task,
  - **Processor request**- refers to number of processors needed to run the task. More the number of processor, faster will be the completion of task.
  - **Importance**- refers to how important the user to a cloud administrator (admin) that is whether the user is old customer to cloud or new customer.
  - **Price**- refers to cost charged by cloud admin to cloud users.

## 10 A Dynamic Resource Allocation Methods for Parallel Data Processing

Nephele is the first data processing framework to explicitly exploit the dynamic and probably heterogeneous. In existing system the resource overload is high. The proposed system increases the efficacy of the scheduling algorithm for the real time cloud computing services. The algorithm utilizes the turnaround time utility efficiently by differentiating it into a gain function for a single task.

The algorithm assigns high priority for early completion task and less priority for abortions/deadlines. Cloud computing performance can be improved by,

- Associate each task with the time utility function (TUF).this is not important to measure the profit when completing a job in time but also account the penalty when a job is aborted or discarded.

In nephele architecture, the client submits the task job manager. Job manager allocate and deallocate VMs. VMS can be differentiated based on the instance type. For example,”m1.small” is a instance type means, it refers 1 cpu core, 1GB RAM, 128GB disk. The task manager receive tasks from the job manager at a time and decides how many and what type of instances job should be executed. The algorithm proves,

- Preemptive scheduling provides the maximum profit than the non-preemptive scheduling.
- Non-Preemptive scheduling provides the maximum penalty than the preemptive scheduling

## 11 Efficient Idle Desktop Consolidation with Partial VM Migration

Idle desktop systems are frequently left powered, often because of applications that maintain network presence. Idle PC consumes up to 60% of its peak power desktop VM often large requiring gigabytes of memory. These VM creates bulk transfer and utilize server memory inefficiently. In existing technique using the ballooning method, this not ensures the quick resume and provides the strain to the network.

Proposed system: Using the partial VM migration technique. This migrates only the working set of an idle VM.it allows user applications to maintain the network presence while the desktop sleeps and to transfer the execution of an Idle VM and it fetches the VM’s memory and disk state on –demand.

Partial migration leverages two insights:

- First, the working set of an idle VM is small, often more than an order of magnitude smaller than the total memory allocated to the VM.
- Second, rather than waiting until all state has been transferred to the server before going to sleep for long durations, the desktop can save energy by micro sleeping early and often, whenever the remote partial VM has no outstanding on-demand request for state.
- Working set migration: when consolidating a VM from the desktop to the server, partial VM migration transfers memory state only as the VM requires for its execution.

## **12** Adaptive Migration Threshold based Resource allocation

Fixed values for the thresholds are unsuitable for an environment with dynamic and unpredictable workloads [42][43], in which different types of applications can share a physical resource. The system should be able to automatically adjust its behavior depending on the workload patterns exhibited by the applications. Therefore, we propose a novel technique for auto-adjustment of the utilization thresholds based on a statistical analysis of the historical data collected during the lifetime of VMs.

Adaptive Migration Threshold based Resource allocation works as follows:

a. Determining when a host is considered as being overloaded requiring migration of one or more VMs from this host;
b. Determining when a host is considered as being under loaded leading to a decision to migrate all VMs from this host and switch the host to the sleep mode;
c. Selection of VMs that should be migrated from an overloaded host;
d. Finding a new placement of the VMs selected for migration from the overloaded and under loaded hosts.
e.

## IX. COMPARISON OF DIFFERENT RAS

| Sr. No. | Resource Allocation strategy | Parameter | | | | |
|---|---|---|---|---|---|---|
| | | Avoids Under Provisioning ? | Avoids Over Provisioning ? | Avoids Resource Contention? | Avoids Resource Scarcity? | Avoids Resource Fragmentation? |
| 1 | Execution time Based RAS | YES | NO | YES | NO | YES |
| 2 | Just in Time RAS | YES | YES | YES | NO | NO |
| 3 | Linear Scheduling RAS | NO | NO | NO | NO | NO |
| 4 | Policy Based/Most-fit Processor Policy | YES | NO | YES | NO | NO |
| 5 | VM based Nephele Architecture | YES | NO | YES | YES | YES |
| 6 | Topology Aware Resource Allocation(TARA) | YES | YES | YES | NO | NO |
| 7 | Robust Cloud Resource Provisioning(RCRP) | YES | NO | YES | YES | NO |
| 8 | Heuristic Based RAS | YES | YES | YES | NO | NO |
| 9 | Priority Based Policy | YES | YES | NO | NO | YES |
| 10 | Adaptive Migration Thresholds based Policy | YES | YES | YES | YES | YES |

## X. 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 t is an evolving technology. Let’s have a comparative look at the advantages and limitations of resource allocation in cloud.

**[A] Advantages:**

1) 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.
2) 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.
3) The user does not need to expend on hardware and software systems.
4) Cloud providers can share their resources over the internet during resource scarcity.

**[B] Limitations**

1) Since users rent resources from remote servers for their purpose, they don’t have control over their resources.
2) 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.
3) 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.
4) 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.
5) 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.

## XI. CONCLUSION

Cloud computing technology is increasingly being used in enterprises and business markets. In cloud paradigm, an effective resource allocation strategy & contention management is required for achieving user satisfaction and maximizing the profit for cloud service providers. With organizations and individuals migrating to the Cloud at an exponential rate, there is constant need to manage/provision the finite, available resources to requesting users such that Cloud providers are able to maximize their profit and at the same time the Quality of Service experienced by users are maintained according to that which is stated in the SLA. This paper basically considers the major resource allocation strategies currently being used and compares them using five parameters: ability to avoid under provisioning, ability to avoid over provisioning, ability to avoid resource contention, ability to avoid resource scarcity and ability to avoid resource fragmentation. Hence this paper summarizes the classification of RAS and its impacts in cloud system. Some of the strategies discussed above mainly focus on CPU, memory resources but are lacking in some factors. Hence this survey paper will hopefully motivate future researchers to come up with smarter and secured optimal resource allocation algorithms and framework to strengthen the cloud computing paradigm.

## REFERENCES

[1] Burford, D. (2010). *Cloud Computing: A Brief Introduction*. LAD Enterprizes Inc.

[2] Rimal, B. P., Choi, E., &Lumb, I. (2009). A Taxonomy and Survey of Cloud Computing Systems. *Fifth International Joint Conference on INC, IMS, and IDC.* IEEE Computer Society. doi:10.1109/NCM.2009.218

[3] Armbusrt, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinski, A., . . .Zaharia, M. (2009). *Above the Clouds: A Berkeley View of Cloud Computing.* Technical Report, University of California at Berkeley, Electrical Engineering and Computer Sciences, Bekeley.

[4] Vaquero, L. M., Rodero-Merino, L., Cacerces, J., &Maik, L. (2009, January). A Break in the Clouds: Towards a Cloud Definition. *ACM Gigcomm Computer Communication Review, 39*(1), 50-55.

[5] Angeles, S. (2014, February 01). *Cloud Computing's Sunny 2014 Forecast*. Retrieved from Business News Daily: http://www.businessnewsdaily.com/5863-cloud-trends-2014.html

[6] *Five Trends Transforming Cloud Computing*. (n.d.). Retrieved from Netsuite: http://www.netsuite.com/portal/resource/articles/cloud-computing-trends.shtml

[7] Weins, K. (2015, February 18). *Cloud Computing Trends: 2015 State of the Cloud Survey*. Retrieved April 2015, from Rightscale: http://www.rightscale.com/blog/cloud-industry-insights/cloud-computing-trends-2015-state-cloud-survey

[8] *Gartner Says Worldwide Cloud Services Revenue Will Grow 21.3 Percent in 2009*. (2009, March 6). Retrieved December 2014, from Gartner: http://www.gartner.com/newsroom/id/920712

[9] Mell, P. &Grance, T. (2009) The NIST Definition of Cloud Computing. Recommendation of the US Department of Commerce's National Institute of Standards and Technology (Draft), available: csrc.nist.gov/group/SNS/cloud-computing/cloud-def-v15.doc

[10] Kuyoro, S. O., Ibikunle, F., &Awodele, O. (2011). Cloud Computing Security Issues and Challenges. *International Journal of Computer Networks (IJCN), 3*(5), 247-255.

[11] Foster, I., Zhao, Y., Raicu, Ioan, & Lu, S. (2008). Cloud Computing and Grid Computing 360-Degree Comapred. *Grid Computing Environments Workshop* (pp. 1-10). Austin: IEEE. doi:10.1109/GCE.2008.4738445

[12] AlZain, M. A., Pardede, E., Soh, B., & Thom, J. A. (2012). Cloud Computing Security: From Single to Multi Clouds, System Science (HICSS). *45th Hawaii International Conference*, (pp. 5490-5499).

[13] Nair, T. R., &Vaidehi, M. (2011). Efficient Resource Arbitration and Allocation Strategies in Cloud Computing through Virtualization. *In the Proceedings of IEEE CCIS*, (pp. 397-401).

[14] Goncalves, G. E., Endo, P. T., Cordeiro, T. C., Palhares, A. V., Sadok, D., Kelner, J., . . . Mangs, J. (2011). Resource Allocation in Clouds: Concepts, Tools and Research Challenges. *SimpósioBrasileiro de Redes de Computadores e SistemasDistribuídos*, (pp. 197- 240).

[15] Patel, R., & Patel, S. (2013, February). Survey on Resource Allocation Strategies in Cloud Computing. *International Journal of Engineering Research and Technology (IJERT), 2*(2), 1-5.

[16] Asha, N., &Rao, G. R. (2013, July). A Review on Various Resource Allocation Strategies in Cloud Computing. *International Journal of Emerging Technology and Advanced Engineering, 3*(7), 177-183. Retrieved March 10, 2014

[17] Vinothina, V. S., &Ganapathi, P. (2012). A Survey of Resource Allocation Strategies in Cloud Computing. *International Journal of Advanced Computer Science and Application, 3*(6), 97-104.

[18] Awasare, V., &Deshmukh, S. (2014). Survey and Comparative Study in Resource Allocation Strategies in Cloud Computing Environment. *IOSR Journal of Computer Engineering, 16*(2), 94-101.

[19] Jiayin, L., Qiu, M., Niu, J., Chen, Y., & Ming, Z. (2010). Adaptive Resource Allocation for Preemptable Jobs in Cloud Systems. *In the Proceedings on the IEEE 10th Conference on Intelligent Systems Design and Applications.*

[20] Costa, R., Brasileiro, F., de Souza Filho, G. L., & Sousa, D. M. (2010). *Just in Time Clouds: Enabling Highly-Elastic Public Clouds over Low Scale Amortized Resources.*Universidade Federal de Campina Grande.

[21] Abirami, S. P., &Shalini, R. (2012). Linear Scheduling Strategy for Resource Allocation in Cloud Environment. *International Journal on Cloud Computing and Architecture, 2*(1).

[22] Huang, K., & Lai, K. (2010). Processor Allocation Policies for Reducing Resource Fragmentation in Multi Cluster Grid and Cloud Environment. *IEEE*, 971-976.

[23] Warneke, D., & Kao, O. (2011). Exploiting Dynamic Resource Allocation for Efficient Parallel Data Processing in the Cloud. *IEEE Transactions on Parallel and Distributed Systems*.

[24] Lee, G., Tolia, N. R., & Katz, R. H. (2010). Topology Aware Resource Allocation for Data-Intensive Workloads. *ACM SIGCOMM Computer Communication Review*, 120-124.

[25] Meera, L., & Mary, L. (2013). Effective Management of Resource Provisioning Cost in Cloud Computing.*International Journal of Advanced Research in Computer Science and Software Engineering, 3*(3), 75-78.

[26] Chaisiri, S., Lee, B. S., &Niyato, D. (2012). Optimization of Resource Provisioning Cost in Cloud. *IEEE Transactions on Services Computing, 5*(2).

[27] Ts'epoMofolo, &Suchithra, R. (2012). Heuristic Based Resource Allocation Strategies using Virtual Machine Migration. *International Refereed Journal of Engineering and Sciences (IRJES)*, 40-45.

[28] Zhang, Q., Cheng, L., &Boutaba, R. (2010). Cloud Computing: state-of-the-art and research challenges. Journal of Internet Services and Applications, 1(1), 7-18. doi:10.1007/s13174-010-0007-6

[29] Kuyoro, S. O., Omotunde, A. A., Ajaegbu, C., &Ibikunle, F. A. (2012). Towards Building a Secure Cloud Computing Environment.*International Journal of Advanced Research in Computer Science, 3*(4), 166-171.

[30] Rodero-Merino, L., Vaquero-Gonzalez, L.-M., Caceres-Exposito, & Juan-Jose, H.-S. (2010, August). SOA and Cloud Technologies, Two Pieces of the Same Puzzle. *Emergint Information Technologies (II), 11*(4), 25-29.

[31] Yau, S. S., &An, H. G. (2009). Adaptive Resource Allocation for Service-Based Systems.*International Journal of Software and Informatics, 3*(4), 483-499. Retrieved from http://www.ijsi.org/1673-7288/3/483.htm

[32] Armstrong, D., &Djemame, K. (2009). Towards Quality of Service in the Cloud. *Proceedings of the 25th UK Performance Engineering Workshop*, (pp. 226-240). Leeds, UK.

[33] Rasmi, K., &Vivek, V. (2013). Resource Management Techniques in Cloud Environment - A Brief Survey.*International Journal of Innovation and Applied Studies, 2*(4), 525-532.

[34] Sivapriyanka, D., &Santhanalakshmi, S. (2014, May). Allocation of Resources Dynamically in Cloud Systems.*International Journal of Engineering Research and Applications, 4*(5), 113-118. Retrieved July 24, 2014

[35] R. Buyya, CS Yeo,S. Venugopal, J. Broberg, I. Brandic, "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility, Future Generation Computer Systems, 2011

[36] Zhang bo, Gao ji, Ai Jieqing, Cloud Loading Balance Algorithm, Information Science and Engineering (ICISE), 2010 2nd International Conference on, 2010.

[37] Mauro Andreolini, Sara Casolari, Michele Colajanni, Dynamic load management of virtual machines in a cloud architecture, Department of Information Engineering, 2010.

[38] Daniel Versick and Djamshid Tavangarian, Reducing Energy Consumption by Load Aggregation with an Optimiazed Dynamic Live Migration of Virtual Machines, International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2010, pp. 164 { 170.

[39] Po-Cheng Chen, Cheng-I Lin, Sheng-Wei Huang, etc., A performance Study of Virtual Machine Migration vs Thread Migration for Grid Systenms, 22th Intenational Conference on Advanced information Networking and Applicayions (IEEE), 2008, pp. 86 { 91.

[40] Timothy Wood, K. K. Ramarkrishnan, CloudNet: Dynamic Pooling of Cloud Resources by Live WAN Migration of Virtual Machines, VEE '11 Proceedings of the 7th ACM SIGPLAN/SIGOPS international Conference on Virtual execution environments (ISBN), 2011, pp. 121 { 132.

[41] Chongguang REN, An Improved Adaptive Dynamic Programming Algorithm for Cloud Storage Resource Allocation, Journal of Computational Information Systems, 2011, pp. 5041 { 5048.

[42] A Beloglazov ,R Buyya, "Energy Efficient Resource Management in Virtualized Cloud Data Centers"10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, 2010 .

[43] A. Beloglazov, R. Buyya, "Energy efficient allocation of virtual machines in cloud data centers". 10th IEEE/ACM Intl. Symp. on Cluster, Cloud and Grid Computing ,2010.