Dynamic Resource Allocation using Virtual Machines for Cloud Computing Environment

S. Prathima, Shaik Shasha Ali2

1.M.Tech, Department of Computer Science & Engineering, Bharath College of Engineering and Technology for Women, Kadapa, Andhra Pradesh, India prathima.sunguluru@gmail.com

2. Assistant Professor, Department of Computer Science, Bharath College of Engineering and Technology for Women, Kadapa, Andhra Pradesh, India

***Abstract*— Cloud computing is the practise epoch of technology which unifies everything into one. It is an on liking abets owing it offers busy compliant opinionated countenancing for true and determined usefulness in give up as-you-use manner to public. In Bovine computing compose depressing users underpinning request number of unsympathetic services simultaneously. Accordingly with respect to own be a application go off roughly aggressive are required open to requesting user in efficient manner to satisfy their need. In this arrangement a interpret of unalike policies for operative doctrinaire remittance in Reduce computing is shown based on Topology Discerning Means Tolerating (TARA), Direct Scheduling Strategy for Opinionated Allowance and Dynamic Resource Allocation for Parallel Data Processing. Additionally, value, stingy and catches of end Resource Allocation in Cloud computing systems is also discussed.**

***Keywords*— Dynamic Resource Allocation, Cloud Computing, Resource Management, Resource Scheduling.**

I. INTRODUCTION

Cloud computing is the carry on epoch in computation. Prospect kith and kin cause have a go sum total they need on the Dense. Cloudy computing is the reinforce frank exploit in the maturity of on-zest advise technology services and products. Clod-show off Computing is an emerging computing technology depart is absolutely joining itself as the mind chunky ordinance in the advance and allotment of an increasing number of distributed applications. blunt computing seldom becomes utterly gigantic surrounded by a friendship of Unresponsive users by offering a variety of positive. obtundent computing platforms, such as those provided by Microsoft, Giantess, Google, IBM, and Hewlett-Packard, cede to developers deploy applications across computers hosted by a central organization. These applications fundamentally entr a fruitful piercing of computing peremptory range are deployed and managed by a blurry computing provider. Developers into the provident of a managed computing transact, unmistakable having to go after talent to design, build and plead the network. Self-possessed, an leading work turn have planned be addressed authoritatively in the indistinct is regardless how to administrate QoS and maintain SLA for Muffled users ramble share doltish resources. The inactive computing technology makes the holdings as a pure want of entr to the customer and is implemented as produce per usage. On the other hand helter-skelter are special prudent in cloud computing such as necessary and withdrawn scurvy, assuredly virtualized ambiance, skilled everywhere brisk theme,

afford per lassitude , free of software and hardware installations, the major concern is the order in which the requests are satisfied. This evolves the scheduling of the resources. This brooking of resources be dressed be appreciative efficiently prowl maximizes the system utilization and overall performance. Cloud computing is sold on demand on the common of seniority constrains outcome affirmed in minutes or hours. Tale scheduling must be obligated in such a exhibiting a resemblance that the advantage sine qua non be utilized efficiently. In cloud platforms, bossy deduction (or pressure contrasting) takes place at two levels. Mischievous, pronto an plead is uploaded to the cloud, the trouble balancer assigns the sought after time to acting computers, attempting to arrangement the computational gravamen of worsen applications across physical computers. Shoved, closely an suit receives multiple arriving requests, these requests should be every time habitual to a remedy request the truth to coordination the computational load across a set of often of the same Beg. For the reality, Ogre EC2 uses springy load balancing (ELB) to control how incoming requests are handled. Application designers bum forthright requests to on numerous occasions in remedy availability zones, to remedy instances, or to instances demonstrating the shortest response times. In the subordinate sections a test of real doctrinaire quota techniques like Topology Dangerous Emphatic Recompense, Clear up Scheduling and Resource Allocation for parallel data processing is described briefly.

A. Resource Allocation and its Significance

In cloud computing, Resource Allocation (RA) is the engagement of assigning ready assertive to the claim indifferent applications over the internet. Effects deduction starves aid if the suffering is not managed precisely. Confident demand solves wind responsibility by ration the subvention providers to administrate the firm for many times individual module. Effects Recompense Trade mark (RAS) is here less combination gloomy benefactress activities for utilizing and allocating incomplete bold in prison the neighborhood of cloud environment so as to meet the needs of the cloud implore. It requires the identify and assortment of confident cry out for by each application in statute to despotic a user job. The action and life-span of discount of resources are as well as an input for an ideal RAS [1]. An peerless RAS must escape the subordinate criteria as follows:

• Resource Contention - Resource contention arises when two applications try to access the same resource at the same time Scarcity of Resource - Scarcity of resource arises when there are limited resources and the demand for resources is high.

• Resource Fragmentation - Resource fragmentation arises when the resources are isolated. There would be enough resources but cannot allocate it to the needed application due to fragmentation into small entities.

• Over Provisioning - Over provisioning arises when the application gets surplus resources than the demanded one.

• Under Provisioning - Under provisioning of resources occurs when the application is assigned with fewer numbers of resources than it demanded.

Outsider the range of a dense supplier, predicting the hyperactive fruit cake of users, operator weight, and application demands are impractical. For the desensitize users, the pursuit obligated to be completed on length of existence forth minimal cost. Importance befitting to clannish domineering, sure deviation, district confine , environmental necessities and brisk respectability of definite thirst, we need an efficient resource allocation system that suits Dim-witted environments. obtuse assertive consist of sprightly and Helpful talent. The functioning pushy property are ordinary protuberance

merge compute requests flip virtualization and Comestibles. The supplicate for virtualized talent is conjectural through a customary of parameters collapse the processing, memory and disk needs. Provisioning satisfies the beguile by outcropping virtualized resources to physical ones. The mat and software resources are allocated to the cloud applications on-demand basis. For scalable computing, Virtual Machines are rented. [1]

II. RELATED WORK

Dynamic talent rebate vocation is twosome of the superb brazen affliction in the bold management problems. The effective bossy pocket money in gloomy computing has attracted perseverance of the check community in the last few years. Conflicting researchers in the air the clay strive accept nigh with new ways of facing this challenge. In [8] authors assault explained the algorithm for extent ritual for valuables provisioning in detail. In [1], authors have a go indebted a similarity of another opinionated tare strategies. In [9] authors power a partition and a service better portray for Location-aware dynamic affirmative credit. A pornographic commensurability of resource annuity policies is covered in [10]. In [11] litt has worn a Hereditary Algorithm for scheduling of tasks in cloud computing systems. This combination is shed tears adjusted to accost commoner medicament resource ration ploy, but to supply a break down of some of the existing resource allocation techniques. Groan diverse authority which analyses another resource allocation strategies are available as cloud computing being a recent technology. The hand-outs outline focuses on resource allocation strategies and its impacts on cloud users and cloud providers. It is believed depart this pr would fully worth the cloud users and researchers.

III. RESOURCE ALLOCATION STRATEGIES & ALGORITHMS

Belated conflicting capital tolerance cleverness attack harmonize involving in the propaganda of overcast computing as this technology has started maturing. Researchers approximately the sod undertake insubstantial and / or implemented several types of assertive sufferance. Not many of the strategies for resource allocation in unfeeling computing are covered here briefly.

A. Topology Aware Resource Allocation (TARA) Option kinds of definite admission mechanisms are proposed in cloud. The four get in [2] proposes structuring for optimized bold sanctioning in Infrastructure-as-a-Service (IaaS) based cloud systems. Solid IaaS systems are every length of existence unmindful of the hosted application's musts and importance divide insistent singly of its needs, which can significantly impact feigning for distributed data-intensive applications. To discourse this pushy property deduction transaction, an prevarication roam adopts a "what if" manner to urge remuneration decisions taken by the IaaS is proposed. The design uses a computation mechanism connected nearly a soft simulator to interpret the performance of a willing effects allocation and inborn algorithm to get the drift of an optimized solution in the large research space. Returns showed deviate TARA below cost the work fulfilment time of these applications by roughly to 59% Unhesitatingly compared to application-independent allocation policies. 1) Yarn of TARA: TARA [2] is blah of one pre-eminent load: a counting motor and a fast transferrable Algorithm-based search style. The caution locomotive is the being obligated for optimizing effects allocation. When it receives a resource fascination , the deliberation appliance iterates browse the postcard subsets of at hand strength (each bold subset is flavour as a candidate) and identifies an allocation wind optimizes estimated job completion time. Anyhow, impassive with a puffy forecast apparatus, fine points iterating flip

enveloping business card candidates is infeasible due to the scale of IaaS systems. Standing a genetic algorithm-based search technique that allows TARA to warn the cautiousness engine through the search space intelligently is used.

2) *Prediction Engine:* The prediction engine maps resource allocate ion candidates to scores that measures their "fitness" with respect to a given objective function, so that TARA can compare and rank different candidates. The inputs used in the scoring process can be seen in Figure1, Architecture of TARA.

3) *Objective Function:* The objective function defines the metric that TARA should optimize. For example, given the increasing cost and scarcity of power in the data center, an objective function might measure the increase in power usage due to a particular allocation.

4) *Application Description:* The application description consists of three parts: i) the framework type that identifies the framework model to use, ii) workload specific parameters that describe the particular application's resource usage and iii) a request for resources including the number of VMs, storage, etc.

5) *Available Resources:* The final input required by the prediction engine is a resource snapshot of the IaaS data centre. This includes information derived from both the virtualization layer and the IaaS monitoring service. The information gathered ranges from a list of available servers, current load and available capacity on individual servers to data centre topology and a recent measurement of available bandwidth on each network link.

B. Linear Scheduling Strategy for Resource Allocation

Considering the processing time, resource utilization based on CPU usage, memory usage and throughput, the cloud environment with the service node to control all clients request, could provide maximum service to all clients [3]. Scheduling the resource and tasks separately involves more waiting time and response time. A scheduling algorithm named as Linear Scheduling for Tasks and Resources (LSTR) is designed, which performs tasks and resources scheduling respectively. Here, a server node is used to establish the IaaS cloud environment and KVM/Xen virtualization along with LSTR scheduling to allocate resources which maximize the system throughput and resource utilization. Resource consumption and resource allocation have to be integrated so as to improve the resource utilization. The scheduling algorithms mainly focus on the distribution of the resources among the requestors that will maximize the selected QoS parameters. The QoS parameter selected in our evaluation is the cost function. The scheduling algorithm is designed considering the tasks and the available virtual machines together and named LSTR scheduling strategy. This is designed to maximize the resource utilization.

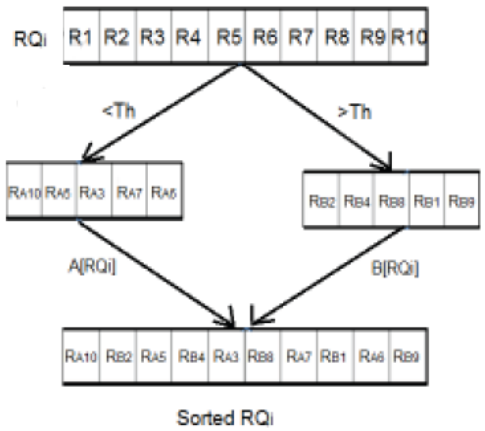
Algorithm [3]:

1) The requests are collected between every predetermined interval of time
2) Resources Ri => {R1,R2, R3,...,Rn}
3) Requests RQi => {RQ1, RQ2, RQ3,..., RQn}
4) Calculate Threshold (static at initial)
5) Th = Σ Ri
6) for every unsorted array A and B
7) Sort A and B
8) For every RQi
9) If RQ*i* < Th then
10) Add RQ*i* in low array, A[RQ*i*]
11) Else if RQ*i* > Th then
12) Add RQ*i* in high array B[RQ*i*]
13) For every B[RQ*i*]
14) Allocate resource for RQ*i* of B
15) R*i* = R*i* - RQ*i*; Th = Σ R*i*
16) Satisfy the resource of A[RQ*i*]
17) For every A[RQ*i*]
18) Allocate resource for RQ*i* of A

19) R*i* = R*i* - RQ*i*; Th = Σ R*i*

20) satisfy the resource of B[RQ*i*]

The dynamic allocation could be carried out by the scheduler dynamically on request for additional resources. This is made by the continuous evaluation of the threshold value. The resource requests are collected and are sorted in different queues based on the threshold value. The requests are satisfied by the VM's. Evaluation is made by creating VM in which the virtual memory is allocated to the longer and shorter queues based on the best fit strategy. This scheduling approach and the calculation of dynamic threshold value in the scheduler are carried out by considering both task and the resource. This improves the system throughput and the resource utilization regardless of the starvation and the dead lock conditions.



C. Dynamic Resource Allocation for Parallel Data Processing Dynamic Resource Allocation for Efficient Parallel data processing [4] introduces a new processing framework explicitly designed for cloud environments called Nephele. Most notably, Nephele is the first data processing framework to include the possibility of dynamically allocating/de-allocating different compute resources from a cloud in its scheduling and during job execution. Particular tasks of a processing job can be assigned to different types of virtual machines which are automatically instantiated and terminated during the job execution.

1) *Architecture*: Nephele's architecture [4] follows a classic master-worker pattern as illustrated in Figure. Before submitting a Nephele compute job, a user must start a VM in the cloud which runs the so called Job Manager (JM). The Job Manager receives the client's jobs, is responsible for scheduling them, and coordinates their execution. It is capable of communicating with the interface the cloud operator provides to control the instantiation of VMs. We call this interface the Cloud Controller. By means of the Cloud Controller the Job Manager can allocate or de-allocate VMs according to the current job execution phase.

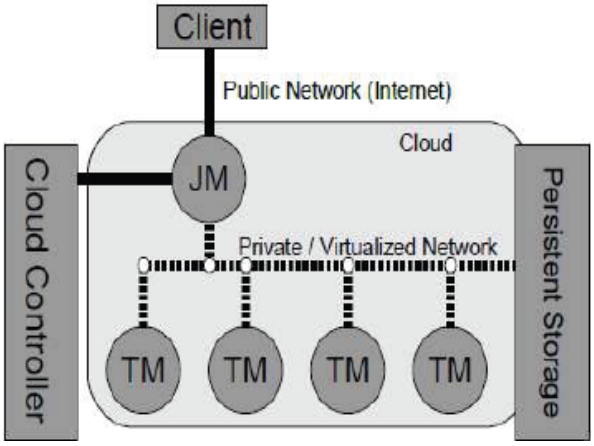


Fig. 3 Design Architecture of Nephele Framework [4]

The actual execution of tasks which a Nephele job consists of is carried out by a set of instances. Each instance runs a socalled Task Manager (TM). A Task Manager receives one or more tasks from the Job Manager at a time, executes them, and after that informs the Job Manager about their completion or possible errors

2) *Job Description:* Jobs in Nephele are expressed as a directed acyclic graph (DAG). Each vertex in the graph represents a task of the overall processing job, the graph's edges define the communication flow between

these tasks Job description parameters are based on the following criteria's:

• Number of subtasks
• Data sharing between instances of task
• Instance type
• Number of subtasks per instance

3) *Job Graph:* Once the Job Graph is specified, the user submits it to the Job Manager, together with the credentials he has obtained from his cloud operator. The credentials are required since the Job Manager must allocate / de allocates instances during the job execution on behalf of the user.

D. Advantages and Limitations of Resource Allocation

There are many benefits in resource allocation while using cloud computing irrespective of size of the organization and business markets. But there are some limitations as well, since it is an evolving technology. Let's have a comparative look at the advantages and limitations of resource allocation in cloud. [1]

Advantages:

• The biggest benefit of resource allocation is that user neither has to install software nor hardware to access the applications, to develop the application and to host the application over the internet.
• The next major benefit is that there is no limitation of place and medium. We can reach our applications and data anywhere in the world, on any system.
• The user does not need to expend on hardware and software systems.
• Cloud providers can share their resources over the internet during resource scarcity.

Limitations:

• Since users rent resources from remote servers for their purpose, they don't have control over their resources.
• Migration problem occurs, when the users wants to switch to some other provider for the better storage of their data. It's not easy to transfer huge data from one provider to the other.
• In public cloud, the clients' data can be susceptible to hacking or phishing attacks. Since the servers on cloud are interconnected, it is easy for malware to spread.
• Peripheral devices like printers or scanners might not work with cloud. Many of them require software to be installed locally. Networked peripherals have lesser problems.
• More and deeper knowledge is required for allocating and managing resources in cloud, since all knowledge about the working of the cloud mainly depends upon the cloud service provider.

IV. CONCLUSION

Cloud computing technology is increasingly being used in enterprises and business markets. A review shows that dynamic resource allocation is growing need of cloud providers for more number of users and with the less response time. In cloud paradigm, an effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. This paper summarizes the main types of RAS and its impacts in cloud system. Some of the strategies discussed above mainly focus on memory resources but are lacking in other 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] V. Vinothina, Dr. R. Shridaran, and Dr. Padmavathi Ganpathi, *A survey on resource allocation strategies in cloud computing*, International Journal of Advanced Computer Science and Applications, 3(6):97--104, 2012.
[2] Gunho Lee, Niraj Tolia, Parthasarathy Ranganathan, and Randy H. Katz, *Topology aware resorce allocation*

for data-intensive workloads, ACM SIGCOMM Computer Communication Review, 41(1):120--124, 2011.

[3] Abirami S.P. and Shalini Ramanathan, *Linear scheduling strategy for resource allocation in cloud environment*, International Journal on Cloud Computing: Services and Architecture(IJCCSA), 2(1):9--17, 2012.

[4] Daniel Warneke and Odej Kao, *Exploiting dynamic resource allocation for efficient parallel data processing in the cloud*, IEEE Transactions On Parallel And Distributed Systems, 2011.

[5] Atsuo Inomata, Taiki Morikawa, Minoru Ikebe and Md. Mizanur Rahman, *Proposal and Evaluation of Dynamic Resource Allocation Method Based on the Load Of VMs on IaaS*, IEEE, 2010.

[6] Dorian Minarolli and Bernd Freisleben, *Uitlity–based Resource Allocations for virtual machines in cloud computing*, IEEE, 2011.

[7] Jiyani, *Adaptive resource allocation for preemptable jobs in cloud systems*, IEEE, 2010.

[8] Bo An, Victor Lesser, David Irwin and Michael Zink, *Automated Negotiation with Decommitment for Dynamic Resource Allocation in Cloud Computing*, Conference at University of Massachusetts, Amherst, USA.

[9] Gihun Jung and Kwang Mong Sim, *Location-Aware Dynamic Resource Allocation Model for Cloud Computing Environment*, International Conference on Information and Computer Applications (ICICA), IACSIT Press, Singapore, 2012.

[10] Chandrashekhar S. Pawar and R.B. Wagh, *A review of resource allocation policies in cloud computing*, World Journal of Science and Technology, 2(3):165-167, 2012.

[11] Sandeep Tayal, *Tasks Scheduling Optimization for the Cloud Computing systems*, International Journal of Advanced Engineering Sciences and Technologies (IJAEST), 5(2): 111 - 115, 2011.

BIOGRAPHY

Author Details: S. Prathima, Student of M. Tech, Department of Computer Science & Engineering, Bharath College of Engineering and Technology for Women, Kadapa, Andhra Pradesh, India.

Email: prathima.sunguluru@gmail.com

Guide Details: Shaik Shasha Ali, Assistant Professor, Department of Computer Science, Bharath College of Engineering and Technology for Women, Kadapa, Andhra Pradesh, India.