International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 3.45 (SJIF-2015), e-ISSN: 2455-2585
Volume 3, Issue 12, December-2017

A Comparative Study of VM Load Balancing Policies in Cloud

Neha Mathur

Dept. of Computer Science & Engineering, Jai Narain Vyas University, Jodhpur

Abstract*— *Cloud computing has become one of the most significant internet-based technology in recent years. Users from almost every sector are demanding various services of the cloud. Cloud computing provides software, platform for creating new applications and hardware or infrastructure as a service. A cloud service provider provides services on the basis of client's requests. Client's requests are processed in the virtualized data centres where a physical machine runs a number of virtual machines on it. An important issue in cloud is load balancing in virtual machines of a data centre. In this paper, the performance of three VM load balancing policies - round robin, throttled and active monitoring/ equally spread execution load is evaluated based on parameters response time and data centre processing time. Cloud Analyst is used as tool. Benchmark data of the users of a social networking site in the world is used for comparative analysis.

Keywords*— *Round-Robin, Throttled, Active Monitoring, Cloud, Load Balancing, VM

I. Introduction

Cloud computing is a computing paradigm for managing and delivering services over the internet and is defined as “a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.” [1]. Cloud computing is an integrated concept of parallel and distributed computing which shares resources like hardware, software, and information to computers or other devices on demand. With the aid of cloud computing and internet facility, the customer can access the aforementioned resources by paying for the duration of use.

Virtual machine (VM) is an execution unit that acts as a foundation for cloud computing technology. Virtualization consists of creation, execution, and management of a hosting environment for various applications and resources. The VMs in the cloud computing environment share resources like processing cores, system bus, and so forth. The computing resources available for each VM are constrained by total processing power.

Since users across the globe are using cloud services at an accelerated rate, the load on cloud data centres and virtual machines is increasing rapidly. Efficient policies are needed to balance load for effective functioning of clouds. Load balancing enables enterprises to handle workload demands by allocating resources among multiple computers, networks or servers. To evaluate the performance of load balancing policies performance metrics need to be considered.

Response time is the time interval between sending a request and receiving its response. It should be minimized to boost the overall performance. Data Center processing time is the total time taken by the data centres in processing a request. The DC processing time should be minimum for user and system satisfaction. The objective of this paper is to evaluate and compare load balancing policies at VM level based on response time and data center processing time.

Organization of rest of the paper is as follow- section -2 focuses on virtualization in cloud, section-3 covers literature survey of VM load balancing algorithms, section -4 describes round robin, throttled and active monitoring VM load balancing polices, section-5 unveils the experimental setup, section 6 presents the observations and result analysis and lastly, the conclusion of this work is discussed.

II. VM Load Balancing Policies in Cloud

The data centres use the load balancer to distribute requests between the available virtual machines. Many VM load balancing policies exist. The most popular ones are as follows:

A) Round Robin

The simplest technique for distributing workloads across vms is round robin load balancing. Beginning from the first VM in the VM list of a DC, the round-robin load balancer forwards a client request to each VM in turn. When it reaches the end of the list, the load balancer loops back and goes down the VM list again. It sends the next request to the first listed VM, the one after that to the second server, and so on. Figure 1 illustrates the working of round-robin load balancer. The first request is sent to VM1, the second to VM2 and so on and the nth request is sent to the nth VM , now when the (n+1)th request arrives and the VM list end is reached, it is sent to the first VM ie VM1.

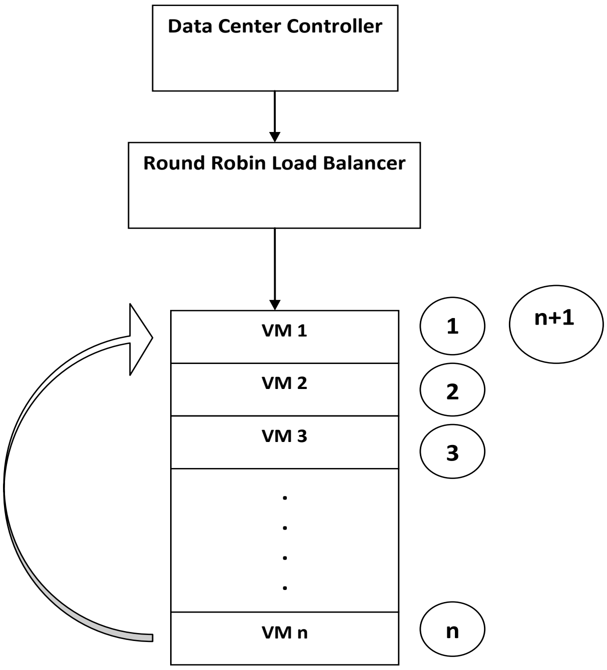

Figure 1: Working of Round Robin Load Balancer

B) Throttled

The load balancer in throttled policy maintains an index table of vms and the state of the VM (BUSY/AVAILABLE). Initially all VM's are available. Whenever the datacenter receives a new request, it queries the throttled load balancer for the next allocation .The load balancer parses the allocation table from top until the first available VM is found or the table is parsed completely. If found, the throttled load balancer returns the VM id to the datacenter. The DC sends the request to the VM identified by that id. The DC notifies the throttled load balancer of the new allocation and updates the allocation table accordingly. If not found, the throttled load balancer returns -1. The DC queues the request. When the VM finishes processing the request, and the DC receives the response cloudlet, it notifies the throttled load balancer of the VM de-allocation. Figure 2 depicts the working of throttled load balancer.

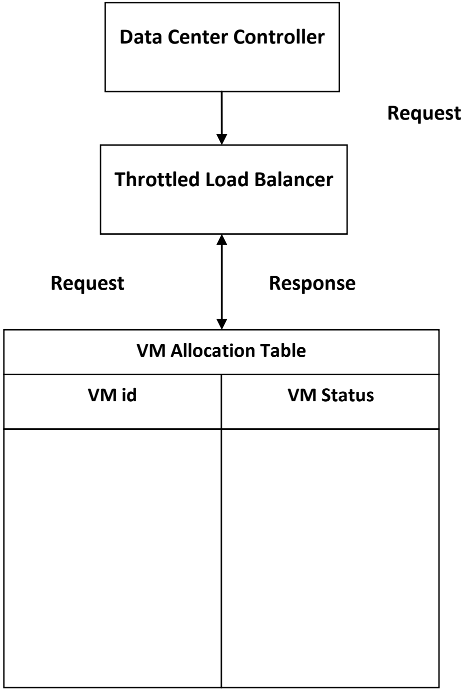

Figure 2: Working of Throttled Load Balancer

C) Active Monitoring

This load balancing policy attempts to maintain equal work loads on all the available vms. The algorithm used is quite similar to the throttled case. Figure 3 presents the working of active monitoring load balancer. The active load balancer maintains an index table of vms and the number of requests currently allocated to the VM. At the start all VM's have 0 allocations. When a request to allocate a new VM from the DC arrives, it parses the table and identifies the least loaded VM. If there are more than one, the first identified is selected and the VM id is returned to the DC. The DC sends the request to the VM identified by that id and notifies the active load balancer of the new allocation. The load balancer updates the allocation table increasing the allocations count for that VM. When the VM finishes processing the request, and the DC receives the response cloudlet, it notifies the active load balancer of the VM de-allocation. The load balancer updates the allocation table by decreasing the allocation count for the VM by one.

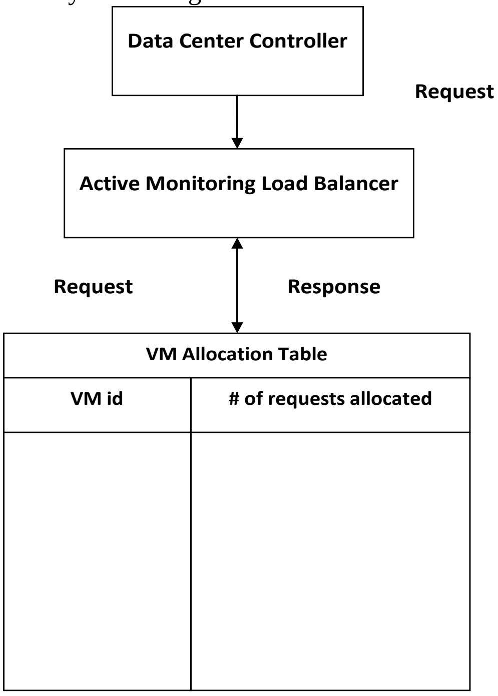

Figure 3: Working of Active Monitoring Load Balancer

III. EXPERIMENTAL SETUP

An experiment was carried out to find out the most appropriate VM Load balancing algorithm. There are three major VM load balancing algorithms: Round-Robin, Throttled and Active Monitoring/ Equally Spread Current Execution as described in section III.

CloudAnalyst is used as the simulator. NetBeans IDE 8.0.2 is used as development environment. In the development environment, the programming language Java is used for coding. The IDE and simulator are setup on a machine which is configured with Intel (R) Core™ i3 CPU M 330 @ 2.16 GHz processor and 2 GBDDR3 RAM, and installed with Windows 7 operating system.

A typical large scale application on the Internet that can benefit from Cloud technology is social networking applications. A popular social networking site has over 200 million registered users worldwide. In June 2010, the approximate distribution of its user base across the globe was the following: North America: 80 million of users; South America: 20 million of users; Europe: 60 million of users; Asia: 27 million of users; Africa: 5 million of users; and Oceania: 8 million of users. [2]

In this experiment, the behavior of social networking application is modeled and CloudAnalyst is used to evaluate cost and performance of various load balancing policies. [3]

Table 1 User bases used in the experiment

User Base	Region	Time Zone	Peak Hours (GMT)	Simultaneous Online Users during Peak Hours	Simultaneous Online Users during Off-Peak Hours
UB1	N. America	GMT -6.00	13:00-15:00	400,000	40,000
UB2	S. America	GMT -4.00	15:00-17:00	100,000	10,000
UB3	Europe	GMT +1.00	20:00-22:00	300,000	30,000
UB4	Asia	GMT +6.00	01:00-03:00	150,000	15,000
UB5	Africa	GMT +2.00	21:00-23:00	50,000	5,000
UB6	Ocenia	GMT +10.00	09:00-11:00	80,000	8,000

A. Simulation Configuration

Six user bases representing the six main regions of the world are defined with parameters described in Table 1.For the simulation, a similar hypothetical application is used at 1/10th of the scale of Facebook. For the sake of simplicity each user base is contained within a single time zone. It is also assumed that 5% of the registered users are online during the peak time simultaneously and only one tenth of that number of users is on line during the off-peak hours.

In terms of the cost of hosting applications in a Cloud, a pricing plan which closely follows the actual pricing plan of Amazon EC2 is assumed. The assumed plan is: Cost per VM per hour (1024Mb, 100MIPS): $ 0.10; Cost per 1 GB of data transfer (from/to Internet): $0.10 [4].

Size of virtual machines used to host applications in the experiment is 100MB. Virtual machines have 1GB of RAM memory and have 10MB of available bandwidth. Simulated hosts have x86 architecture, virtual machine monitor Xen and Linux operating system. Machines have 2 GB of RAM and 100GB of storage. Each machine has 4 CPUs, and each CPU has a capacity power of 10000 MIPS. A time-shared policy is used to schedule resources to VMs. Users are grouped by a factor of 10, and requests are grouped by a factor of 10. Each user request requires 100 instructions to be executed. User bases used in the experiments are described in Table 1.

B. Simulation Parameters

Table 2 illustrates the simulations parameters used in the experiments.

Table 2: Simulation Parameters

Parameter	**Value**
Datacenter Architecture	x86
OS	Linux
Virtual Machine Manager VMM	Xen
Cost per VM/Hr $	$0.10
Data Transfer Cost $/GB	$0.10
Physical HW Units (Machines) per Datacenter	2
No. Of Processors Per Machine	4
Processor Speed	10000 MIPS
VM Policy	TIME SHARED
DC Level Load Balancing /Service Broker Policy	Optimized Response Time

C. Simulation Scenarios

Three scenarios are considered in this work. All policies are applied in all three scenarios to perform comparison of the three. Optimized Response Time service broker policy is used for load balancing at data center level.

1. Classic Configuration

This is simplest one which consists of modeling the case where a single Cloud Data Center is used to host the social network application in each region. Data center has 100 virtual machines allocated to the application.

2. Homogeneous Configuration

In this, two data centers, each with 50 VMs dedicated to the application are used.

3. Heterogeneous Configuration

In this three data centers have different amount of virtual machines, each with 20, 30 and 50 VMs.

Each of these scenarios was evaluated with execution of workload previously described for all load balancing policies- round robin, throttled and active monitoring. Parameters used for comparative analysis of the policies are response time and time spent for processing a request by a data center. Results are discussed next.

IV. OBSERVATIONS AND RESULT ANALYSIS

V.

Table 2 depicts the results of the simulation for Round Robin VM Load balancing policy.

Table 3: Round-Robin Experiment Results

Scenario	Overall Response Time (ms)	Overall DC Processing Time (ms)
Classic	75.57	24.72
Homogeneous	73.5	23.92
Heterogeneous	72.15	21.38

Table 3 depicts the results of the simulation for Throttled VM Load balancing policy.

Table 4: Throttled Experiment Results

Scenario	Overall Response Time (ms)	Overall DC Processing Time (ms)
Classic	86.35	25.69
Homogeneous	67.89	17.15
Heterogeneous	61.78	10.84

Table 4 depicts the results of the simulation for Active Monitoring VM Load balancing policy.

Table 5: Active Monitoring Experiment Results

Scenario	Overall Response Time (ms)	Overall DC Processing Time (ms)
Classic	110.61	49.02
Homogeneous	83.95	33.24
Heterogeneous	72.06	21.3

A. Classic Scenario

Table 6 illustrates the results of classic scenario.

Table 6 : Results of Classic Scenario

VM Load Balancing Policy	Overall Response Time (ms)	Overall DC Processing Time (ms)
Round Robin	75.57	24.72
Throttled	86.35	25.69
Active Monitoring	110.61	49.02

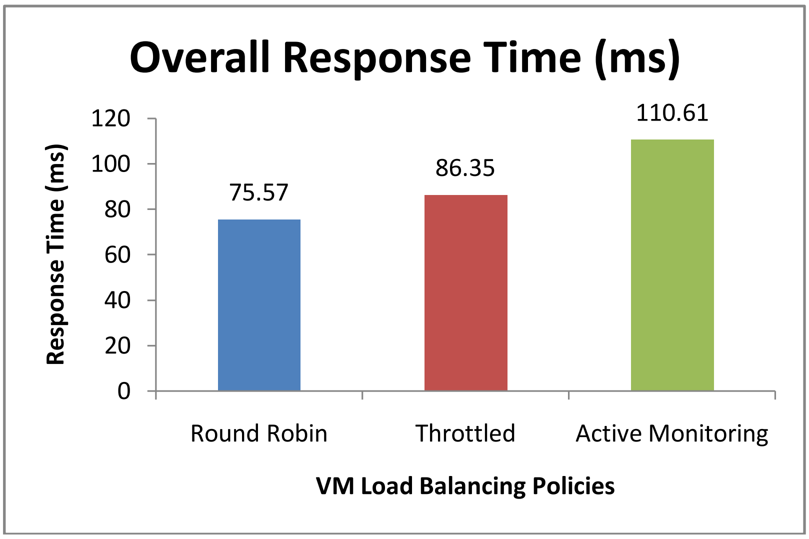

Figure 4: Overall Response Time Classic Scenario

Figure 4 presents the overall response time of all policies in classic scenario. It can be observed that the response time delivered by round robin policy is the lowest and active monitoring is highest.

Figure 5 presents the overall DC processing time of all policies in classic scenario. It can be observed that the Dc processing time delivered by round robin policy is the lowest and active monitoring is highest.

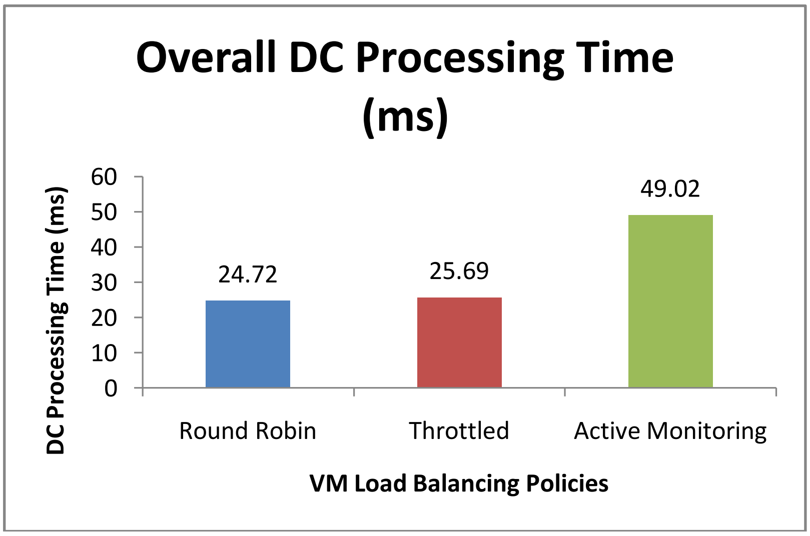

Figure 5: Overall DC Processing Time - Classic Scenario

Figure 6 presents the cumulative value of overall response time and overall DC processing time (Turnaround Time) of all policies in classic scenario. It can be observed that the turnaround time delivered by round robin policy is the lowest and active monitoring is highest.

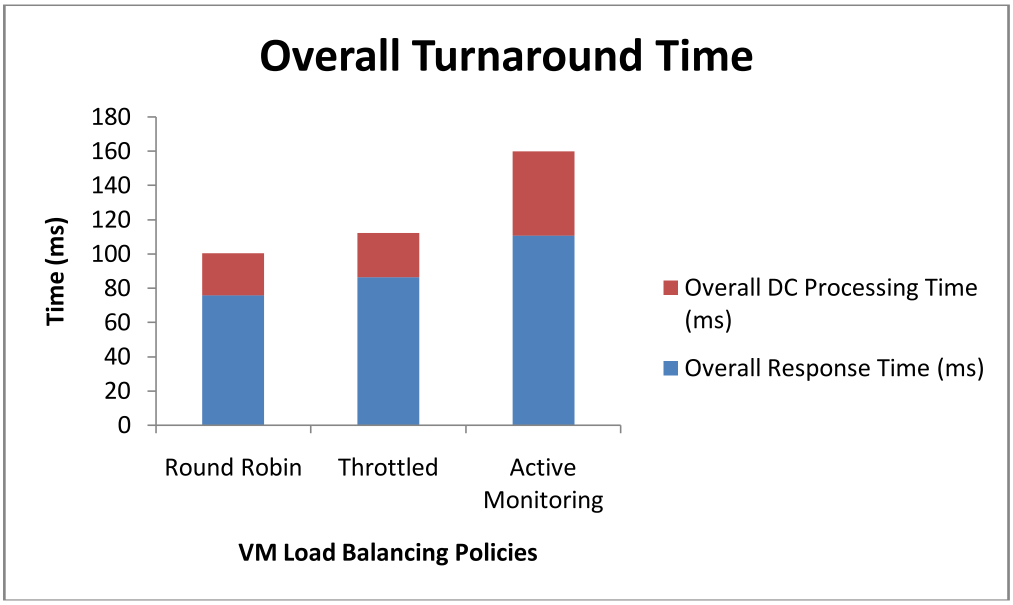

Figure 6: Overall Turnaround Time in Classic Scenario

B. Homogeneous Scenario:

Table 7 illustrates the results of homogeneous scenario.

VM Load Balancing Policy	Overall Response Time (ms)	Overall DC Processing Time (ms)
Round Robin	73.5	23.92
Throttled	67.89	17.15
Active Monitoring	83.95	33.24

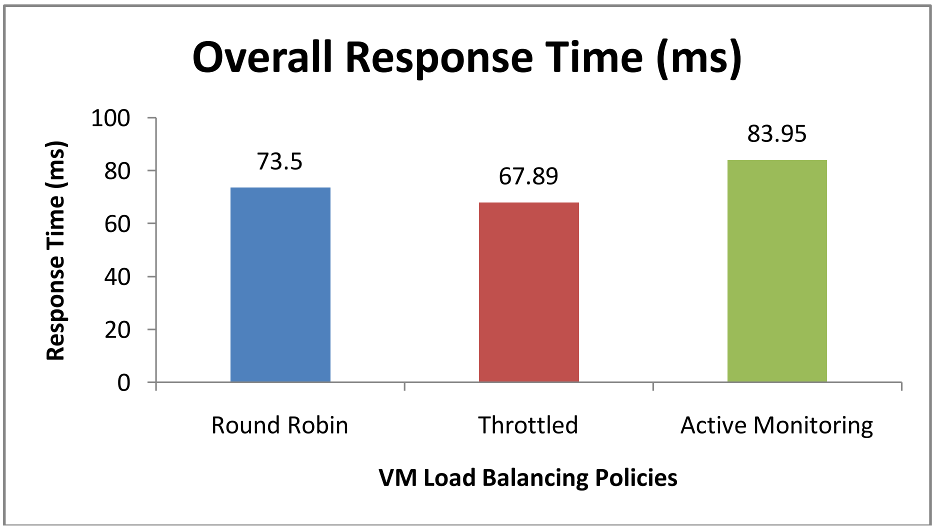

Figure 7: Overall Response Time - Homogeneous Scenario

Figure 7 presents the overall response time of all policies in homogeneous scenario. It can be observed that the response time delivered by throttled policy is the lowest and active monitoring is highest.

Figure 8 presents the overall DC processing time of all policies in homogeneous scenario. It can be observed that the DC processing time delivered by throttled policy is the lowest and active monitoring is highest.

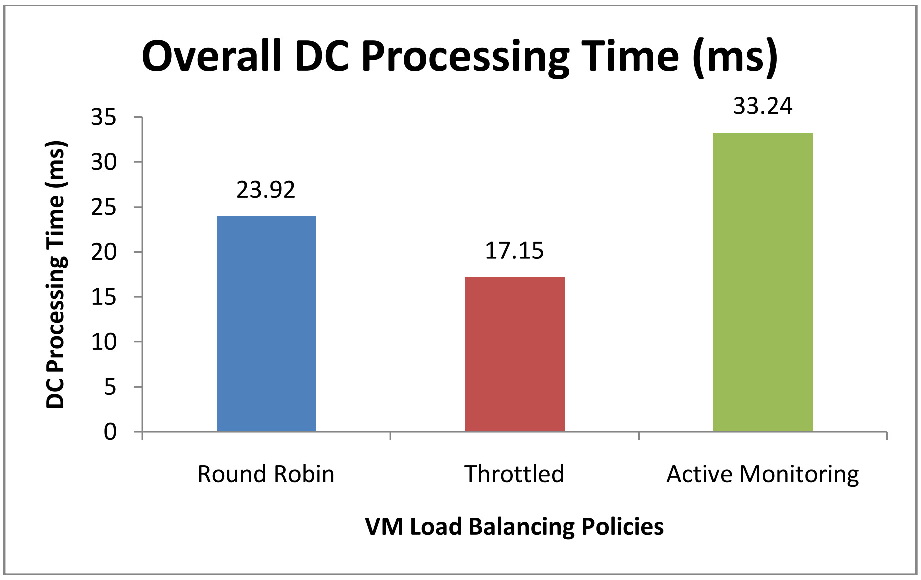

Figure 8: Overall DC Processing Time- Homogeneous Scenario

Figure 9 presents the cumulative value of overall response time and overall DC processing time (Turnaround Time) of all policies in homogeneous scenario. It can be observed that the turnaround time delivered by throttled policy is the lowest and active monitoring is highest.

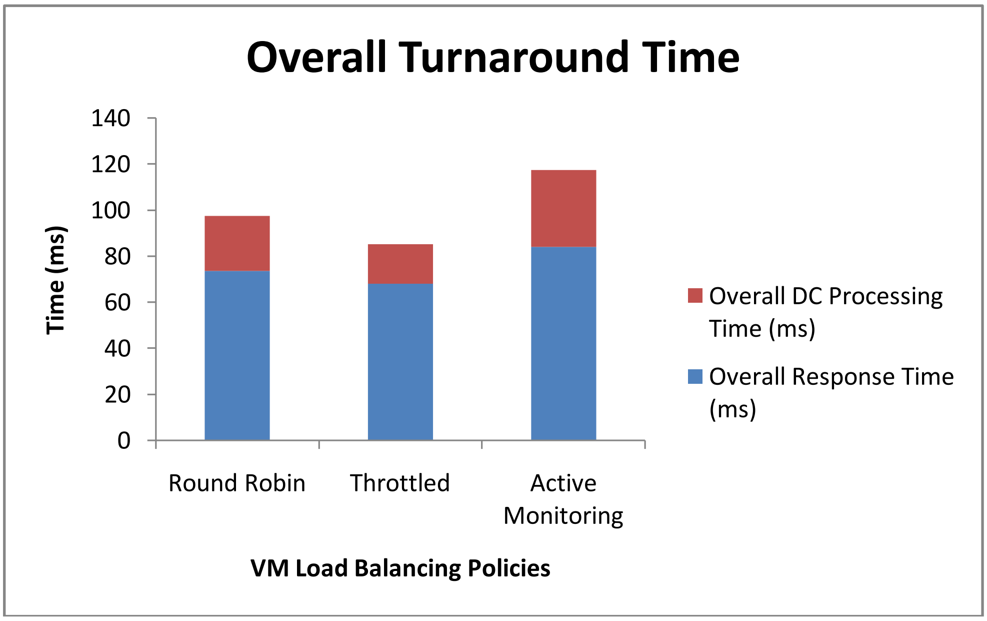

Figure 9: Overall Turnaround Time - Homogeneous Scenario

C. Heterogeneous Scenario:
Table 8 illustrates the results of heterogeneous scenario.

VM Load Balancing Policy	Overall Response Time (ms)	Overall DC Processing Time (ms)
Round Robin	72.15	21.38
Throttled	61.78	17.15
Active Monitoring	72.06	21.30

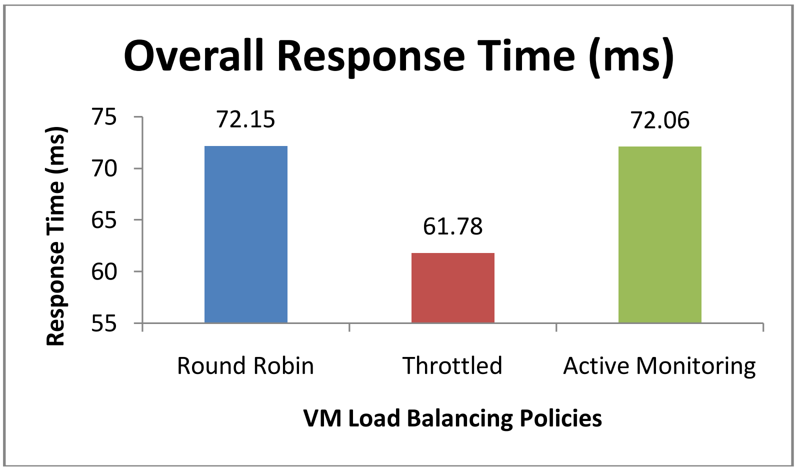

Figure 10: Overall Response Time - Heterogeneous Scenario

Figure 10 presents the overall response time of all policies in heterogeneous scenario. It can be observed that the response time delivered by throttled policy is the lowest and round robin is highest.

Figure 11 presents the overall DC processing time of all policies in heterogeneous scenario. It can be observed that the DC processing time delivered by throttled policy is the lowest and round robin is highest.

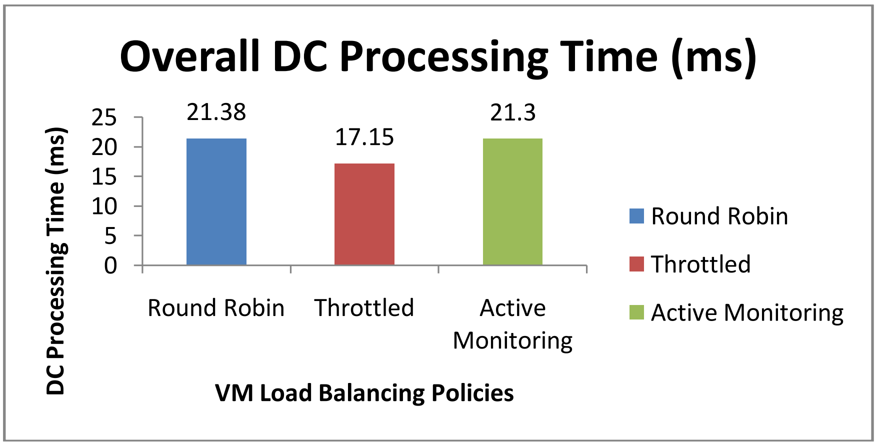

Figure 11: Overall DC Processing Time - Heterogeneous Scenario

Figure 12 presents the cumulative value of overall response time and overall DC processing time (Turnaround Time) of all policies in heterogeneous scenario. It can be observed that the turnaround time delivered by throttled policy is the lowest and round robin is highest.

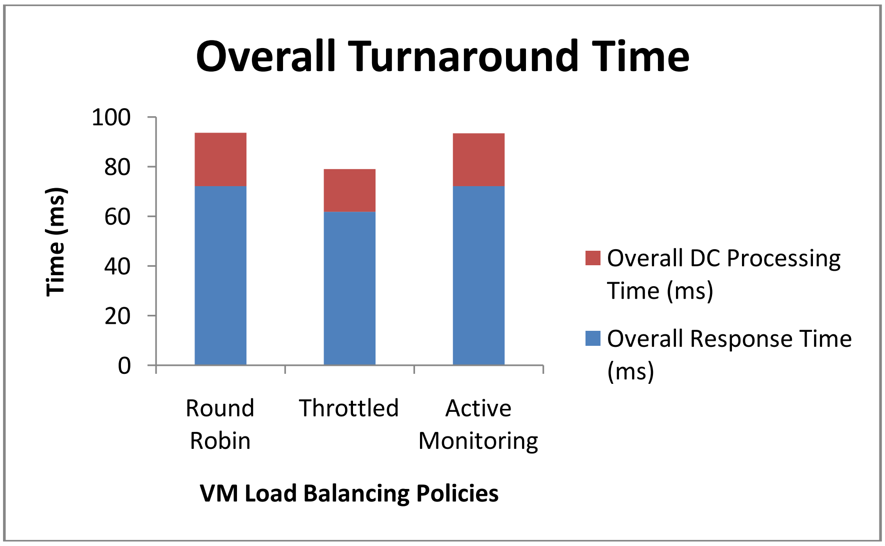

Figure 12: Overall Turnaround Time - Heterogeneous Scenario

VI. CONCLUSION

Cloud computing is a new emerging trend in computer technology that has influenced every other entity in the entire industry, whether it is in the public sector or private sector. With the advent of cloud, new possibilities are opening up on how application can be built and how different services can be offered to the end user through virtualization, over the internet. In cloud computing load balancing is used for distributing the load on virtual machine and cloud resources. It minimizes the total waiting time for resources.

In this work, a comparative study of VM load balancing policies is done. In classic scenario, the round robin policy performs better than throttled and active monitoring policies. The throttled load balancing policy is better than round-robin and active monitoring load balancing policies in homogeneous and heterogeneous scenario.

The response time should be as less as possible for user satisfaction. The throttled load balancing policy yield lower response times in homogeneous and heterogeneous scenarios. The overall DC processing time should be as less as possible for better throughput. For system and user satisfaction DC processing time should be low. The throttled load balancing policy performs better than the other two policy yielding lower DC processing times in homogeneous and heterogeneous scenarios.

REFERENCES

[1] Z. Xiao, W. Song, and Q. Chen, “Dynamic resource allocation using virtual machines for cloud computing environment,” IEEE Transactions on Parallel and Distributed Systems, vol. 24, no. 6, pp. 1107–1117, 2013.
[2] www.internetstats.com [Last Retrieved on 20-12-2017]
[3] B. Wickremansinghe , R. N. Calheiros and R. Buyya , “CloudAnalyst: A CloudSim- based Visual Modeler for Analyzing Cloud Computing Environments and Applications”, IEEE Computer Society, 2010, pp. 446-452.
[4] https://aws.amazon.com/ [Last Accessed On: 22-12-17]