Performance Analysis of Multi Scale Applications on Cloud Systems

Vijaya Kumbhar, Ajay Kumar

Abstract: Cloud Computing is the most adoptable technology in use recently where bigger and complex applications are deployed and used on cloud server. Multi scaled applications are such applications which have multiple attributes (data size, number of requests, number of concurrent users etc.) that can cause to poor performance of its application service. For better performance, it needs to be measured, analyzed and optimized using tools, strategies or algorithms. The multi scale application’s performance is measured at end user level, network level and data center level. In this paper, the focus is measuring the performance of multi scale applications in terms of response time and throughput using CloudAnalyst simulator. Multiple scenarios like varying data size, varying number of concurrent users and varying number of requests are simulated for multiscale application and tested. The results obtained show that response time and throughput of application gets reflected mostly by Data Size, number of requests per user and number of concurrent users. It is also observed that the most significant factor is data size along with huge volume of concurrent users, impacts the response time of multi scale application. Hence this research recommends that multi scale data i.e. data type and data size plays vital role in cloud based multi scale application’s performance measurement.

Keywords: cloud computing, Multi Scale Cloud Application, Performance measurement, CloudAnalyst, Request-Response Time, System Throughput.

I. INTRODUCTION
The cloud computing is the current methodology of deploying applications over internet. It is the practice of using managed services that are hosted on network of remote servers over the internet to store, manage, and process data and its deal with all transaction as way the local server or personal computer deal [38]. Generally, applications running in cloud environment are designed to deal with small to extremely large number of users, is termed as scalability. Multi Scale Applications in cloud consists of variable payload i.e. the applications whose payload gets changed with time scale[1,2]. Usability of cloud applications is increasing day by day and QoS is an important factor to understand the cloud application’s performance. The service response time, throughput, bandwidth, resources utilization, waiting time, processing time are the factors on which performance of cloud application is defined and QoS is measured to complement the performance for what the cloud is intended. Hence, on increase in number of users & service concurrency, it is important to keep eye on the performance of cloud application to justify the cloud characteristics. The most important factor for application performance in cloud environment is measured in terms of ‘processing-time’ and ‘response-time’. The variable as data size(DS), concurrency of users(NU) and number of requests(NR) are to be applied to understand application performance in cloud environment. The response time of some applications is very sensitive and crucial such as banking applications, any online service oriented applications, real time applications, share market applications, high performance computing (HPC) applications and high throughput computing (HTC) applications. The best and acceptable response time for Share Market applications is less than or equal to 1 sec. whereas Generate-Order-Transaction applications, acceptable response time is 4 to 6 sec.[27]. Similarly E-Commerce, banking and retail applications expects the response time of max. 2 sec.[19]. Considering some applications where response time is important but not very much sensitive for which max. 10 sec. response time is acceptable [20]. Response time is highly substantial factor in cloud environment. It can be defined as time required for applications responsiveness for a given transaction. Normally it is measured as how fast the application responds to particular request. As per SLA guidelines of cloud, response time is the difference of the time of cloud service customer initiated event (stimulus) and cloud service provider initiated event in response to that stimulus [45]. Jacob Neilson [18] evaluated ideal response time values as 0.1 ms (min) to 10 sec for any web applications. For the applications running in cloud environment, the ideal values are generally given by cloud providers according the type of service in Service Level Agreement (SLA). It is slightly variable depending on type of application, its necessity and urgency of the responsiveness. Although throughput, latency and data processing time are also most significant metrics, but good response time reflects good throughput and latency [44]. Hence it is necessary to study under what circumstances the response time gets changed. In this research paper, the aim is to measure at what variability of response time and throughput is better than the best performance while applying different variable attributes of multiscale applications.

II. PROBLEM STATEMENT & OBJECTIVES

When multiscale applications are executed in cloud environment, different types of data, data size of the application, number of requests and the concurrency of its users affect overall performance [38]. Performance bottlenecks of such applications need to be evaluated before its actual deployment.

The objectives of this research are:

- To measure the performance of multi scale applications
in virtualized cloud environment in terms of response time and throughput.

- To determine the significant factors affecting multi scale applications’ performance at end user level.

### III. RELATED WORK

Dr. Neeraj Bhargava [12] analyzed how number of distributed clients in social networking applications of cloud can affect the performance of overall application. Rajkumar Buyya et al. [6,7] analyzed performance of large scale social networking application using the scenarios of concurrent user on different configured infrastructure. Samip Raut et al. [37] analyzed extremely large data size and concluded that for better performance of big data application in cloud, it need to change the infrastructure or at application designing level. Simon Ostermann et al. [32] analyzed performance of large scaled scientific applications and various performance metrics are evaluated but fixed and large size data had been considered for experimental work. Arshdeep Bahga et al. [5] analyzed performance of multi-tier applications and concluded that performance is sensitive to attributes of application (such as database read/write intensive workload etc.). characteristics of workload (such as session length, inter-session interval, think-time, workload mix etc.). Such applications can give better performance for high memory and high CPU capacity. Roberto R. Expósito et al. [15] studied scientific applications and its characteristics for analyzing high performance and concluded that such applications are sensitive and can give better response for high speed internet characteristics.

Cloud Computing is the most adoptable technology in use recently where bigger and complex applications are deployed and used on cloud server. Multi scaled applications are such applications which have multiple attributes (data size, number of requests, number of concurrent users etc.) that can cause to poor performance of its application service. For better performance, it needs to be measured, analyzed and optimized using tools, strategies or algorithms. The multi scale application’s performance is measured at end user level, network level and data center level. In this paper, the focus is to measure the performance of multi scale applications in terms of response time and throughput through CloudAnalyst simulator. The results obtained shown that response time and throughput of application gets reflected mostly by Data Size, number of requests per user and number of concurrent users. The most significant factor is data size along with huge volume of concurrent users impacts the response time of multi scale application.

B. Santhosh Kumar [8], Sandip Patel [38], Soumya Ray [39] explained importance of load balancing algorithms used in VM allocation. Authors in [10,11,22,31,36,41,42] compared and analyzed all existing load balancing algorithms while in [4], whereas authors in [10,11,16,22,28,33,41] analyzed importance of service broker policies. Various researcher work shows that the application’s overall performance gets affected by the payload of the application, type of data incurred in it and concurrency of the application in the cloud. In the conclusion of above researchers indicate that for social networking applications and multimedia based applications in cloud, the performance gets affected by payload of application & concurrency of application. Hence it needs to be measured, analyzed and improved by changing network characteristics by upgrading the infrastructure of servers or by enhancing the algorithm.

### IV. PERFORMANCE METRICS FOR EVALUATION

The performance of multiscale application’s is measured and analyzed with response time and throughput [6,7,32]. The response time is the time that requires to response the requests of applications i.e. the time interval from submission of request till response received [5][15]. It is measured by taking average of the response time of all users and requests. Throughput is defined as number of tasks completed per unit time [5][15].

Overall Response time = average response time of all users and all requests

Throughput = \( \frac{\text{number of requests} \times \text{number of concurrent users}}{\text{overall response time}} \)

### V. SIMULATION PARAMETERS SETUP

CloudAnalyst v1.0 Beta is used for the simulation of cloud application to measure the performance on varying data size, requests and concurrent users. Figure-1 represents diagrammatic view of simulated cloud application that is being used in this research [6,7].

---

Figure-1: Pictorial View of the simulated experiment for cloud application performance measurement
VI. EXPERIMENT OF PERFORMANCE MEASUREMENT

The simulation experiment is performed to understand what data size, number of users, and number of requests of multiscale applications deviates the response time. The users traffic i.e. number of users from all over the world has been considered. The entire earth is divided into six regions. Hence numbers of users spreads non-homogeneously in different regions. It is necessary to check performance of multiscale application on minimum infrastructure, hence number of data center is configured as one, number of servers per data center is configured as two and number of virtual machine per server is configured as five. Each virtual machine has a waiting queue of incoming requests to be processed. Detailed parameter description is in following tables.

Table-1 shows configuration of distributed number of users, table -2 shows configuration of user bases, table-3 shows configuration for data center of application deployed, table-4 shows the configuration of Virtual Machines in each data center and table-5 shows the algorithm details used in cloudlet and data center controllers.

Table-6: Multiscale applications response time and throughput obtained by varying the data size

<table>
<thead>
<tr>
<th>Data Size</th>
<th>Avg. Response Time (in ms)</th>
<th>Throughput (tasks per sec)</th>
<th>Increment in Response Time</th>
<th>Deduction in Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 Bytes</td>
<td>292.11</td>
<td>205402</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1 KB</td>
<td>292.17</td>
<td>205360</td>
<td>0.02 %</td>
<td>0.01 %</td>
</tr>
<tr>
<td>10 KB</td>
<td>293.40</td>
<td>204499</td>
<td>0.42 %</td>
<td>0.01 %</td>
</tr>
<tr>
<td>100 KB</td>
<td>306.09</td>
<td>196021</td>
<td>4.33 %</td>
<td>0.04%</td>
</tr>
<tr>
<td>1 MB</td>
<td>420.16</td>
<td>138838</td>
<td>41.19 %</td>
<td>0.41%</td>
</tr>
<tr>
<td>10 MB</td>
<td>1692.22</td>
<td>35456</td>
<td>291.57 %</td>
<td>2.92 %</td>
</tr>
<tr>
<td>100 MB</td>
<td>14240.32</td>
<td>4213</td>
<td>741.52 %</td>
<td>7.42 %</td>
</tr>
<tr>
<td>1 GB</td>
<td>95961.23</td>
<td>66</td>
<td>832.17 %</td>
<td>63.33 %</td>
</tr>
<tr>
<td>10 GB</td>
<td>2050898.96</td>
<td>29</td>
<td>931.80 %</td>
<td>1.24 %</td>
</tr>
</tbody>
</table>

VII. SIMULATION RESULTS AND ANALYSIS

The number of experiments is conducted to understand the performance of multiscale applications in cloud environment by varying data size(DS), number of concurrent users(NU) and number requests per user(NR).

Experiment-1: Performance of multi scale applications by varying DS and by keeping constant NU and NR- In this scenario, data size is scaled from 100 bytes to 10 GB i.e. small to large keeping the intension that each user having different requirement of using cloud application, with no. of highest requests as 60 by considering that each user can give maximum 60 requests per hour and number of concurrent users as 1 billion keeping intention that the cloud application is much popular and used by maximum number of users from all over the world. The throughput is calculated as total number of tasks/requests processed per second. The result is tabulated in table-6 and presented in figure-2.

Observation:
The figure-2 and table-6 shows that, up to 1 GB data size, the response time is growing slightly and less than 100 sec. for the entire application and less than 0.1 sec on an average for each user. Hence the

![Figure-2: Multiscale applications response time and throughput by varying the data size](image-url)
response time upto 1 GB is excellent but after 1 GB, it grows up to 35 min. for the entire application which indicates degradation in performance. In case of throughput obtained, it is observed that, the throughput is excellent till 100 KB but it starts degrading for large data size.

Experiment-2: Performance of multi scale application by varying NU and by keeping DS and NR constant:

In this scenario, number of concurrent users is scaled from 1000 to 1 billion i.e. less to large volume of number of users, keeping the intention that users from all regions are using the applications, with no. of highest requests as 60 by considering that each user can give maximum 60 requests per hour and largest data size as 10 GB. The result is tabulated in table-7 and presented in figure-3.

<table>
<thead>
<tr>
<th>Number of Concurrent users</th>
<th>Response Time (in ms)</th>
<th>Throughput</th>
<th>Increment in Response Time</th>
<th>Deduction in Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^3</td>
<td>2208719.03</td>
<td>37</td>
<td>--</td>
<td>0 %</td>
</tr>
<tr>
<td>10^4</td>
<td>2235041.98</td>
<td>37</td>
<td>1.02 %</td>
<td>0 %</td>
</tr>
<tr>
<td>10^5</td>
<td>2273265.6</td>
<td>38</td>
<td>1.02 %</td>
<td>0 %</td>
</tr>
<tr>
<td>10^6</td>
<td>2278131.94</td>
<td>38</td>
<td>1.01 %</td>
<td>0 %</td>
</tr>
<tr>
<td>10^7</td>
<td>2290669.19</td>
<td>38</td>
<td>1.01 %</td>
<td>0 %</td>
</tr>
<tr>
<td>10^8</td>
<td>2300773.88</td>
<td>38</td>
<td>1.00 %</td>
<td>0 %</td>
</tr>
<tr>
<td>10^9</td>
<td>2346165.19</td>
<td>39</td>
<td>1.02 %</td>
<td>0 %</td>
</tr>
<tr>
<td>10^10</td>
<td>2388003.69</td>
<td>40</td>
<td>1.02 %</td>
<td>0 %</td>
</tr>
</tbody>
</table>

Table-7: Multiscale applications response time and throughput obtained by varying the users

<table>
<thead>
<tr>
<th>No. of Requests</th>
<th>Response Time (ms)</th>
<th>Throughput (tasks per second)</th>
<th>Increment in Response Time</th>
<th>Deduction in Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2030898.87</td>
<td>295</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>20</td>
<td>2187899.56</td>
<td>274</td>
<td>1.08 %</td>
<td>0.08 %</td>
</tr>
<tr>
<td>30</td>
<td>2223458.94</td>
<td>270</td>
<td>1.02 %</td>
<td>0.02 %</td>
</tr>
<tr>
<td>40</td>
<td>2234558.36</td>
<td>269</td>
<td>1.00 %</td>
<td>0.00 %</td>
</tr>
<tr>
<td>50</td>
<td>2345458.92</td>
<td>256</td>
<td>1.05 %</td>
<td>0.05 %</td>
</tr>
<tr>
<td>60</td>
<td>2456898.16</td>
<td>244</td>
<td>1.05 %</td>
<td>0.05 %</td>
</tr>
</tbody>
</table>

Figure-3: Multiscale applications response time and throughput obtained by varying the concurrent users

Observation:
From the figure-3 and table-7, it is observed that, when the application’s number of concurrent users is increased, the overall response time after 1000 users till 10 billion users has been increased slightly. In case of throughput also there is growth with increase in number of concurrent users.

In case of throughput obtained, it is observed that, the throughput is much affected by number of users but it is beyond accepted values.

Experiment-3: Performance of multi scale applications by varying NR and by keeping DS and NU constant:

In this scenario, requests are is scaled from 10 to 6 keeping the intention that each user hitting the requests maximum 60 times to same application, with no. of highest data size as 10GB an number of concurrent users as 1 billion keeping intention that the cloud application is much popular and used by maximum number of users in its peak hour from all over the world. The result is tabulated in table-8 and presented in figure-4.

<table>
<thead>
<tr>
<th>No. of Requests</th>
<th>Response Time (ms)</th>
<th>Throughput (tasks per second)</th>
<th>Increment in Response Time</th>
<th>Deduction in Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2030898.87</td>
<td>295</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>20</td>
<td>2187899.56</td>
<td>274</td>
<td>1.08 %</td>
<td>0.08 %</td>
</tr>
<tr>
<td>30</td>
<td>2223458.94</td>
<td>270</td>
<td>1.02 %</td>
<td>0.02 %</td>
</tr>
<tr>
<td>40</td>
<td>2234558.36</td>
<td>269</td>
<td>1.00 %</td>
<td>0.00 %</td>
</tr>
<tr>
<td>50</td>
<td>2345458.92</td>
<td>256</td>
<td>1.05 %</td>
<td>0.05 %</td>
</tr>
<tr>
<td>60</td>
<td>2456898.16</td>
<td>244</td>
<td>1.05 %</td>
<td>0.05 %</td>
</tr>
</tbody>
</table>
Figure-4: Multiscale applications response time and throughput obtained by varying the number of requests

Observation:
The figure-4 and table-8 shows that when the application’s number of concurrent users is increased, the overall response time gets increased slowly. But it is high due to large number of users. In case of throughput obtained, it is observed that, the throughput is much affected by number of users but it is beyond accepted values.

Experiment 4- Performance of multi scale applications by varying DS, NR and NU:
In this scenario, Data size is scaled from 100 bytes to 10 KB and number of concurrent users increased from 1000 to 1 billion) and number of requests are increased from 60 to 15360, assuming that the load of the application is gradually increasing in peak time. The result is tabulated in table-9 and presented in figure-5.

Table-9: Multiscale applications response time and throughput obtained by varying the data size, users and requests

<table>
<thead>
<tr>
<th>Data Size</th>
<th>No. Of Requests</th>
<th>No. of users</th>
<th>Response Time (ms)</th>
<th>Throughput (tasks per second)</th>
<th>Increase in Response Time</th>
<th>% deduction in Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 Bytes</td>
<td>60</td>
<td>10^3</td>
<td>249.09</td>
<td>240.88</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1 KB</td>
<td>120</td>
<td>10^3</td>
<td>294.28</td>
<td>203.89</td>
<td>0.18 %</td>
<td>0.18 %</td>
</tr>
<tr>
<td>10 KB</td>
<td>240</td>
<td>10^3</td>
<td>296.27</td>
<td>200.59</td>
<td>0.01 %</td>
<td>0.02 %</td>
</tr>
<tr>
<td>100 KB</td>
<td>480</td>
<td>10^3</td>
<td>306.08</td>
<td>196.03</td>
<td>0.03 %</td>
<td>0.02 %</td>
</tr>
<tr>
<td>1 MB</td>
<td>960</td>
<td>10^3</td>
<td>445.22</td>
<td>134.76</td>
<td>0.45 %</td>
<td>0.45 %</td>
</tr>
<tr>
<td>10 MB</td>
<td>1920</td>
<td>10^3</td>
<td>13758.47</td>
<td>4.36</td>
<td>29.90 %</td>
<td>29.91 %</td>
</tr>
<tr>
<td>100 MB</td>
<td>3840</td>
<td>10^3</td>
<td>14240.32</td>
<td>4.21</td>
<td>0.04 %</td>
<td>0.04 %</td>
</tr>
<tr>
<td>1 GB</td>
<td>7680</td>
<td>10^3</td>
<td>95961.23</td>
<td>0.63</td>
<td>5.74 %</td>
<td>5.68 %</td>
</tr>
<tr>
<td>10 GB</td>
<td>15360</td>
<td>10^3</td>
<td>2050898.96</td>
<td>0.03</td>
<td>20.37 %</td>
<td>20.00 %</td>
</tr>
</tbody>
</table>

Figure-5: Multiscale applications response time and throughput obtained by varying the data size, users and requests

Observation:
From the figure-5 and table-9, it is observed that, the overall response time upto 1 MB with requests960 and 10 Core concurrent users was slightly increasing but after 100MB data size and requests3840 and 1 Billion users, it is drastically increasing and is not adequate.

In case of throughput obtained, it is seen that, when the application load in terms of data size, requests and concurrent users is changed and increased, the throughput obtained was excellent till 100 KB but it is starting reducing when application’s overall load is increased slowly and it drops to low level i.e. to 0.03 ms.

VIII. DISCUSSION

The above results interpret that to process a request in its peak time, it will take less than 0.1 sec. on an average but if request is submitted with large volume of data even if less number of users are there, the response time taken will be more. If the data size is more or less, number of user doesn’t reflect response time.

In multiscale applications, when data size is increasing to large, the response time is acceptable and excellent and at par the ideal cloud application. But as data size is huge and bulky, the response time drops to large extent. If the infrastructure is not scaled, the throughput also gets decreased due to large data size per request but not that significantly as that of response time. The reason behind this is when large data size in involved, the load per request gets increased, which ultimately causes increase in load per packet. Though number of concurrent users and number of requests are constant, to transfer heavy data i.e. requests and packets, the processing time required is more.
When there are huge number of concurrent users, response time and throughput both get changed with increase in number of users but the change in both values is least significant. The reason behind this is the scaled architecture of web and cloud applications.

When number of request are more and large, response time and throughput both get changed but are of least significant but more effective than number of concurrent users. The reason behind this is again increasing the payload of requests and ultimately numbers of packets. Hence to transfer more number of packets, the required time is more.

When the payload of the application gets changed by changing all the attributes involved, the response time gets increased and throughput gets decreased. The reason behind this can be heavy load per request, per packet and huge number of concurrent users.

Above scenarios can be found in many real applications. In banking applications, transactions like generating OTP or payment module, the response time will be acceptable but if for new account creation; customer photo and signature updating may take more time. In multimedia or animation based applications, the response time is affected by larger requests. In stock marketing applications, the transactions are based on small sized text; response time is not much affected, even though there is large number of users. In E-Commerce applications, payment module may run faster but displaying product image can be slower if huge numbers of users are hitting the application.

IX. CONCLUSION

Now a day, the high performance computing (HPC) applications tend to evolve towards heterogeneous architectures of cloud systems. A multi scale cloud application contains multiple components interacting with each other, which has different impact with respect to change in space and time scales. Hence in current research, change in space and time parameters is considered while measuring multi scale cloud application performance.

The simulation experiments are performed to understand the factors affecting multi scale application performance by applying increase in data size, number of users and number of requests. Research test result analysis proved that the increase in data size, increases response time drastically but throughput is not affected significantly. The increase in users and requests does not affect significantly on response time and throughput for the multiscale applications in cloud environment. It is observed that the most influencing factor is data size because its consequence was more on response time and throughput; comparatively number of requests is less significant because less impressive on response time and throughput. Thenumber of concurrent users is least significant factor for multiscale applications because of scaled nature of all applications over internet.

REFERENCES

3. Alex Conninno,”Emerging Issues: Cloud Computing”, Association for Progressive Communications (APC) in partnership with SANGEONET.
8. B. Santhosh Kumar, ”An Implementation Of Load Balancing Policy For Virtual Machines Associated With A Data Center”, International Journal Of Computer Science & Engineering Technology (IJCSET), Issn : 2229-3345 Vol. 5 No. 03 Mar 2014
17. Hetal V. Patel1, Ritesh Patel Cloud Analyst: An Insight of Service Broker Policy, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 1, January 2015, ISSN (Online) : 2278-1021ISSN (Print) : 2319-5940


**AUTHORS PROFILE**

**VijayaKumbhar** is a Research Student in Computer Science Department, SavitribaiPhule Pune University(SPPU) and Assistant Professor in Indira College of Commerce and Science , Pune ,India. She received Master of Computer Science (MCS) degree in 2002 from SavitribaiPhule Pune University, Pune, MS, India. Her research interests are Algorithms Analysis and Cloud Computing.

**Dr. Ajay H. Kumar** is a Research Guide in Computer Science Department, SavitribaiPhule Pune University and Director of Jaywant Technical Campus. His research interests are Computer Networks, Data Warehousing, Cloud Computing and performance of communication systems.