Performance Evaluation by Throughput Analysis in Private Cloud

S. Jagannatha [#1], Niranjanamurthy [*2],K Venkatesh [#3]
[#] *Dept of Computer Applications, M. S. Ramaiah Institute of Technology*
Bangalore – 54

Abstract

Analysis of throughput in private cloud is an important consideration. Sub task of each use case in a scenario is assigned to a resources for commutations. Predicting number of machines are in operational is determined using markov chain model. This can be analysed based on the failure rate and machine switching/repair/allocate new machine. In this paper failure of physical machines is considered by making use of the markov-chain model to do a probabilistic prediction of failure and compute the throughput of all these machine which are in working. Second is effective use of these resources by computing turnaround time by proper assign tasks to those resources which are in operational. We propose a general algorithm and illustrated with a case study by simulation.

Keywords

UML, Use case, Throughput, SAM, Markov chain

I. INTRODUCTION

The private cloud is developed for internal network for a selected users in computing services offered either over the Internet. The advantage of using private cloud is for scalability, elasticity, security and self-service needs of computing within the organization. In case any machine is failed, it replace a failed machine, by additional machine or machine is repaired by repair person. The data usage within the organization ensure operations and sensitive data are not accessible. The company's IT department is held responsible for the cost and accountability of managing the private cloud. These resources are managed by separate staff for manage, maintenance and maintenance expenses as traditional datacenter.[22].

Cloud give range of services to customers. Behind all this ease of scalability and managing of resources present in the cloud easy to use. They are the key problem that is faced by cloud. Efficient management of resources failure led to reduce the throughput. Resource management in cloud as we know is very complex because of the scale of the cloud and also the inclusion of vast number of components that make the cloud. we cannot get the exact state of the whole system.

The Markov chain model the transitions takes place from one state to another. The probability of the next state depends only on the current state. Throughput in general means the rate at which something can be processed. The processing capacity of the machines that is present in the datacenter. The throughput in terms of handling requests per second per machine in the datacenter. Based on the throughput to determine the optimal number of machines are in working private cloud. The probability of the state depends only on the current state and not on the events that preceded it. The same concept is used for the machines in the datacenter, which have operational state and failed state. We are trying to make a probabilistic prediction of failure of set of machines in the datacenter by assuming a failure rate for a machine.

II. RELATED WORK

In Daniel A Menanse et el addresses throughput analysis of internet data center (IDC). The author also computing the total processing capacity based on machines and number of requests are processed by each machine. He analyses failed machine and repair rate for operating data center. Repaired machine is, it returns to the queue of operational machines [25]. In the work done by seematai S Patil and koganti the concept expresses different dimensional resource utilization of the server is considered. By minimizing the skewness different types of workloads that can be combined together. These results improving the overall utilization of the server resources. They have also developed a set of heuristics that prevent system overload while saving energy used [7]. In [5] the author discusses resource allocation strategies they include argument, scarcity, fragmentation, over-provisioning and under-provisioning, input parameters that should be considered from the CSP (cloud service provider) and customer perspective of the resources. The author talked about the merits and demerits of the various resource allocation strategies. In paper [8] gossip based co-operative virtual management allocation and cost management is used for the organizations. These results in co-operate and share the available resources to reduce the cost. In paper [5] to improve the utilization of resources, optimization of job scheduler is proposed.

The jobs can be classified by resource dependency as CPU-bound, Network I/O bound, Disk I/O bound and memory bound. This MJO scheduler can detect the type of the jobs and parallel jobs of different category.

In the work done in [3] by Gunho Lee, he talks about the resource allocation in cloud keeping in mind the unpredictability factor. Author classifies the users of cloud one is the cloud service provider and another is the user or customer who makes use of these resources. Proper resource allocation and utilization will only happen when both the parties are exchanging some information between them like the cloud service provider giving the architectural details. TARA (topology aware resource allocation) is used to do the optimal resource allocation, results show that TARA is more efficient when compared to other systems.

In [6] author mainly talks about the allocation of resource to tasks by considering the priority of the task and also considers the dynamic nature of the cloud into consideration. Authors in [8] have proposed a new approach that adopt a distributed architecture where resource management is decomposed into independent tasks each of which is performed by node agents. And these agents carry out configurations in parallel using the Prmnethee method. Authors in paper [9] have discussed about the service level agreement, based resource allocation problem for multi-tier applications in cloud computing. The author proposed an algorithm based on force directed search to solve the problem. Authors in paper [10] proposes the network resource allocation strategies and its possible applicability in cloud computing. The authors have focused on the aspect of network awareness and consistent optimization of network resource allocation strategies. In paper [11] the author proposes a utility mechanism design to allocate resources among virtual machines' efficiently. They also have made use of stochastic approximation to get stochastic solution for allocation outcomes. In paper [12] the authors have presented a framework for the adaptive resource co allocation management in virtualization-based datacenters. They have also developed an algorithm that can be adaptively find out an optimal resource configuration scheme step by step procedure in each interval.

In paper [13] authors have proposed an auction mechanism dynamically to the allocation probe m of computation capacity in cloud computing environment. They have also introduced auctioning mechanism into the resource allocation problem. Authors in paper [14] have introduced and compared four automated resource allocation strategies relying on the expertise that can be captured in workflow-based applications. In paper [15] authors discussed the process of resource allocation in distributed clouds where in application developers can selectively lease geographically distributed resources. They have mainly focused on challenges inherent to the resource allocation process particular to distributed clouds which offers a stepwise view of this process that covers the initial modeling through to the optimization phase. Authors in paper [16] have proposed an algorithm for resource allocation algorithm for the cloud system with pre emptily tasks. The algorithm adjusts the resource allocation adaptively based on the updated of the actual task executions. In paper [17] the authors have proposed resource allocation method based on the load of the virtual machine and Infrastructure as services. This method has enabled the users to dynamically scale up and scale down instances on the basis of load and conditions specified by the user. Authors in paper [18] have presented a new systematic approach to predict the resource needs in cloud based on the past usage. The approach analyzes the resource allocation logs of virtual server for resource prediction. In paper [19] author presents a load balancing algorithm balancing workload using virtual machine. Monitoring to different performance parameters for the clouds of different sizes is presented in this paper. Authors in [20] discuss and highlight the cases where there are problems and also how that can be improved. In [6] author has proposed a new approach an algorithm that allocates resources with maximizes usages and maximum profit. The algorithm is based on parameters of cloud like time, cost, no of CPU's required. In the work done by jiani, actual task execution time, pre-emptive scheduling is considered for resource allocation. It has overcome the problem of resource contention and increases the resource utilization by using different modes

In this paper we are try to analyze throughput in different machines with N number of repair person or switching time and We here only concentrate on the compute throughput and the cost involved and also the failure of machines. In the following sections III. We will explain about the algorithm that is devised for throughput analysis and resource utilization. In section IV we look at parts of the implementation and how they are interpreted. In section V conclusion to the work and finally references are incorporated.

III.BASIC CONCEPTS OF THROUGHPUT COMPUTING METHODOLOGY

The private cloud in an organization is composed of many resources, which can be utilized effectively. Failure of one machine immediately repaired and made operational. Throughput analysis is an important aspect of private cloud. Allocating resources is like providing the tasks that execute on the cloud with the required resources on demand. Machines as we normally know have two states a) operational state b) failed state. Machines in a cloud the number of machines present is very huge [23], [25].

Markov-chain model which mainly is used to predict the future states of a system by knowing the present state of the system. This model to find the probability of machine failing without considering

the events that led to the failure. The machine failure that we are concentrating on can be found by using the theory of probability. To calculate throughput we need to know how many machines in operational. For this we need certain inputs such as the number of machines present in the private cloud, the number of repair persons, the failure rate of the machines, the average time it takes to repair a machine or allocating new machine in private cloud. The system (Failure – recovery model for IDC and formulae for computing throughput is referred in [23] [25]. Using formulae we will compute the throughput of the M machine in private cloud and a set of repairpersons or switching or allocate new machine. Throughput is generated by varying the number of machine and repairperson or switching time in a private environment. In the computed results of throughput values we find the optimal throughput for a private cloud. The number of machines for which we get the optimal throughput will be considered for resource allocation. The throughput values of varying M and N values of private cloud datacenter, one can determine an optimal throughput for set of M machines. Now we consider these M machines as a resource and we will try to allocate these resources among jobs using the server allocation (SAM) methodology [25].

IV. PROPOSED METHODOLOGY

Algorithm for computing throughput of resources of different machines in Private cloud.

Step 1. Consider M number of machines present in the cloud. Assume that each machine is capable of £ requests/second.

Step 2. Consider the N number of repair persons required for repairing the machine

Step 3. Take the average failure (**λ**) rate of the individual machine in the cloud

Step 4. Take the average switching state/allocate duplicate system in place of repair machine/ repair rate/ (μ) per machine in the cloud

Step 5. Compute the initial probability Po using the formulae referred in [25]

Step 6. Compute P_k using the formulae[25]with the values of M, N, λ, μ for k machines with k ranging from k=1 to N

Step 7. Compute P_k using the formulae[25]with the values of M, N, λ, μ for k machines with k ranging from k ranging from k=n+1 to M.

Step 8. The sum of the probability of P_k k ranging from 1 to M is equal to 1 or nearing 1.

Step 9. Compute the throughput of the datacenter using the formulae[25]

Step 10. Repeat the steps 1 through 9 by varying the value of N linearly until N reaches 1/4th of the M value

Step 11. Gather the throughput results using a data structure and find the maximum value of the throughput.

Step 12. Take the value of M for which the throughput value is high

Step 13. Generate Comparison of correlation coefficients r (x, y)with variable M and N values

Step 14. Analyze the results with collrelation coefficients r

Step 15. Algorithm to determine the optimal allocation by resource allocation presented in [25]

V. ILLUSTRATION WITH CASE STUDY

In this section we will start with the illustration of throughput computation using markov chain model. The steps for computation is as follows:

1. We assumed that a value of 500 minutes for mean time to failure (MTTF) so that λ = 1/500 i.e. 0.002 failure per minute.
2. We will assume the average repair rate (μ) per machine in the datacenter to be 20 minutes. So the repair rate is given by 1/20 = 0.05 repair per minute/switching state
3. We will consider the value of total number of machines (M) to start with 30 up till 210 with a step count of 30. Number of requests processed by a machine is assumed to be 50 requests/sec.
4. We will consider the value of total number of repairpersons to start with 1 up till 10 with a step count of 1.
5. For these values of M and N we will compute both P_0 and P_k
6. We will verify that that the total sum of P_k is 1. In this case we have computed the sum and found the value to be 0.9999 or 1.
7. We will make use of all the P_k values for computing the throughput. Below we have given the computed throughput values for all the values of M and N in table 1

TABLE I COMPUTED THROUGHPUT

N/M	30	60	90	120	150	180	210
1	**1105.34**	1250.0	1250.0	1250.0	1250.0	1250.0	1250.0
2	1016.60	2406.90	2499.99	1250.0	1250.0	1250.0	1250.0
3	984.326	2610.60	3689.19	3749.99	3750.0	3750.0	3750.0
4	980.33	2602.61	4122.98	4956.81	4999.99	5000.0	5000.0
5	979.87	2599.81	4178.67	5546.63	6217.76	6249.98	6250.0
6	979.83	2599.41	4190.02	5668.87	69922.67	7475.49	7499.98
7	979.83	2599.41	4193.41	5699.72	7109.24	8273.17	8731.36
8	979.83	2599.42	4194.52	5709.65	7161.56	8524.00	9607.21
9	979.83	2599.43	4194.88	5713.15	7179.58	8600.16	9925.35
10	979.83	2599.43	4194.99	5714.39	7186.41	8627.51	10027.3
Absolute Maximum Throughput	1500	3000	4500	6000	7500	9000	105000

The average throughput percentile is presented in table 2 for private cloud of different sizes and varying number of repair persons/switching time.

TABLE II. COMPUTED THROUGHPUT IN PERCENTILE

N	30	60	90	120	150	180	210
1	73.69	41.67	27.78	20.83	16.67	13.87	11.90
2	67.77	80.23	55.56	41.67	33.33	27.78	23.81
3	65.62	87.02	81.98	62.50	50.00	41.67	35.71
4	65.36	86.75	91.62	82.61	66.67	55.56	47.62
5	65.33	86.66	92.86	92.44	82.90	69.44	59.52
6	65.32	86.65	93.11	94.48	92.30	83.06	71.43
7	65.32	86.65	93.19	95.00	94.79	91.92	83.16
8	65.32	86.65	93.21	95.16	95.49	94.72	91.50
9	65.32	86.65	93.22	95.22	95.73	95.56	94.53
10	65.32	86.65	93.22	95.24	95.82	95.86	95.50

From the above table 1 and 2 we observe that the optimal maximum throughput value are indicated in bold. In case M = 120. The number of repair persons N= 4 the percent throughput is 82.61%. It increases to 92.44% for N= 5. Further even the number of repair person increased to N = 10 there is only about 3% increases the throughput. It will be colossal waste of human resource if 10 repair persons have to be used for private cloud of 120 machines to achieve maximum throughput in percentile. On the other hand, N= 5 achives optimal maximum throughput for private cloud and it also uses optimal human resources. Similarly the optimal maximum throughput for cloud of size of 90 is corresponding to the number of repair person N= 4.

The number of repair persons needed to achieve optimal throughput for different size of cloud is summarized in table 3.

TABLE III REPAIR PERSONS NEEDED TO ACHIEVE OPTIMAL THROUGHPUT

M	30	60	90	120	150	180	120
N	1	3	4	5	6	7	8

Table 3. The number of repair persons needed to achieve maximum throughput for different size of cloud is summarized in table 4.

TABLE IV. REPAIR PERSONS NEEDED TO ACHIEVE MAXIMUM THROUGHPUT.

M	30	60	90	120	150	180	120
N	1	3	10	10	10	10	10

The correlation coefficients r is computed in table 3 and table 4. The variable M and N are represented on the x and y axes respectively. The correlation coefficients is computed.

TABLE V CORRELATION COEFFICIENTS R

Computer r	Correlation coefficient	Correlation coefficient percent
Table 3	0.9923	99.23%
Table 4	0.8677	86.77%

The strong correlation in the case of data in table 3 supports the fixing of number of repair persons to achieve optimal maximum throughput for the private cloud of different size

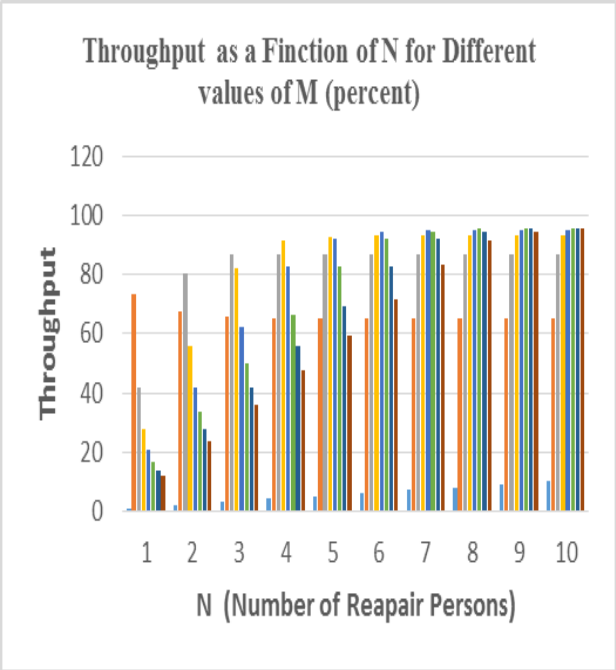

Fig 1: Throughput for different values of M and N

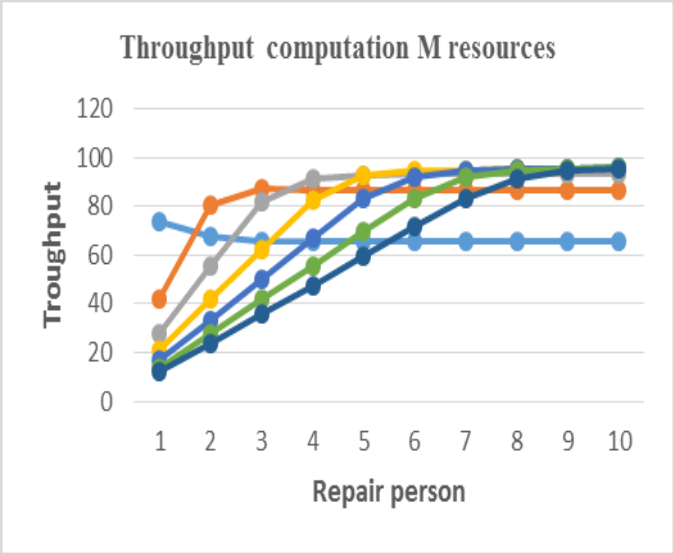

Fig 2: Throughput computation

From the graph in figure 2 indicates that the through put in the case of private cloud with 30 machines shows a decreasing trend as N, the number of repair persons increased. This shows that one repair person is the optimal requirement for the cloud with 30 machines. This decreased trend of throughput of private of 30 machines indicates that the Morkov chain model behaves differently for smaller sample size. Using an optimum number of servers one can find the optimum usage of these resources.

VI. CONCLUSIONS

In this paper, we use the markov-chain model for determining optimum repair person is required to operate private cloud of M resources. We have also proposed a general methodology for optimal resource are in operational state and compute the throughput of these resources. Finally we determine optimum number of required repair personal to maximum usage of resources by analyze the repair time/switching time of failed resource. Finally we simulated and results are discussed. The correlations coefficient is computed and results are validated.

REFERENCES

[1] Dr. S. jagannatha, Dr. D.E Geetha, Dr. T. V. Suresh kumar, K. Rajini Kanth, "Load Balancing in Distributed Database System using Resource Allocation Approach" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 7, July 2013.

[2] Dr. S. Jagannatha, "Performance Analysis of Datacenters using Markov-chain Model" National confrene at R V college of Engineering

[3] Gunho lee, "Resource Allocation and Scheduling in Heterogeneous Cloud Environments" may 10, 2012,university of california at berkley.

[4] Chandrashekar S. Pawar, Rajinikant b. Wagh, "Priority Based Dynamic

[5] TResource Allocation in Cloud Computing", IJACSA

[6] V.Vinothina, Dr. S. Sridaran, Dr. PadmavathiGanapathi, "A Survey on resource Allocation Stratergies in Cloud Computing", IJACSA, volume – 3, Issue-6 , 2012

[7] K. C. Gouda, Radhika T. V, Akshatha M, "Priority Based resource Allocation Model for Cloud Computing", IJSETR,Volume-2,Issue-3,January – 2013.

[8] Seematai S.Patil, Koganti, "Dynamic Resource Allocation using Virtual machines for Cloud Computing Environment" IJAET, Volume-3, Issue 6, August-2014,

[9] Ya˜gız Onat Yazır, Chris Matthews, Roozbeh Farahbod, "Dynamic Resource Allocation in Computing Clouds using Distributed Multiple Criteria Decision Analysis", 2010 IEEE 3rd International Conference on Cloud Computing.

[10] Hadi Goudarzi and Massoud Pedram, University of Southern California, Los Angeles, "Multi-dimensional SLA-based Resource Allocation for Multi-tier Cloud Computing Systems", 2011 IEEE 4th International Conference on Cloud Computing.

[11] M. Asad Arfeen, Krzysztof Pawlikowski, Andreas Willig, "A Framework for Resource Allocation Strategies in Cloud Computing Environment", 2011 35th IEEE Annual Computer Software and Applications Conference Workshops.

[12] Zhen Kong, Cheng-Zhong Xu, Minyi Guo, "Mechanism Design for Stochastic Virtual Resource Allocation in Non-Cooperative Cloud Systems", 2011 IEEE 4th International Conference on Cloud Computing.

[13] Xiaoying Wang, Hui Xie, Rui Wang, Zhihui Du, Li Jin, "Design and Implementation of Adaptive Resource Co-allocation Approaches for Cloud Service Environments", 2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE).

[14] Wei-Yu Lin, Guan-Yu Lin, Hung-Yu Wei, "Dynamic Auction Mechanism for Cloud Resource Allocation", 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing.

[15] Tram Truong Huu, Johan Montagnat, "Virtual resources allocation for workflow-based applications distribution on a cloud infrastructure", 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing.

[16] Patricia Takako Endo, André Vitor de Almeida Palhares, Nadilma Nunes Pereira, "Resource Allocation for

Distributed Cloud: Concepts and Research Challenges", IEEE Network, July/August 2011.
[17] Jiayin Li, Meikang Qiu, Jian-Wei Niu, Yu Chen, Zhong Ming, "Adaptive Resource Allocation for Preemptable Jobs in Cloud Systems", 978-1-4 244 -813 6-1/10/ 2010 IEEE.
[18] Atsuo Inomata, Taiki Morikawa, Minoru Ikebe, Sk. Md. Mizanur Rahman, "Proposal and Evaluation of a Dynamic Resource Allocation Method based on the Load of VMs on IaaS", 978-1-4244-8704-2/11/2011 IEEE.
[19] B.V.V.S Prasad, Sheba Angel, "Predicting Future Resource Requirement for Efficient Resource Management in Cloud", Internationa Journal of Computer Applications (0975-8887) volume 101-No.15, September 2014.
[20] Bhaskar. R, Deepu. S.R, Dr. B.S. Shylaja, "dynamic allocation method for efficient load balancing in virtual machines for cloud computing environment", Advanced Computing: An International Journal (ACIJ), Vol.3, No.5, September 2012.
[21] Rupali Shelke, Rakesh Rajani, "Dynamic resource allocation in Cloud Computing", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 2 Issue 10, October – 2013.
[22] Jagannatha, S., N. S. Shravan, and S. Kavya. "Cost performance analysis: Usage of resources in cloud using Markov-chain model." Advanced Computing and Communication Systems (ICACCS), 2017 4th International Conference on. IEEE, 2017.
[23] https://azure.microsoft.com/en-in/overview/what-is-a-private-cloud/
[24] Jagannatha, T.V.Suresh Kumar, Rajanikanth Algorithm of Performance Prediction by Resource Sharing in Distributed Database International Journal of Computer Applications (0975 – 8887) Volume 66– No.11, March 2013.
[25] S Jagannatha, TV Suresh Kumar, DE Geetha, K Rajani Kanth, Assessment of Workload Using Shapely Value in Distributed Database, Proceedings of International Conference on Advances in Computing,31- 40,Publisher, Springer India, 2012.
[26] Daniel A. Menascé Performance and Availability of Internet Data Centers Published by the IEEE Computer Society 1089-7801/04/$20.00 © 2004 IEEE INTERNET COMPUTING