EVALUATION OF SYSTEM UTILIZATION IN DATA CENTER BY INTEGRATING DMS-PSO IN QUEUING MODEL

[1]Shruthi.P.S, [2]Dr.D.R Umesh

[1]Associate Professor, [2]Assistant professor, [1]PESCE, Mandya[2],PESCE,Mandya

Abstract : Now-a-days an increasing demand for internet service as reached its extremities such that productive resource allocation and utilization had the knock the minds of cloud service providers. However many works have been put forth the queuing models for servicing the request, in our work we have considered M/M/c/K(c servers and queue K to hold client request) for resource allocation and appended the Dynamic MultiSwarm-Particle Swarm Optimization(DMS-PSO).Applying DMS-PSO, waiting time servers is collected to find local best and regrouped based on it, finally least global best swarm(least waiting time swarm) is selected and the request is sent to it continuously till the servers become busy. By collaborating M/M/c/K and DMS-PSO over head on distribution calculation is reduced which leads pavement for efficient system utilization.

IndexTerms **- Cloud datacenter, queuing model, Dynamic MultiSwarm-Particle Swarm Optimization, waiting time and power efficiency.**

I. INTRODUCTION

The increasing demand for service from the internet providers and cloud based computation has exponentially increased the count of servers in huge data centers. The electric power consumption in these data centers has reached the maximum capacity. Even at the low service rate datacenters require about 65% of the total power consumption **[1][2]**. The reason behind this is that the datacenters are designed for high traffic even though the average service load is 60% of the maximum service load **[3]**.To maintain the efficient power utilization the unutilized servers will be turned off **[4][5]**, however this operation have major drawback with concern to performance factors of the servers**[4]**. Therefore, the issue of managing power consumption of the servers to reach outstanding performance is important. In real time applications of data centers defined set of backup servers will be turned on/off dynamically. In our proposed work the server groups are dynamically selected based on the average waiting time of the group of servers. The financial cost and time duration for switching the servers cannot be neglected **[6]**, because too many backup servers may reduce the service rate and increase the power utilization. Identifying an optimal backup collection of servers and number of servers in each collection is a challenge in achieving efficient power consumption. Considering the above situation, our work collectively correlates the power consumption of all the server groups and service request waiting time in dynamic environment. Firstly all the servicing servers are grouped into different groups based the worst average waiting time and best average waiting time dynamically based on Dynamic Multi Particle Swarm optimization **[7]**. The queuing of service is established to maintain the load balancing between the different groups of servers which are dynamically grouped. In the goal of identifying the optimal server switching mechanism, we emphasize on minimizing the power consumption by considering the performance factor in terms of average service waiting time. The paper is organized as follows: In Section **2** the related work of discuss about cloud data center, DynamicMultiSwarm-Particle Swarm Optimization(DMS-PSO) and queuing model. Section **3,** have the proposed work which explains the integration of queuing model with DMS-PSO. Section **4** projects the formulated results for waiting time in the system and its utilization percentage with queuing model and integration of queuing model and DMS-PSO. Section **5** concludes our view about the system utilization evolution by considering queuing model alliance with DMS-PSO.

II. RELATED WORK

2.1 CLOUD DATACENTER:

A cloud is collection of resources like development software, storage services, middleware and virtual infrastructures required for computation on the cloud paradigm, which pays the way for designing, developing,

implementing and managing cloud applications. The data centers on cloud are geographically distributed based on requirements by the cloud service providers. Sometimes the workload on these data centers are more than their capacity but the service providers should optimally distribute the service request among the data centers with the goal of increasing their revenue. Reaching these objectives is quite difficult due some of the incident that leads to cloud service breakdown such as power rampage, security breakdown and natural calamities etc. Therefore cloud service providers are experiencing these kind of serious issues in providing continues service which is directly proportional to their revenue, were in providers should have efficacious strategy to handle straggle service availability. A positive motivation of research work has been under process to focus on optimizing the provisioning of effective cloud service by decreasing its operational cost **[8].** Usually the modern servers operates between 10% to 50% of maximum possible utilization. Later at this degree of utilization, the energy efficiency of the gradually reduces **[9].** Inspite of low average utilization there exists upgradation of system to reach the essentials of service-level-agreements (SLAs), services providers forcibly allocating huge amount of resources, which leads to energy inefficiency **[10].**

Productive power and performance issues in huge data centers has attracted in recent years. It can be differentiated into dynamic resource management and static resources management. In case of static management enormous energy is consumed for provisioning efficacious performance necessities under high traffic. Considering this criteria recent investigations have come up with dynamicity which switch on and off servers according to real time load **[10]**.In **[11]**, author have proposed a power consumption model and workload allocation model in order to check the trade-off between them.

2.2 DMS_PSO:

Particle swarm optimizer imitates flocking of birds, school of fish, herd of animal to resolve optimization complications. The PSO was proposed by Kennedy and Eberhart in 1995 **[12][13].** Too many single objective optimization issues can be denoted as

$$Min\ f(x), x = [x_1, x_2, \ldots\ldots, x_D] \quad x \in [x_{min}, x_{max}] \qquad \textbf{(1)}$$

Where, D total count of parameters to be optimized. X_{min} and x_{max} are upper extremity and lower extremity of the search area. In PSO, each solution carrying node is regarded as particle. Every Particle move through the D dimensional parameter space while observing the previous data collected in the process of search. The solution carrying nodes have incline towards better search area throughout the course of search activity **[14]**. In dynamic multi swarm particle swarm optimization (DMS-PSO) the swarms are tiny group of nodes. The whole set of nodes are grouped into smaller swarms, these groups of swarms are regrouped based on different regrouping schedules and later data is exchanged themselves. It is more efficient than any other PSO type for multi nodal issues **[15].**

2.3 QUEUING MODEL:

In distributed environment like cloud, the request from the client is sent to the application which is provided in data center. These user requests are distributed among the servers in the data center by the load balancer or efficient resource allocating algorithms. By this the resources in the data center are utilized up to their mark, in turn it reduces waiting time which decreases the power consumption of the system. Once the service request is issued from the client the response time is calculated with help of Queuing Theory (QT) **[16]**.

III. PROPOSED WORK

In this section we propose our work for less power consumption strategy by implementing queuing theory and considering average waiting of the service request. The proposed model adjust the number of servers using Dynamic multiswarm- particle swarm optimization, the swarm of servers are regrouped every time based on best average waiting time and worst average waiting time. A queuing model is illustrated in figure 1, which shows the M/M/c/K queuing layout.

According to M/M/c/K framework, (λ) a passion entry for service request, the time difference between requests is exponential with the rate μ, the buffer size (K) of the queue is finite, so the if it is full the request is rejected. The queue follows the First-In-First-Out algorithm **[17]**.

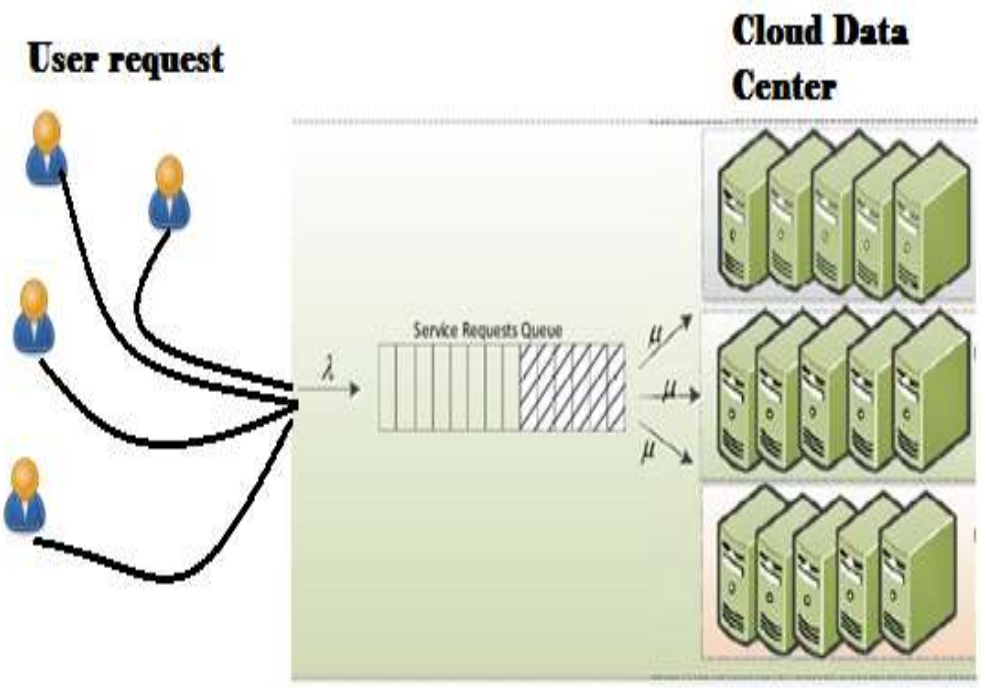

Fig 1: Service quest queuing model

Figure1 explains the queuing model for infinite user request, with K be the queue buffer size and C be the number of servers in the system.

λ→Service request arrival rate.
μ → Average service rate.
L→Average number of customers in the server
Lq→ Average number of customers in the queue
W→average time spent in the server
W_q →average waiting time in the queue
$\boldsymbol{\rho_q}$ →Congestion rate at queue.
$\boldsymbol{\rho}$ → Congestion rate at server.

According to Little's queuing rules **[18]**: $\mathbf{L = \lambda W}$ **(2)**

s $\mathbf{Lq = \lambda Wq}$ **(3)**

$$\mathbf{W = W_q + \frac{1}{\mu}} \quad \textbf{(4)}$$

Let c denote number of identical server

$$\boldsymbol{\rho = \frac{\lambda}{c\mu}} \quad \textbf{(5)}$$

K be the buffer size of queue

$$\boldsymbol{\rho_q = \frac{\lambda}{K\mu}} \quad \textbf{(6)}$$

For M/M/c queuing model:

$$\boldsymbol{L_q = \frac{P_0\left(\frac{\lambda}{\mu}\right)^c \rho_S}{c!(1-\rho_S)^2}} \quad \textbf{(7)}$$

Where

$$p_0 = 1/\left[\sum_{m=0}^{c-1} \frac{(c\rho)^m}{m!} + \frac{(c\rho)^c}{c!(1-\rho)}\right] \quad \textbf{(8)}$$

P_0 donates the probability that there is no request in the system.

$$W_q = \frac{L_q}{\lambda} \quad \textbf{(9)}$$

$$W = W_q + \frac{1}{\mu} \quad \textbf{(10)}$$

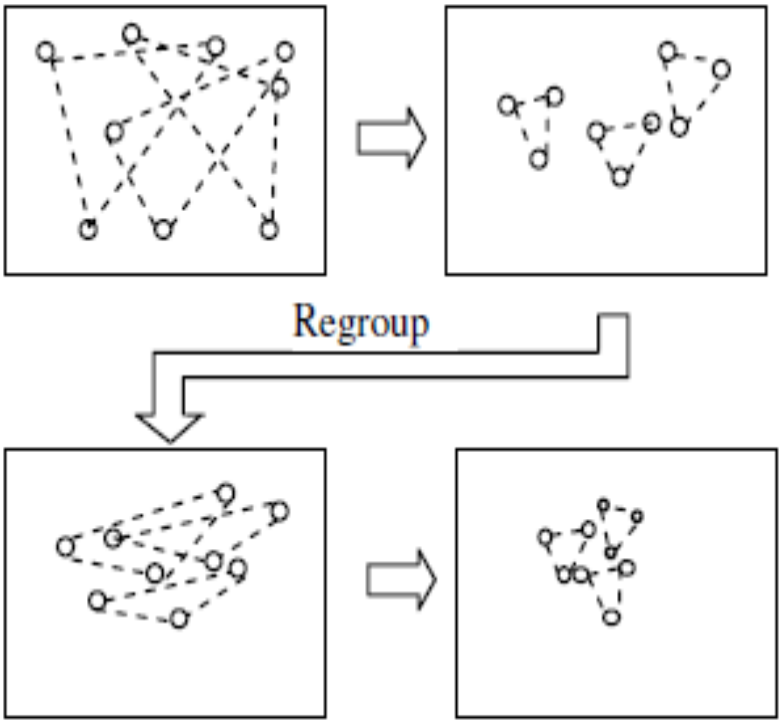

Fig 2: DMO-PSO regrouping

Initially the nodes are randomly divided into number of swarms and the request is fed into those servers that are idle. As when the request rate fluctuate the server becomes busy and idle every now and then, switching on and off server based on request rate will add initialization cost and time so Dynamic Multiswarm- Particle Swarm Optimization (DM-PSO) is used to gain power efficiency. In our work we have consider a set of nodes, assume they are grouped randomly into some three swarms. Let each swarm consist of servers, the waiting time in each server calculated as local best (LB). The local best of each node in the swarm is compared with its neighbor node in the same swarm to identify the servers which is having least average waiting time and the highest average waiting time.

According to DMS-PSO the swarm is regrouped into different swarms dynamically based on some factors. Fig 2 depicts Dynamic Multiswarm –Particle Swarm Optimization. Appling DMS-PSO, in our work the servers of different swarms having least range of average waiting time are dynamically group to single swarm, the servers of different swarm having highest range of waiting time will be grouped into one more different swarm and later the local best in between least and highest range will be grouped in to one group.

The swarms created by least range of average waiting time will be global best. So as the request load on server varies, switching the server between on and off state maximizes the power consumption. Instead client request should be serviced by the swarm having global best. This way global best swarm will be fed with further requests and make the best utilization of the server efficiency.

IV. EXPERIMENTAL SETUP AND RESULTS

To work with the proposed idea, we have consider 12 servers initially they are grouped into swarm of 4 nodes and assume queue size be 3. With the different pattern of arrival rate (λ) and the service rate be (μ=5) the system is formulated to find the waiting time in each server.

The system waiting time and utilization is tabulated below shows the expected results for resource allocation in the cloud data center for efficient power utilization by combining M/M/c/K queuing model and DMS-PSO.

Table 4.1: Waiting time and system utilization at cloud data center with M/M/c/K queuing model

Arrival rate (λ)	**service waiting time**	**System utilization(%)**
19	-18.75	0.66
20	-17.85	0.66
21	8.56	0.57
22	2.45	0.63
23	0.845	0.65
24	0.376	0.69

Negative system waiting time indicates unstable system but later is reaches 0.3 and system utilization is graduallsy increased. The system utilization shows too variations initially however later it projected with idle values.

Table 4.2: Waiting time and system utilization combine M/M/c/K queuing model and DMS-PSO

Arrival rate (λ)	service waiting time	System utilization(%)
19	0.1515	0.80
20	0.149	0.833
21	0.1459	0.85
22	0.1426	0.867
23	0.1391	0.883
24	0.1355	0.9

The results shows the service waiting time as gradually decreased by appending DMS-PSO with queuing model for efficient power utilization.

V. CONCLUSION

In this paper, we have analyzed the issue of service request allocation among servers in data center. As the increase in service request the cloud data centre is featured with queuing model for efficient resource allocation. The M/M/c/K queuing model is applied for allocating request for servers. In DMS-PSO the particles are divided into sub swarms based on particle best and the global best is calculated by regrouping the swarms, finally the request is sent to swarm with lowest global best value which means swarm with lowest waiting time. Collaborating M/M/c/K queuing model and DMS-PSO, once the swarm of servers is identified with global best the next upcoming request will be sent to them directly without any lay back, till those servers become busy. Similarly further regrouping takes places, so that for certain amount of time the distribution calculation can be with held. In our future work we are trying to implement on

realistic approach by conducting simulation and develop algorithm to find optimization solution mathematically, with large number of heterogeneous servers.

REFERENCES:

[1] US EPA, Report to congress on server and data center Energy efficiency, 2007.
[2] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud : research problem in data center networks", SIGCOMM Computer Communication Revised ,vol 39,no 1,pp-89-73,2009.
[3] T.Benson,A.Akella and D.Maltz,"Network traffic data center in the wild", in internet measurement conferences(IMC), Melbourne Australia, November 2010.
[4] M. Mazzucco, D. Dyachuk and R.Deters,"Maximizing cloud providers revenue via energy aware allocation policies", IEEE internaitional conference on cloud computing, Miami, Florida, July 2010, pp. 131-138.
[5] D.Dyachuk and M.Mazzucco,"On allocation policies for power and performance" 11th ACM/IEEE International Conference on Grid Computing (Grid 2010) - Energy Efficient Grids, Clouds and Clusters Workshop (E2GC2-2010), pp.313-320,Brussels, Belgium, October 2010.
[6] A. Gandhi, V. Gupta, M. Harchol-Balter, M. Kozuch." Optimality analysis of energy-performance trade-off for server farm management". In: Proceedings of the 28th Performance, 2010.
[7] S. Z. Zhao1, J. J. Liang1, P. N. Suganthan1, Snr Member, IEEE and M. F. Tasgetiren " Dynamic Multi-Swarm Particle Swarm Optimizer with Local Search for Large Scale Global Optimization" vol-978-1-4244-1823-7, IEEE Congress on Evolutionary Computation(CEC) 2008.]
[8]MahmoudAl-Ayyoub,MuneeraAlzuraan,YaserJararweh,ElhadjBenkhelifa,SalimHariri, "Resilient service provisioning in cloud based data centers" , journal -Future Generation Computer Systems Volume 86, September 2018, Pages 765-774.
[9] FUCHS, H SHEHABI, A GANESHALINGAM, M "CHARACTERISTICS AND ENERGY USE OF VOLUME SERVERS IN THE UNITED STATES", POWERED BY UNIVERSITY OF CALIFORNIA, ON 2017-11-01
[10] SPASH MITTAL," POWER MANAGEMENT TECHNIQUES FOR DATA CENTERS: A SURVEY", AT RESEARCHGATE ,JULY 10 2014.
[11] JIANG ET AL.,"ENERGY COST MINIMIZATION FOR INTERNET DATA CENTERS CONSIDERING POWER OUTAGES", SPRINGER-VERLAG BERLIN HEIDELBERG, ENERGY MANAGEMENT OF INTERNET DATA CENTERS IN SMART GRID, GREEN ENERGY AND TECHNOLOGY, 2015.
[12] R. C. EBERHART, AND J. KENNEDY, "A NEW OPTIMIZER USING PARTICLE SWARM THEORY", PROC. OF THE SIXTH INT. SYMPOSIUM ON MICROMACHINE AND HUMAN SCIENCE, NAGOYA, JAPAN. PP. 39-43, 1995.
[13] J. Kennedy, and R. C. Eberhart, "Particle swarm optimization ".Proceedings of IEEE International Conference on Neural Networks,Piscataway, NJ. pp. 1942-1948, 1995.
[14] S. Z. Zhao , J. J. Liang , P. N. Suganthan , Snr Member, IEEE and M. F. Tasgetiren," Dynamic Multi-Swarm Particle Swarm Optimizer with Local Search for Large Scale Global Optimization", 978-1-4244-1823-7/08, IEEE World Congress on computational Intelligence, 2008.
[15] Liang, J. J., & Suganthan, P. N. (n.d.)." Dynamic multi-swarm particle swarm optimizer", Proceedings 2005 IEEE Swarm Intelligence Symposium, 2005. SIS 2005. doi:10.1109/sis.2005.
[16] Vetha S and Vimala Devi K ," Dynamic resource alloxation in cloud using Queuing Model", Jr. of Industrial Pollution Control 33(2)(2017) pp 1547-1554,2017.
[17] Dan Liao, Ke Li, Gang Sun, Vishal Anand, Yu Gong, Zhi Tan," Energy and performance management in large data centers: A queuing theory perspective" ,IEEE Internaltional conference on computing , Networking and communications(ICNC), Workshop on computing , networking and communications , 2015
[18] Vetha.s and Vimala Devi K," Dynamic Resource Allocation in Cloud using Queuing Model", Journal of Industrial Pollution Control, 33(2)(2017) pp 1547-1554