Sindh Univ. Res. Jour. (Sci. Ser.) Vol.49(3) 529-534 (2017)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Analytical Model of Delayed Server System (DSS) for Energy Conservation

S. SARWAR, H. KHALID, L. ASLAM, W. U. QOUNAIN, M. M. YOUSAF

Punjab University College of Information Technology (PUCIT) University of the Punjab, Lahore

Received 7th March 2016 and Revised 9rd May 2017

Abstract: This paper proposes a delayed server system (DSS) which regulates the number of servers in ON-state, and corresponding OFF-state servers, on the basis of queue occupancy for the sake of energy conservation. Upon arrival of demand, server allocation is delayed if predefined servers are already busy and, thus, customer is going to wait in the queue before it gets service. Hence, DSS allows us to keep some servers at least OFF for fraction of total operational, time resulting in energy conservation. We have developed the Markov chain models for DSS and M/M/n-s systems. Using the Markov chain models, the numerical evaluation has been carried out for the key performance metrics such as blocking probability and average energy conservation. Furthermore, our study has investigated and figured out the optimal configurations of delayed server system in order to maximize the energy conservation.

Keywords: Energy Conservation; DPM; DVFS; Virtulization; Analytical Model; Markov Chain Model; Performance Evaluation

1. INTRODUCTION

In recent years, electricity consumption has grown exponentially due to phenomenal growth in the sectors of Information and Communication Technology (ICT). So far worldwide, the consumption of ICT sector is roughly 3% of the total production of electricity. Also, it is expected that the consumption of electricity in ICT would be increasing at least at the rate of 20% every year. (Koomey 2008)

Besides electricity consumption and its associated cost, it is implied that ICT infrastructure is responsible for depleting both natural resources and the ozone layer. In the process of electricity generation CO_2 is emitted that attacks and destroys the Ozone layer. A less energy consumption simply means a small carbon footprint or vice versa. Thus, energy conservation mechanisms are in the focus in order to mitigate the potential negative impact on the natural resources and the Ozone layer. In the data centers, besides these environmental effects, the operational cost is high due to the excessive cooling requirements.

Electricity consumption is the major operational cost of ICT server- and network-infrastructure. For instance, data centers are an essential component of today's ICT which have grown in terms of size and population density resulting in increased electricity consumption. Typically, data centers are meant to support cloud computing and to host servers of Internet applications which are designed using client-server architecture. Usually, large-scale data centers are designed according to three-tier high-speed architecture that entails core-, aggregation-, and access-layer. A maximum of 4000 servers can be there in a L2 layer domain. Additionally, in order to provide fault tolerance, multiple communication links and switches between every pair of connected layers and back-up internal servers are used, at each layer. High energy consumption translates into more heat dissipation which in turn requires excessive cooling. Furthermore, inside densely populated data centers hefty cooling units keep temperature to an acceptable level at the price of higher energy consumption and carbon footprint. In this scenario, the cost of cooling escalates to fifty percent of total cost of energy consumed in a data center (Wang *et al.,* 2011) In order to reduce this cost of cooling various proposals have been made in the literature (Wang *et al.,* 2011) (Wang *et al.,* 2010)

In several studies of data centers, it has been projected that strategies should be developed for the sake of electricity conservation.

2. BACKGROUND AND RELATED WORK

During the design phase of a data center the aspect of energy efficiency receive little or no attention while focus remains on performance metrics. The energy conservation strategies can be applied at data centers with small overhead of application. Energy efficiency techniques can be classified broadly in three categories a) dynamic methodology of power management (DPM), b) dynamic voltage and frequency scaling (DVFS), and c) virtualization (Masanet *et al.,* 2011) In the following text, a brief description of DPM, DVFS, and virtualization is presented.

++Corresponding Author: Email: {s.sarwar, hafsa, laeeq.aslam, swjaffry, murtaza}@pucit.edu.pk

Dynamic Power Management (DPM)

The DPM methodology provides a conceptual framework for designing a dynamic reconfigurable system to cater desired level of service and performance by using minimum resources. To achieve energy efficiency, numerous techniques of turning off idle system components constitute the DPM methodology. A theoretical and experimental framework for hierarchical autonomic power management in large scale data center has been proposed in (Khargharia *et al.,* 2007) The design techniques of system-level DPM have been collected in (Benini *et al.,* 2000). The basic premise is that system experiences non-uniform workload during its operation. In order to employ DPM, future workload needs to be computed in advance which can be estimated using probabilistic techniques using the history of workload. Two major drawbacks can be attributed to DPM a) future workload is estimated with some degree of uncertainty b) workload observation and estimation consumes significant energy.

Dynamic Voltage/Frequency Scaling (DVFS)

This technique is based on the fact that power consumption of a processing-chip depends on supplied voltage and frequency of clock. Power consumption is described by $P = V^2f$, such as V and f represent supplied voltage and frequency of clock, respectively. Therefore, power consumption can be reduced by downscaling supply voltage or clock frequency. Most of the manufacturers have adopted the OS-independent advanced-configuration and power-interface (ACPI) standard. In literature, a DVFS technique has been proposed for multi-tier web server architecture (Horvath *et al.,* 2007) Where, higher-tier servers invoke services of lower-tier servers and DVFS technique reduces 30% electricity consumption. It is assumed that constraints of end-to-end delay are soft. The DVFS can be applied until the end-to-end delay surpasses a threshold value.

The down side of the DVFS consists of three factors those are a) reduction in frequency only applies to the power of CPU whereas memory, bus, and disk are independent of CPU frequency, b) support of hardware is mandatory, and c) in DVFS schemes, most are dependent on the implementation of the ACPI standard.

Virtualization

Virtualization allows sharing of a single physical server among multiple virtual machines (VMs). Virtualization enables us to serve a set of different applications on each virtual machine and CPU and memory resources can be on-demand dynamically provisioned for a virtual machine, according to the desired level of performance for an application. Thus, virtualization is best suited for energy conservation in a data center.

For the sake of consolidation of partially loaded physical server, (Beloglazov *et al.,* 2010). have proposed that live migration of VMs that can be used for converging jobs on minimum servers so that remaining idle servers can be put in energy saving mode (Beloglazov and Buyya (2010) Live migration of VMs between servers generates significant amount of traffic which burdens the network resources (Stage and Setzer (2009) However, a VM placement strategy has been proposed that minimizes the distance between servers participating in live migration (Meng *et al.,* 2010) Similarly, the migration manager optimally schedules the live migrations without causing congestion in the communication network.

Collectively, the ICT equipment and the data center face an average workload of 30%. Here the percentage is of the peak workload (Liu *et al.,* 2009) Thus, it clearly indicates that a technique for energy conservation needs to be developed that does not a) rely on observation and estimate of traffic load, b) require hardware support, and c) increase network utilization.

In this paper, we propose an energy-conservative *delayed server system* (DSS). Upon arrival of a demand, the server allocation may be delayed until the waiting demands reach to a certain threshold in the queue. It is assumed that delay constraint is soft and delay caused by inter-server communication is negligible. In other words, DSS regulates access to servers based on the queue occupancy. The DSS is a modified form of the M/M/n-s system,

3. SYSTEM DISCRIPTION, QUEUING DIAGRAM, MARKOV CHAIN MODEL, AND PERFORMANCE METRICS

Let us consider a multi-server delay-loss system to cater total offered traffic *A*. The system is comprised of a single queue with capacity of holding *s* demands and there are *n* server-nodes. The system configuration is presented in the **(Fig. 1).** The server-nodes would be referred as servers in the remaining text. A First-In First-Out (FIFO) queue is maintained. Furthermore, waiting demands are assumed to be non-reneging, meaning they do not give up before the service.

In this paper, we propose an energy-conservative *delay server system* (DSS) which regulates access to servers based on the queue occupancy. Let the queue occupancy, lower-threshold, and higher-threshold are represented by *qo*, th_1, and th_2, respectively. It is assumed that $th_1 < th_2$.

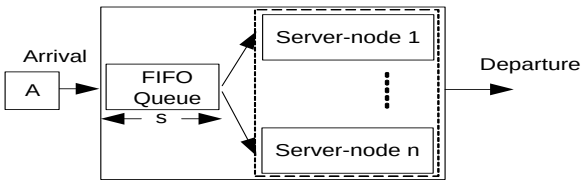


Fig. 1: System model

If queue occupancy is less than the *higher-threshold*, only a fraction of *n* servers are in ON-state. Otherwise, all servers are in ON-state. However, after reaching to *higher-threshold*, if queue occupancy drops back to *lower-threshold*, only limited servers would remain in ON-state for servicing the present demands.

The M/M/n-s is a delay-loss system; all *n* servers are accessible to the demands of offered load A, without any restriction of queue occupancy. In other words, server is immediately allocated on the arrival of a demand, if at least one server is idle. Such systems have been studied widely and are well known as M/M/n-s queuing systems. In the following text, for both M/M/n-s and DSS, the queuing diagram and corresponding Markov chain model have been developed and presented.

A. Queuing Diagram

In the Fig. 2, the queuing diagrams of M/M/n-s and DSS are presented. A Poisson traffic source is assumed with mean service time h and λ as total mean arrival rate. The λ is negative exponentially distributed. Also, the population of demands is infinite. The mean service rate is represented by μ such that $\mu = 1/h$.

M/M/n-s System

This is a delay-loss system for traffic flow A with full access to all *n* number of servers and *s* waiting places in the queue. Demands of the traffic flow can only be served immediately if at least one server is idle, out of *n* servers in **(Fig. 2(a).** Otherwise, arriving demands are queued. However, if queue is full then the new arrivals are blocked and consequently are lost.

Delayed Server System (DSS)

Initially, the proposed system is in energy conservation mode, when only limited servers (kn, $0<k<1$) are in ON-state. Initially, kn servers can be allocated to cater the arriving demands, as shown in the **(Fig. 2 (b).** Demands of the traffic flow can only be served immediately if out of kn servers at least one server is idle. Otherwise, arriving demands are queued and it is interesting to note that some of the servers are still unoccupied.

If the queue occupancy reaches to the higher threshold th_2, remaining servers ($n–kn$) are additionally made available, as shown in the **(Fig. 2 (b).** That means, now all the *n* servers are available to serve the incoming customers. Arriving customer would instantly occupy a server if at least one server is idle out of ($n–kn$) ON-state servers. Otherwise, waiting places would be occupied by the arriving customers. While the customers in the queue are being served by limited (kn) servers. If all n servers are busy, waiting places would be occupied by the arriving demands. When the queue is full, the new arrivals are blocked.

Afterwards, in contrast, if queue occupancy drops to the lower-threshold th_1, the additional servers can be switched off, one-by-one. This would again bring the system into the initial energy conservation mode where only limited (kn) servers are in ON-state and queue occupancy is less than equal to the lower-threshold th_1.

B. Markov Chain Model

In this section, the Markov chain models are presented for M/M/n-s and DSS. The implementation of these models would enable us to study the system for various performance metrics.

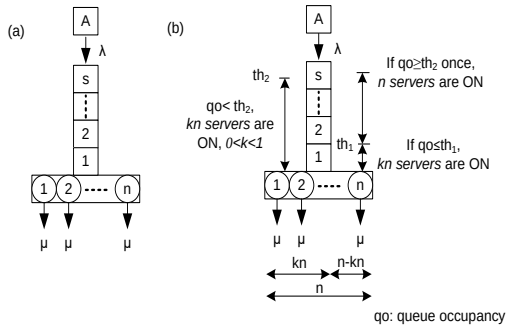


Fig.2. Queuing diagram
(a) M/M/n-s and
(b) Delayed server system (DSS)

M/M/n-s

According to the queuing diagram presented in Fig. 2 (a), a state of the system is defined by two parameters. A system state is represented as (y_1, y_2). The number of busy servers is denoted by y_1, $0 \le y_1 \le n$. The number of waiting places occupied by arrivals of flow *A* is denoted by y_2, $0 \le y_2 \le s$. The waiting places are occupied by demands of *A* if no server is idle. That means states do not exist with $y_1 < n$ and $y_2 > 0$.

The one-dimensional Markov chain model of M/M/n-s is shown schematically in the **(Fig. 3(a)** along with its states, transitions, and their rates. The steady state probability for an existing state is denoted by $p(y_1, y_2)$.

Delayed Server System (DSS)

According to the queuing diagram of DSS, (Fig. 2 (b), the states of the system is defined with three parameters. Each state can be represented as state (z_1, z_2, z_3). The number of busy servers is denoted by z_1, $0 \le z_1 \le n$. The number of waiting places occupied by arrivals of flow *A* is denoted by z_2, $0 \le z_2 \le s$. The waiting places would be occupied only by demands of *A*, if busy servers are at least kn. Thus, system states are non-existent where $z_1 < kn$ *and* $z_2 > 0$. However, if queue occupancy reaches the threshold th_2, additional servers, equal to ($n–kn$), are made available to serve the arriving customers. If queue occupancy is greater than the threshold th_2 ($z_2 \ge th_2$ and $z_1 = n$) then the value of third parameter is one ($z_3 = 1$) represents that queue

occupancy has surpassed the threshold th_2 and all n servers will remain in ON-state unless queue occupancy drops down to the threshold th_1. The corresponding two-dimensional Markov chain model is shown schematically in the **(Fig. 3(a,b).** Further, let p(z_1, z_2, z_3) represents steady state probability.

The steady state probabilities have been computed by solving the system of linear equations using the iterative successive-over-relaxation (SOR) (Collins 2003): with the help of software implemented in C++, which takes as input states, permissible transitions, and transition rates of the Markov chain model.

C. Performance Measures

The significant performance measures are the blocking probability and average energy conservation.

Blocking Probability

Let B is the blocking probability of traffic flow A. According to the system description, blocking occurs when one arriving demand neither can occupy a server nor a waiting place in the queue. The blocking probabilities can be obtained for M/M/n-s and DSS using the Equations 1(a) and 1(b), respectively.

$$B_{M/M/n-s} = p(n, s) \qquad 1(a)$$

$$B_{DSS} = p(n, s, 1) \qquad 1(b)$$

Energy Conservation

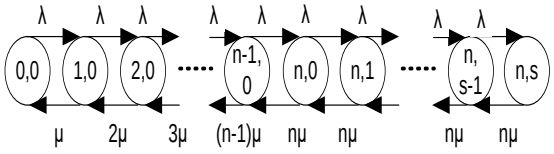
For DSS, it is evident that total energy conservation is comprised of two factors. First, when limited servers are in ON-state and additional servers are in OFF-state, thus, taking product of probability of only limited servers in ON-state and number of OFF-state servers gives us expected number of servers that would remain OFF, denoted by $E[EC^{\mathrm{kn}}]$, and correspondingly it is equivalent to energy conservation, (Fig. 4). Second, when some of additional servers ($n–kn$) are also in ON-state, the product of probability that at least m servers are in ON-states, $kn+1 \le m \le n–1$ and count of additional servers in OFF-state that is ($n–m$), is denoted by $E\left[EC^{\mathrm{kn+1}}\right]$, **(Fig. 4).**

Let energy conservation is denoted by EC. For a given load, EC and can be obtained by Equation 2.

$$EC_{GivenLoad} = E[EC^{kn}] + E[EC^{kn+1}] \qquad 2$$

$$E[EC^{kn}] = \sum_{x_1=0}^{kn} \sum_{x_2=0}^{th2} \frac{100(n-kn)}{n} p(x_1, x_2, 0) \qquad 3$$

Such that $E\left[EC^{\mathrm{kn}}\right]$ and $E\left[EC^{\mathrm{kn+1}}\right]$ can be obtained by the Equations 3 and 4, respectively. (a)



(b)

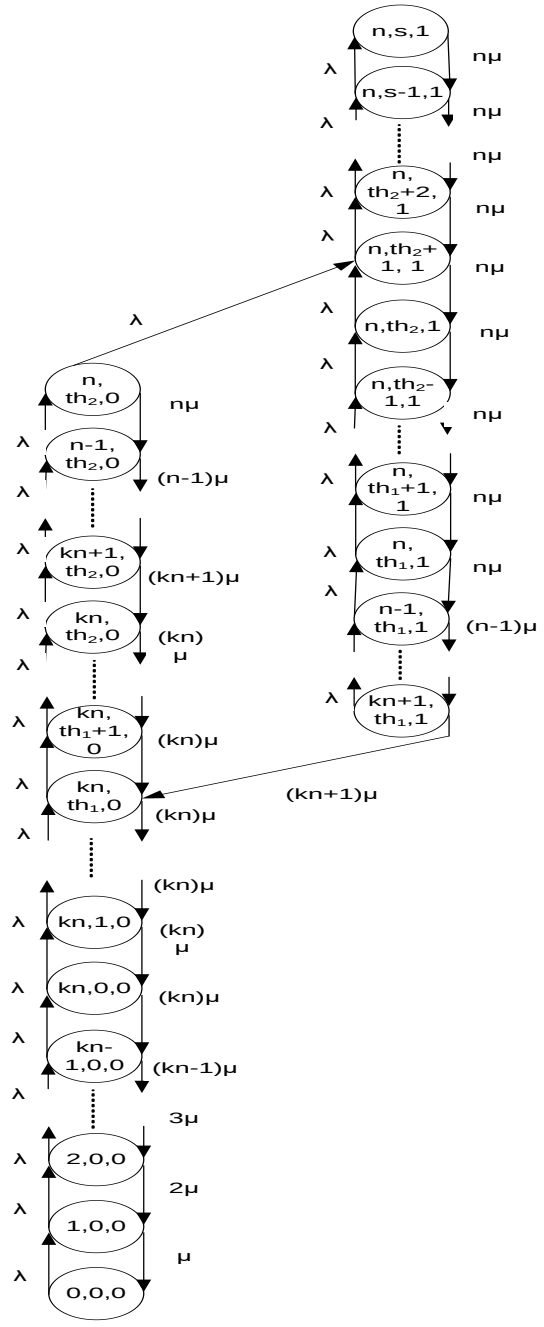


Fig. 3. Markov chain models (a) M/M/n-s and (b) delayed server system (DSS).

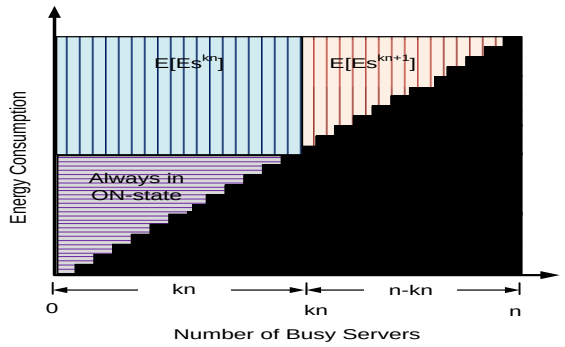


Fig. 4. Number of busy servers versus energy consumption

For a given system configuration and under the assumption that traffic load uniformly varies from 0.1 to 1.0 Erlang, average energy conserved in terms of percentage of total energy consumed is denoted by AEC and given by the Equation 5. It can be assumed that some energy penalty has to pay in order to bring a server in ON-state from OFF-state. Thus, energy conservation obtained by Equation 2 would be an overestimate if the energy penalty is not accommodated. We justify this energy penalty by reducing the energy conservation equivalent of expected energy conservation in case there are at least kn+1 or more servers are in ON-state that is $E[EC^{kn+1}]$, as shown in the Equation 4. It is trivial that energy penalty cannot be more than the energy conserved by keeping some servers in OFF-state.

$$AEC = \frac{EC_{0.1}+EC_{0.2}+\ldots+EC_{1.0}}{DataSamplesCount} \quad 5$$

4. NUMERICAL RESULTS AND DISCUSSION

The modification of M/M/n–s system and transforming it into delayed server system (DSS) has led us to raise the most pertinent following two questions.

Table 1: System configuration for DSS

Scenario no. 1, Question no. 1								
Number of servers = n = 50 and Size of the Queue = s = 1.2n Percentage of ON/OFF servers = 50/50								
th_1	th_2	*Number of servers = n*						
0.5s	0.75s	10	20	30	40	50		
Scenario no. 2, Question no. 2								
Number of servers = n = 50 and Size of Queue = s = 1.2n = 60								
th_1	th_2	Percentage of ON/OFF servers = (100*kn/100(1-kn)*)						
0.5s	0.75s	20/80	30/70	40/60	50/50	60/40	70/30	80/20

Question no. 1. How many numbers of servers (n) should be there, more or less?
Question no. 2. What should be the percentage of servers that are initially in ON and OFF states, represented as ON/OFF?

The answers of these questions would enable us to configure the system for achieving optimal energy conservation. In order to answer these questions, we have devised two scenarios corresponding to these questions. For numerical evaluation, values of system parameters are given in **(Table 1).**

In order to find the affect of increasing number of servers on the blocking probability, difference of blocking probabilities of delayed server system (DSS) and M/M/n-s systems is presented in **(Fig. 5).** Delayed server system (DSS) offers less additional blocking in comparison to the M/M/n-s if there are more servers, that is n = 50. Simply, increase in the number of servers suggests that DSS starts behaving close to M/M/n-s.

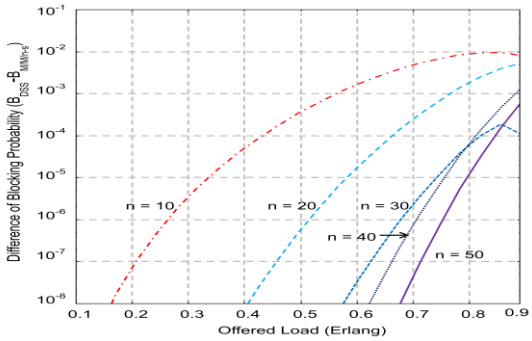


Fig. 5. Difference of blocking probabilities of M/M/n-s and delayed server system (DSS) under varying offered load

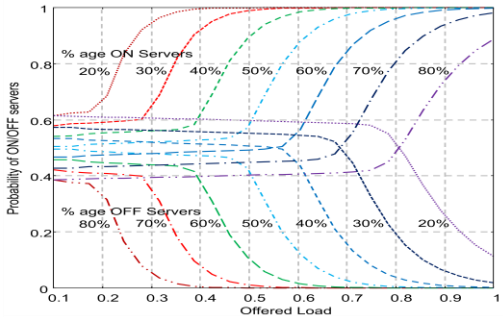
However, one question needs to be answered that what percentage of ON/OFF servers would yield the highest energy conservation. Firstly, probabilities of limited ON/OFF servers are computed using the Equation 6(b), secondly, energy conservation under a given load is obtained using the Equation 6(a) and are shown in the **(Figs. 6(a) and 6(b)),** respectively. Average energy conservation (AEC) is obtained using the Equation 7 and is presented in the **(Fig. 6(c)).**

In the Fig. 6(a), it is observed that by increasing the offered load probabilities of ON-state and OFF-state servers tend to move in opposite direction. Similarly, it is observed that probability of bringing the OFF servers into ON-state increases if offered load is increased.

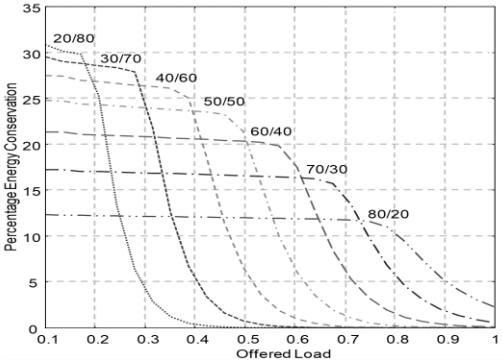
In the Fig. 6(b), it is evident that energy conservation is inversely proportional to the offered load, as given in Equation 5. It can be concluded that DSS conserves more energy under low arrival rate of traffic. However, it would be unrealistic to assume that traffic load remains always low. Thus, it is assumed that traffic load varies in the range from low to high, 0.1 to 1.0 Erlang.

Further, the average energy conservation is obtained by the Equation 8 and presented in Fig. 6(c), for various

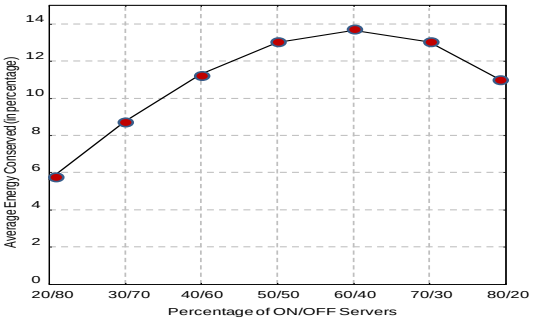
(a)



(b)



(c)



(Fig. 6. (a) Probability of ON/OFF servers versus offered load under varying percentage of ON/OFF servers (b) Energy conservation $EC_{GivenLoad}$ versus offered load for various percentage of ON/OFF servers (c) Average energy conserved versus percentage of ON/OFF servers system configurations in terms of percentage of ON/OFF servers. It is observed that the configuration with 60%-ON and 40%-OFF servers results in the maximum average energy conservation; see Fig. 6(c). Therefore, it is concluded that 60/40 in the most beneficial configuration, which answers the Question no. 2, as well.

5. CONCLUSIONS

For multi-server delay-loss system, we have studied the behavior of delayed server system inspired from the M/M/n-s queue. This study has revealed answers to several important questions related to the optimal configuration of the system. It has been established that by increasing the number of servers the delayed server system would offer blocking probability similar to the M/M/n-s system. Further, this study has showed that the optimal ratio of ON/OFF servers is 60/40.

REFERENCES:

Benini L, A. Bogliolo, G. De Micheli (2000) A survey of design techniques for system-level dynamic power management. IEEE transac. on very large scale integration (VLSI) systems, 8(3), 299–316

Beloglazov A., R Buyya (2010) Energy efficient resource management in virtualized cloud data centers. 10th IEEE/ACM international conference on cluster, cloud and grid computing (CCGrid), 826–831

Collins II. G. W., (2003) Fundamental Numerical Methods and Data Analysis, Harvard-Smithsonian Center for Astrophysics, USA.

Horvath T, T. Abdelzaher, K. Skadron X. Liu (2007) Dynamic voltage scaling in multitier web servers with end-to-end delay control. IEEE Trans Comput 56(4):444–458.

Koomey J. G. (2008) Worldwide electricity used in data centers. Environ Res Lett 3(3):1–8 (IOPscience)

Khargharia B., S. Hariri, F. Szidarovszky, M, El-Houri H, Rewini I. Ahmad, M Yousif (2007) Autonomic power performance management for large-scale data centers. In: IEEE international symposium on parallel and distributed processing symposium, IPDPS 1–8

Liu J., F. Zhao, X. Liu, W. He (2009) Challenges towards elastic power management in internet data centers. In: 29th IEEE international conference on distributed computing systems workshops, ICDCS workshops '09, 65–72.

Masanet E, R Brown, B. Nordman (2011) Estimating the energy use and efficiency potential of U.S. data centers. Proc IEEE 99(8):1440–1453

Stage A., T. Setzer (2009) Network-aware migration control and scheduling of differentiated virtual machine workloads. Proc. of the 2009 ICSE workshop on software engineering challenges of cloud computing, IEEE Computer Society, Washington, DC, 9–14.

Wang L, S. Khan J. Dayal (2011) Thermal aware workload placement with task-temperature profiles in a data center. J Supercomput 61:780–803

Wang L, G.von Laszewski F. Huang J. Dayal T. Frulani G. Fox (2011) Task scheduling with ann-based temperature prediction in a data center: a simulation-based study. Eng Comput (Lond) 27(4):381–391

Wang L, C. Fu (2010) Research advances in modern cyberinfrastructure. N Gener Comput 28:111–112

Wang L, G. von Laszewski M. Kunze J. Tao (2010) Cloud computing: a perspective study. N Gener Comput 28(2):137–146