Journal of ICT, 10, pp: 1–13

OPTIMIZING WORKLOAD ALLOCATION IN A NETWORK OF HETEROGENEOUS COMPUTERS

Rahela Rahim[1] and Ku Ruhana Ku-Mahamud[2]

Division of Physical Sciences and Graduate Department of Computer Science
College of Arts and Sciences
Universiti Utara Malaysia

rahela@uum.edu.my[1]
ruhana@uum.edu.my[2]

ABSTRACT

The allocation of workload to a network of computers is investigated. A new workload allocation model based on Generalized Exponential (GE) distribution is proposed for user-level performance measures. The criterion used for effective workload allocation is the one that minimizes the expected response time in systems to which jobs are routed. A closed-loop expression for workload arrival to minimize systems means queue length and response time are derived using the optimization technique. Results are presented with numerical examples and sensitivity analysis with respect to changes of total workload. Results are verified using the simulation technique.

Keywords: Workload allocation, Multi server queuing system, Optimization, Generalized exponential distribution.

INTRODUCTION

Recently process improvement has been given a lot of attention. Since then many modelling techniques and tools have been used to support the effort. However most of the currently available tools use static models such as diagrams to model such processes. Some are quite dynamic where the functional aspects of the process have been modelled using simulation. Furthermore feedback to the designer concerning process functional and alternative design option should be done at early the process design. At this stage, analytical modelling provides quantitative properties, whereby these will provide the global indication of the expected performance. In the final stage,

more accurate predictions may be required to fine-tune the designs; therefore, the analytical modelling proposed here is at the highest abstraction level of the process design, i.e. to get the initial idea on the process performance.

In this paper, we stress the quantitative measures of the processes in a network of computers to those of the concurrent discrete-event system. Based on this, we show that by using quantitative modelling, arrival to computers can be reallocated to get optimal performance measures. We focus on the issue of job allocation in a network of computers where different computers have different job processing times. The optimization criterion studied here is to minimize the expected job-response time in the systems to which jobs are allocated. Jobs arrive at a scheduler that allocates jobs to the computers according to a pre-calculated arrival rate using the optimization method.

RELATED WORK

The problem of workload allocation is common to a variety of communication systems especially when it involves a network of computers. Workload allocation seeks to allocate job arrival among computers as evenly as possible. In a parallel setting, where jobs may have many possible paths at the job's scheduler, the job allocation problem is of interest. For each job entering the scheduler, a path is assigned to optimize the allocation of workload. Many studies considered developing a closed-loop expression for service rate (Hsiao & Lazar, 1990; Harrison & Patel, 2000; Gunther, 2000), and studies concerned with optimizing the allocation based on the total arrival rate at the service centers are quite recent (Rahim & Ku-Mahamud, 2006; 2008). For networks of computers, the workload-allocation problem is of particular interest, since there are several ways to affect the distribution of workload among computers. In general, network traffic is assumed exponentially distributed (Gelenbe & Mitrani, 1980; Bennani & Menasc´e, 2005). The general exponential (GE) distributions for traffic arrival and service time have been considered (Rahim & Ku-Mahamud, 2002; 2006; 2008; 2010), as these types of distribution posses flexible parameters.

Queueing network models have been recognized as powerful tools for evaluating the performance of computer systems (Allen, 1990; Smith & Williams, 2001) and the communication network (Lazar, 1982; Koavatsos & Othman, 1989a; 1989b; Koole, 1999; Boxma, 1995). These analytical models have become very important tools for predicting the behaviour of new designs or proposed changes to existing systems (Koavatsos, 1985; Menasce & Almeida, 2000; Urgaonkar, Pacifici, Shenoy, Spreitzer & Tantawi, 2005). Most queueing network models are used either by making assumptions to assure

exact numerical solution or by employing approximate methods (Kobayashi, 1974; Lazar, 1983; Koavatsos & Othman, 1989b). The control of arrivals to a network of queues with the objective of maximizing throughput subject to a response-time constraint has been considered (Kleinrock, 1975; Ross & Yao, 1991; Combe & Boxma, 1991; Hsiao & Lazar, 1991). A throughput time-delay function based on an optimality criterion has been developed (Kleinrock, 1975; Hsiao & Lazar, 1991) where the arrival that maximizes the throughput under the constraint of the average response time will not exceed a preassigned value. Then Ku-Mahamud, (1993) continued with the problem of random routing. All these literature have been devoted to the probabilistic analysis of the queueing system; their optimization is somewhat lagging behind. Only recently, optimization problems related to the network of queues have been studied for instance by Lazar (1981; 1984), Tantawi and Towsley (1985), Harrison and Patel (1992), Koole (1999), Liu (1999), Jongh (2002), Srikant (2004), Felegyhazi and Hubaux (2006), and Rahim, Ibrahim, Syed Yahaya and Khalid (2010). Most of the studies focus on reducing the amount of waiting time in a system with several servers either parallel or serial. However none of the studies consider the impact of jobs inter-arrival and service-time variation (CV's) in modelling the systems performance. Without considering the effect of variation in measuring, systems performance may lead to inaccurate results. This has somehow motivated this study, that is, to develop a predictive analytical model which can optimize the systems performance and consider data variation.

MULTISERVER QUEUEING SYSTEM MODEL

When several users compete for the use of a common resource, the limited capacity of the resource can give rise to congestion, hence queueing is a common phenomena. Queueing occurs normally when the demand exceeds the service capacity of the resource and even when the otherwise occurs. This is due to the fact that the inter-arrival times of the users, and their required service times, are generally not fixed; therefore, a mathematical model of congestion phenomena represents the inter-arrival and the service-times of the users by random variables. The Queueing Theory is devoted to the description, analysis and optimization of such a queueing system (Lazar, 1981). It focuses on a few key performance measures, like queue lengths and waiting times. Due to the stochastic nature of the arrival and service processes, and of the routing process of jobs through a network of queues, the main performance measures are also random variables. With this in mind we use the multiple-queue multiple server model to represent a central job routing system which is shown in Figure 1.

In using this model, hardware resources are represented by service centers at which jobs queue and compete for service. The workload is modelled as a single stream of jobs (file request), with total arrival ϕ. Each newly-arrived job, is assigned to computer *i* according to a new arrival rate λ_i which is a fraction of the total arrivals. We consider the set of computers to be heterogeneous as this is common in real systems and also it can be generalized to homogeneous servers. In the context of general queueing network models, the generalized exponential (GE) distributional model is of the form;

$$f_S(t) = \left[\frac{C^2 - 1}{C^2 + 1}\right] u_0(t) + \frac{4\mu}{\left(C^2 + 1\right)^2} \exp\left[\frac{-2\mu t}{C^2 + 1}\right], \qquad t \geq 0 \tag{1}$$

Where μ is the mean service rate, *C* is the coefficient of variation and $u_0(t)$ is the unit impulse function, which has been used to represent the inter-arrival and service-time distributions. This model is robust and versatile due to it memoryless properties and has been shown to maximize the entropy function subject to mean value constraints. Furthermore it can be shown that the exact mean number of jobs in the GE/GE/1 queue as given by Liu (1999):

$$L = \frac{\rho}{2}\left(1 + \frac{C_a^2 + \rho C_s^2}{1 - \rho}\right), \quad \text{for} \quad \frac{1 - C_a^2}{1 - C_s^2} \leq \rho < 1 \tag{2}$$

where C_a^2, C_s^2 are the squared coefficients of variation for the inter-arrival and service-time distributions (CV's) respectively. This means the queue length function will be used as an objective in the optimization model.

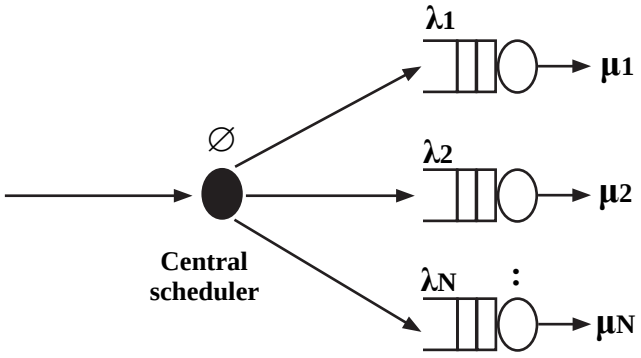

Figure 1. Multiple Queue Multiple-Server Model

OPTIMIZATION MODEL USING GENERALIZED EXPONENTIAL (GE) DISTRIBUTION

In this section, a workload allocation model for the GE type distribution system is proposed. In this case, an optimization problem of the queueing system can be generalized to a number of arrival and service distributions by configuring the value of coefficient of variation for inter-arrival and service time.

We formulated an optimization problem of the N GE/GE/1 queueing system as below:

P1 Min

$$\sum_{i=1}^{N} L_i = \sum_{i=1}^{N} D_i\left(\frac{\lambda_i\beta_i}{1-\lambda_i\beta_i}\right)\left(\lambda_i\left(\frac{C_{si}^2-1}{2}\right)+\frac{C_{ai}^2+1}{2}\right) \quad (3.1)$$

s.t
$$\sum_{i=1}^{N}\lambda_i = \phi \quad (3.2) \quad (3.2)$$

$$0 \le \lambda_i \le \frac{1}{\beta_i}, \quad i = 1,\dots,N. \quad (3.3) \quad (3.3)$$

$$\lambda_i \ge 0 \quad (3.4) \quad (3.4)$$

$$\beta_i \ge 0 \quad (3.5) \quad (3.5)$$

where $\rho = \frac{\lambda}{\mu}$ and $\beta = \frac{1}{\mu}$

Problem P1 allows an analytical solution. Using Lagrange multiplier techniques we obtain with δ the Lagrange multiplier, the following first order Kuhn-Tucker constraints:

$$\frac{d}{d\lambda_i}\left\{D_i\left(\frac{\lambda_i\beta_i}{1-\lambda_i\beta_i}\right)\left(\lambda_i\left(\frac{C_{si}^2-1}{2}\right)+\frac{C_{ai}^2+1}{2}\right)\right\} = \delta$$

$$i = 1,\dots,N. \quad (3.6)$$

$$\sum_{i=1}^{N}\lambda_i - \phi = 0 \quad (3.7)$$

From (3.6) we find the unique optimal values

$$\lambda_i^* = \frac{1}{\beta_i}\left(1 - \left(\frac{C_{si}^2 + C_{ai}^2}{C_{si}^2 - 1 + \frac{2}{\beta_i}\delta}\right)^{1/2}\right) \tag{3.8}$$

and the Lagrange multiplier is derived by solving the constraint equation below:

$$\sum_{i=1}^{N} \frac{1}{\beta_i}\left(1 - \left(\frac{C_{si}^2 + C_{ai}^2}{C_{si}^2 - 1 + \frac{2}{\beta_i}\delta}\right)^{1/2}\right) = \phi \tag{3.9}$$

When C_{ai} and C_{si} the GE workload expression is reduced to the N-M/M/1 model. D_i is the cost associated with having one job in queue and for simplicity we assign the value of 1.

COMPUTATIONAL RESULTS

In this section, numerical results are presented to assess the credibility of the GE distribution used. For result validation, simulation models were developed to simulate the proposed arrival and service rate. The mean queue length and the mean response time obtained using simulation were compared with the results obtained using the proposed analytical model. Two configurations are shown. For the first configuration, the service rate of the tasks is assumed to be:

$$\mu_1 = 3, \mu_2 = 4, \quad C_{a_1} = 0.5, C_{a_2} = 0.3, C_{s_1} = 0.2, C_{s_2} = 0.4 \cdot$$

For the second configuration, the service rate of the tasks is assumed to be:

$$\mu_1 = 3, \mu_2 = 4, C_{a_1} = 0.1, C_{a_2} = 0.2, C_{s_1} = 0.4, C_{s_2} = 0.3 \cdot$$

Table 1

Results of the Proposed and Classical Approaches of Queuing System Mean Queue Length, W : Mean Response Time

Classical		Proposed		Classical		Proposed	
λ_1	λ_2	λ_1	λ_2	*L*	*W*	*L*	*W*
1.6	2.1	1.578	2.122	1.158	0.313	1.153	0.312
1.8	2.4	1.776	2.424	1.47	0.35	1.465	0.349
2.0	2.7	1.981	2.719	1.896	0.403	1.891	0.402
2.2	2.9	2.15	2.95	2.396	0.47	2.377	0.466
2.4	3.2	2.367	3.233	3.36	0.6	3.342	0.597
2.6	3.4	2.544	3.456	4.86	0.81	4.775	0.796

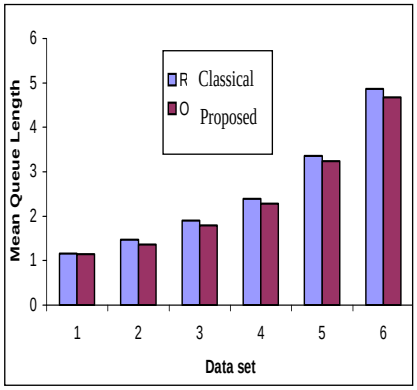

Figure 2. Performance Improvement of Mean Queue Length Simulation Result

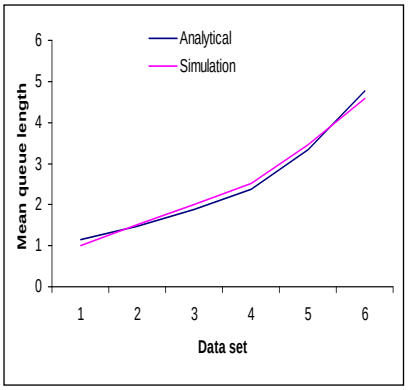

Figure 3. Analytical Versus for a Dual GE/GE/Queuing System

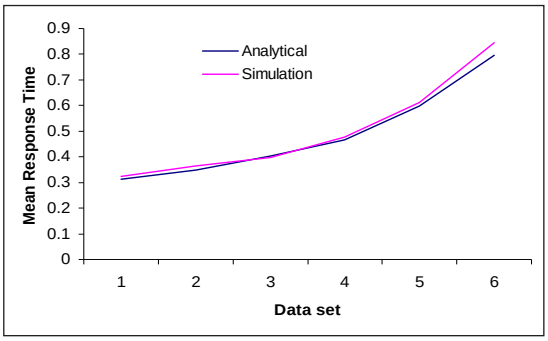

Figure 4. Analytical Versus Simulation Result

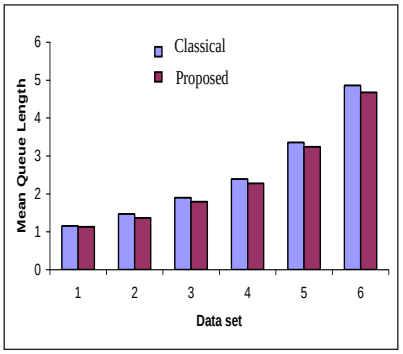

Figure 5. Performance Improvement of Mean Response Time for a Dual GE/GE/1 Queueing System

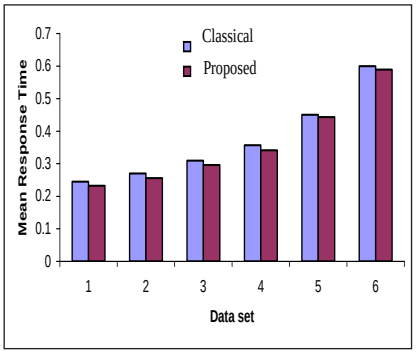

Figure 6. Performance Improvement of Mean Queue Length

Table 2

Results of the Classical and Proposed Approaches of 2-GE/GE/1 Queuing System

Classical		Proposed		Classical		Proposed	
λ_1	λ_2	λ_1	λ_2	*L*	*W*	*L*	*W*
1.6	2.1	1.465	2.235	0.906	0.245	0.897	0.212
1.8	2.4	1.698	2.502	1.14	0.27	1.132	0.253
2.0	2.7	1.931	2.769	1.455	0.31	1.449	0.298
2.2	2.9	2.117	2.983	1.82	0.357	1.805	0.342
2.4	3.2	2.35	3.25	2.52	0.45	2.507	0.446
2.6	3.4	2.536	3.464	3.599	0.6	3.539	0.598

Further analysis for sample cases of a number of computers, $N = \{3,4,5,6\}$, are shown below.

A sample of parameters for three queueing system.

μ_i = (3,2,1) C_{ai} = (0.1,0.2,0.3) C_{si} = (0.2,0.3,0.1)

A sample of parameters for four queueing system.

μ_i = (4,3,2,1) C_{ai} = (0.1,0.2,0.3,0.4) C_{si} = (0.2,0.4,0.3,0.1)

A sample of parameters for five queueing system.

μ_i = (5,4,3,2,1) C_{ai} = (0.1,0.2,0.5,0.4,0.3) C_{si} = (0.3,0.4,0.1,0.2,0.5)

A sample of parameters for six queueing system.

μ_i = (6,5,4,3,2,1) C_{ai} = (0.1,0.2,0.5,0.4,0.3,0.6) C_{si} = (0.6,0.3,0.4,0.1,0.2,0.5)

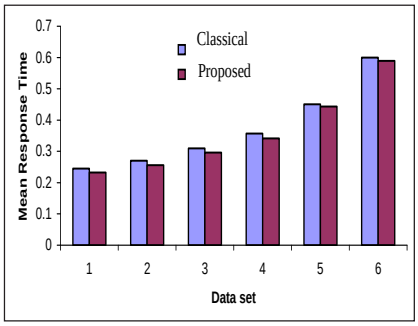

Figure 8. Performance Improvement of Mean Response Time

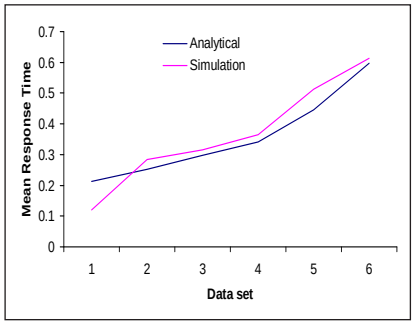

Figure 9. Analytical Versus Simulation Result for a Dual GE/ GE/1 Queueing System

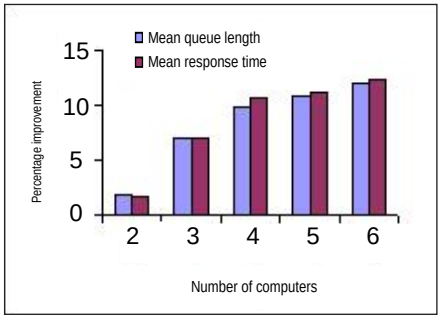

Figure 10. Performance Improvement for a Sample Number of Computers Where ρ = 0.9

The analysis shows that a larger range for the service rates and CV's results in a greater percentage of improvements of our aggregate objectives. The result of the analysis for 2, 3, 4, 5 and 6 computers is summarized in Figure 10. The results clearly show that the mean queue length and the mean response

time have improved for a network of more than 3 computers. However this study requires more number of computers for results generalization. The improvement in the system's performance can be seen in Figures 2, 5, 6, and 8. From the result, we can conclude that the optimal arrival rate improved the queue's performance by reducing the mean number of jobs and the mean response time in the system. One factor to note here is the performance improvement is achieved by increasing the rate of arrival to task with a higher service rate and reducing the rate of arrival to task with a lower service rate. Simulation models were developed to validate the proposed analytical results. Similar generic data as used in the proposed analytical model was used in the simulation for model validation. The simulation results were obtained from the simulation models run at 500 replications using ARENA. The results of the proposed model were compared with the results from the simulation, which are depicted in Figures 3, 4, 7 and 9.

CONCLUSION

In this paper, a new optimization model of allocating arrivals to a network of computers on Generalized Exponential arrival and service-time distribution has been proposed. A closed loop-expression to obtain the routing rate was constructed. An analytical model and simulation approaches were used to show that the classical queueing allocation of total arrivals among parallel systems with the same utilization rate does not provide an efficient performance result. A sample of results for up to six computers is shown to view the improvement. The GE distribution has been used as it could represent exponential and other general distributions. There are several directions to extend the applicability of this allocation model such as different performance objective functions, other arrival and service distribution and arrival with different types of jobs. These examples would involve interesting mathematical problems and could be the subject of future research.

REFERENCES

Allen, A. (1990). *Probability, statistics, and queuing theory with computer science applications* (2nd ed.). San Diego: Academic Press.

Bennani, M. N., & Menasc´e, D. A. (2005). Resource allocation for autonomic data centers using analytic performance models. *IEEE, International Conference on Autonomic Computing*, 229–240.

Boxma, O. (1995). Static optimization of queuing systems. *CWI Report*. BS-R 9302.

Harrison, P., & Patel, N. (1992), *Performance modeling of communication networks and computer architectures*. Addison-Wesley.

Felegyhazi, M., & Hubaux, J. (2006), Game theory in wireless networks: A tutorial. *Technical Report LCA-REPORT-2006-002*. EPFL Switzerland.

Gelenbe, E., & Mitrani, I. (1980). *Analysis and synthesis of computer systems*: London: Academic Press.

Gunther, N. (2000). *The practical performance analyst.* McGraw-Hill.

Hsiao, M., & Lazar, A. (1991). Optimal decentralized flow control of Markovian queuing networks with multiple controllers. *Performance Evaluation, 13*(3),181–204.

Hsiao, M., & Lazar, A. (1990). Optimal flow control of multiclass queuing networks with partial information. *IEEE Transaction on Automatic Control, 35*(7), 855–860.

Jongh, J. (1999). Share scheduling in distributed system (Unpublished doctoral dissertation). Netherland: University of Technische.

Kleinrock, L. (1975). *Queuing systems volume 1*: *Theory*. John Wiley.

Kobayashi, H. (1974). Application of the diffusion approximation to queuing networks I: Equilibrium queue distributions. *Journal of the Association for Computing Machinery, 21*(2), 316–328.

Koole, G. (1999). On the static assignment to parallel servers. *IEEE Transactions on Automatic Control, 44,* 1588–1592.

Kouvatsos, D., & Othman, A. (1989a). Optimal flow control of end-to-end packet switched network with random routing. *IEE Proceedings-Computers and Digital Techniques, 136*(2), 90–100.

Kouvatsos, D., & Othman, A. (1989b). Optimal flow control of a G/G/1 queue. *International Journal of Systems Science, 20* (2), 251–265.

Kouvatsos, D. (1985). A maximum entrophy queue length distribution for a G/G/1 finite capacity queue. *Journal of ACM*, 224–236.

Menascé, D. & Almeida, V. (2000). *Scaling for e-business*. Prentice Hall.

Ku-Mahamud, K. (1993). *Analysis and decentralized optimal flow control of heterogeneous computer communication network models* (Unpublished doctoral dissertation). Universiti Pertanian Malaysia.

Lazar, A. (1982). Centralized optimal control of a Jacksonian network. *Proceedings of the 16th Annual Conference on Information Sciences and Systems*, 316–324.

Lazar, A. (1981). Optimal control of an M/M/1 queue. In *Proceedings. 19th Allerton Conference on Communication, Control and Computing*, 279–289.

Lazar, A. (1984). Optimal control of an M/M/m queue. *Journal of the Association for Computing Machinery, 31,* 86–98.

Lazar, A. (1983). The throughput time delay function of an M/M/1 queue. *IEEE Transaction on Information Theory, 6,* 1001–1007.

Liu, J. (1999) A multilevel load balancing algorithm in a distributed system. *Proceedings of the 19th Annual Conference on Computer Science*, 35–142.

Rahim, R., Ku-Mahamud, K. R., & Othman, A. T. (2002). Performance modeling of e-procurement workflow using generalised stochastic petri net (GSPN). *The Journal of Information and Communication Technology*, *1*(1), 55–68.

Rahim, R., & Ku-Mahamud, K. R. (2006). Analytical modeling and analysis of workload allocation in a network of service centers. *Jurnal Teknologi Maklumat dan Multimedia*, *3*(1), 17–26.

Rahim, R., & Ku-Mahamud, K. R. (2008). Optimal workload allocation in a network of computers with single class job. *Journal of Modern Applied Science*, *2*(2) 101–107.

Rahim, R., Ibrahim, H., Syed Yahaya, S. S., & Khalid, K. (2010). Measurement and analysis of web portal's performance: A case study in UUM. *Journal of Quality Measurement and Analysis*, *6*(2),17–22.

Srikant, R. (2004). *The mathematics of internet congestion control*. Birkhouser.

Tantawi, A., & Towsley, D. (1985). Optimal static load balancing in distributed computer systems, *Journal ACM*, *32*(2), 445–465.

Urgaonkar, B., Pacifici, G., Shenoy, P., Spreitzer, M., & Tantawi, A. (2005). An analytical model for multi-tier internet services and its application. In *Proceeding of the ACMSIFMETRICS'2005*.