Jurnal Teknologi Maklumat & Multimedia **3**(2006): 17-29

Analytical Modeling and Analysis of Workload Allocation in a Network of Service Centers

RAHELA RAHIM & KU RUHANA KU MAHAMUD

ABSTRACT

Redesign of business processes is currently attracting a lot of interest. Although many tools and technique have been used to model such processes, less attention is given to the analytical aspect to support or optimize the redesign process. In other application areas, where a network of server systems are concerned such as computer and communication systems, telecommunication systems and manufacturing systems, many analytical techniques have been developed to analyze the temporal aspect of the systems. The similarities exist between these kinds of system and business processes. An analytic approach is presented that considers the performance of business process designs. A close-loop expression is derived from a non-linear optimization problem to obtain an optimal value for jobs arrival.

Keywords: Analytical modeling; optimization and queueing system; workload allocation

ABSTRAK

Ketika ini reka bentuk semula bagi proses kerja menarik minat banyak pihak. Walaupun banyak alatan dan teknik digunakan untuk memodel proses, kurang perhatian diberikan pada aspek analitikal untuk menyokong atau mengoptimumkan reka bentuk semula proses. Dalam bidang aplikasi selain daripada bidang di mana suatu rangkaian sistem-sistem pelayan seperti sistem pengkomputeran dan komunikasi, sistem telekomunikasi dan sistem perindustrian, banyak teknik analitikal telah dibangunkan untuk menganalisis aspek masa bagi sistem. Terdapat kesamaan di antara sistem-sistem ini dengan proses kerja. Pendekatan analitik dicadangkan bagi mempertimbangkan prestasi proses kerja pada peringkat reka bentuk proses. Suatu ungkapan gelung tertutup ditahkikkan daripada masalah pengoptimuman bukan-linear untuk mendapatkan nilai optimum bagi ketibaan kerja.

Kata kunci: Permodelan beranalitikal; pengoptimuman dan sistem baris; penuntukan beban kerja

INTRODUCTION

Recently business process redesigns have been given a lot of attention. Since then many modeling techniques and tools have been used to support the redesign process. However, most of the currently available tools using static model such as diagram to model process. Some are quite dynamic where the functional aspect of the process have been modeled using simulation. However hardly any studies are found in the literature that use analytical model to study the quantitative behavior and optimization of business process.

Generally for business processes, the cost of the process depends on its time behavior. Customer satisfactions also often depend on time-based performance measure such as response time or turnaround time. For these reasons, it is important to optimize the quantitative performance of business process. An important advantage of analytical modeling is that they allow for the off-line analysis of the effect of redesign process, without disturbing actual processes. To obtain the effective design model, feedback to the designer concerning process functional and alternative design option should be done early in the design process. At this stage, analytical modeling 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 design. Therefore, the analytical modeling proposed here is at the highest abstraction level of design process, i.e. to get the first idea on how the process should behave.

The quantitative measures of business processes to those of concurrent discrete-event system have been focused. Based on the quantitative measure of business proceses to those of concurrent discrete-event system, shows that by using quantitative modeling, arrival to resource center can be reallocated to get the optimal performance measure. The issue of job allocation in a network of service centers where different service centers have different job processing time have been focused. 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 allocate jobs to the service center according to a calculated arrival rate computed using Lagrange multiplier theorem. The paper is organized as follows: next section shows the related work in this area followed by the proposed queueing and optimization model. Later numerical results and models verification are presented. Finally concluding remarks and directions for further research are described.

RELATED WORK

The problem of workload allocation is common to a variety of business systems especially when it involves a network of service centers. Workload allocation seeks to allocate job arrival among service centers as evenly as possible. In a parallel setting, and particularly for multi service center systems where jobs may have many possible paths at the job's scheduler, jobs allocation problem is of interest, each job entering the scheduler, a path is assigned to optimize the allocation of workload (Lin & Kumar 1984; Ni & Hwang 1985; Liu 1999). The workload allocation problem is of particular interest for networks of service centers, since there are several ways that affect the distribution of workload among the centers (Ross & Yao 1991).

Queueing network models have been recognized as powerful tools for evaluating performance of computer systems (Gelenbe & Mitrani 1980; Allen 1990) and communication network (Kleinrock 1975; Harrison & Patel 1992; Bell & Williams 1999; Gunther 2000; Menasce & Almeida 2000). This analytical model has become a very important tool for predicting the behaviour of new designs or proposed changes to existing systems (Kouvatsos & Othman 1986, 1989; Hsiao & Lazar, 1990, 1991; Smith & Williams 2001). Most queueing network models are used either by making assumptions to assure exact numerical solution or by employing approximate method (Kobayashi 1974; Chandy et al. 1975; Harrison & Patel 1992). Lazar (1981, 1983, 1984), considered the control of arrivals to a network of queues with the objective of maximizing throughput subject to a response time constraint. He developed a throughput time delay function based on an optimality criterion where the arrival that maximize the throughput under the constraint of the average response time will not exceed a preassigned value. Then continued with the problem of random routing (Kouvatsos & Othman 1989). All these literature have been devoted to the probabilistic analysis of queueing system, their optimization is somewhat lagging behind. Network of queues in parallel are a natural way to model problems of resource and traffic allocation and many optimization problems based on them have been studied for instance in Kleinrock (1975); Ni & Hwang (1985); Boxma (1995); Chombe & Boxma (1995); Koole (1999); and de Jongh (2002). 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 considered the impact of jobs inter arrival and service time variation in modeling the systems performance. Although analytical workload allocation models are widely applied in area of computer and communication network, they are also applicable in the case of business processes especially by the increasing application of workflow management systems.

QUEUEING SYSTEM MODEL

When several users request for a job (e.g customer request for product or service) at the same time, competition for the use of a common resource and 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 occur. This is due to the fact that the inter arrival times of the request, and their required service times, are generally not fixed, therefore a mathematical model of congestion represents inter arrival and service times of users by random variables. Queueing Theory is devoted to the description, analysis and optimization of such queueing system (Boxma 1995). 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 allocation process of jobs through a network of queues, the main performance measures are also random variables. With this in mind, we use multiple queue multiple server model to represent a central job allocation system as shown in Figure 1.

In using this model, resources that provide services are represented by

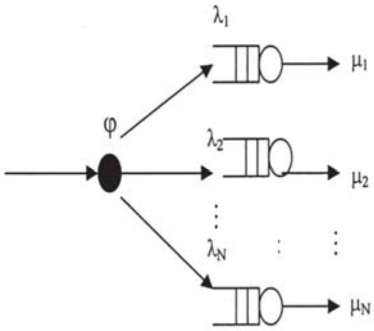

FIGURE 1. Multiple queue multiple server model

service centers at which jobs queue and compete for service. The workload is modeled as a single stream of job. A common example of workload is order request. Total workload rate arrived is given by φ. Each newly arrived job, it is assigned to a node or server i with rate λ_i, according to the scheduling policy used. We consider the set of service centers to be heterogeneous which is common cases in real systems and also it can be generalized to homogeneous servers. In the contex of general queueing network models, the generalized exponential (*GE*) distributional model of the form (Kouvatsos & Othman 1986, 1989).

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

Where μ is the mean service rate, C is the coefficient of variation and u_0 (t) is the unit impulse function, has been used to represent the inter arrival and service time distribution. 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 has been shown in (Kouvatsos 1986; Ku Mahamud 1993) that the exact mean number of jobs in the GE/GE/1 queue is given by

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

Where C_a^2, C_s^2 are the squared coefficients of variation of the inter arrival and service time distributions respectively. This mean number of jobs function will be used as the objective function in the optimization method.

OPTIMIZATION MODEL USING GENERALIZED EXPONENTIAL (*GE*) DISTRIBUTION

In this section, the use of Lagrange method to *GE* type distributional model is proposed. In this case an optimization problem of queueing system can be generalized to a number of arrival and service distribution by configuring the value of coefficient of variation for inter arrival and service time. We formulated an optimization problem of N -GE/GE/1 where N indicates a number of queueing system, as below:

Notation:

λ_i	:	mean arrival rate of job-*i* at service center *i*
μ_i	:	mean service rate of service center *i*
β_i	:	mean service time of service center *i*
ϕ	:	total arrival rate
L_i	:	mean jobs queue length
C_{si}	:	mean coefficient of variation in service time
C_{ai}	:	mean coefficient of variation in inter arrival time
ρ_i	:	utilization of service center *i*

P1 Min

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

s.t

$$\sum_{i=1}^{N} \lambda_i = \phi \tag{2}$$

$$0 \le \lambda_i \le \frac{1}{\beta_i}, \quad i = 1,\ldots, N. \tag{3}$$

$$\lambda_i \infty 0 \tag{4}$$

$$\beta_i \infty 0 \tag{5}$$

where $\rho_i = \frac{\lambda_i}{\mu_i}$ and $\beta_i = \frac{1}{\mu_i}$

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

$$\frac{d}{d\lambda_i}\left\{D_i\left(\frac{\lambda_i\beta_i}{1-\lambda_i\beta_i}\right)\left(\lambda_i\beta_i\left(\frac{C_{si}^2-1}{2}\right)+\frac{C_{ai}^2+1}{2}\right)\right\}=\delta \qquad i = 1,\ldots,N \tag{6}$$

$$\sum_{i=1}^{N}\lambda_i=\phi=0 \tag{7}$$

From (6) and (7) we find the unique optimal values

$$\lambda_i{*}=\frac{1}{\beta_i}\left(1-\left(\frac{C_{si}^2+C_{ai}^2}{C_{si}^2-1+2\mu_i\delta}\right)^{1/2}\right) \tag{8}$$

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

$$\sum_{i=1}^{N}\frac{1}{\beta_i}\left(1-\left(\frac{C_{si}^2+C_{ai}^2}{C_{si}^2-1+2\mu_i\delta}\right)^{1/\,2}\right)=\phi \tag{9}$$

When C_{ai}=1 and , C_{si}=1, the *GE* traffic optimal expression is reduced to exponential optimal expression. D_i is the cost associated with having one job in queue and for simplicity, we assign the value of 1.

COMPUTATIONAL RESULT

In this section, numerical results are presented to assess the credibility of the *GE* distribution used. Two configurations will be shown. For the first configuration, service rate of the tasks are assumed to be $\mu_1 = 3, \mu_2 = 4, C_{a_1} = 0.5, C_{a_2} = 0.3, C_{s_1} = 0.2, C_{s_2} = 0.4$. The improvement of the performance measures is presented in Fig. 2 and 4. To verify the results, we use simulation and the comparative results are presented in Figure 3 and 5.

Table 1 provides the result of mean queue length and mean response time obtained from the classical and proposed allocation scheme of two GE/GE/1 queueing systems with service rate $\mu_1 = 3$, $\mu_2 = 4$, and variation parameters of inter arrival and service time, $Ca_1^2 = 0.5$, $Ca_2^2 = 0.3$ and $Cs_1^2 = 0.2$, $Cs_2^2 = 0.4$.

TABLE 1. Result of classical and proposed approaches of 2-GE/GE/1 queueing system, L: 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

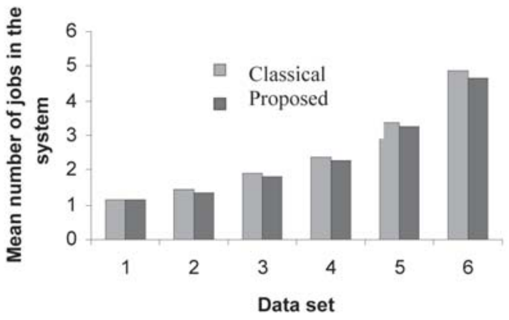

FIGURE 2. Performance improvement of mean queue length for a dual GE/GE/1 queueing system

Figure 2 and 4 shows the comparative improvement in the mean number of jobs and mean response time in the 2-GE/GE/1 systems obtained in Table 1. The maximum improvement for the given dataset is given by 13.5%.

The results obtained in Table 1 are validated using simulation model whereby the same parameters setting have been used. The maximum comparative error of 12.08% is given in Figure 3 and 5.

Other parameter setting of two GE/GE/1 queueing systems with service rate $\mu_1 = 3$, $\mu_2 = 4$, and variation parameters of inter arrival and service time, $Ca_1^2 = 0.1$, $Ca_2^2 = 0.2$ and $Cs_1^2 = 0.4$, $Cs_2^2 = 0.3$ is presented in Table 2.

Figure 6 and 7 shows the comparative improvement in the mean queue length and mean response time of the 2-GE/GE/1 systems obtained in Table 2. The maximum improvement for the given dataset is given by 7.14%.

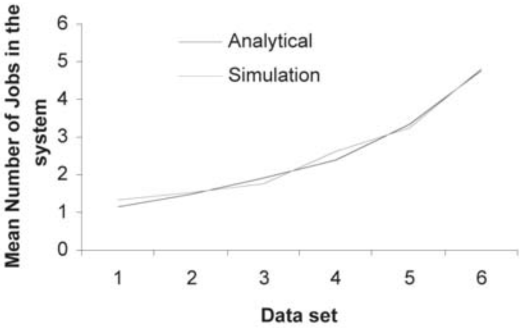

FIGURE 3. Analytical versus simulation result

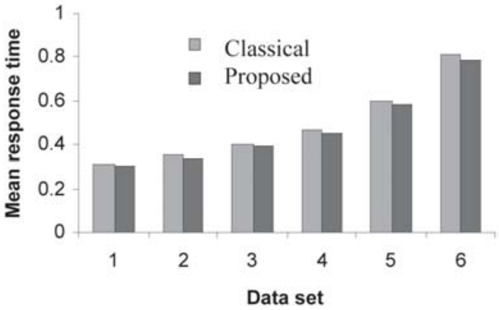

FIGURE 4. Performance improvement of mean response time for a dual GE/GE/1 queueing system

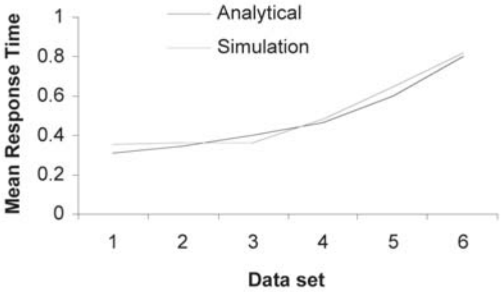

FIGURE 5. Analytical versus simulation result

TABLE 2. Result of classical and proposed approaches of 2 GE/GE/1 queueing 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

The results obtained in Table 2 are validated using simulation model whereby the same parameters setting have been used. The maximum comparative error of 13.06% is given in Figure 7 and 9.

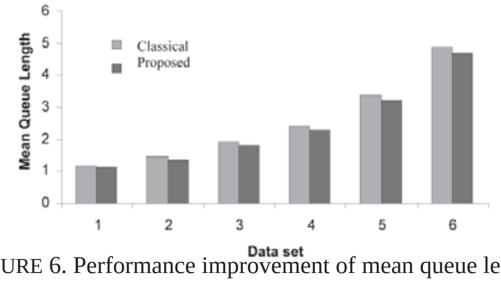

FIGURE 6. Performance improvement of mean queue length for a dual GE/GE/1 queueing system

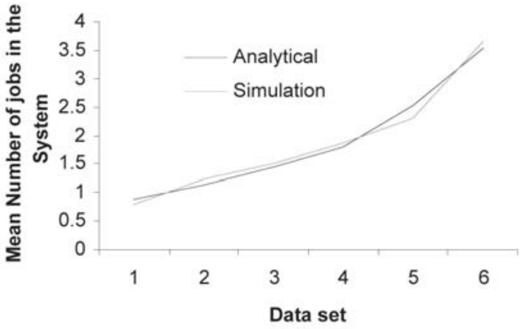

FIGURE 7. Analytical versus simulation result

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

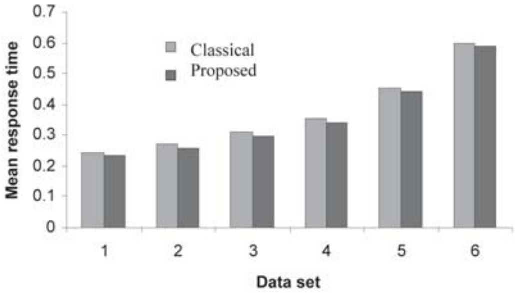

FIGURE 8. Performance improvement of mean response time for a dual GE/GE/1 queueing system

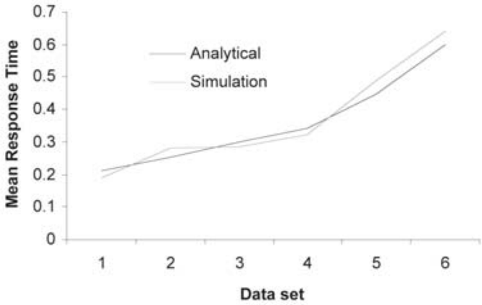

FIGURE 9. Analytical versus simulation result

Numerical Result for Three Queueing Systems:

$$\mu_i = (3,2,1) \quad C_{a_i} = (0.1, 0.2, 0.3) \quad C_{s_i} = (0.2, 0.3, 0.1)$$

Numerical Result for Four Queueing Systems

$$\mu_i = (4,3,2,1) \quad C_{a_i} = (0.1, 0.2, 0.3, 0.4) \quad C_{s_i} = (0.2, 0.4, 0.3, 0.1)$$

Numerical Result for Five Queueing Systems

$$\mu_i = (5,4,3,2,1) \quad C_{a_i} = (0.1, 0.2, 0.3, 0.4, 0.5) \quad C_{s_i} = (0.3, 0.4, 0.1, 0.2, 0.5)$$

Numerical Result for Six Queueing Systems

$$\mu_i = (6,5,4,3,2,1) \quad C_{a_i} = (0.1, 0.2, 0.0.5, 0.4, 0.3, 0.6)$$

$$C_{s_i} = (0.6, 0.3, 0.4, 0.1, 0.2, 0.5)$$

The analysis shows that a larger range for the service rates results in greater percentage improvements of our aggregate objectives. The result of the analysis for the queueing systems is summarized in Figure 10 below. Utilization of 90% has been assumed as this indicates that the service centers are almost fully utilized.

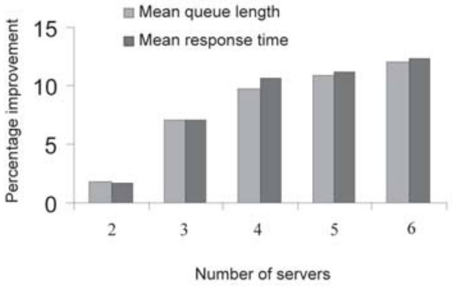

FIGURE 10. Performance improvement for a sample number of servers where $\rho = 0.9$.

CONCLUSION

A new optimization model of allocating job arrivals to a network of service centers based on Generalized Exponential arrival and service time distribution has been proposed. A closed-loop expression to obtain the routing rate was described. Analytical model and simulation approaches were used to show that the classical allocation of total arrivals between two service centers with the same utilization rate does not provide an efficient performance result for the queueing systems. The result for a number of service centers was shown to portray the improvement. The *GE* distribution has been employed because it could represent exponential and other general distribution. The analytical modeling proposed can be used to guide the business process redesign in some parts where networks of service centers are concerned. There are several directions to extend the applicability of this allocation algorithm such as different performance objective function, other arrival and service distribution and arrival with different type of jobs. These examples would involve interesting mathematical problem and could be the subject of future research.

REFERENCES

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

Bell, S. L. & Williams, R. J. 1999. Dynamic scheduling of a system with two parallel servers: Asymptotic optimality of a continuous review threshold policy in heavy Trac. *In Proceedings of the 38th Conference on Decision and Control, Pheonix, Arizona*: 1743-1748.

Boxma, O. J. 1995. Static Optimization of Queueing Systems. *CWI Report*. BS-R 9302.

Chombe, M. B. & Boxma, O. J. 1995. Optimization of static traffic allocation policies. *Theoretical Computer Science* 125: 17-43.

Chandy, K. M., Herzog, U. & Woo, L. 1975. Parametric analysis of queueing networks. *IBM J. Research and Development* 19(1): 36-42.

De Jongh, J. F. C. M. 2002. *Share scheduling in distributed system*. PhD Thesis University of Technische, Netherland.

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

Gunther, N. J. 2000. *The practical performance analyst*. New York: McGraw Hill.

Harrison, P. G. & Patel, N. M. 1992 *Performance modeling of communication networks and computer architecture.* Addison Wesley.

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

Hsiao, M. T. & Lazar, A. A. 1991. Optimal decentralized flow control of markovian queueing networks with multiple controllers. *Performance Evaluation* 13(3): 181-204.

Klienrock, L. 1975. *Queueing systems volume 1: Theory*. New York: John Wiley Inc.

Kobayashi, H. 1974. Application of the diffusion approximation to queueing networks I: Equilibrium queue distribution. *Journal of the ACM* 21(2): 316-328.

Koole, G. 1999. On the static assignment to parallel servers. *IEEE Transaction on Automatic Control* 44: 1588-1592.

Kouvatsos, D. D. 1986. A maximum entropy queue length distribution for A G/G/1 finite capacity queue. *Performance Evaluation Review* 4: 224-236.

Kouvatsos, D. D. & Othman, A. T. 1989. Optimal flow control of a G/G/1 queue. *International J. of Systems Science* 20(2): 251-265.

Kouvatsos, D. D. & Othman, A. T. 1986. Optimal flow control of a G/G/C finite capacity queue. *J. Operational Research Society* 40(7): 659-670.

Kouvatsos, D. D. & Othman, A. T. 1989. Optimal flow control of end to end packet switched network with random routing. *IEE Proceedings* Part E-136(2): 90-100.

Ku Mahamud, K. R. 1993. *Analysis and decentralized optimal flow control of heterogeneous computer communication network models*, PhD thesis, Universiti Pertanian Malaysia.

Lazar, A. A. 1981. Optimal control of an M/M/1 queue. *In Proc. 19th Allerton Conf. On Communication, Control and Computing*: 279-289.

Lazar, A. A. 1982. Centralized optimal control of a Jacksonian network. *In Proceedings of the Sixteenth Conference on Information Science and Systems*, 17-19 March 1982, New Jersey: 316-324.

Lazar, A. A. 1983. The throughput time delay function of an M/M/1 Queue. *IEEE Transaction on Information Theory* 29(6): 1001-1007.

Lazar, A. A. 1984. Optimal control of an M/M/m Queue. *Journal of the ACM*. 31(1): 86-98.

Lin, W. & Kumar, A. 1984. Optimal control of a queueing system with two heterogeneous servers. *IEEE Trans Automatic Control* 29(8): 696-703.

Liu, J. B. 1999. A multilevel load balancing algorithm in a distributed system. *Proceedings of the 19th annual conference on Computer Science*: 35-142.

Menasce, D. A. & Almeida, V. A. F. 2000. *Scaling for E-Business*. Prentice Hall.

Ni, L. M. & Hwang, K. 1985. Optimal load balancing in a multiple processor system with many job classes. *IEEE Trans. Software Engineering*: 491-496.

Ross, K. W. & Yao, D. D. 1991. Optimal load balancing and scheduling in a distributed computer system. *Journal of the ACM* 38(3): 679-690.

Smith, C. U. & Williams, L. G. 2001. *Performance solutions, a practical guide to creating responsive, scalable software*. Pearson education.

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

Rahela Rahim
Fakulti Sains Kuantitatif
06010 UUM Sintok
Kedah
rahela@uum.edu.my

Ku Ruhana Ku Mahamud
Pengarah
Pusat Pengajian Siswazah
06010 UUM Sintok
Kedah
ruhana@uum.edu.my