A STUDY OF OPTIMAL DIMENSIONING OF QUEUES WITH RESPECT TO SOCIAL AND INDIVIDUAL PROFIT

G.H. Shahkar and H.R. Tareghian

Department of Statistics, Faculty of Sciences, Ferdowsi University of Mashhad, Mashhad, Islamic Republic of Iran

Abstract

In this paper, a system of GI/G/1/K queue is considered. The optimal system's capacity (K), when the system is optimized with respect to the benefit of the entire system (*social optimization*) and when the criterion for optimality is individual gains (*individual optimization*), is determined and compared. In social optimization, the system capacity is obtained through maximization of the system's profit. However, when individual gains is the criteria, the system capacity is determined as a result of customers not joining the system because they estimate that getting service from the system does not yield them any profit and therefore leave the system. It is shown by simulation that irrespective of the traffic intensity, ρ and arrival and service time distributions with different failure and acceleration rates, the K obtained from social optimization of the system is always equal or less than the K obtained from individual optimization. Thus, the social optimization of the system not only maximizes the system's profit but also preserves individual gains and prevents the customer from leaving the system. In this paper, some other interesting results are also reported.

Introduction

For any queueing system in which customers renege, the system itself reaches a certain size, due to reneging customers. This self determined capacity (K) may not even satisfy the intentions of customers about their waiting time in the system. Since, in the short period of time during which customers are observing the behaviour of the queue, they can not correctly guess the waiting time distribution, and hence their estimate of their waiting time may be incorrect. In addition, it is not clear whether the server's capabilities are optimally utilized, if the queueing system is designed with the capacity equal to (K).

Keywords: Failure rate functions; Optimization; Queueing theory; Simulation

It is natural for customers to stay in the system for service as long as their estimate of their waiting time is tolerable. In other words, it is profitable for them to stay and get served. Now if we optimize a queueing system with respect to the system's profit without considering the customer's waiting time, do we still satisfy the individual customer's benefit? And does not the waiting time of the customer in the *optimized system* discourage him from staying in and getting served? Rue and Rosenshine [6] have shown that if an M/M/1/K system is designed, such that the system's profit is maximized, then the benefit of the individual cutomer is also satisfied. In this paper, the above questions are considered for the general queueing system of GI/G/l/K.

In a GI/G/l/K queueing system, we consider the following objective function:

$$Z = (R\lambda' - CL - DK) \tag{1.1}$$

in which

Z= Expected system profit per unit of time.
R= A fixed fee charged per customer.
λ'= The effective arrival rate, i.e., the mean rate of customers actually entering the system.
C= The cost incurred per customer per unit time.
L= The expected system size.
D= The cost of providing one unit of space, and
K= The system capacity.

The value of K that maximizes Z, is referred to as *social optimization*, that is, it maximizes benefits to the entire system, rather than the personal gain of an individual customer.

Now consider the case where customers may renege. If an arriving customer finds N customers ahead of himself in the system, and if he estimates the mean service time as $1/\mu$, then his estimate of his waiting time is $(N+1)/\mu$. Therefore, from an individual customer's standpoint we get the following objective function:

$$Z = R - \frac{C(N+1)}{\mu} \tag{1.2}$$

The customer will serve his own interest (positive gain) by joining the queue as long as $C(N+1)/\mu \leq R$, or equivalently, $N \leq R\mu/C-1$. Hence, if every customer uses this strategy, a GI/G/l/K queue results, where $K=[R\,\mu/C]$, (*individual optimization*).

Rue and Rosenshine have shown that for the M/M/l/K queueing system, Z is concave in K. They have also shown that the K obtained from social optimization of the system is always less than the K obtained when the criterion is individual gains.

In the rest of this paper, we will try to examine the possibility of generalizing Rue and Roshenshine's remarks to the general system of GI/G/l/K. The stimulus for this study stems from the fact that the implications of the results obtained from this study can be valuable in the design and control of queueing systems. We will first consider the M/G/l/K system.

For M/G/l/K, like any other queueing system in steady state, the effective arrival rate, λ', is equal to the effective departure rate, i.e.:

$$\lambda' = \lambda(1-q_k) = \mu(1-p_0) \tag{1.3}$$

where q_k is the probability that an arriving customer finds K customers in the system, p_0 is the probability of the system being empty, and λ is the arrival rate.

It can be shown that (see [2] and [3]):

$$p_0 = \frac{\pi_0^*}{\pi_0^* + \rho} \tag{1.4}$$

where generally π_n^* is the steady state probability of n customers in the system at a departure point.

In order to optimize Z, we need L, the average system size which is obtained from:

$$L = \sum_{n=0}^{k-1} n\pi_n^* \tag{1.5}$$

where π_n^* can be obtained by solving K equations of:

$$\pi_i^* = \begin{cases} \pi_0^* k_i + \sum_{j=1}^{i+1} \pi_j^* k_{i-j+1} & (i=0,1,2,\ldots,K\text{-}2) \\ 1 - \sum_{n=0}^{k-2} \pi_n^* & (i=K\text{-}1) \end{cases} \tag{1.6}$$

and:

$$k_n = \Pr\{\text{ n arrivals during a service time } S=t\}$$

Although λ' and L, needed to optimize (1.1) for an M/G/l/K, can be obtained theoretically from (1.3) and (1.5) respectively, even for the simplest service process, achieving numerical results for M/G/l/K needs a lot of complicated mathematical work [2].

In M/G/l/K the interarrival time is exponential which has the Markovian property. In some cases, the memoryless property of exponential distribution may not be appropriate. For these cases, we will consider some distributions that have different characteristics. One particularly useful characteristic is *failure rate function*.

Under the assumption of a continuous cumulative distribution function, B(u), *instantaneous failure rate*, h(u), is defined as (see for example [2]):

$$h(u) = \lim_{\Delta u \to 0}\left[\frac{1}{\Delta u}\Pr\{\text{Service or arrival of age u will be completed i } (u, u+\Delta u)\}\right]$$

$$= \frac{B'(u)}{1-B(u)} \qquad (1.7)$$

The hazard or failure rate function, h(u), can be increasing irr u (called an Increasing Failure Rate-IFR), decreasing in u (DFR), constant, or a combination of these cases. The constant case implies the memoryless property.

It can be shown that Erlang distribution, E_k, has IFR with decreasing acceleration when $k > 1$, and a mixture of exponentials (H_k) has DFR (see for example [2] pp. 392-393). For *Weibull distribution* [with $1-F(t) = \exp(-(t/\beta)^{\alpha})$] we can select the shape parameter α such that we obtain an IFR with decreasing acceleration, constant acceleration, and increasing acceleration as well as obtaining even a DFR.

The pdf of the Weibull distribution is:

$$f(t) = \alpha\beta^{-\alpha}t^{\alpha-1}\exp\left[-\left(\frac{t}{\beta}\right)^{\alpha}\right] \qquad (1.8)$$

Thus

$$h(t) = \alpha\beta^{-\alpha}t^{\alpha-1} \qquad (1.9)$$

It is easy to see that for $\alpha = 1/2$ say, the Weibull distribution is DFR with decreasing acceleration, for $\alpha = 2$ and $\alpha = 3$ it is an IFR with constant and increasing acceleration respectively.

Since analytical study of the queueing models with general arrival and service processes seems to be impossible, the queueing models considered in this paper are analyzed by simulation. By simulating various GI/G/l/K systems, we show that irrespective of the system's parameters the K obtained from social optimization of the system is always equal or less than the K obtained from individual optimization.

For the purpose of the simulation study, the following combinations of distributions are considered and simulated. Erlang distribution is used because of its flexibility, while Weibull distribution is used because by changing its shape parameter one can obtain a distribution with IFR, DFR and even increasing or decreasing acceleration, and finally generalized Erlang distribution (GE) is used because it is considered a good approximation to any distribution. As a matter of fact, it can be shown that any cumulative distribution function can be approximated to almost any degree of accuracy by the convolution of exponential distributions (see e.g. [2]).

Simulation

The reason why the system under study is analyzed by simulation rather than analytical methods is twofold. First, because deriving formulae for the system performance measures (e.g. expected profit rate or mean waiting time) becomes too complicated when the arrival and service time distributions are as we have considered, and in addition, the analysis of some of the distribution functions (e.g. convolution of exponential distributions) requires numerical approaches. Second, we wanted to develop an interactive model to enable us to easily test the effect of various input parameters on the system performance.

The system performance measures are collected when the system is in steady state. In order to determine the time period at which the system reaches steady state, Welch's method [4] is used. In this paper based on Welch's method, 30 independent replications of the simulation model are made and a plot of mean transit time (W) is drawn against time. The results of two such simulation runs are shown in Figure 2.1. It is observed that irrespective of the values of ρ and K, steady state is reached at approximately 675 units of time.

Each simulation run consists of the arrival and departure of 1500 served customers. The sample size required to establish a given statistical significance at a given level of precision is obtained using *central limit theorem* and sequential sampling [8].

The simulation model is constructed based on the event processing approach [5]. To simulate the interarrival and service times, samples are drawn from the corresponding distributions. The method of *inverse transformation* [1] is used for this purpose when applicable. For the Erlang distribution, we use the method of *convolution* [4], and for the GE distribution used in this paper, i.e.

$$G(x) = (F_1 * F_2)(x) = \int_0^x F_1(x-y)\, dF_2(y) \qquad (2.1)$$

where

$$F_i(x) = 1 - e^{-\mu_i x}, \quad i = 1, 2. \qquad (2.2)$$

hence,

$$G(x) = \int_0^x \mu_2 e^{-\mu_2 y}\left[1 - e^{-\mu_1 (x-y)}\right] dy, \qquad (2.3)$$

$$G(x) = 1 - \frac{\mu_1}{\mu_1 - \mu_2}e^{-\mu_2 x} + \frac{\mu_2}{\mu_1 - \mu_2}e^{-\mu_1 x} \quad (\mu_1 \neq \mu_2)$$

$$=1-e^{-\mu x}\,(1+\mu x) \quad (\mu_1=\mu_2=\mu) \qquad (2.4)$$

we have

$$1-RN=\frac{\mu_1}{\mu_1-\mu_2}e^{-\mu_2 x}-\frac{\mu_2}{\mu_1-\mu_2}e^{-\mu_1 x} \quad (\mu_1\neq\mu_2) \qquad (2.5)$$

where RN $\simeq$ U(0,1). To generate random numbers, we use the well-tested acceptable *prime modulus multiplicative linear congruential generator* ($R_i = aR_{i-1}$(mod m)), with m = 2^{31}-1, a = 630360016 and R_0 as any integer from 1 through m-1 [4]. From (2.4) and by taking y= e^{-x}, we get

$$y=\exp\left(\frac{1}{\mu_2}\mathrm{Ln}\left(\frac{\mu_2}{\mu_1}y^{\mu_1}+\frac{RN\,(\mu_1-\mu_2)}{\mu_1}\right)\right) \qquad (2.6)$$

Samples from GE are drawn by solving (2.6) using numerical methods, for instance fixed-point or Newton-Raphson method [10]. The tolerance considered is 0.001.

Model Validation

Various authors [7,8] suggest that for model validation the three steps of 1) face validation, 2) testing of model assumptions and 3) testing of input-output (I/O) transformation be made.

In this study, after making sure that the logic is correct, we analyze the model's I/O relationship to determine whether the internal behaviour of the system is reasonable. In the second stage, all the samples drawn from the various distributions are tested to determine if they are drawn from the corresponding distributions. For the final step, the results obtained from the M/M/l/K simulation model is statistically compared with the corresponding analytical results. In the M/M/l/K model, the following assumptions are made (λ = 1; R = 300; C = 15; D = 1). Two values of ρ (0.55 and 0.99) are also considered. As it can be seen from Figures 3.1 and 3.2, the true values of Z and W are always contained within their respective 95% confidence intervals.

Any queueing model that has been considered in this study is fundamentally based on the validated M/M/l/K model. The only difference is that for each model the interarrival and service times are sampled from the corresponding distributions. In addition, for each queueing model the fact that for K= 1; W= 1/μ has been checked.

Discussion of Results

The simulation model is run with different arrival and service processes, and also with various values of ρ, R, C, and D. The results of some of these runs are presented in Figure 4.1. In analyzing the results, it is observed that irrespective of the aforementioned parameters, W and Z are always concave. Therefore, in comparing the social and individual optimization, it is not deemed necessary to change the values of R, C and D.

The results of simulating different queueing models with different traffic intensities for R = 300, C = 15 and D = 1 are given in Tables 4.1, 4.2 and 4.3. Comparing mean transit times in these tables, one can observe that irrespective of ρ and queueing model the social optimization of the system always preserves the customer's benefits more than the individual optimization. The only exception is M/M/l/K in light traffic, where transit times

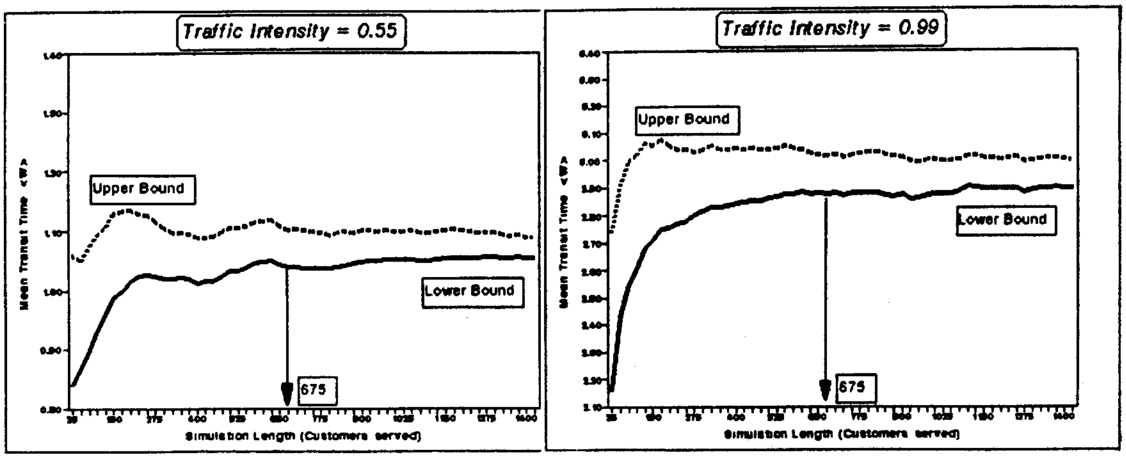

Figure 2.1 Simulation initial transient

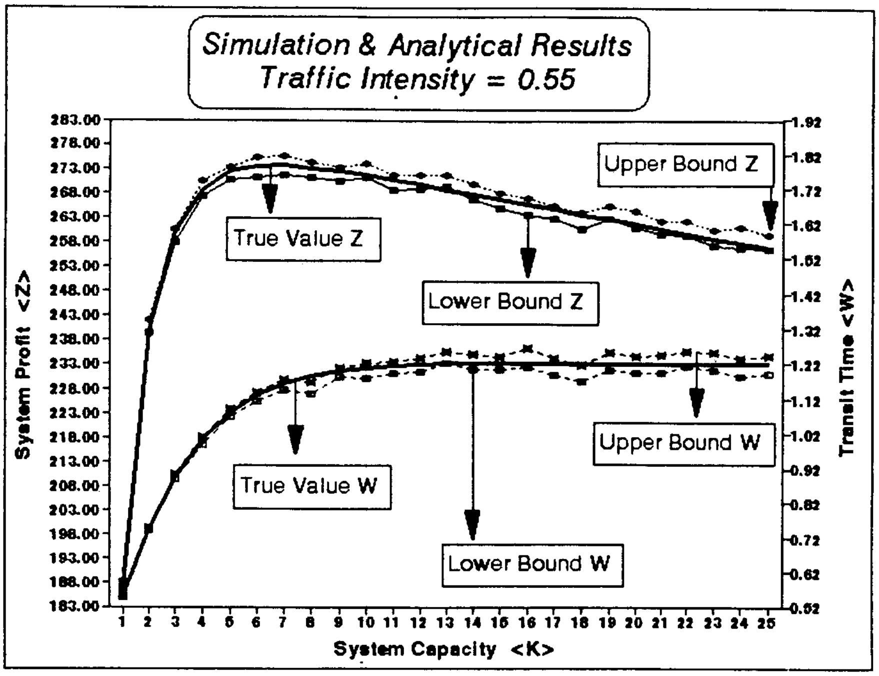

Figure 3.1. Simulation and analytical results for ρ= 0.55

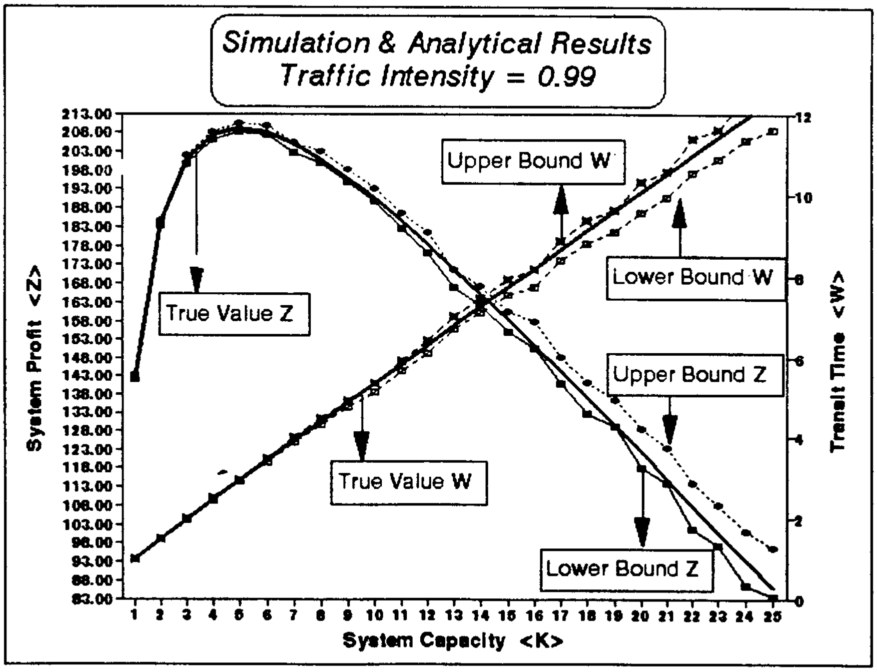

Figure 3.2. Simulation and analytical results for ρ= 0.99

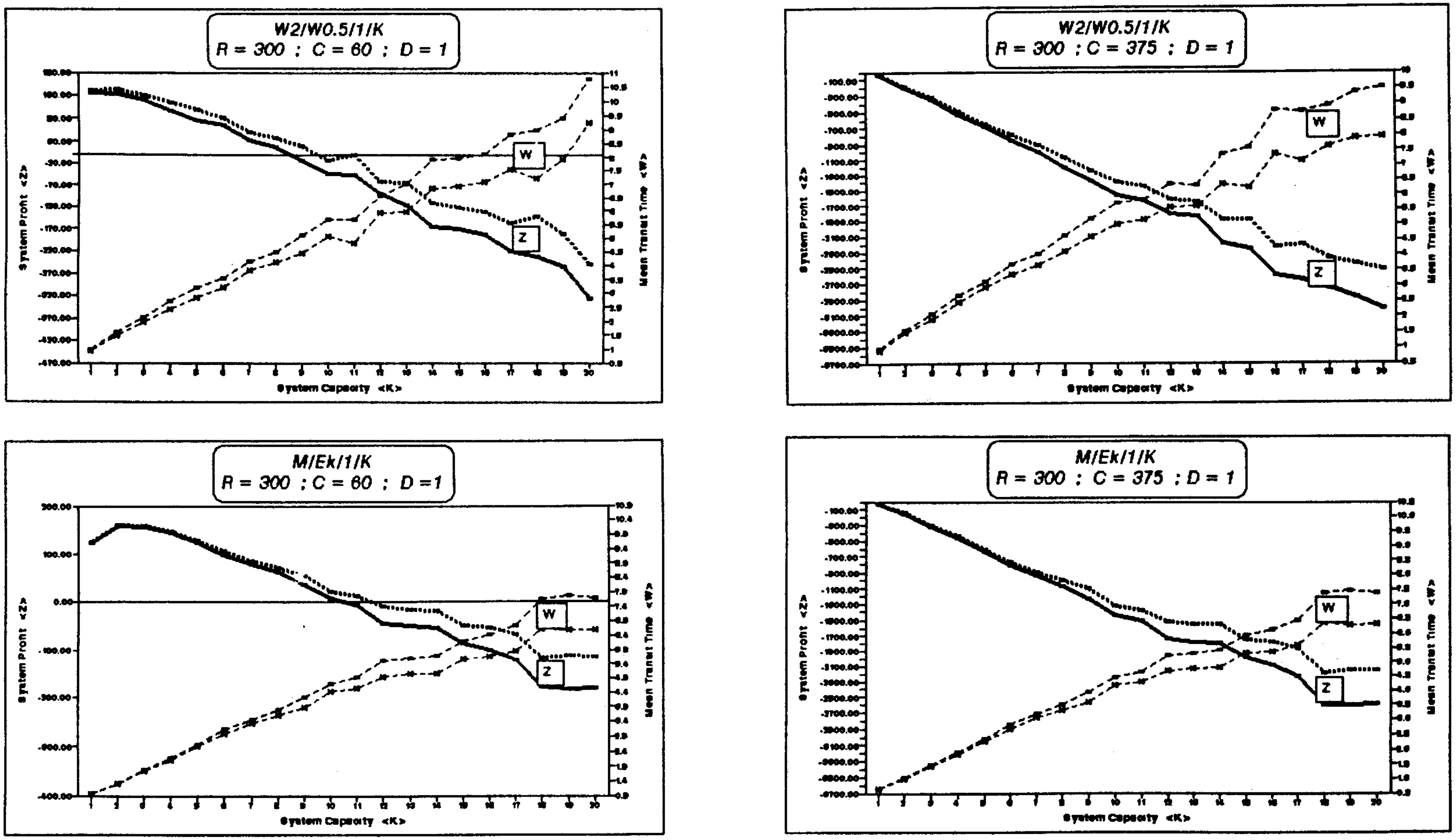

Figure 4.1. Upper and lower bounds of Z and W from differ ent models with different values of R and C for ρ= 0.95

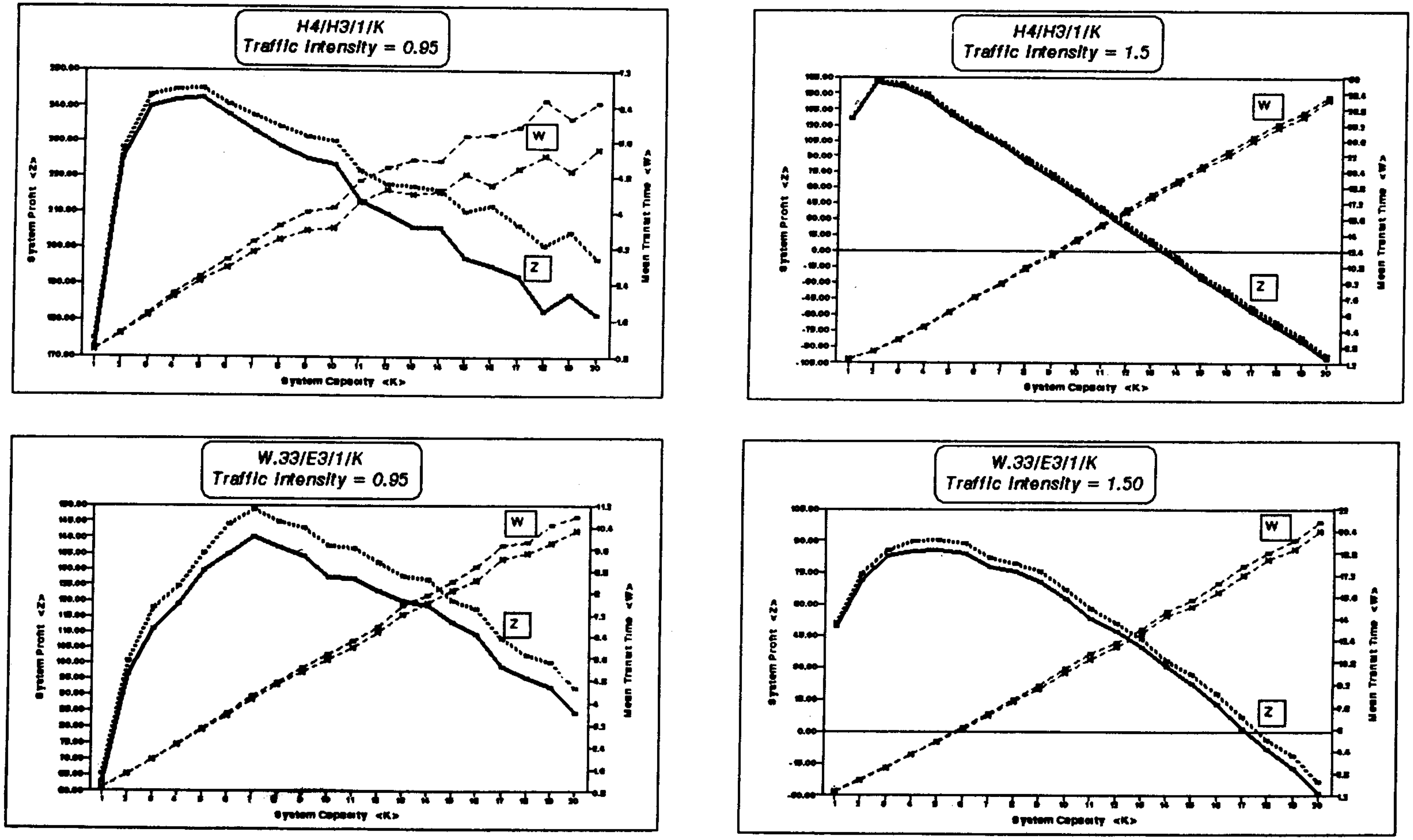

Figure 4.2. Upper and lower bounds of Z and W from different models with ρ= 0.95 and ρ= 1.5

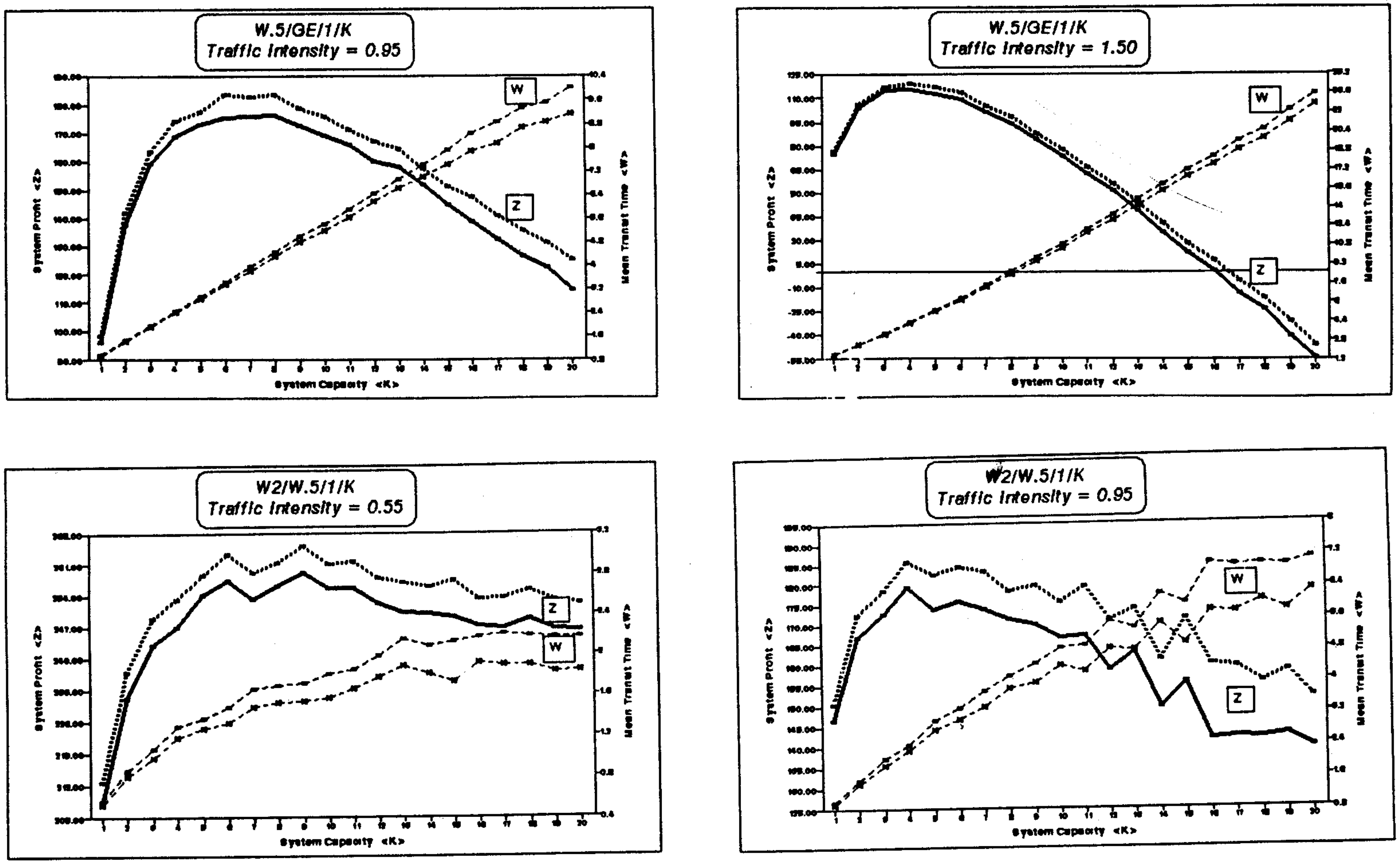

Figure 4.3. Upper and lower bounds of Z and W from different models with ρ= 0.95 and ρ= 1.5

Table 4.1. Comparison of results obtained from social and individual optimization *Light Traffic* ($\rho = 0.55$).

MODEL		Variance		System Capacity		Mean Trans. Time		Mean Syst. Profit	
		Arrv.	Serv.	Indv.	Soci.	Indv.	Soci.	Indv.	Soci.
$M/E_k/1/K$	k= 1	1.00	0.30	36	9	1.22	1.22	242.8	275.3
	k= 5	1.00	0.06	36	5	0.95	0.88	248.3	281.2
	k= ∞	1.00	0.00	36	5	0.89	0.86	251.5	282.7
$H_4/H_3/1/K$		2.50	1.55	36	4	0.72	0.71	254.2	285.3
$E_3/W_3/1/K$		0.33	0.11	36	3	0.66	0.64	254.0	286.6
$W_{1/3}/E_3/1/K$		19.00	0.10	36	11	5.30	2.89	184.9	213.4
$W_{1/2}/GE/1/K$		5.38	0.16	36	10	2.62	2.08	227.6	253.1
$W_2/W_{1/2}/1/K$		0.60	0.18	36	9	2.01	1.62	234.3	262.2

Table 4.2. Comparison of results obtained from social and individual optimization *Heavy Traffic* ($\rho = 0.95$).

MODEL		Variance		System Capacity		Mean Trans. Time		Mean Syst. Profit	
		Arrv.	Serv.	Indv.	Soci.	Indv.	Soci.	Indv.	Soci.
$M/E_k/1/K$	k= 1	1.00	0.90	21	6	8.43	3.15	147.1	214.6
	k= 5	1.00	0.20	21	5	7.24	2.60	166.7	230.6
	k= ∞	1.00	0.00	21	5	7.19	2.56	169.1	236.1
$H_4/H_3/1/K$		2.50	4.63	21	4	5.99	2.50	187.5	243.5
$E_3/W_3/1/K$		0.33	0.33	21	4	4.73	2.07	206.5	257.5
$W_{1/3}/E_3/1/K$		19.00	0.33	21	7	10.99	4.16	83.2	144.2
$W_{1/2}/GE/1/K$		5.38	0.51	21	8	10.08	4.38	112.5	179.8
$W_2/W_{1/2}/1/K$		0.60	4.86	21	4	6.92	2.32	142.3	182.4

Table 4.3. Comparison of results obtained from social and individual optimization Non-stationary ($\rho = 1.50$).

MODEL		Variance		System Capacity		Mean Trans. Time		Mean Syst. Profit	
		Arrv.	Serv.	Indv.	Soci.	Indv.	Soci.	Indv.	Soci.
$M/E_k/1/K$	k= 1	1.00	2.25	13	3	16.34	3.42	23.7	142.1
	k= 5	1.00	0.45	13	3	17.24	3.32	14.6	152.3
	k= ∞	1.00	0.00	13	3	17.43	3.32	12.6	156.1
$H_4/H_3/1/K$		2.50	11.53	13	2	17.90	2.34	8.1	160.4
$E_3/W_3/1/K$		0.33	0.81	13	2	18.23	2.33	4.7	165.7
$W_{1/3}/E_3/1/K$		19.00	0.75	13	5	13.05	5.14	41.5	87.6
$W_{1/2}/GE/1/K$		5.38	1.17	13	4	14.4	4.21	40.6	117.1
$W_2/W_{1/2}/1/K$		0.60	12.11	13	3	13.21	3.43	53.28	127.6

are the same. Although in light traffic one does not observe a significant difference between some of the individual or social transit times (see column 4 of Table 4.1), in individual optimization, the system's profit is significantly reduced (see column 5 of Table 4.1). In heavy traffic and nonstationary queues, the differences between transit times and the system's profits, when optimized individually or socially, are more significant.

As the customers' estimate of their gains is based only on $1/\mu$ and not on the arrival and service processes, in some cases they might even incure some losses in entering the system. In light traffic, the difference between two mean transit times in individual and social optimization may not be significant in some cases, but in heavy and non-stationary queues the difference is significant (see column 4 of the tables).

In columns 2, we have computed the variances of interarrival and service time distributions. The mean of Weibull distribution is

$$E(X) = \beta\,\Gamma\left(\frac{1}{\alpha} + 1\right) \text{ and by using Stirling's formula}$$

$$\cong \beta \frac{\sqrt{\frac{2\pi}{\alpha}}}{(\alpha e)^{1/\alpha}} \qquad (4.1)$$

In order to obtain a better approximation for β in using Stirling's formula, β is first calculated from (4.1) considering the given value of E(x). Then the simulation model is run with different values of β, around its calculated value. By using the fact that for $K = 1$, the mean transit time is equal to mean service time, the approximation is thus improved.

Considering the mean transit times and the variances of service times in $M/E_k/1/K$ queueing models, one can observe that for light and heavy traffic, even when the system's capacity is finite, the trend in Pollaczek-Khintchin's formula [3] is observed. This is also true for the cases where interarrival process is non-Markovian. For instance, for $W_2/W_{1/2}/1/K$ when variance of service times are 0.18, 4.86 and 12.11, the mean transit times are 1.62, 2.32 and 3.29 respectively.

Considering the graphs in Figures 4.2 and 4.3, it is observed that as K increases, the confidence intervals of Z and W widens. This means that the variance of transit time is an increasing function of the system capacity.

Conclusion

From this study and for the queueing models that were considered the following conclusions are drawn:

1. Irrespective of traffic intensity and queueing models, Z and W are concave in K.
2. The K obtained from social optimization is always equal or less than the K obtained from individual optimization.
3. The social optimization preserves the individual gains and prevents the customer from leaving the system.
4. The trend in Pollaczek-Khintchin's formula is observed irrespective of the arrival process and traffic intensity.
5. The variance of transit time is an increasing function of the system capacity.

Analysis of the sensitivity of the service and interarrival time distributions' moments on the distribution of transit times is an interesting topic which the authors are working on at present.

References

1. Fishman, G.S. *Principles of discrete event simulation*, p. 263. John Wiley, New York, (1978).
2. Gross, D. and Harris, C.M. *Fundamentals of queueing theory*, p. 263. John Wiley, New York, (1985).
3. Heyman, D.P. and Sobel, M.J. *Stochastic models in operations research*, Vol. 1. McGraw-Hill, New York (1982).
4. Law, A.M. and Kelton, W.D. *Simulation modeling and analysis*, pp. 428-431. McGraw Hill, (1991).
5. Pidd, M. *Computer simulation in management science*. John Wiley, New York, (1986).
6. Rue, R.C. and Rosenshine, M. Some properties of optimal control policies for entries to an M/M/1 queue. *Nav. Res. Log. Quart.*, **28**, 225-232, (1981).
7. Sargent, R.G. *Validating simulation models*. Proceedings of the 1983 Winter Simulation Conference. Syracuse Univ., Syracuse, New York, (1983).
8. Shannon, R.E. *Systems simulation: the art and science*. Prentice Hall, (1975).
9. Stidham, S. Jr. *Optimal control of arrivals to queues and network of queues*. 21st IEEE Conf. on Decision / Control, (1982).
10. Stoer, J. and Bulirsch, R. *Introduction to numerical analysis*. Springer Verlag, Berlin, Heidlberg, New York, (1980).