A Reprint from

ISRAEL
Journal of
TECHNOLOGY

The Weizmann Science Press of Israel

Analysis of the Telephone Information Service

URI YECHIALI

Tel Aviv University, Tel Aviv, Israel
and
Israel Ministry of Communications

Recieved April 27, 1975

ABSTRACT

A queueing-analysis of the Telephone Information Service is carried out. The system is first considered as an $M/E_2/s/N$ queue and system-parameters are derived. It is then compared with the simpler $M/M/s/N$ queue and it is discovered that for low and medium offered loads the latter model may serve as a good approximation for the former. Numerical calculations and several graphs are added to illustrate the analytical results and their implications.

1. INTRODUCTION

The problem of determining the number of operators required to handle customer requests for service is central in any design of a queueing system. In most service systems this number varies as a function of time. In particular, the number of telephone operators required on duty at switchboards in the Telephone Information Service fluctuates widely during the day, where increased demand is experienced during certain busy hours. The actual number of operators required at any given hour is generally determined by the arrival rate of customers, the distribution of holding times and a service criterion that specifies, for example, that no more than α per cent of customers be delayed more than β units of time.

Recently, M. Segal (1974) considered the problem of assigning and scheduling operators to "trics" (= working shifts that contain several relief periods) and proposed a method of solution based on the out-of-kilter algorithm. However, Segal—as well as many other authors — assumes that the holding times are distributed according to a negative-exponential distribution. This assumption — although not so realistic — is usually made in queueing studies for case of analysis. Clearly, one would like to know how good this assumption is and if it can be justified by other arguments.

Statistical studies of the distribution of holding times of subscribers' requests for assistance from the Information Service in Tel Aviv have revealed (See *Yechiali et al.*, 1973) that these times are best approximated by the Erlang distribution of order two rather than by the Exponential distribution. It thus raised the problem of analyzing such a system in order to be able to assign and schedule operators to shifts in an efficient way. Such an analysis is also needed for purpose of comparison with the easier — and most frequently used — model that assumes exponential holding times.

In this paper we study a system denoted as an $M/E_2/s/N$ queue. This mathematical model is a representation of the following physical system that exists in the Telephone Information Service in Israel.

Subscribers request information (service) according to a Poisson stream, s operators are assigned to the switchboards to handle the subscribers' requests. The holding (service) time of requests possesses the Erlang distribution of order two. There are N waiting positions in the queue, and the queue discipline is FIFO.

Let n denote the number of calls in the system at instant of arrival. A new call enters service with no delay if it finds $n < s$ subscribers in the system. A customer is blocked and delayed if he finds $s \leqq n < s + N$ subscribers ahead of him. In such a case the delayed customer hears a pre-recorded announcement informing him that all the operators are busy. A call that arrives when all s servers are busy and all N waiting positions are occupied is blocked and lost.

Analytic study of the $M/E_k/s/\infty$ system has appeared in the literature (*Heffer*, 1969) but the results there are not easily available for numerical calculations. Such calculations are needed for purpose of managing and controlling the system and therefore we present here a method which is much simpler and is readily available for detailed calculations. At this point we would like to stress that in reality unbounded queues do not exist. The unbounded queue is strictly a mathematically convenient abstraction. In practical situations there is always a finite waiting room, usually with loss of those customers who find the waiting room full.

The numerical results obtained for the $M/E_2/s/N$ queue are then compared with numerical results derived by using the well known formulas for the $M/M/s/N$ queue. It is then discovered that for low and medium loads the numerical results for both models differ from each other by only a small percentage. This implies that in most cases the simple and well-known formulas for the $M/M/s/N$ queue may serve as good approximations for the more complex $M/E_2/s/N$ queue, and can be used to determine the number of operators required to handle customers' requests.

2. THE $M/E_2/s/N$ QUEUE

Consider a service facility with s channels in parallel, into which customers arrive according to a homogeneous Poisson process with rate $\lambda > 0$ and are served on a first-come, first-served basis. If all channels are occupied customers form a queue whose length is limited by a finite waiting room of size N. Customers that find $s + N$ units in the system are lost. Suppose the service time for each channel is a random variable distributed according to the Erlang density function

$$f(t) = \theta^2 t e^{-\theta t}, \qquad t \geqq 0, \theta > 0. \tag{1}$$

Such a service time is distributed as a sum of two independent and identically distributed random variables, each with the exponential density function

$$g(t) = \theta e^{-\theta t}, \qquad t \geqq 0. \tag{2}$$

Consequently, we can consider each customer as undergoing two consecutive *phases* of service, remaining in each phase for a duration that is distributed according to the density $g(\cdot)$. A channel becomes *free* only upon *completion* of the *second* phase of sevice.

At any time t the system can be described by the state space $\{(i,k): i = 0, 1, 2, \cdots, s + N;\ k = i, i + 1, \cdots, 2i$ for $i \leqq s$ and $k = 2i - s, 2i - s + 1, \cdots, 2i$ for $i > s\}$, where i denotes the number of customers in the system (waiting and in service) and k denotes the total number of phases remaining to complete the service of the i customers present in the system at time t.

This description enables us to consider the system as a continuous-time Markov chain of birth-and-death type. Following standard methods of writing the Chapman-Kolmogorov differential equations and passing into limit as $t \to \infty$ we obtain the steady state equations of the system. Let P_{ik} denote the limiting probability that the system is in state (i,k), then the steady state equations may be written in a compact form as

$$\pi A = 0 \tag{3}$$

where

$$\begin{aligned}\pi = (&P_{00}; P_{11}, P_{12}; P_{22}, P_{23}, P_{24}; \cdots; P_{ss}, \cdots, P_{s,2s};\\ &P_{s+1,s+2}, \cdots, P_{s+1,2(s+1)}; P_{s+2,s+4}, \cdots,\\ &P_{s+2,2(s+2)}; \cdots; P_{s+N,2(s+N)-s}, \cdots, P_{s+N,2(s+N)})\end{aligned}$$

is the vector of the probabilities $\{P_{ik}\}$ and A is the infinitesimal generator matrix that its elements $a[(i,k),(i',k')]$ are given as follows:

For $0 \leqq i \leqq s$ and $i \leqq k \leqq 2i$

$$a[(i,k),(i+1,k+2)] = \lambda$$

$$a[(i,k),(i,k)] = -(\lambda + i\theta) \tag{4a}$$

For $1 \leqq i \leqq s$

$$a[i,k),(i,k-1) = (k-i)\theta, \quad k = i+1, i+2, \cdots, 2i$$

$$a[(i,k),(i-1,k-1)] = (2i-k)\theta, \quad k = i, i+1, \cdots, 2i \tag{4b}$$

For $s < i \leqq s+N-1$ and $2i-s \leqq k \leqq 2i$

$$a[(i,k),(i+1,k+2)] = \lambda$$

$$a[(i,k),(i,k)] = -(\lambda + s\theta) \tag{4c}$$

$$a[(i,k),(i,k-1)] = (s+k-2i)\theta$$

$$a[(i,k),(i-1,k-1)] = (2i-k)\theta$$

For $i = N+s$ and $2i-s \leqq k \leqq 2i$

$$a[(i,k),(i,k)] = -s\theta$$

$$a[(i,k),(i,k-1)] = (s+k-2i)\theta \tag{4d}$$

$$a[(i,k),(i-1,k-1) = (2i-k)\theta$$

The solution of the set (3) where one of the equations is replaced by

$$\sum_i \sum_k P_{ik} = 1 \tag{5}$$

yields a solution to the unknown probabilities $\{P_{ik}\}$ for any combination of numerical values of λ and θ. Note that, because of the limited waiting room and the fact that customers who find $s+N$ units in system are blocked and cleared, the system attains its steady state regime no matter what the relative values of λ and θ are.

Of interest is the distribution $\{P_{i\cdot}\}$ of the number of customers in the system. For any numerical solution of $\{P_{ik}\}$ these probabilities are calculated by

$$\begin{aligned} P_{i\cdot} &= \sum_{k=i}^{2i} P_{ik}, & 0 \leqq i \leqq s \\ P_{i\cdot} &= \sum_{k=2i-s}^{2i} P_{ik}, & s \leqq i \leqq s+N. \end{aligned} \tag{6}$$

(In the following, for ease of notation, we will write P_i instead of $P_{i\cdot}$).

The set of probabilities $\{P_i\}$ enables us to calculate various parameters of the system.

The mean number of units in system is given by

$$L = \sum_{i=1}^{s+N} i P_i \tag{7}$$

The mean number of subscribers who hear the pre-recorded announcement is

$$L_q = \sum_{j=1}^{N} j P_{s+j} \tag{8}$$

The fraction of time that a server is busy is

$$R = \frac{L - L_q}{s} \tag{9}$$

The probability of receiving a busy signal is $P_{s+N} = B$.

The probability of delay (hearing a pre-recorded announcement) is

$$P_D = \sum_{i=s}^{s+N-1} P_i \tag{10}$$

The mean number of unsuccessful attempts per unit time is $\lambda(1 - P_{s+N})$.

The probability of attaining service with no delay is $1 - P_D - P_{s+N}$.

The conditional mean delay of a customer who has been delayed is

$$[W_q \mid W_q > 0] = \frac{L_q}{\lambda P_D}. \tag{11}$$

3. THE $M/M/s/N$ QUEUE

For completeness of exposition we describe briefly this well known model and give the formulas for the parameters of interest.

The stream of customers is a homogeneous Poisson process with intensity λ. The service times are exponentially distributed with mean μ^{-1}. There are s identical servers and the waiting room is limited to N positions so that a call that finds $s+N$ subscribers in the system is blocked and cleared. The delay procedure is identical to the one described for the $M/E_2/s/N$ system.

Denote by P_i the steady state probability for having i units in the system ($i = 0, 1, 2, \cdots, s+N$).

Then,

$$P_i = \frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i P_0, \qquad i \leqq s \qquad P_{s+j} = \left(\frac{\lambda}{s\mu}\right)^j P_s, \quad 1 \leqq j \leqq s + N \tag{12}$$

$$P_0 = \left[\sum_{i=0}^{s}\frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s\left(\frac{\rho - \rho^{N+1}}{1-\rho}\right)\right]^{-1}$$

where

$$\rho = \frac{\lambda}{s\mu}.$$

$$L_q = \sum_{i=s+1}^{s+N}(i-s)P_i = P_0\frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s\frac{\rho}{(1-\rho)^2}[1-(N+1)\rho^N + N\rho^{N+1}] \tag{13}$$

$$L = \sum_{i=1}^{s+N} iP_i = \sum_{i=1}^{s} iP_i + \sum_{i=s+1}^{s+N}(i-s)P_i + s\sum_{i=s+1}^{s+N}P_i$$

$$= \sum_{i=1}^{s} iP_i + L_q + P_0\frac{1}{(s-1)!}\left(\frac{\lambda}{\mu}\right)^s\left(\frac{\rho-\rho^{N+1}}{1-\rho}\right) \tag{14}$$

$$[W_q \mid W_q > 0] = \frac{1}{P_D}\left[\frac{1}{\mu s}\sum_{i=s}^{s+N-1}(i-s+1)P_i\right] = \frac{1}{P_D}\frac{1}{\mu s}\sum_{j=1}^{N}\frac{1}{\rho}jP_{s+j} = \frac{L_q}{\lambda P_D} \tag{15}$$

$$R = \frac{L-L_q}{s} = \frac{1}{s}\left[\sum_{i=1}^{s} iP_i + s\sum_{i=s+1}^{s+N}P_i\right] = \frac{1}{s}\left[\frac{\lambda}{\mu}\sum_{i=1}^{s}\frac{(\lambda/\mu)^{i-1}}{(i-1)!}P_0\right.$$

$$\left. + \frac{\lambda}{\mu}\sum_{i=s+1}^{s+N}P_{i-1}\right] = \rho(1-P_{s+N}). \tag{16}$$

4. NUMERICAL RESULTS

Detailed numerical calculations and computer programs for the above two models are available (*Yechiali et al.*, 1973) We choose to present here a few illustrative graphs and tables that will give a better understanding of the previous results and their implications.

As it is evident from the analysis of the $M/E_2/s/N$ and $M/M/s/N$ models, all the quantities such as $\{P_i\}, L, L_q$, etc. are functions of the load $a \equiv \lambda E$ (service time) offered to the system. Clearly, $a = \lambda(2/\theta)$ for the Erlangian model and $a = \lambda/\mu$ for the Exponential case. Thus, all the figures and tables in the sequel will be constructed as functions of a.

We start with Fig. 1 which gives the probability of receiving a busy signal, B, as a function of s and a when $N = 3$. It is clear that when the offered load is not too high the results for the two models differ from each other by a small percentage only.

The implication of Fig. 1 (and similar figures for various values of s and N) is that the simpler $M/M/s/N$ model can be used for actually assigning

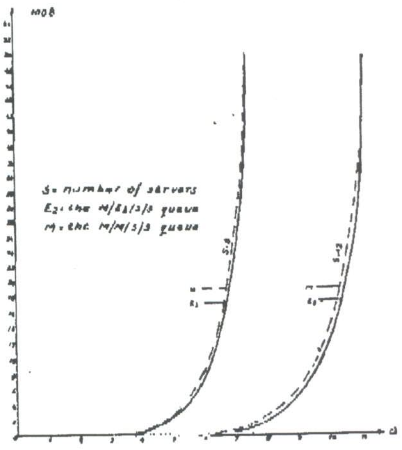

Fig. 1. The probability of receiving a busy signal as a function of the offered load a.

operators to shifts depending on the offered load and the service criteria.

Figure 2 shows how B is changed with s for the $M/M/s/3$ queue.

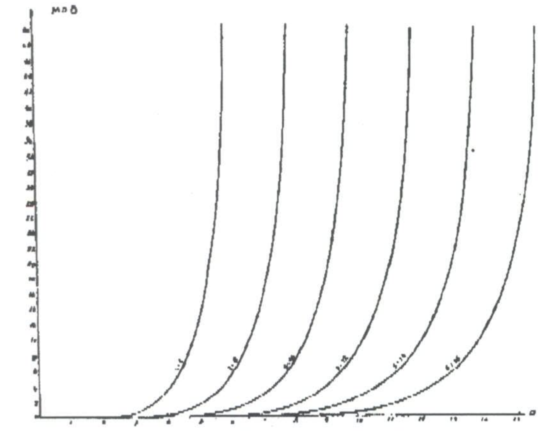

Fig. 2. The dependance of the probability of a busy signal on the number of operators s for the $M/M/s/3$ model.

Figure 3 illustrates that practically the busy fraction of a server is a linear function of the offered load a.

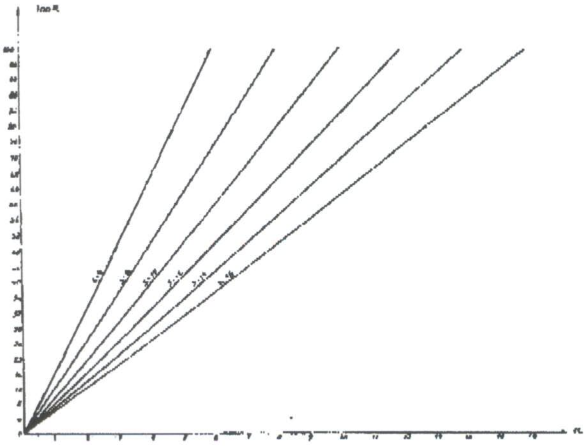

Fig. 3. The busy fraction of each server in the $M/M/s/3$ model.

Figure 4 demonstrates the sensitivity of B to the changes in the number of waiting positions N.

If one is interested in the maximal load that may be offered to a group of servers such that the probability of a busy signal will not exceed a preassigned grade of service, then for the $M/M/s/3$ model Table I is constructed. This table may be used for actually assigning operators to shifts.

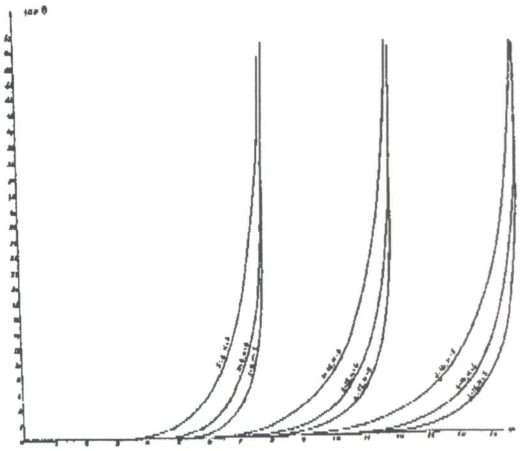

Fig. 4. The probability of a busy signal as a function of the number of waiting positions N for the $M/M/s/N$ model.

TABLE I

MAXIMAL OFFERED LOAD (IN ERLANGS) IN THE $M/M/s/3$ MODEL FOR DIFFERENT VALUES OF GRADE OF SERVICE

B S	6	8	10	12	14	16
0.01	3.20	4.48	6.08	7.57	8.93	10.67
0.02	3.57	4.93	6.67	8.27	9.81	11.52
0.03	3.87	5.31	7.04	8.69	10.35	12.11
0.05	4.27	6.35	7.55	9.28	11.04	12.85
0.10	4.85	6.56	8.35	10.19	12.05	13.95
0.20	5.44	6.69	9.20	11.12	13.01	14.99
0.30	5.76	7.60	9.63	11.55	13.52	15.49

For completeness of exposition we conclude by calculating Table II for the $M/E_2/s/N$ model were $s = 8$ and $N = 3$. Equations (7), (8), (9) and (10) are used for this purpose.

The values for $[W_q \mid W_q > 0]$ may be calculated from Table II, using Eq. (11), for any combination of λ and θ. For example, if $a = 12.0$ Erlang such that $\lambda = 300$ requests per hour (i.e., $\theta = 2\lambda/a = 50$ subscribers per hour) then

$$[W_q \mid W_q > 0] = \frac{L_q}{\lambda P_D} = \frac{1.73}{300 * 0.533} \text{ h} \sim 40 \text{ s}.$$

TABLE II

VARIOUS SYSTEM PARAMETERS FOR THE $M/E_2/8/3$ QUEUE AND THEIR DEPENDENCE ON THE OFFERED LOAD a

a	L	L_q	R	B	P_D
1.33	1.33	0.00	0.167	0.000	0.000
2.67	2.67	0.00	0.333	0.000	0.006
4.00	4.03	0.04	0.499	0.003	0.052
5.33	5.41	0.18	0.654	0.019	0.168
6.67	6.72	0.46	0.782	0.061	0.321
8.00	7.80	0.83	0.872	0.128	0.446
9.33	8.60	1.19	0.926	0.206	0.516
10.67	9.15	1.49	0,957	0.282	0.538
12.00	9.53	1.73	0.974	0.351	0.533
13.33	9.79	1.92	0.984	0.410	0.513
14.67	9.98	2.07	0.990	0.460	0.487
16.00	10.13	2.18	0.993	0.503	0.459

ACKNOWLEDGEMENT

We wish to thank Dr. M. Resh for suggesting the problem, to Ms. Z. Tavory and Mr. Y. Gon for actually performing the numerical calculations and to Mr. S. Krokin for many valuable and fruitful discussions.

REFERENCES

J. C. HEFFER, 1969. Steady-state solution of the $M/E_k/c$ (∞, FIFO) queueing system, *Can. Oper. Res. Soc.*, **7**, 16–30.

M. SEGAL, 1974. The operator-scheduling problem: A network-flow approach, *Oper. Res.*, **22**, 808–823.

U. YECHIALI, Y. GON, Z. TAVORI, AND S. KROKIN, 1973. Analysis of information-service "14" in Tel Aviv, Internal Report, Israel Ministry of Communication (in Hebrew).