ABSTRACT

Performance Analysis of Queueing Systems with Synchronous Server

by
Wonmu Hur

The performance measures are derived for a queueing system which is characterized by its *synchronous* server operations. This property is the one which normal queueing systems lack. A system of this nature contains a single common server to which multiple number of buffers are connected in parallel, where the server operates ***only when*** all system buffers are occupied.

In this thesis, the analysis is carried out focusing on the simplest system involving only two buffers. Throughout the analysis, a symmetricity of the system is assumed(i.e., system is symmetrical in terms of arriving customer statistics.). Also assumed are the Poisson arrival and the exponential service time distribution.

As a first step, the behaviour of the original system is investigated and, as a result, it is converted into an equivalent M/G/1 with the new probability density function of the arriving packets. Once we obtain the equivalent model, we apply M/G/1 analyzing method to the model and derive the probability of queue length in a closed form. The deviation of the result from that of M/M/1 are considered. All the performance measures directly follow from the probability of queue length equation.

PERFORMANCE ANALYSIS OF QUEUEING SYSTEMS WITH SYNCHRONOUS SERVER

by
Wonmu Hur

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science
Department of Computer and Information Science
October 1992

APPROVAL PAGE

Performance Analysis of
Queueing Systems with Syschronous Server

by
Wonmu Hur

Dr. Daniel Y. Chao, Thesis Adviser
Assistant Professor of Computer and Information Science
Department, New Jersey Institute of Technology

BIOGRAPHICAL SKETCH

Author: Wonmu Hur

Degree: Master of Science in Computer and Information Science

Date: October, 1992

Education:

- Master of Science in Computer and Information Science, New Jersey Institute of Technology, Newark, New Jersey, 1992
- Bachelor of Science in Electrical Engineering, Seoul National University, Seoul, Korea, 1985

Major: Computer and Information Science

This thesis is dedicated to
K.M.C.

ACKNOWLEDGMENT

The author wishes to thank Dr. Daniel Chao, his thesis advisor, for his guidance and help throughout this research. The author is also grateful to Dr. David Wang for his reviewing the material.

TABLE OF CONTENTS

Page

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

When the property of synchronization is incorporated into a queueing system, it may model a processing system in which it is required that customers(data packets) of each different class be processed simultaneously and the processing be halted if a customer of *any* class is not available in the system. In the simplest case, which involves two buffers connected in parallel to a common server, the server operates *only when* both buffers are occupied. If any of them is empty at any time, the server stops its operation and remains idle until that empty buffer is occupied again. The system is drawn in figure (1).

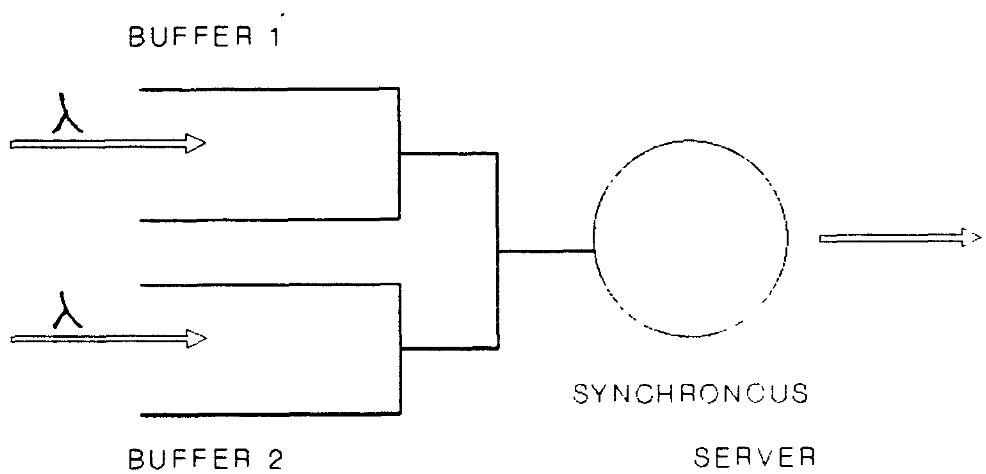

Figure 1: A queueing system with a sinchronous server

Refering to the above picture, the system's operating and non-operating conditions are summarized as follows.

operating condition

1. both buffer 1 and buffer 2 are occupied

non-operating conditions

1. buffer 1 is empty and buffer 2 is occupied
2. buffer 2 is empty and buffer 1 is occupied
3. both buffer 1 and buffer 2 are empty

Due to the presence of the property of synchronization, the performance characteristics of these systems are quite deviated from those of normal queueing systems. In this paper, we derive performance measures such as average queue length, average waiting time, average delay and throughput for the simplest system. Throughout the analysis, we assume;

1. Infinite buffer capacity
2. Poisson arrival process(exponential inter-arrival distribution)
3. Exponential service time distribution
4. Symmetrical system – each buffer is identical in terms of arriving data packet statistics.

In queueing system analysis, a state transition diagram may be constructed to derive a set of balence equations. In some limited simple cases, this set of balence equations leads to the closed form solution to the probability of queue length, from which, in turn, performance measures are drawn. As the dimension or complexity of the diagram grows, however, this approach becomes infeasible. In our case, the state transition diagram has infinite dimensions in both horizontal and vertical directions. Thus, instead of using state transition diagram, we take an alternative approach.

From our assumption of the system's symmetricity, we conclude that it suffices to limit ourselves to *either* of the two queues of the system and keep track of the packets arriving into it. Since the server operates only when both buffers are occupied, it is easily conceivable that *some portion of data packets get blocked when they attempt to enter the server.* This phenominon occurs when the queue on the other side happens to be empty at that moment. In this connection, we define the following terms.

- **server blocking** : The phenominon as stated above. (We call it so to distingush it from *blocking* in an ordinary sense which takes place due to a finite capacity of a buffer.)
- **server blocking period** : the interval between the instant a packet at the front of a buffer finds the server empty and the instant it actually enters the server.

Obviously, in ordinary M/M/1 systems, server blockings do not happen and therefore the duration of server blocking period is always zero.

CHAPTER 2
M/G/1 CONVERSION

To begin with, we convert the original system into the one which is equivalent to M/G/1. Then we derive the *probability density function*(pdf) of the service time for this equivalent system. This pdf is used later when we derive the probability of queue length using *Pollaczec-Khinchin transform equation.* We start with introducing the concept of *effective service time.*

2.1 Effective Sevice Time

The existence of the server blocking phenomina makes it cumbersome to analyze the system directly. To get around this difficulty, we eliminate these server blocking phenomina by taking a different viewpoint of the phenomina as described below.

To facilitate our analysis, we divide the whole incoming data packets into the following two categories.

Type 1 packets : packets which do not suffer *server blockings.*

Type 2 packets : packets which suffer *server blocking blockings.*

Note that this classification is *not* based on the intrinsic nature of a particular packet. As a matter of fact, *any* packet could fall into either Type 1 or Type 2 category. It is a sheer luck which category an arriving packet eventually falls into.

According to this classification, a Type 2 packet is a packet which experiences blocking at the time when it tries to enter the idle server. It therefore must wait an extra amount of time (*server blocking period*) until there arrives a packet into the other buffer which was empty.

The key point in our analysis is to *treat* this server blocking period *as if it were an extension of the service time associated with the packet.* Then we regard this packet as having a longer service time than its original one. This is just a different way of interpretation of the *same* phenomina, preserving the original properties of the system. As far as the performance evaluation is concerned, this interpration works because it clearly doesn't affect the queue length characteristics of the system. (Remember that a queueing system is fully described by its probability of queue length.)

This way of interpretation yields the concept of *effective service time.* i.e.,

$$l_e = l_o + l_b \tag{1}$$

where

l_e is the *effective* service time of a packet, l_o is the original(actual) service time of a packet and l_b is the length of server blocking period. ($l_b = 0$ for Type 1 packets and $l_b > 0$ for Type 2 packets.)

The time diagram of the typical behaviours of Type 1 and Type 2 packets are illustrated in Figure (2). The importance of l_e in our analysis is that if we use l_e intead of l_o, we will no longer have to consider the server blocking effect separately in our analysis.

From (1), however, it is easily seen that the new probability distribution of l_e is no longer exponential (as is l_o) due to the inclusion of l_b. Instead, it becomes a general distribution and consequently, the system can be analyzed as if it were an M/G/1.

2.2 PDF of Effective Service Time

In order to obtain the pdf of l_e, the followings must be known;

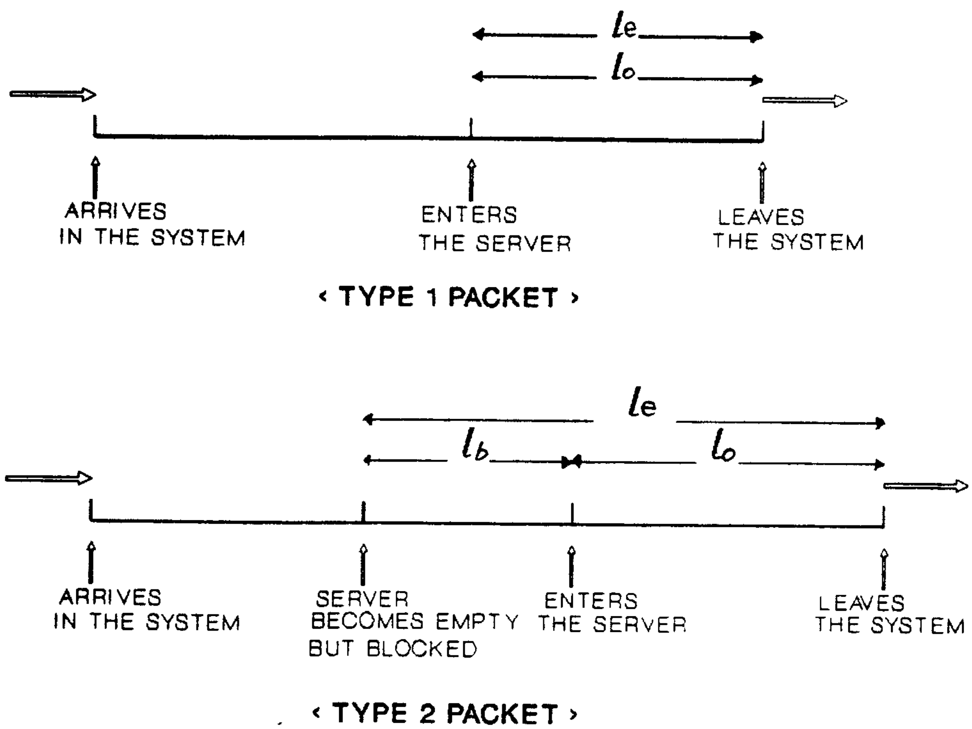

Figure 2: Time diagram for Type 1 and Type 2 packets

1. The probability that a packet suffers a server blocking(or, it falls into Type 2 category)

2. The pdf of the server blocking period

Considering that a server blocking on one side of the system is due to an idleness of the opposite side and it durates until that idle period ends, we can draw the following reasonable conclusions.

1. *The probability that a packet suffers a server blocking is equal to the probabilty that a queue is idle.*

2. *The pdf of server blocking period is equal to the pdf of idle period.*

If we denote the probability of a queue being idle as p_0, the first conclusion is equivalent to saying that p_0 times the total number of packets are classified as Type 2 and the remaining $(1 - p_0)$ times the total number of packets are classified as Type 1 packets. Consequently, the pdf of the effective service time can be represented as the following weighted sum.

$$b(x) = (1 - p_0) \cdot b_1(x) + p_0 \cdot b_2(x) \tag{2}$$

where $b_1(x)$ and $b_2(x)$ are the pdfs of effective service times for Type 1 packets and Type 2 packets respectively.

$b_1(x)$, the pdf of effective service time for Type 1 packets

Since a Type 1 packet's effective service time is equal to its actual service time,

$$b_1(x) = \mu e^{-\mu x} \tag{3}$$

$b_2(x)$, the pdf of effective service time for Type 2 packets

From our second conclusion(*the pdf of server blocking period is equal to the pdf of idle period*) and equation (1), the effective service time of a Type 2 packet can be rewritten as

$$\begin{aligned} l_e &= l_o + l_b \\ &= l_o + l_\imath \end{aligned} \tag{4}$$

where l_i denotes the length of the idle period(of the queue on the opposite side).

The above equation suggests that $b_2(x)$ can be derived from the pdfs of l_o and $l_\imath$. Let us denote them by $b_{2_o}(x)$ and $b_{2_\imath}(x)$ respectively. Obviously $b_{2_o}(x) = \mu e^{-\mu x}$. And, $b_{2i}(x) = \lambda e^{-\lambda x}$ since, in M/G/1, the pdf of idle period is given by $\lambda e^{-\lambda t}$.

We now turn to deriving $b_2(x)$, the composite pdf of those for l_o and l_e. Refering to equation (4), we can say;

the probability of l_e having the value of x is equal to the probability of l_o having the value of τ <u>and</u> the probability of l_i having the value of $x-\tau$, where $0 \le \tau \le x$.

Mathematically,

$$P(l_e = x) = \sum_{all\ \tau} P(l_o = \tau) \cdot P(l_i = x - \tau) \tag{5}$$

However, the summation on the right side of the above equation must be carried out over infinitely many combinations of $P(l_o = \tau) \cdot P(l_i = x - \tau)$ because τ is a continuous variable taking any value between 0 and x. Consequently, $b_2(x)$ can be obtained by the following integration;

$$\begin{aligned} b_2(x) &= \int_0^x b_{2o}(\tau) \cdot b_{2i}(x-\tau)\, d\tau \\ &= \int_0^x \mu e^{-\mu\tau} \cdot \lambda e^{-\lambda(x-\tau)}\, d\tau \\ &= \frac{\mu\lambda}{\mu-\lambda}(e^{-\lambda x} - e^{-\mu x}) \end{aligned} \tag{6}$$

$b(x)$, the pdf of effective service time

Substituting (3) and (6) into (2), we obtain

$$\begin{aligned} b(x) &= (1-p_0)\cdot \mu e^{-\mu x} + p_0 \cdot \frac{\mu\lambda}{\mu-\lambda}(e^{-\lambda x} - e^{-\mu x}) \\ &= (\mu - \frac{p_0\mu^2}{\mu-\lambda})e^{-\mu x} + \frac{p_0\mu\lambda}{\mu-\lambda}e^{-\lambda x} \\ &= \mu\left((1 - p_0\frac{1}{1-\rho})e^{-\mu x} + p_0\frac{\rho}{1-\rho}e^{-\lambda x}\right) \end{aligned} \tag{7}$$

with $\rho \equiv \lambda/\mu$.

The *effective* average length of the packet is then

$$\begin{aligned} \frac{1}{\mu_e} &= \int_0^\infty x \cdot b(x) dx \\ &= \int_0^\infty x \cdot \left[\mu\left((1 - p_0 \cdot \frac{1}{1-\rho})e^{-\mu x} + p_0\frac{\rho}{1-\rho}e^{-\lambda x}\right)\right] dx \\ &= \frac{1}{\mu}(1 + p_0 \cdot \frac{1}{\rho}) \end{aligned} \tag{8}$$

and the *effective* average traffic intensity is

$$\begin{aligned} \rho_e &= \lambda \cdot \frac{1}{\mu_e} \\ &= \lambda \cdot \frac{1}{\mu}(1 + p_0 \cdot \frac{1}{\rho}) \\ &= \rho + p_0 \end{aligned} \tag{9}$$

On the other hand, in an M/G/1 system, the following relationship holds;

$$p_0 = 1 - \rho_e \tag{10}$$

Solving (9) and (10) simultaneously, we obtain

$$\rho_e = \frac{1+\rho}{2} \tag{11}$$

and,

$$p_0 = \frac{1-\rho}{2} \tag{12}$$

Substituting (12) into (7) leads to

$$b(x) = \frac{\mu}{2}e^{-\mu x} + \frac{\lambda}{2}e^{-\lambda x} \tag{13}$$

This is the desired probability density function of the service time for our M/G/1 equivalent system.

In conclusion, the original system with average arrival rate λ, probability distribution of service time $\mu e^{-\mu t}$ and traffic intensity ρ has been converted into an M/G/1 system with the same average arrival rate λ, probability distribution $(\mu/2)e^{-\mu x} + (\lambda/2)e^{-\lambda x}$ and traffic intensity $(1+\rho)/2$. (Again, it must be noted that the server blocking effect in the original system has been removed as a result of the conversion.) This is our object system for which various performance evaluations are to be made.

CHAPTER 3
PERFORMANCE MEASURES

Since we obtained $b(x)$, we are now ready to derive the system performance measures — average queue length, average waiting time and average delay. Our first step is to find the $p(n)$, the probability of queue length. For this purpose, the *Pollaczec-Khinchin transform equation* is used. i.e.,

$$Q(z) = B(\lambda - \lambda z)\frac{(1-\rho_e)(1-z)}{B(\lambda - \lambda z) - z} \tag{14}$$

where $Q(z)$ is the z-tranform of $p(n)$, $B(s)$ is the Laplace transform of $b(x)$, which was obtained in (13), and ρ_e is the traffic intensity in our M/G/1 equivalent system which was obtained in (11). Once $Q(z)$ is known, $p(n)$ can be obtained from it by taking inverse z-tranformation.

3.1 Probability of Queue Length

To begin with, $B(s)$ must be found. From the definition of Laplace tranformation,

$$\begin{aligned} B(s) &= \int_0^\infty e^{-sx}\, b(x)\, dx \\ &= \int_0^\infty e^{-sx}\left(\frac{\mu}{2}e^{-\mu x} + \frac{\lambda}{2}e^{-\lambda x}\right) dx \\ &= \frac{(\mu+\lambda)s + 2\mu\lambda}{2(s+\mu)(s+\lambda)} \end{aligned} \tag{15}$$

Therefore,

$$B(\lambda - \lambda z) = \frac{(\mu+\lambda)(\lambda - \lambda z) + 2\mu\lambda}{2(\lambda - \lambda z + \mu)(\lambda - \lambda z + \lambda)} \tag{16}$$

Substituting this result into equation (14), we obtain

$$\begin{aligned} Q(z) &= \frac{(\mu+\lambda)(\lambda-\lambda z) + 2\mu\lambda}{2(\lambda - \lambda z + \mu)(\lambda - \lambda z + \lambda)} \cdot \frac{(1-\rho_e)(1-z)}{\dfrac{(\mu+\lambda)(\lambda-\lambda z)+2\mu\lambda}{2(\lambda-\lambda z+\mu)(\lambda-\lambda z+\lambda)} - z} \\ &= \frac{[(\mu+\lambda)(\lambda-\lambda z) + 2\mu\lambda]\,(1-\rho_e)(1-x)}{(\mu+\lambda)(\lambda-\lambda z) + 2\mu\lambda - 2z(\lambda-\lambda z+\mu)(\lambda-\lambda z+\lambda)} \end{aligned} \tag{17}$$

After dividing both numerator and denominator by $\mu\lambda$, the above equation becomes

$$Q(z) = \frac{N(z)}{D(z)} = \frac{[(1+\rho)(1-z)+2]\,(1-\rho_e)(1-z)}{(1+\rho)(1-z)+2-2z(\rho-\rho z+1)(2-z)} \tag{18}$$

Note that $N(1) = D(1) = 0$, which means the numerator and denominator polynomials have the common factor $(1-z)$. (i.e., $N(z) = (1-z)N_1(z)$, $D(z) = (1-z)D_1(z)$) Eliminating this common factor from $N(z)$ and $E(z)$, we obtain

$$Q(z) = \frac{[(1+\rho)(1-z)+2]\,(1-\rho_e)}{(1+\rho)(1-z)-2\rho z(2-z)+2(2-z)} \tag{19}$$

Substituting the result of (11) and after some algebraic manipulations, the above equation finally becomes

$$Q(z) = \frac{1-\rho}{4\rho} \cdot \frac{(3+\rho)-(1+\rho)z}{z^2 - \left(2+\frac{1}{\rho}\right)z + \left(\frac{3+\rho}{2\rho}\right)} \tag{20}$$

with $\rho \equiv \lambda/\mu$.

In order to apply inverse z-transform, we decompose the above equation as the following form;

$$Q(z) = \frac{1-\rho}{4\rho}\left(\frac{C_1}{z-\alpha} + \frac{C_2}{z-\beta}\right) \tag{21}$$

where α and β are the roots of the denominator of equation (20), and, C_1 and C_2 are appropriate constants.

Once C_1, C_2, α and β are known, the inverse z-transform of (21) takes the form of

$$p(n) = \frac{1-\rho}{4\rho}\left[-\frac{C_1}{\alpha}\left(\frac{1}{\alpha}\right)^n - \frac{C_2}{\beta}\left(\frac{1}{\beta}\right)^n\right] \tag{22}$$

Actual computation leads to;

$$\begin{aligned}
\alpha &= \frac{2\rho+1+\sqrt{2\rho^2-2\rho+1}}{2\rho} \\
\beta &= \frac{2\rho+1-\sqrt{2\rho^2-2\rho+1}}{2\rho} \\
C_1 &= -\frac{(2\rho+1+\sqrt{2\rho^2-2\rho+1})(-\rho+\sqrt{2\rho^2-2\rho+1})}{2\sqrt{2\rho^2-2\rho+1}} \\
C_2 &= -\frac{(2\rho+1-\sqrt{2\rho^2-2\rho+1})(\rho+\sqrt{2\rho^2-2\rho+1})}{2\sqrt{2\rho^2-2\rho+1}}
\end{aligned}$$

Substituting the above results into (22), we obtain

$$p(n) = \frac{(1-\rho)(-\rho+\sqrt{2\rho^2-2\rho+1})}{4\sqrt{2\rho^2-2\rho+1}}\left(\frac{2\rho}{2\rho+1+\sqrt{2\rho^2-2\rho+1}}\right)^n + \frac{(1-\rho)(\rho+\sqrt{2\rho^2-2\rho+1})}{4\sqrt{2\rho^2-2\rho+1}}\left(\frac{2\rho}{2\rho+1-\sqrt{2\rho^2-2\rho+1}}\right)^n \quad (23)$$

It is of interest to compare the above result with $p(n) = \rho(1-\rho)^n$, the probability of queue length in M/M/1. Figure (3) shows the comparison using the particular value of $\rho = 0.5$. With $\rho = 0.5$, equation (23) becomes $p(n) = 0.0366(0.369)^n + 0.2134(0.773)^n$ and $p(n) = 0.5(0.5)^n$ for M/M/1. As illustrated in Figure (3), the probability of the system having n customers is lager than that of M/M/1 for $n > 1$. This is expected result since the server blocking effect causes the packets to build up in the buffer, thus increases the probability of queue length for large n.

3.2 Average Queue Length

$E(n)$ can directly be derived from (20) in the following way.

Since $Q(z) \equiv \sum_{n=0}^{\infty} z^n p(n) = z^0 p(0) + z^1 p(1) + z^2 p(2) + z^3 p(3) + \ldots$

$$\frac{dQ(z)}{dz} = \sum_{n=0}^{\infty} n \cdot z^{n-1} p(n) = 0 \cdot p(0) + 1 \cdot p(1) + 2z \cdot p(2) + 3z^2 \cdot p(3) + \ldots \quad (24)$$

On the other hand

$$E(n) = \sum_{n=0}^{\infty} n \cdot p(n) = 0 \cdot p(0) + 1 \cdot p(1) + 2 \cdot p(2) + 3 \cdot p(3) + \ldots \quad (25)$$

Comparing (24) and (25) we conclude

$$\begin{aligned} E(n) &= \left.\frac{dQ(z)}{dz}\right|_{z=1} \\ &= \left.\frac{d}{dz}\left(\frac{1-\rho}{4\rho} \cdot \frac{(3+\rho)-(1+\rho)z}{z^2 - \left(2+\frac{1}{\rho}\right)z + \left(\frac{3+\rho}{2\rho}\right)}\right)\right|_{z=1} \\ &= \frac{\rho^2+3}{2(1-\rho)} \end{aligned} \quad (26)$$

This result is illustrated in figure (4) along with $E(n) = \rho/(1-\rho)$ for M/M/1. It is noted from the above equation that the system is stable for $0 < \rho < 1$. Also note that even when $\rho = 0$, there are likely to be 1.5 customers in the queue. Simulation shows that there are non-zero customers in the queue even when $\rho = 0$

3.3 Average Delay

From Little's formula,

$$\begin{aligned} E(t) &= \frac{1}{\lambda} \cdot E(n) \\ &= \frac{1}{\lambda} \cdot \frac{\rho^2+3}{2(1-\rho)} \\ &= \frac{1}{\mu} \cdot \frac{\rho^2+3}{2\rho(1-\rho)} \end{aligned} \tag{27}$$

3.4 Average Waiting Time

Average waiting time is defined as the delay minus actual service time. It measures a packet's pure waiting time in the buffer on the average. i.e.,

$$\begin{aligned} E(w) &= E(t) - \frac{1}{\mu} \\ &= \frac{1}{\mu} \cdot \frac{\rho^2+3}{2\rho(1-\rho)} - \frac{1}{\mu} \\ &= \frac{1}{\mu} \cdot \frac{3\rho^2-2\rho+3}{2\rho(1-\rho)} \end{aligned}$$

or after normalizing

$$\mu E(w) = \frac{3\rho^2-2\rho+3}{2\rho(1-\rho)} \tag{28}$$

This result is illustrated in figure (5) along with $\mu E(w) = \rho/1-\rho$ for M/M/1. The most surprising deviation of the system's behaviour from that of M/M/1 is evident in this figure. When ρ approaches zero, the waiting time becomes infinity rather than becomes zero. The simulation results support this peculiar result.

3.5 Throughput

Since we assumed the buffers of infinite capacity, no incoming packets get lost and therefore the total arrival rate must be equal to the system throughput in equilibrium. i.e.,

$$\nu = 2\lambda \tag{29}$$

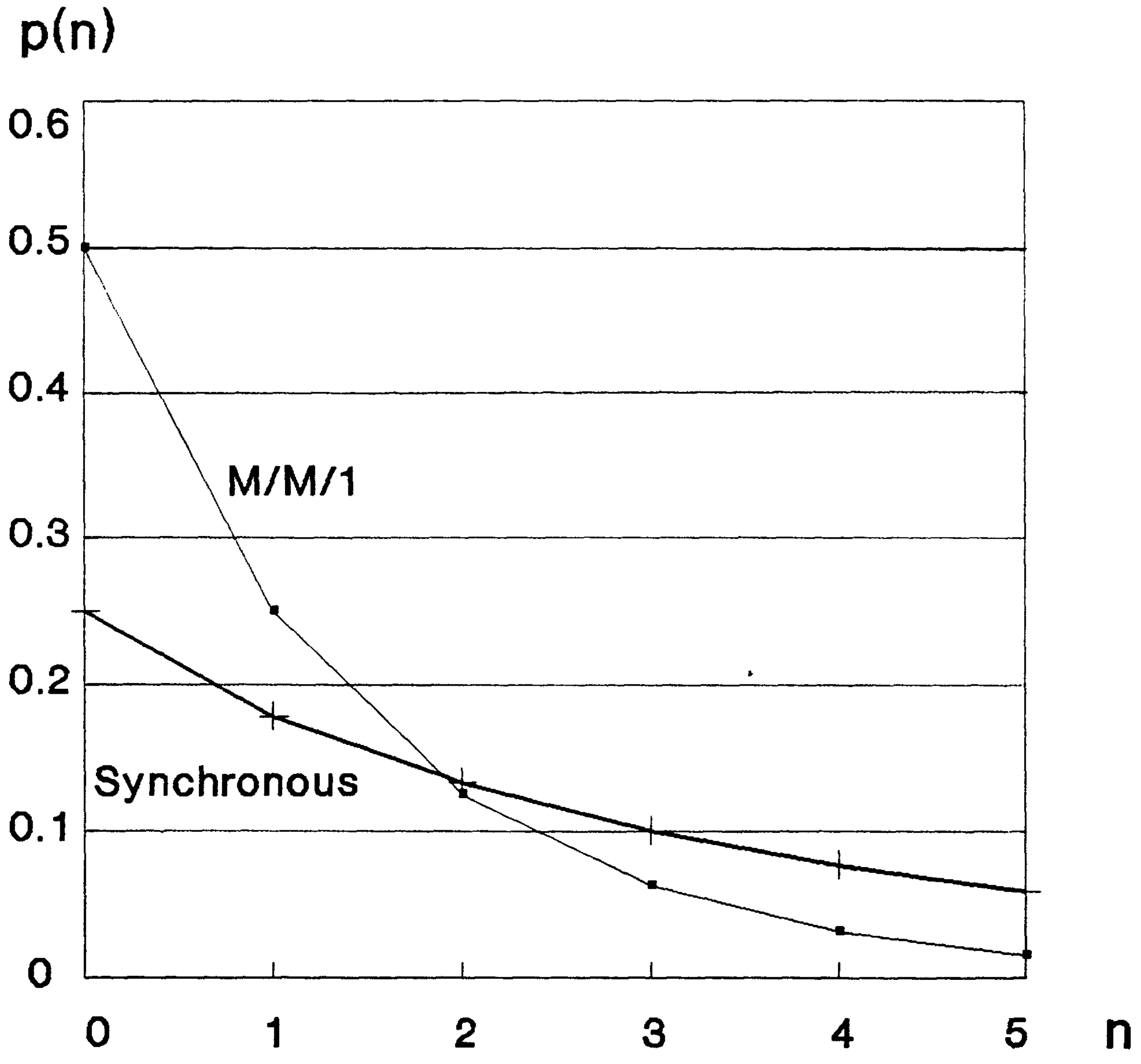

Figure 3: probability of queue length, $p(n)$, for $\rho = 0.5$

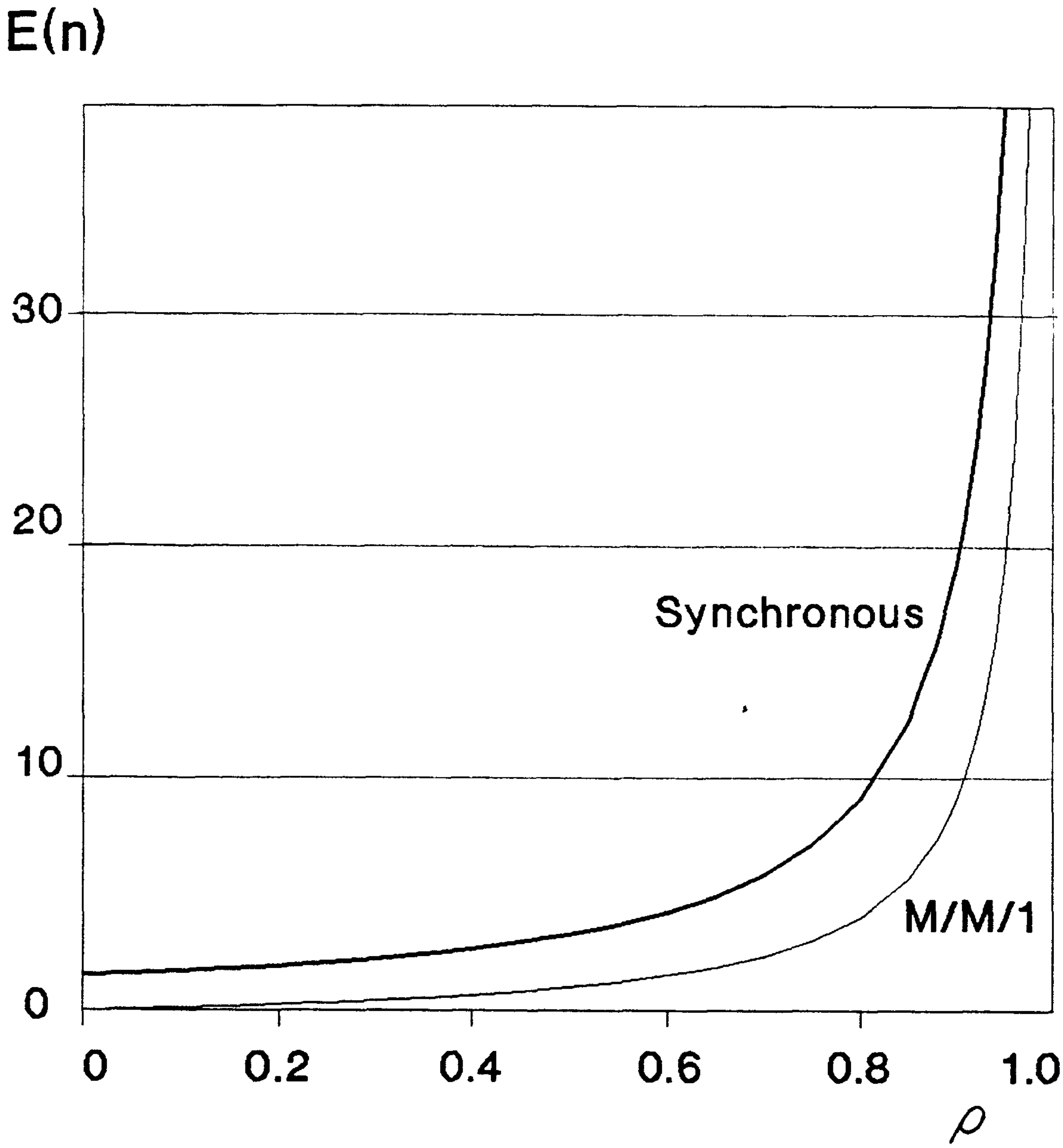

Figure 4: Average queue length, $E(n)$

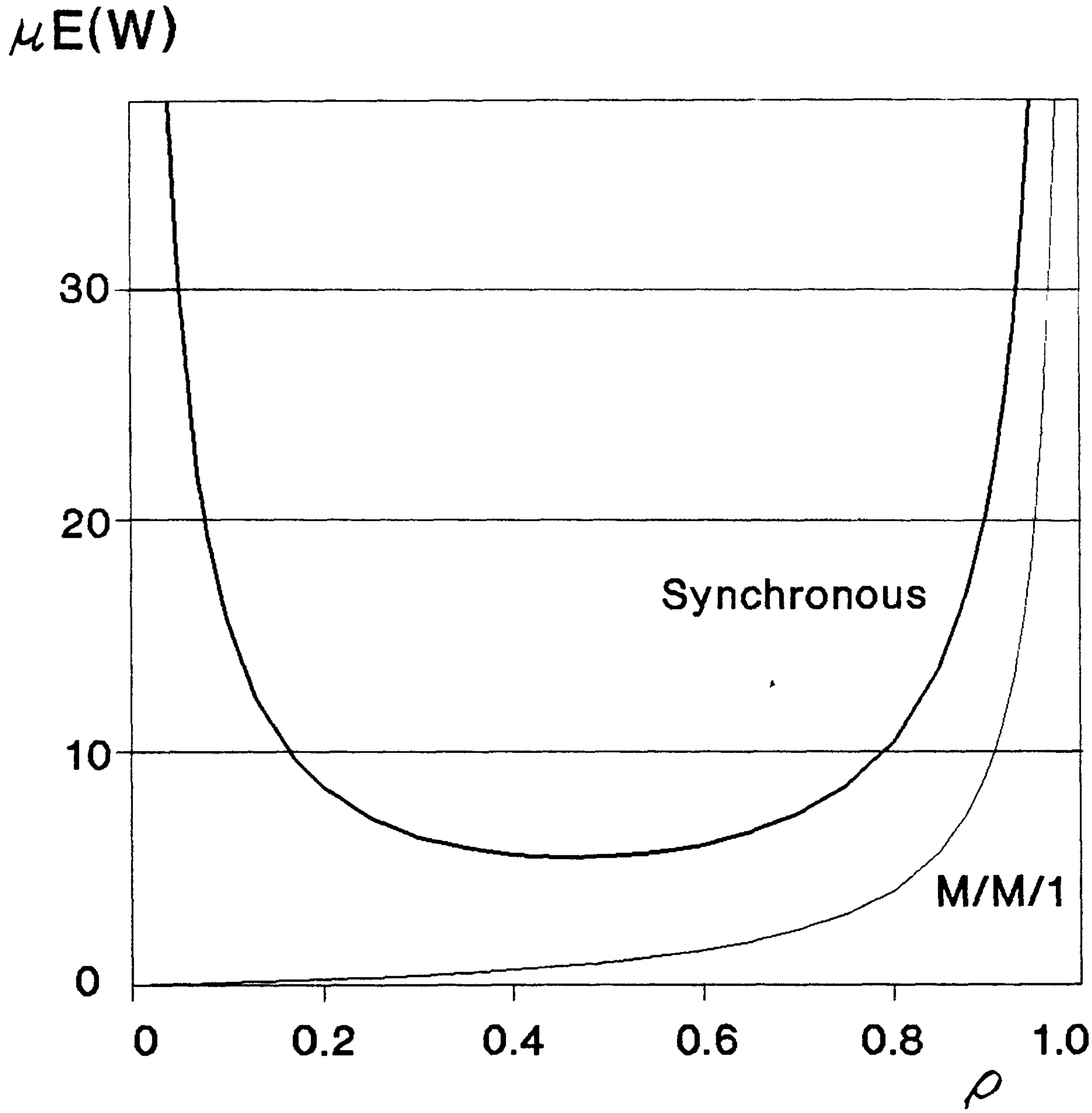

Figure 5: Average normalized waiting time, $\mu E(w)$

CHAPTER 4
CONCLUSION

The performance measures were derived in the queueing system where the server operates in a synchronous fashion. Even though our analysis was performed for the simplest system involving two buffers only, it gives an insight into the more general cases where more than two buffers are involved. The simulation results show good agreement with our theoretical works.

Our analysis was also based upon the assumption of the symmetricity of the system. Eliminating this assumption results in an asymmetric system where different buffer has different packet statistics. In particular, when the system has different average arrival rates for each buffer, it can easily be proven that the system is unstable. If the system has different average service times for each buffer while having the same average arrival rates, the analysis would be almost identical to our foregoing one with slight modifications.

BIBLIOGRAPHY

Kleinrock, L. "Queueing Systems, Volume 1 : Theory." John Wiley & Sons, New York, 1975.

Schwartz, Mischa. "Telecommunication Networks, Protocol, Modeling and Analysis." Addison-Wesley publishing company.

Schriber, Thomas J. "An Introduction to Simulation Using GPSS." John Wiley & Sons, New York, 1991.