POWER AND DETERMINISTIC RULES OF THUMB FOR PROBABILISTIC PROBLEMS IN COMPUTER COMMUNICATIONS*

Leonard Kleinrock

Computer Science Department
University of California
Los Angeles, California 90024

ABSTRACT

By applying the tools from queueing theory to a number of problems in computer communications and other multi-access systems, we have found a number of useful and intuitively pleasing rules of thumb which predict system behavior and which are easily calculated by deterministic reasoning. These rules of thumb allow one to calculate the "proper" operating point in these systems where the basic tradeoff is usually efficiency versus either delay (due to queueing) or loss (due to blocking) or some combination thereof. Using a previously defined notion of power as applied to delay systems, we extend that definition to combined loss and delay systems and then are able to define the optimal system operating point as that which maximizes this defined power. These results take advantage of the smoothing effect of the law of large numbers as applied to computer communication systems.

1. INTRODUCTION

Queueing theory is hard. Most interesting queueing models cannot be solved exactly and this leaves the systems analyst in a difficult position, especially in the initial design phase when he is simply trying to size the system and yet may lack the engineering intuition required for that analysis. Indeed, in the past we have often seen gross errors in prediction when queueing effects have been neglected; it is only in the last few years that analysts have come to appreciate the need for including the effects that queues can have in degrading system performance. It is the purpose of this paper to provide some engineering rules of thumb which may be used in such situations. Indeed we will show that deterministic reasoning (i.e., a "fluid" approximation [KLEI 76]) is usually a dangerous approach in real system performance evaluation unless we find that the system obeys the law of large numbers; in such a limit we may use deterministic reasoning (as opposed to stochastic or queueing calculations) which is quite accurate and which leads to the deterministic conclusion that systems can be driven close to 100% of their capacity and still perform well. This last statement is a mortal error in most queueing systems since typically we find that the average response time T varies with system utilization ρ according to the formula:

$$T = \frac{f}{1-\rho} \qquad (1.1)$$

In this equation we observe that as $\rho \to 1$, then the system deteriorates in that both the response time and the backlog grow to infinity. In general ρ represents a measure of the efficiency with which the system resources are utilized (and is often called the utilization factor); we see therefore that one is prohibited from approaching an efficiency of 100% due to Eq. (1.1). However f, a function which usually tends to vary slowly with ρ, may depend critically on other system parameters and may save the day in the case when $\rho \to 1$ as we shall see. Of course where there is no randomness due to queueing, then we know that proper scheduling would permit ρ to approach one without any backlog forming at all and it is such deterministic reasoning which we explore in this paper.

Now for some definitions. We will use terminology from computer communications although that is clearly not neccessary. We consider a computer communication system as a queueing system in which messages arrive, spend some time passing through the system (hopefully being routed and transmitted) and finally leave the system. Let:

T = average time spent in the system by a message (also known as the average system *response* time)

$\bar{x}$ = average system response time when there is no interfering traffic from other messages

W = $T-\bar{x}$ (W is the average wasted time or waiting time in the system)

λ = arrival rate of messages to the system

B = Blocking probability (i.e., the probability that an arriving message is rejected by the system at the input)

Since there is competition within the system for access to the system's resources (typically transmission resources) we find that a message spends on the average T seconds in the system rather than the minimum time which is $\bar{x}$ seconds; therefore we see that a message spends $T/\bar{x}$ times as long in the system as it should if the system's resources were available for that message's sole use. We will have occasion to use this normalized response time in our discussion below. Furthermore since the system may reject messages (a fraction B of the time) then we define γ to be the traffic (messages per second) actually carried by the network which must be equal to

$$\gamma = \lambda(1-B) \qquad (1.2)$$

Lastly we will be using the following well-known result (Little's result) in our development

$$\bar{N} = \gamma T \qquad (1.3)$$

where $\bar{N}$ is simply the average number of messages in the system.

Let us now introduce the notion of *efficiency* of a system resource. For purposes of this paper we simply define it to be the utilization factor of the servers and in the case of an m-server system carrying a total traffic γ we have that the utilization factor (hence the efficiency) is simply given by

*This research was supported by the Advanced Research Projects Agency of the Department of Defense under contract MDA 903-77-C-0272.

$$\rho = \gamma\bar{x}/m \qquad (1.4)$$

Here we are assuming that a single transmission is all the service required by a message. Clearly these notions can be extended to multi-hop systems but such extensions will not be considered in this paper. In the case of a multiple-server system we see that ρ is simply the average fraction of busy servers and therefore corresponds to the *efficiency* of the system resources.

The terms we have defined are clearly terms of interest to systems analysts: response time, throughput, efficiency, and loss. In the next section we discuss how these terms may be combined into a single measure of system performance. In the balance of the paper we discuss the behavior of some example systems and show the way in which deterministic rules of thumb may be used in evaluating system performance.

2. THE POWER OF A SYSTEM

In a system with no loss, there are two performance measures which compete with each other: *response time* and *throughput*. In [GIES 78], these two performance measures were combined into a single measure known as power, P', defined as follows

$$P' = \frac{\gamma}{T} \qquad (2.1)$$

With this measure we see that an increase in throughput or a decrease in response time increases the power. Throughout this paper we will use the symbol * to denote variables which are optimized with respect to power. In [KLEI 78] the author examined this function and found that for any system described by a delay-throughput function $T(\gamma)$, power will be maximized at that value of throughput where a ray out of the origin of the T, γ plane is tangent to the $T(\gamma)$ function. See Figure 2.1. For example in an M/M/1 (see [KLEI 75] for an explanation of the */*/* notation for queueing systems) queueing system, power is maximized at that optimal delay (denoted by T^*) and optimal throughput (γ^*) combination such that the delay is *twice* the minimum delay and the throughput is ½ the maximum throughput, that is, $T^*(\gamma^*)=2T(0)=2\bar{x}$ and $\gamma^*=\gamma_{\max}/2$ where $\gamma_{\max}$ is that throughput which drives the system into saturation, namely when $\rho_{\max}=\gamma_{\max}\bar{x}=1$.

We wish to introduce a more useful notion of power which includes the effect of blocking in a loss system and also normalizes the performance parameters in a suitable fashion. This *new* definition of power, P, is given as follows

$$P = \frac{\rho(1-B)}{T/\bar{x}} \qquad (2.2)$$

We note that

$$P = P'\frac{(\bar{x})^2}{m}(1-B) \qquad (2.3)$$

The introduction of the term $(\bar{x})^2/m$ is simply to convert the throughput γ to the efficiency ρ through the relationship $\rho=\gamma\bar{x}/m$ and to convert the response time T to its normalized version $T/\bar{x}$; these normalizations are convenient but not especially significant. Of more importance of course is the introduction of the factor $1-B$ which represents the fraction of applied traffic λ which is actually carried by the network; see Eq. (1.2). Thus our generalized definition of power is simply the fraction of traffic which is carried times the system efficiency divided by the normalized response time. Such a measure is intuitively appealing and behaves in the right direction with respect to all of our performance variables, namely, efficiency (or throughput), response time, and loss. More importantly, as we show below, it identifies the "proper" operating point for our system; this proper operating point will turn out to satisfy our intuitive expectation.

Let us rewrite Little's result as follows

$$\begin{aligned} \bar{N} &= \gamma T \\ &= (\gamma\bar{x}/m)mT/\bar{x} \\ &= \rho m(T/\bar{x}) \qquad (2.4) \end{aligned}$$

Now in the case of a system with no loss, ($B=0$), then we have

$$\bar{N} = \rho^2\frac{m}{P} \qquad (2.5)$$

Let us now observe for the system M/M/1 that we have, at optimal power, $\rho = 1/2$, $P = 1/4$ and therefore an average number of messages in the system equal to unity, that is:

$$\bar{N}^* = 1 \quad \text{for M/M/1} \qquad (2.6)$$

This is an interesting result and says that an M/M/1 system has maximum power when on the average there is only one message in the system; this is intuitively pleasing since it corresponds to our deterministic reasoning that the proper operating point for a single server system is exactly when only one customer is being served in the system and no others are waiting for service at the same time.

It turns out that this last result is general and holds also for the system M/G/1; that is, we have the following theorem:

Theorem 2.1

For any M/G/1 queueing system, power, as defined in Eq. (2.2), (with B=0), is maximized when

$$\bar{N}^* = 1 \qquad (2.7)$$

PROOF.

The proof follows simply when one uses the observation made in [KLEI 78] that power is maximized for any $T(\gamma)$ function when

$$\frac{dT(\gamma)}{d\gamma} = \frac{T(\gamma)}{\gamma} \qquad (2.8)$$

If we now substitute in the $P-K$ equation for M/G/1 [KLEI 75, Eq. (5.7)] we find that it will satisfy this last equation when $\bar{N}^*=1$.

This result is shown in Figure 2.1 in which we plot the normalized response time versus system utilization for M/G/1. (The ratio of standard deviation to mean service time is C_b, the

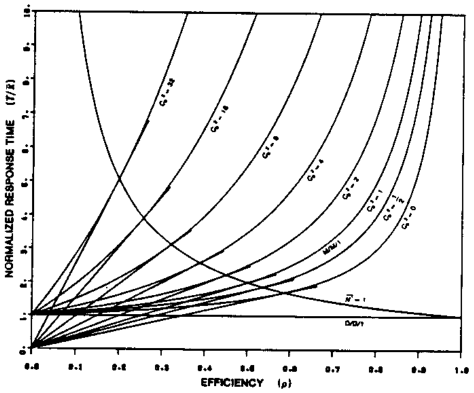

Figure 2.1 Optimal Power

coefficient of variation of the service time distribution.) We have also shown the loci for which the equation $\bar{N}=1$ is true and we find that it passes through the tangent point from a ray out of the origin to the response time function as predicted. Here again our intuition is satisfied in that it corresponds to our deterministic reasoning. However we now wish to extend this reasoning to the law of large numbers and also later to extend this notion of power.

3. PURE DELAY SYSTEMS

In this section we study the queueing system M/M/m as a representative queueing system with m servers. An infinite queue is permitted and so the loss probability is 0 (that is, $B = 0$). See Figure 3.1. In this system a free server (an idle channel) will transmit that message at the head of the common queue being served by these m channels. The behavior of the M/M/m system is an approximation to a large class of delay systems and we believe that the limiting behavior we describe below is characteristic of these other systems. By means of this pure delay system we will demonstrate two resource-sharing principles which come about when we deal with large numbers. The first of these is simply a *scaling* effect and we will discuss it only in this section although its effects can be observed in the other systems of this paper. The second, and more important effect which is the main subject of this paper, is the gain to be had when the law of large numbers takes effect; this characteristic is that which we will study later in this section and in the remaining sections of the paper.

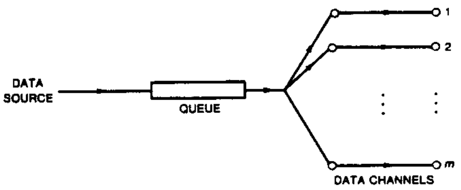

Figure 3.1 The Queueing System M/M/m

The effect of scaling was studied in [KLEI 77]. The principle has two useful forms as stated below. Consider an M/M/m system which represents, say, a data communication system with m communication channels. Let each channel have a speed of C bits per seconds, let each message have an average length $\bar{l}$ bits per message and let $T(m, \gamma, m\,C)$ represents the mean response time of the system containing m channels, each of capacity C, and supporting a total throughput of γ messages per second. We note that the average service time in the system is simply $\bar{x}=\bar{l}/C$ seconds and that the efficiency of each channel is given by $\rho=\gamma \bar{l}/m\,C$. For such a system the scaling principle, as given in [KLEI 77], may be stated in two useful forms as follows:

The Scaling Principle for Resource Sharing Systems

(i) A delay system whose throughput γ is scaled up by a factor h and whose capacity is also scaled up by the same factor h (while maintaining a constant number of data channels (m) and a constant average message length $\bar{l}$) has a response time which is h times less than the response time of the original system, that is

$$T(m, h\,\gamma, h\,m\,C) = \frac{1}{h} T(m, \gamma, m\,C) \qquad (3.1)$$

(ii) Alternatively, scaling up the capacity C more slowly than is the throughput scaled, results in a system which maintains a constant mean response time, and with an increasing efficiency; that is, one can achieve arbitrarily high efficiency as the throughput and capacity are each scaled up in a way which maintains constant performance.

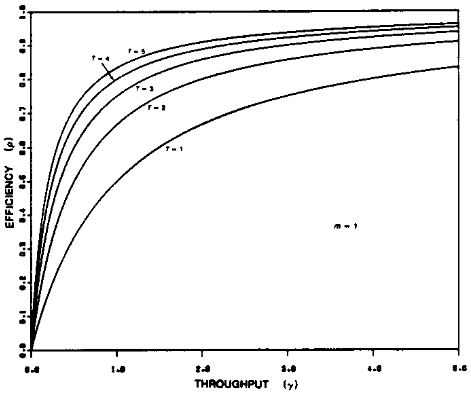

Figure 3.2 The Tradeoff for Efficiency, Throughput and Response Time $(m=1)$

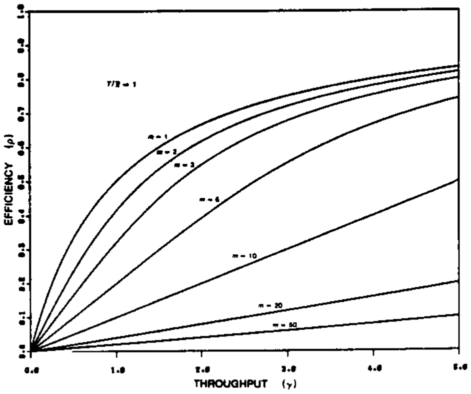

Figure 3.3 The Tradeoff for Efficiency, Throughput and Number of Data Channels $(T/\bar{x}=1)$

The second form of this scaling principle is demonstrated in Figure 3.2 as originally given in [KLEI 77]. This figure is for the case $m=1$; similar curves may be drawn for other values of m as shown in Figure 3.3. Both of these figures are shown for the system M/M/m but, as proven in [KLEI 77], similar results hold for the system G/G/m. The first form of the scaling principle is really quite remarkable and is not a generally known fact. Indeed it seems to defy one's intuition, so let us take a moment to discuss the phenomenon. Since $\bar{N}$, the average number in system, is a dimensionless quantity, then it reasonable to expect that it can only depend upon the dimensionless quantity ρ and not upon γ, $\bar{l}$ or C by themselves. Now since $\rho=\gamma\bar{l}/m\,C$ then if both γ and C are each scaled up by the same factor h, it is clear that ρ will not change and therefore one would expect $\bar{N}$ to remain constant; this is consistent with one's intuition. Now, however, applying Little's result we see that $T=\bar{N}/\gamma$ and therefore scaling both γ and C by the factor h will result in a mean response time T which is reduced by the same factor h. Q.E.D.

The scaling principle is really quite important in system design but it is distinct from our second principle which we now discuss and which depends upon the law of large numbers. Many of the beautiful results in the theory of systems subject to random phenomena are due to the smoothing effect of the law of large numbers; perhaps Shannon's noisy coding theorem is the most outstanding of these results [SHAN 49]. The principle is as follows:

The Smoothing Principle for Large Shared Systems

A large population presents a total demand which is equal to the sum of the average demands of each member of the population (as opposed to the sum of the peak demands of each).

This is simply the law of large numbers [KLEI 75] which states that the sum of n independent random variables, when divided by n, takes on a value which is predictable to any degree of precision as n gets large and this value is simply equal to the average value of each of the random variables (in its simplest form we assume that the random variables are identically distributed). This principle is the key to our deterministic rules of thumb for probabilistic problems in computer communications. Simply stated, if we have a large enough population of independent demands then we can neglect the random behavior of each member of that population and treat each as if it were perfectly deterministic. In such a case it is quite simple to determine system throughput, response time, blocking probability, efficiency, etc. For example, in the case of M/M/m the normalized response time is given by

$$\frac{T}{\bar{x}} = 1+\frac{P_m}{m(1-\rho)} \tag{3.2}$$

where

$$P_m = \frac{(m\rho)^m/[(1-\rho)m!]}{\frac{(m\rho)^m}{(1-\rho)m!}+\sum_{k=0}^{m-1}\frac{(\gamma\bar{x})^k}{k!}} \tag{3.3}$$

If we now plot this normalized response time as a function of system efficiency we obtain Figure 3.4. In this figure we see that as the number of channels increases, the normalized response time decreases and in the limit we observe that the behavior is exactly the same as D/D/1. The system D/D/1 is a system with no random effects and is capable of achieving an efficiency approaching unity with no increase in delay. This is simply the case in which we perfectly schedule arrivals such that the previous arrival departs from a channel just prior to the next arrival. One might be surprised that this author is suggesting that a multiple channel system is superior to a single channel system due to his comments in [KLEI 77] in which he stated that a single channel was superior to multiple channels. These two seemingly contradictory statements are nevertheless consistent since in the current case we are increasing the total system capacity and are holding $\bar{x}$ constant whereas in the discussion of [KLEI 77], we were maintaining the total system capacity fixed and dividing it equally among the m channels, thereby increasing $\bar{x}$ by the factor m.

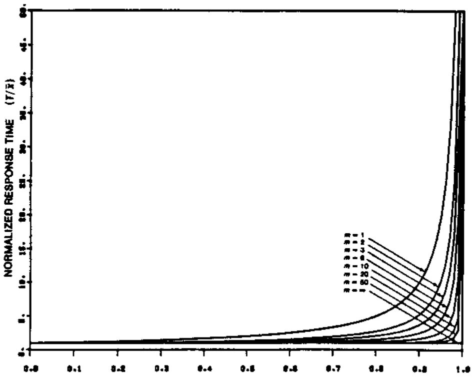

Figure 3.4 Response Curves for M/M/m

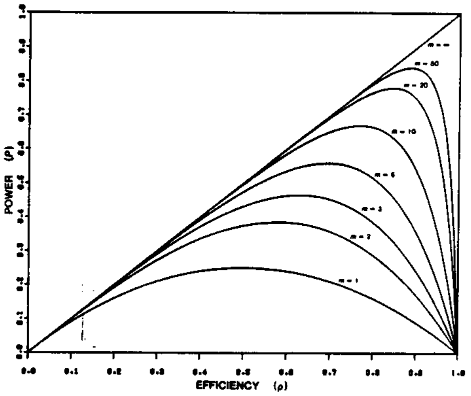

Figure 3.5 Power for M/M/m

Let us now apply the notion of *power* introduced in Section 2 to this pure delay system. In this case of pure delay we remind the reader that $B=0$ and so from Eq. (2.2) we have simply

$$P = \frac{\rho}{(T/\bar{x})} \tag{3.4}$$

This function is plotted versus efficiency in Figure 3.5. Note for $m=1$, that the maximum power occurs at $\rho=1/2$ as observed in Section 2 for the system M/M/1. Of more interest is the behavior of the power function for larger m, in particular we observe that as

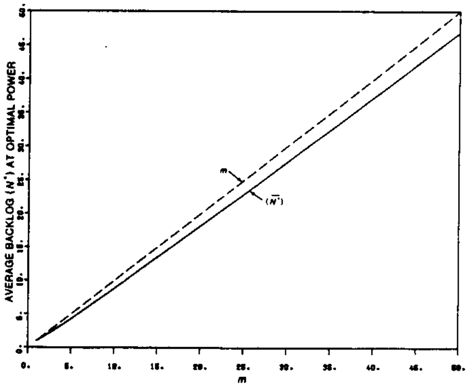

Figure 3.6 Average Backlog at Optimal Power

$m \to \infty$, we have that the power is directly proportional to the throughput so long as $0 \leqslant \rho < 1$. If we substitute Eq. (3.2) into Eq. (3.4) we easily see that at maximum power, $\overline{N}^* = 1$ for $m = 1$; similarly we observe that $\overline{N}^* \approx m$ for $m \to \infty$. One wonders if, in general, $\overline{N}^* = m$ for all m. Unfortunately this is not true; for example for $m = 2$ we find that the optimal power is achieved when $\overline{N}^* = \sqrt{3} = 1.732...$ However, as can be seen from Figure 3.6, in which we have plotted the value of $\overline{N}^*$ which is achieved at maximum power as a function of the number of channels, m, we find that $\overline{N}^* \approx m$ is quite a good approximation (the dotted line represents m itself). The intuition here is similar to that expressed at the end of Section 2, namely, one should have a number of messages in the system such that each channel has on the average one message in transmission and no other messages waiting; in the deterministic case this is an exact statement and here we see that in the random case M/M/m, that as long as m gets large we find this deterministic rule of thumb is a good approximation. In the next section we will see a dramatic manifestation of the smoothing principle just described.

4. PURE LOSS SYSTEMS

In this section we study the pure loss system M/G/m/m which consists of m channels with no storage space for queued messages. Any message which arrives when all channels are busy will be rejected by this system; the probability that a message is rejected is, as defined above, B. The system structure is shown in Figure 4.1. Here we see the m data channels where again we assume that the average time required to transmit a message is given by $\bar{x}$ seconds. Furthermore, we have decomposed the Poisson input into M independent Poisson sources each of which generates traffic at a rate α messages per second; thus the total input rate is simply

$$\lambda = M\alpha \qquad (4.1)$$

Of course the total traffic from the M sources is equivalent to a single source generating traffic at a rate λ. The product $\alpha\bar{x} = a$ is often referred to in the field of telephony as the number of Erlangs of traffic offered by each of the sources. Further, let us define the total applied load as

$$A = \lambda\bar{x} = \mathrm{M}\alpha\bar{x} = \mathrm{M}a$$

Now let us suppose that $\alpha = 1/2$, $\bar{x} = 1$ and $m = 1000$. One wonders how many sources, M, this system of 1,000 channels can support? Indeed what is the proper number M for "good" system operation. This is clearly an undefined question since we have not defined the notion of "good". However the tradeoff is clear, namely, if we wish to have a low loss system then M should be small (but this results in the inefficient use of the data channels); on the other hand if we wish to have highly utilized channels, then M should be large (however in this case the loss probability will be unacceptably large). Since $a = 1/2$ we see that each source really requires the use of a data channel half the time so one might expect that the "proper" number of sources which the system can support is equal to twice the number of channels, namely 2,000. Clearly this is deterministic reasoning since if each source required the use of the channel *exactly* half the time and if it required this use in a deterministic way, then we could exactly schedule two sources on a single channel in a way which would produce no interference between these sources, thereby supporting exactly 2,000 sources. If we attempted to drive the system harder than that, then beyond 2,000 sources there would be a clear interference whereas prior to that point there would be no blocking at all; indeed at exactly M=2,000, we have a system with zero blocking and 100% utilization of the channels! Below we show that one can achieve this best of all possible situations even with non-deterministic inputs as long as the number of channels is large enough so that the smoothing principle comes into effect (remember that this principle guarantees that a large collection of random sources will present a total load which appears deterministic in nature).

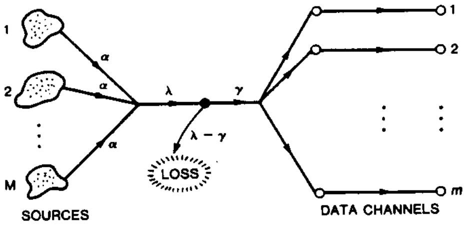

Figure 4.1 The Pure Loss System M/G/m/m

The performance variables we wish to examine in this case of pure loss are as follows: ρ, the efficiency of each channel; B, the loss probability for an arriving message; and P, the power of the system as defined in Eq. (2.2). Since this is a system with pure loss, then any message which is accepted by the system will spend on the average an amount of time in the system equal to the average transmission time of a message, that is $T/\bar{x} = 1$. Therefore the appropriate definition of power in the pure loss case is simply

$$P = \rho(1-B) \qquad (4.2)$$

Furthermore from Eq. (1.2) we have

$$\rho = \lambda(1-B)\bar{x}/m = A(1-B)/m \qquad (4.3).$$

Thus we see that the three performance functions ρ, B and P are all determined by the by the blocking probability B. For the system M/G/m/m it is well-known [GROS 74] that:

$$B = \frac{A^m/m!}{\sum_{k=0}^{m} A^k/k!} \qquad (4.4)$$

This is Erlang's famous blocking formula. Unfortunately, for finite values of m, the sum in the denominator cannot be expressed in a simple form and so one finds this function tabulated in most telephony handbooks. Of course for $m = 1$ we have the simple expression

$$B = \frac{A}{1+A} \qquad (m=1) \qquad (4.5)$$

Whereas B remains as complex as Eq. (4.4) for finite values of m, we find for very large values of m that a very simple behavior maintains which is related to the behavior at $m = 1$. Indeed the critical variable to consider is M/m which is simply the ratio of *sources to channels.* M/m will be used to describe the input to our system in the remainder of this paper. The limiting behavior here is given in the following theorem (see the Appendix for a proof of the theorem).

Theorem 4.1

In the limit as the number of data channels approaches infinity, we have

$$\lim_{m\to\infty} B = \begin{cases} 0 & 0 \leqslant \frac{\mathrm{M}}{m} < \frac{1}{a} \\ 1 - \frac{1}{a\mathrm{M}/m} & \frac{1}{a} \leqslant \frac{\mathrm{M}}{m} \end{cases} \qquad (4.6)$$

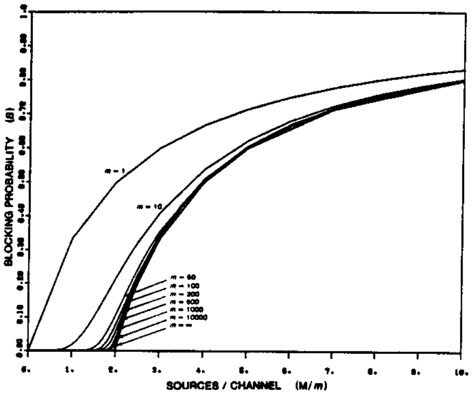

Figure 4.2 The Blocking Probability

B is plotted versus the critical parameter M/m, the number of sources per channel, for various values of m in Figure 4.2. The important observation to make is that the blocking probability rather quickly drops to zero as long as $a(\mathrm{M}/m)<1$ and beyond this point appears to behave like an $m=1$ system as far as the blocking probability is concerned. We may rewrite this condition simply as $A<m$. What this tells us is that for a large number of data channels, this system is behaving *deterministically* in that it has perfectly scheduled the large number of sources generating messages (remember the smoothing principle is now in effect) such that no mutual interference occurs; this is true as long as the number of channels can support the number of sources *on the average*. However when the number of sources exceeds this limit, then the system begins to block as if all additional traffic were being fed into a single channel system.

From Theorem 4.1 follows two obvious corollaries:

Corollary 4.1

In the limit as the number of channels approaches infinity, the efficiency of each channel is given by

$$\lim_{m\to\infty} \rho = \begin{cases} a\dfrac{\mathrm{M}}{m} & 0 \leqslant \dfrac{\mathrm{M}}{m} < \dfrac{1}{a} \\ 1 & \dfrac{1}{a} \leqslant \dfrac{\mathrm{M}}{m} \end{cases} \tag{4.7}$$

The behavior here is shown in Figure 4.3 in which we plot the channel efficiency versus M/m. We see that in the limit of a large number of channels, the channel efficiency grows linearly at a slope a with the ratio of sources to channels until the efficiency reaches 100% at which point it remains at this value as the load increases.

Corollary 4.2

As the number of channels approaches infinity the power of the system behaves as follows

$$\lim_{m\to\infty} P = \begin{cases} a\dfrac{\mathrm{M}}{m} & 0 \leqslant \dfrac{\mathrm{M}}{m} < \dfrac{1}{a} \\ \dfrac{1}{a\mathrm{M}/m} & \dfrac{1}{a} \leqslant \dfrac{\mathrm{M}}{m} \end{cases} \tag{4.8}$$

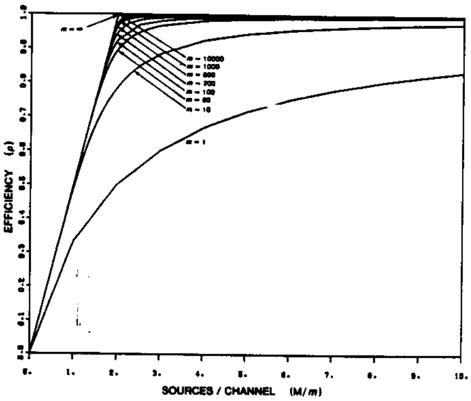

Figure 4.3 Channel Efficiencv

Here we see the beauty of our definition of power. As shown in Figure 4.4, it has peaked at exactly the right point, namely, when $a(\mathrm{M}/m)=1$. Below this point the channels are underutilized. Above this point blocking begins to set in. For finite values of m, we see that these statements are approximately true and that the maximum power point occurs at smaller values of the source to channel ratio; however, at $m=1$ we see that the power once again peaks at this same critical point namely when $a(\mathrm{M}/m)=1$. The proofs for these two corollaries easily follow from Theorem 4.1 and are given in the Appendix.

In this section we have seen that in the limit of a large number of data channels, the behavior is easily predicted by deterministic reasoning due to the smoothing principle stated in the previous section.

5. COMBINED LOSS AND DELAY SYSTEMS

In this section we study the system depicted in Figure 5.1 which combines both losş and delay. This is the system M/M/m/K which consists of m channels and space for $K-m$ queued messages;

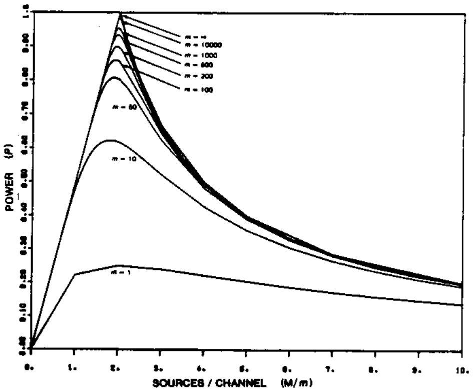

Figure 4.4 Power

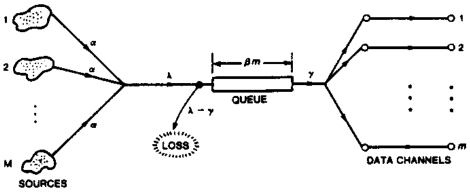

Figure 5.1 The Loss Plus Delay System M/M/m/K

for convenience of scaling we select $K=(1+\beta)m$ (in the examples below we will choose $\beta=1$). Thus the system can hold at most K customers of which at most m will be in the process of transmission. If a message arrives when all storage spaces are full, then that message will be rejected by the system; again the probability that a message is rejected is B. As in the previous sections we have M sources each generating traffic independently from a Poisson process at a rate α yielding a total applied traffic λ as given in Eq. (4.1). Recall the critical parameter $a=\alpha\bar{x}$ where $\bar{x}$ = average service time. Below we show that the proper operating point for this system, when m is large, is such that the ratio of the number of sources to the number of channels, namely M/m , is selected so that

$$\frac{\mathrm{M}}{m}=\frac{1}{a} \qquad (5.1)$$

This is the smoothing principle again and is the result one would obtain if the entire load were deterministic.

Using the techniques from elementary queueing theory [KLEI 75] we readily obtain the following expression for the loss probability

$$B=\frac{A^K/(m!m^{K-m})}{\dfrac{A^m[1-(a\mathrm{M}/m)^{K-m+1}]}{m!(1-a\mathrm{M}/m)}+\sum_{k=0}^{m-1}\dfrac{A^k}{k!}} \qquad (5.2)$$

As with Eq. (4.4) the sum in the denominator cannot be expressed in a simple form. For $m=1$ we have the simple expression

$$B=A^K\frac{1-A}{1-A^{K+1}} \quad (m=1) \qquad (5.3)$$

As in the previous section we find that the complex expression for the loss probability given in Eq. (5.2) takes on a rather simple form for very large values of m. Again the appropriate variable to consider is M/m, the ratio of sources to channels, and the limiting behavior is given in the following theorem

Theorem 5.1

In the limit as the number of data channels approaches infinity, we have

$$\lim_{m\to\infty} B=\begin{cases} 0 & 0\leqslant\dfrac{\mathrm{M}}{m}<\dfrac{1}{a} \\ 1-\dfrac{1}{a\mathrm{M}/m} & \dfrac{1}{a}\leqslant\dfrac{\mathrm{M}}{m}\end{cases} \qquad (5.4)$$

We note that Theorem 5.1 is the same as Theorem 4.1 showing that the provision of a finite storage capacity does not affect the limiting behavior of the blocking probability. The proof of this theorem is similar to that of Theorem 4.1 and is not given here. B is plotted versus the critical parameter M/m for various values of m in Figure 5.2. The behavior is similar to (and better than) the behavior of the blocking probability in the pure loss system of the previous section.

From Theorem 5.1 we have the two following corollaries

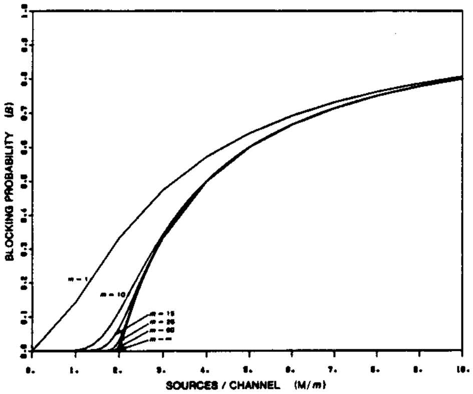

Figure 5.2 The Blocking Probability

Corollary 5.1

In the limit as the number of channels approaches infinity, the efficiency of each channel is given by

$$\lim_{m\to\infty}\rho=\begin{cases} a\dfrac{\mathrm{M}}{m} & 0\leqslant\dfrac{M}{m}<\dfrac{1}{a} \\ 1 & \dfrac{1}{a}\leqslant\dfrac{\mathrm{M}}{m}\end{cases} \qquad (5.5)$$

The behavior of the efficiency is given in Figure 5.3.

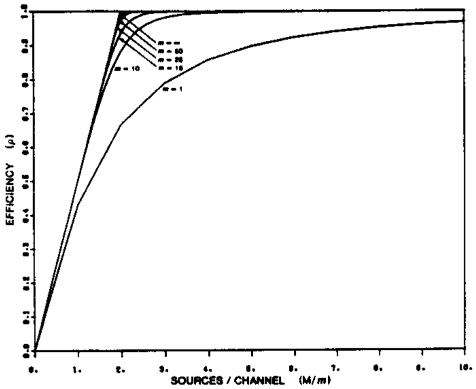

Figure 5.3 Channel Efficiency

Corollary 5.2

As the number of channels approaches infinity, the power of the system behaves as follows

$$\lim_{m\to\infty}P=\begin{cases} a\dfrac{\mathrm{M}}{m} & 0\leqslant\dfrac{\mathrm{M}}{m}<\dfrac{1}{a} \\ \dfrac{1}{(1+\beta)a\mathrm{M}/m} & \dfrac{1}{a}\leqslant\dfrac{\mathrm{M}}{m}\end{cases} \qquad (5.6)$$

Again we see the beauty of our definition of power. As seen in Figure 5.4, it peaks at exactly the right point, namely, when

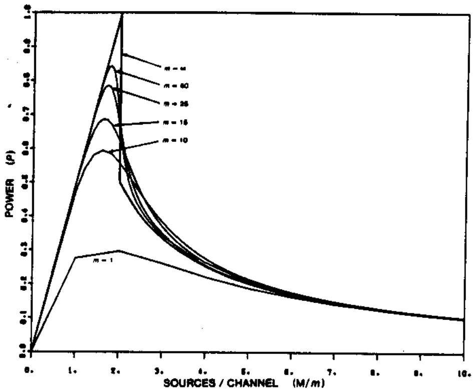

Figure 5.4 Power

$a\mathrm{M}/m = 1$. Below this point the channels are underutilized and above this point blocking sets in. The proofs for these two corollaries easily follow from Theorem 5.1 and from the definition of power. We note that the normalized delay $T/\bar{x}$ is, in the limit,

$$\lim_{m\to\infty} T/\bar{x} = \begin{cases} 1 & 0 \leqslant \dfrac{\mathrm{M}}{m} < \dfrac{1}{a} \\ 1+\beta & \dfrac{1}{a} \leqslant \dfrac{\mathrm{M}}{m} \end{cases} \tag{5.7}$$

With no queues the mean response time is simply equal to an average service time whereas with an essentially full queue, the time in system is equal to the time to empty the queue ($\beta\bar{x}$) seconds plus one average service time. The behavior of this normalized delay as a function of M/m is given in Figure 5.5; here we see the dramatic transition in response time as described in Eq. (5.7); this is the familiar "zero-one" behavior so often seen when the law of large numbers comes into effect. The mean number in system $\bar{N}$ may also be normalized with respect to the number of channels, that is $\bar{N}/m$, and this is plotted in Figure 5.6. The limiting behavior is given by

$$\lim_{m\to\infty} \frac{\bar{N}}{m} = \begin{cases} a\dfrac{\mathrm{M}}{m} & 0 \leqslant \dfrac{\mathrm{M}}{m} < \dfrac{1}{a} \\ 1+\beta & \dfrac{1}{a} \leqslant \dfrac{\mathrm{M}}{m} \end{cases} \tag{5.8}$$

This last follows directly from Little's result as given in Eq. (2.4), that is,

$$\frac{\bar{N}}{m} = \rho(T/\bar{x}) \tag{5.9}$$

and so Eq. (5.8) follows from Eq. (5.5) and Eq. (5.7).

Thus we see as in Section 4, that the behavior in the limit of a large number of data channels is predictable by simple deterministic reasoning.

6. GENERALIZED POWER

In the discussion of Section 2, we showed some interesting properties of power (as defined in Section 2.2). In Sections 3,4 and 5, we demonstrated the importance of this definition of power in the limiting case as the number of data channels increased to infinity. However, the critical reader might well complain that the definition as given in Eq. (2.2) forces one to accept the relative importance of efficiency, blocking, and response time to be that

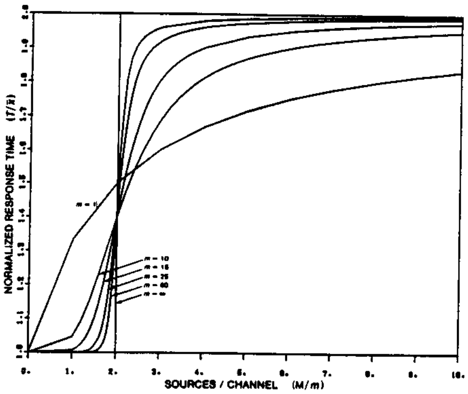

Figure 5.5 Response Time

given in that simple equation. In this section we offer a more generalized definition of power (applied to pure delay systems for purposes of this paper) which gives the reader the opportunity to emphasize the relative importance of throughput versus response time in any fashion he deems appropriate.

Our generalization is to introduce a nonnegative real variable r and to redefine power as

$$P = \frac{\rho^r}{T/\bar{x}} \tag{6.1}$$

(the obvious generalization which includes the blocking probability as well will be the subject of a forthcoming paper). By the introduction of this new variable, we see that the system efficiency can be favored more heavily over system response time simply by increasing the parameter r. Indeed, this parameter permits one to redefine the location of the knee in the curve which describes the response time as a function of throughput (or efficiency).

It is not difficult to show the remarkable result as given in the following theorem

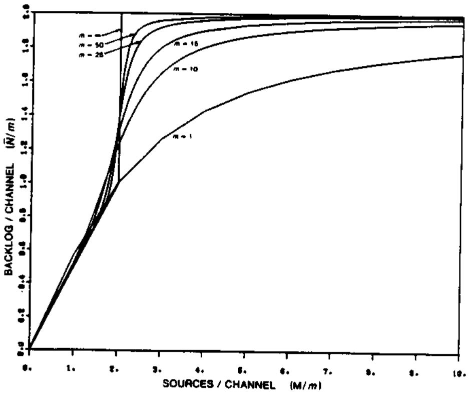

Figure 5.6 Backlog

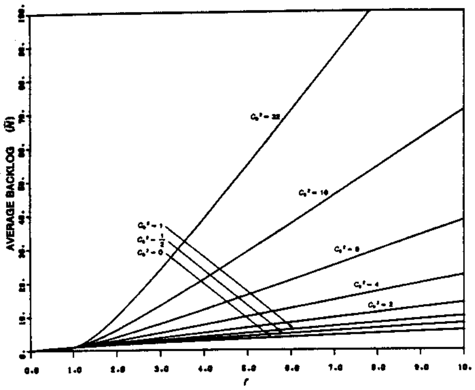

Figure 6.1 Behavior of Optimal Average Backlog for Various r

Theorem 6.1

For the system M/M/1 we find that power is maximized when

$$\overline{N^*} = r \tag{6.2}$$

This is a most pleasing result due to its extreme simplicity. In the case of M/G/1, we are not quite so fortunate and we find that at maximum power the average number in system is a more complicated function both of the parameter r and of the coefficient of variation, C_b, of the service time distribution (recall that C_b is simply the ratio of the standard deviation to the mean service time). Indeed we have

Theorem 6.2

For the system M/G/1 we find that at maximum power, the average number in system, is given by

$$\overline{N^*} = \frac{2(r-1)^2x^2+4x(r^2-r+2)+8(r+1)+2R\,[x(r-1)-2]}{4r\,[(x+2)(r+1)-R\,]} \tag{6.3}$$

where

$$R = \sqrt{(r-1)^2x^2+4x(r^2+1)+4(r+1)^2} \tag{6.4}$$

and

$$x = C_b{}^2-1 \tag{6.5}$$

Note, for $r=1$, that $\overline{N^*}=1$ which simply is Theorem 2.1 again; note also that for $x=0$ (which corresponds to the case $C_b{}^2=1$ which implies that we are dealing with the system M/M/1) that $\overline{N^*}=r$ which is simply Theorem 6.1.

In Figure 6.1 we plot $\overline{N^*}$ as a function of r and in Figure 6.2 we give the same plot on expanded scales to show the behavior below $r=1$. It is easy to show the following

Theorem 6.3

For large r we have the following limiting behavior

$$\lim_{r\to\infty}\frac{\overline{N^*}}{r} = \frac{1+C_b{}^2}{2} \tag{6.6}$$

This behavior is easily seen in Figure 6.1.

We may also describe ρ^* which is the efficiency obtained at maximum power. This is given in

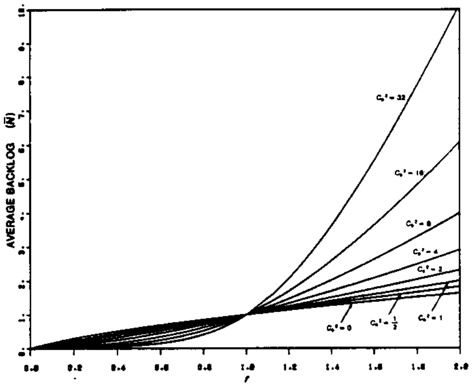

Figure 6.2 Behavior of Optimal Average Backlog for Various r
(expanded scale)

Theorem 6.4

For optimal power we have that the efficiency is given by

$$\rho^* = \frac{(r-1)x-2(r+1)+R}{2rx} \tag{6.7}$$

and for large r we find

$$\rho^* \to 1-\frac{1}{r} \tag{6.8}$$

Let us now discuss the power for $m\to\infty$. From Theorem 5.1 and Corollary 5.1, we see that B remains at zero for $0 \leq (M/m) < (1/a)$ and that ρ rises continuously in the same range. Further, for $\beta<\infty$, we see that the normalized response time, $T/\bar{x}$, remains constant in this range. For $(M/m) > (1/a)$, the blocking increases, the efficiency no longer climbs and the delay takes a stepwise increase at this boundary. Therefore *any* definition of power, say

$$P = f(\rho)g(B)h(T)$$

such that f is increasing and both g and h are decreasing functions of their arguments, *must* (in the limit as $m\to\infty$) peak at $(M/m)=(1/a)$. The expression given in Eq. (6.1) is simply one such example.

7. CONCLUSION

In this paper we have tried to establish the validity of using deterministic reasoning to evaluate the performance of computer - communication systems. We have shown that the smoothing effect of the law of large numbers does indeed permit such reasoning in the case of many shared resources. The general result is that one should operate a system at that load which just saturates the system resources; in this calculation, one may assume that the load is deterministic and perfectly scheduled.

By defining power in terms of efficiency, blocking and response time, we were able to show that, in the limit, power is maximized at this saturated load. A generalized definition of power was also given and this too peaked at the saturation load for a large class of power functions.

ACKNOWLEDGEMENTS

The author takes pleasure in acknowledging the assistance of Frank S. Heath in the numerical evaluation and in the creation of the many computer-generated curves shown herein, as well as Lou Nelson and Leon Lemons in transcribing the draft into the lovely phototypesetter output shown here.

REFERENCES

[GIES 78] Giessler, A., J. Hänle, A. König and E. Pade, "Free Buffer Allocation - An Investigation by Simulation," *Computer Networks*, Vol. 1, No. 3, July 1978, pp. 191-204.

[GROS 74] Gross, D. and C. M. Harris, *Fundamentals of Queueing Theory*, John Wiley and Sons, New York, 1974 - See Sections 3.4 and 5.2.2.

[KLEI 75] Kleinrock, L., *Queueing Systems, Vol. I: Theory*, Wiley-Interscience, New York, 1975.

[KLEI 76] Kleinrock, L., *Queueing Systems, Vol. II: Computer Applications*, Wiley-Interscience, New York, 1976.

[KLEI 77] Kleinrock, L., "Performance of Distributed Multi-Access Computer-Communication Systems," *Proc. of IFIP Congress 77*, August 1977, pp. 547-552.

[KLEI 78] Kleinrock, L., "On Flow Control in Computer Networks," *Proc. of the International Conference on Communications*, Vol. 2, June 1978, pp. 27.2.1-27.2.5.

[SHAN 49] Shannon, C. E. and W. Weaver, *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana, Ill., 1949.

APPENDIX:

Proof of Theorem 4.1

We study the behavior of the loss probability B at fixed values of M/m as $m \rightarrow \infty$; thus, we let $\mathrm{M} = pm/a$ for fixed values of p. We also define

$$f_k = \frac{A^k}{k!}$$

Let us consider the two regions $(0 \leqslant \frac{\mathrm{M}}{m} < \frac{1}{a}$ and $\frac{1}{a} \leqslant \frac{\mathrm{M}}{m})$ separately:

Case 1: $0 \leqslant \frac{\mathrm{M}}{m} < \frac{1}{a} \quad (0 \leqslant p < 1)$

Let us assume (at no loss of generality in the limit as $A \rightarrow \infty$) that pm is an integer; then $pm = \mathrm{M}a = A$ is an integer Now consider the ratio

$$R = \frac{f_A}{f_{yA}} = \frac{A^A/A!}{A^{yA}/(yA)!}$$

where yA is a non-negative integer and $y \neq 1$. Using Stirling's approximation, we have

$$R \approx e^{A(1-y)} y^{yA} \sqrt{y} = e^{A[1-y+y\log y]} \sqrt{y}$$

But, as is well-known, for $y \neq 1$, $y > 0$,

$$\log y > 1 - \frac{1}{y}$$

(equality holds for $y=1$). Thus

$$1-y+y\log y > 0$$

and so

$$\lim_{A \rightarrow \infty} R = \lim_{A \rightarrow \infty} e^{A[1-y+y\log y]} \sqrt{y} = \infty$$

This result simply shows that the term f_A dominates all other terms f_k for $k \neq A$.

Now since $p = \mathrm{M}a/m = A/m < 1$ (i.e., $m > A$), we have

$$0 \leqslant B = \frac{f_m}{\sum\limits_{k=0}^{m} f_k} \leqslant \frac{f_m}{f_A} = \frac{1}{R}$$

with $y = m/A > 1$. Clearly, since $A = mp$ for fixed p,

$$0 \leqslant \lim_{m \rightarrow \infty} B \leqslant \lim_{A \rightarrow \infty} \frac{1}{R} = 0 \qquad \text{Q.E.D. (Case 1)}$$

Case 2: $\frac{1}{a} \leqslant \frac{M}{m} \quad (p \geqslant 1)$

From Eq. (4.4) we see that

$$1-B = \frac{\sum\limits_{k=0}^{m-1} f_k}{\sum\limits_{k=0}^{m} f_k} = \frac{\sum\limits_{k=0}^{m-1} f_k/f_m}{\sum\limits_{k=0}^{m} f_k/f_m} = \frac{\sum\limits_{k=0}^{m-1} \frac{m!}{k!} A^{k-m}}{\sum\limits_{k=0}^{m} \frac{m!}{k!} A^{k-m}} = \frac{\sum\limits_{k=1}^{m} \frac{m!}{(m-k)!} A^{-k}}{\sum\limits_{k=0}^{m} \frac{m!}{(m-k)!} A^{-k}}$$

But, by Stirling's approximation,

$$\frac{m!}{(m-k)!} \approx \left(\frac{m}{e}\right)^k \left(1-\frac{k}{m}\right)^{-m+k-\frac{1}{2}}$$

For k fixed, we then have that for m large

$$\frac{m!}{(m-k)!} \approx m^k$$

and so, for m large and $p \geqslant 1$ we have (using $p = A/m$)

$$1-B \approx \frac{\sum\limits_{k=1}^{m} p^{-k}}{\sum\limits_{k=0}^{m} p^{-k}} = \frac{p^{-1} - p^{-(m+1)}}{1 - p^{-(m+1)}}$$

Thus $\lim\limits_{m \rightarrow \infty} 1-B = \frac{1}{p} = \frac{m}{A}$ Q.E.D. (Case 2)

Proof of Corollary 4.1

From Eq. (4.3) we have

$$\rho = \lambda\,(1-B)\overline{x}/m = \frac{A}{m}(1-B) = p(1-B)$$

Thus, from Theorem 4.1,

$$\lim_{m \rightarrow \infty} \rho = \begin{cases} p = \frac{\mathrm{M}}{m} a & 0 \leqslant \frac{M}{m} < \frac{1}{a} \\ \frac{p}{p} = 1 & \frac{1}{a} \leqslant \frac{M}{m} \end{cases}$$

Proof of Corollary 4.2

From Eq. (4.2) we have

$$P = \rho(1-B) = p(1-B)^2$$

Thus, from Theorem 4.1,

$$\lim_{m \rightarrow \infty} P = \begin{cases} p = a\frac{\mathrm{M}}{m} & 0 \leqslant \frac{M}{m} < \frac{1}{a} \\ \frac{p}{p^2} = \frac{1}{a\mathrm{M}/m} & \frac{1}{a} \leqslant \frac{M}{m} \end{cases}$$