Simon S. Lam, *Packet Switching in a Multi-Access Broadcast Channel with Application to Satellite Communication in a Computer Network*, Ph.D. Dissertation, Computer Science Department, University of California at Los Angeles, March 1974; published as Technical Report UCLA-ENG-7429 (ARPA), School of Engineering and Applied Science, UCLA, April 1974, 306 pages.

Available in eleven .pdf files:

- TRcovers.pdf
- Abstract+ToC.pdf
- Chapter1.pdf
- Chapter2.pdf
- Chapter3.pdf
- Chapter4.pdf
- Chapter5.pdf
- Chapter6.pdf
- Chapters7-8.pdf
- Bibliography.pdf
- Appendices.pdf

CHAPTER 5

CHANNEL STABILITY

The slotted ALOHA random access method enables a multi-access channel to be statistically multiplexed in an efficient way by a large number of users. Such a system was studied in Chapter 3 as an infinite population model; equilibrium results on the channel throughput-delay performance were given. However, simulations have shown that the assumption of channel equilibrium may not always be valid. In fact, the channel, after some finite time period of quasi-stationary conditions, will drift into saturation with probability one. Thus, we realize that the equilibrium throughput-delay results are not sufficient to characterize the performance of the infinite population model. A more representative measure of channel performance in this case is the stability-throughput-delay tradeoff. To do so, we must first define channel stability and a stability measure.

We consider in this chapter a slotted ALOHA channel supporting a total of M users. The variable M is assumed to be large, but finite. We show below that the exact value of M is an important stability parameter. The purpose of this chapter is to characterize the instability phenomenon in the following ways:

- We define stable and unstable channels

- We show that in a stable channel, equilibrium throughput-delay results presented in Chapter 3 are achievable over an infinite time horizon. In an unstable channel, such channel performance is achievable only for some finite time period before the channel goes into saturation
- For an unstable channel, we define a stability measure and give an efficient computational procedure for its calculation
- Using the above stability measure, we examine the stability-throughput-delay tradeoff for an unstable channel

5.1 The Model

In the last chapter, we realized that the source of our difficulty in analysis lies in the complexity of the state description. Below we first define a mathematical model which characterizes the channel state by a single variable. Practical considerations and the model approximations to a physical system will then be examined. This mathematical model will also be adopted in the next chapter.

Our model is similar to the linear feedback model studied by Metcalfe who gave a steady-state analysis of the system behavior [METC 73A]. Lu [LU 73] studied the same model and characterized the time-dependent channel behavior through a set of linear difference equations. However, his approach (like our results in Section 4.1) cannot be easily applied to a system with many states (i.e., channel users).

5.1.1 The Mathematical Model

We consider a slotted ALOHA channel with a user population consisting of M small users (see Section 2.3). Each such user can be in one of two states: blocked or thinking. In the thinking state, a small user generates and transmits a new packet in a time slot with probability σ. A packet which had a channel collision and is waiting for retransmission is said to be backlogged. A small user with a backlogged packet is blocked in the sense that he cannot generate (or accept from his input source) a new packet for transmission. The retransmission delay RD of each backlogged packet is assumed to be geometrically distributed, i.e., each backlogged packet retransmits* in the current time slot with probability p.

Let N^t be a random variable (called channel backlog) representing the total number of backlogged packets at time t. Given that $N^t = n$, the channel input rate at time t is $S^t = (M - n)\sigma$. Note that S^t decreases linearly as n increases. Thus, this will also be referred to as the linear feedback model. The vector (N^t, S^t) will be denoted as the channel state vector. In this context, both M and σ may be functions of time. We shall assume M and σ to be time-invariant unless stated otherwise. In this case, N^t is a Markov process (chain) with stationary transition probabilities and serves as the state description for the system. The state space will now consist of the set of integers $\{0, 1, 2, \ldots. M\}$. The one-step state transition probabilities of N^t are, for $i = 0, 1, 2, \ldots.$

*Assuming bursty users, we must have $p >> \sigma$.

$$p_{ij} = \text{Prob}[N^{t+1} = j \mid N^t = i]$$

$$= \begin{cases} 0 & j \le i - 2 \\ ip(1-p)^{i-1}(1-\sigma)^{M-i} & j = i - 1 \\ (1-p)^i (M-i)\sigma(1-\sigma)^{M-i-1} & \\ \quad + [1 - ip(1-p)^{i-1}](1-\sigma)^{M-i} & j = i \\ (M-i)\sigma(1-\sigma)^{M-i-1}[1-(1-p)^i] & j = i+1 \\ \binom{M-i}{j-i}\sigma^{j-i}(1-\sigma)^{M-j} & j \ge i+2 \end{cases} \tag{5.1}$$

For the infinite population model in which $M \to \infty$ and $\sigma \to 0$ such that $M\sigma = S$ which is constant and finite, the above equation becomes

$$p_{ij} = \begin{cases} 0 & j \le i - 2 \\ ip(1-p)^{i-1} e^{-S} & j = i - 1 \\ (1-p)^i Se^{-S} + [1 - ip(1-p)^{i-1}]e^{-S} & j = i \\ Se^{-S}[1-(1-p)^i] & j = i+1 \\ \frac{S^{j-i}}{(j-i)!} e^{-S} & j \ge i+2 \end{cases} \tag{5.2}$$

5.1.2 Practical Considerations

The above mathematical model approximates a physical system in several ways. First, M and σ will be assumed to be time-

invariant. However, if we distinguish active and inactive channel users such that only active users will generate packets for transmission over the channel (with probability σ), the variable M in our model actually corresponds to the number of active users. In a real system, M will most probably vary during the day with alternate periods of heavy and light "load." Since such time periods are usually extremely large with respect to our time scale (a packet transmission time), M can be regarded as time-invariant during each period. A good rule of thumb in the system design is to optimize the channel performance under the assumption of a heavy load since the performance of a lightly loaded channel is relatively insensitive to the system design. This will be our philosophy in this chapter and the next. Most of our numerical examples are based upon the assumption of a heavily loaded channel. If we consider the average user think time to be 1-30 seconds in an interactive computer communications environment [JACK 69]. Our range of interest to be assumed for the number* of active channel users is between $M = 10$ to $M = 500$.

The mathematical model assumption that RD is geometrically distributed permits the use of a single variable for the state

---

*The user think time as defined in our model represents quantities such as the real thinking and typing time of an interactive terminal user or computer interburst time in the data stream model of Jackson and Stubbs [JACK 69]. The upper estimate $M = 500$ is obtained as follows. From our assumptions in Section 2.3.1 for a 50 KBPS channel, there are 44.4 time slots in one second. For an average user think time of 30 seconds, $\sigma = 1/(30 \times 44.4)$. From $M\sigma \leq 0.37$, we get $M \leq 0.37 \times 30 \times 44.4 \cong 500$. Note that our assumption of a 50 KBPS channel was quite arbitrary. If a higher channel data rate is considered (say 5 MBPS), we may want to assume different average think times to reflect a different type of users.

description. This assumption implies zero deterministic delay ($R = 0$). In a satellite channel this obviously represents an approximation. However, it is physically realizable in radio communications over short distances in which channel propagation delays are negligible compared to a packet transmission time. In this case, the duration of each channel time slot can be made longer to include R.

A satellite channel (such as considered in Section 2.3) has a round trip propagation delay of 0.27 seconds, which necessitates a state description consisting of the channel history for at least R consecutive time slots. The difficulty in mathematical analysis using such a state description was illustrated in the last chapter. Moreover, it was shown that the channel recovery time following an input pulse depends only upon the channel input rate and the channel backlog size. This observation provided the motivation for the current mathematical model. Below we show by simulations that the mathematical model also gives excellent prediction of the throughput-delay performance of a channel with nonzero R. The conclusion is that in most cases of interest, the slotted ALOHA channel performance is dependent primarily upon the expected value of the retransmission delay ($\overline{RD}$) and quite insensitive to the exact probability distributions considered.

In order to use the analytic results of the mathematical model to predict the throughput-delay performance of a slotted ALOHA channel with nonzero R, it is necessary to use a value of p in the mathematical model which gives the same $\overline{RD}$. For example, to approximate a slotted ALOHA channel with uniform retransmission

randomization, we must let

$$p = \frac{1}{R + (K + 1)/2} \tag{5.3}$$

such that $\overline{RD} = R + (K + 1)/2$ in both cases.

We define the length of time for which a packet is backlogged to be the backlog time of the packet and denote the <u>average backlog time</u> by D_b. To obtain the average packet delay as defined in Section 2.3 and illustrated in Fig. 2-2, we must add to D_b, $R + 1$ time slots, which represent the delay incurred by each successful transmission. Thus, we have

$$D = D_b + R + 1 \tag{5.4}$$

In the mathematical model $N^t = n$ implies that in the t^{th} time slot $(M - n)$ users are in the thinking state, each of which may generate and transmit a new packet with probability σ. Hence the channel input rate is $S^t = (M - n)\sigma$. However, when R is nonzero, the number of thinking users may be less than $(M - n)$, since some users may have had a successful transmission, but R time slots must pass by before they learn that "successful transmission occurred" (see Section 2.3). Suppose the channel throughput rate is S_{out}. By Little's result [LITT 61], there are on the average $S_{out} \cdot R$ such users (approximately equal to 4.5 for $R = 12$ and $S_{out} = \frac{1}{e}$) which is negligible when M is large (say a few hundreds). Moreover, the value of M can be adjusted to reflect this average value. For our purposes, this discrepancy will be ignored.

To show that the throughput-delay performance of a slotted ALOHA channel is dependent primarily upon $\overline{RD}$ and quite insensitive to its exact probability distribution, we performed a simulation experiment with the following probability distributions for RD :

(1) R = 12 and uniform randomization

(2) R = 0 and uniform randomization

(3) R = 12 and geometric randomization

(4) R = 0 and geometric randomization

The number of channel users M was assumed to be 200. The duration of each simulation run was 8000 slots. As in Chapter 3, only those simulation runs which satisfied our channel equilibrium criterion were considered. Two values of $\overline{RD}$ were used for each of the four cases: a large $\overline{RD}$ corresponding to K = 60 in case (1) and a small $\overline{RD}$ corresponding to K = 10 in case (1). Equivalent values of K and p giving the same $\overline{RD}$ were used for the other three cases. In cases (2) and (4), Eq. (5.4) was used to compute the average packet delay. The throughput-delay tradeoffs for all four cases at each of the two values of $\overline{RD}$ are shown in Fig. 5-1. Within the accuracy of the simulation experiment, all four cases give practically the same throughput-delay performance, lending validity to our claim that the channel throughput-delay performance is dependent upon the expected value rather than the exact probability distribution of RD . (Of course, in certain uninteresting situations such as K = 1 or 2 in case (1) or p very close to one in case (3), our claim is obviously invalid.)

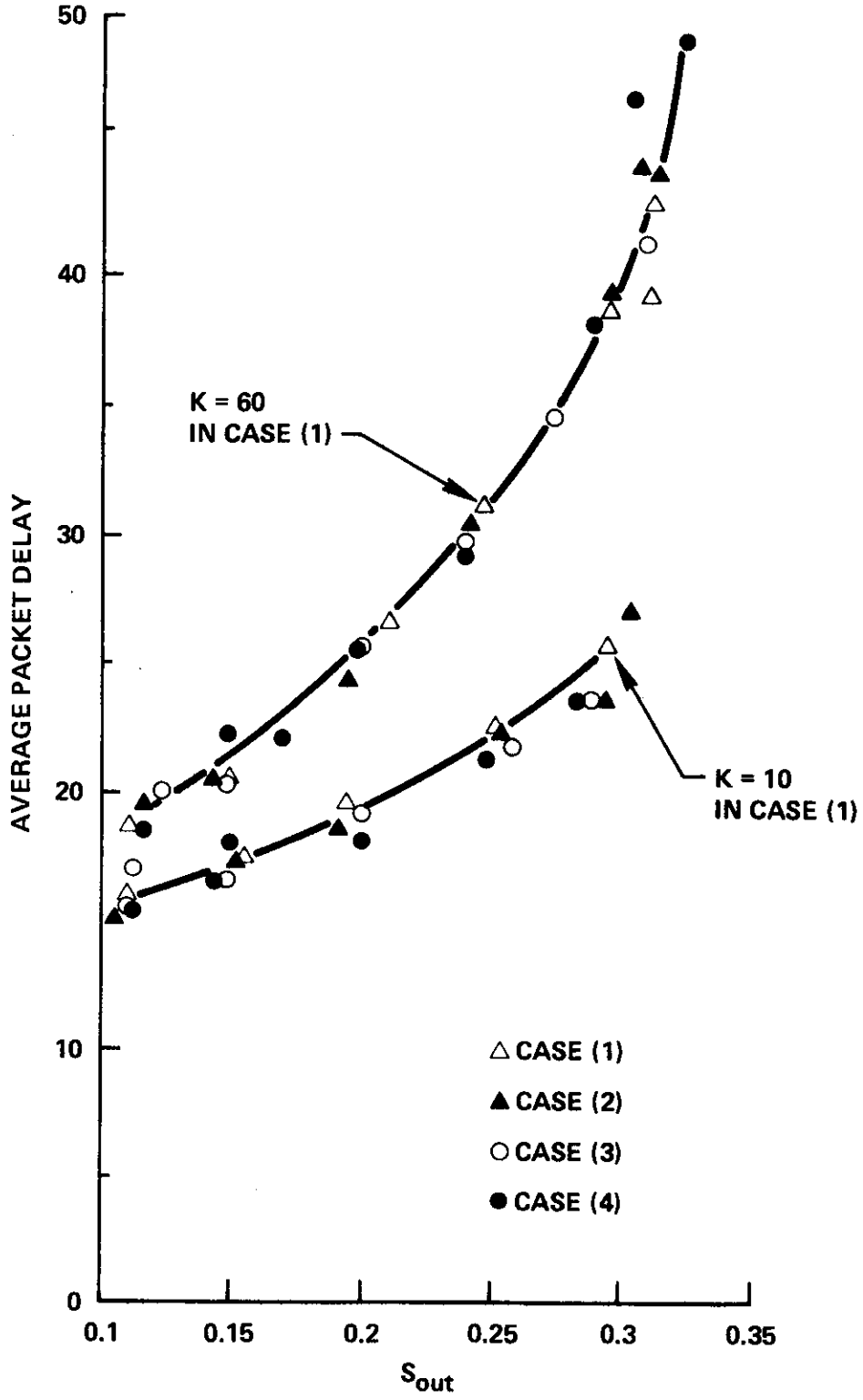

Figure 5-1. Comparison of Four RD Probability Distributions.

In this chapter and the next, the mathematical model as defined in the previous section will be studied. Use of Eqs. (5.3) and (5.4) enables the numerical results to be expressed in terms of K and compared with previous results on the slotted ALOHA channel performance for nonzero R and uniform retransmission randomization.

5.1.3 Channel Throughput

Conditioning on $N^t = n$, the expected channel throughput $S_{out}(n,\sigma)$ is the probability of exactly one packet transmission in the t^{th} time slot. Thus,

$$S_{out}(n,\sigma) = (1 - p)^n (M - n)\sigma(1 - \sigma)^{M-n-1} + np(1 - p)^{n-1}(1 - \sigma)^{M-n} \tag{5.5}$$

For the infinite population model, i.e., in the limit as $M \uparrow \infty$ and $\sigma \downarrow 0$ such that $M\sigma = S$ is finite and the channel input is Poisson distributed at the constant rate S, the above equation reduces to

$$S_{out}(n,S) = (1 - p)^n S e^{-S} + np(1 - p)^{n-1} e^{-S} \tag{5.6}$$

This expression is very accurate even for finite M if $\sigma << 1$ and if we replace $S = M\sigma$ by $S = (M - n)\sigma$. We assume that the condition $\sigma << 1$ is always satisfied in problems of interest to us.

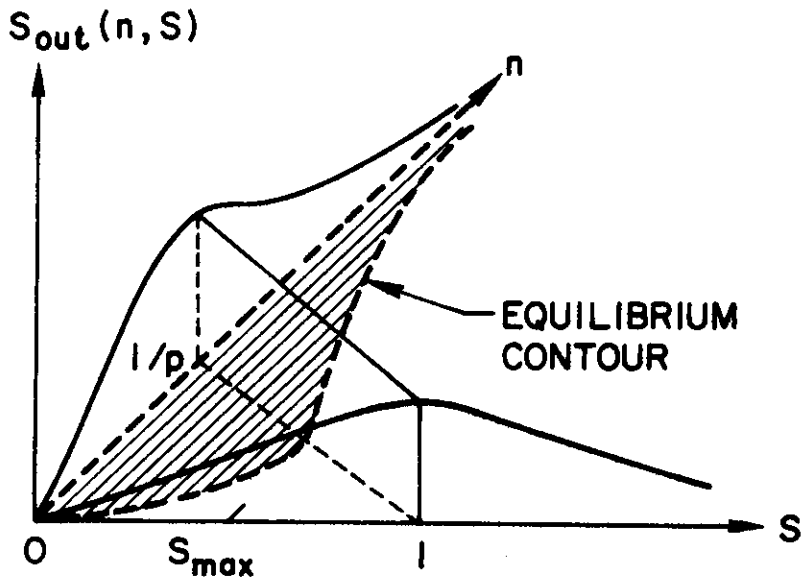

Figure 5-2. Channel Throughput Surface on the (n, S) Plane.

5.1.4 Equilibrium Contours

In Fig. 5-2, for a fixed K we show $S_{out}(n,S)$ as a 3-dimensional surface on the (n,S) plane given by Eq. (5.6). Note that there is an equilibrium contour in the (n,S) plane on which the channel input rate S is equal to the expected channel throughput $S_{out}(n,S)$. In the crosshatched region enclosed by the equilibrium contour, $S_{out}(n,S)$ exceeds S ; elsewhere, S is greater than $S_{out}(n,S)$ (the system capacity is exceeded!). In Fig. 5-3, a family of equilibrium contours for various K are displayed. We see that if we increase the average retransmission delay (by increasing K or equivalently decreasing p), these equilibrium contours move upwards. We show below that these equilibrium contours play a crucial role in determining the stability behavior of the channel.

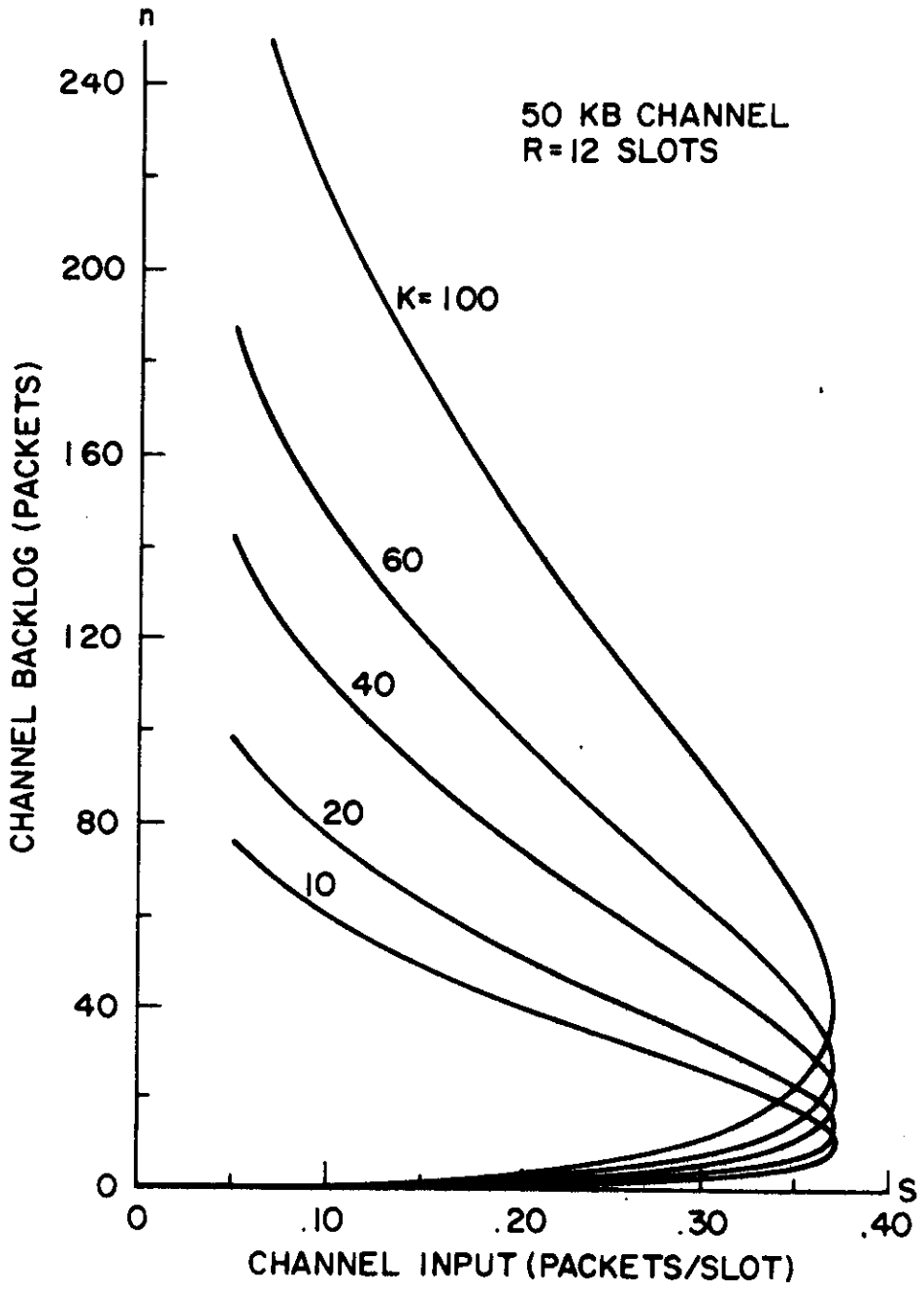

Figure 5-3. Equilibrium Contours on the (n, S) Plane.

Given M and σ and suppose a stationary probability distribution $\{P_n\}_{n=0}^{M}$ exists for N^t. Let $\bar{N} = \sum_{n=0}^{M} nP_n$. The stationary channel throughput rate S_{out} must be equal to the stationary channel input rate. That is,

$$S_{out} = \sum_{n=0}^{M} S_{out}(n,\sigma)\ P_n = \sum_{n=0}^{M} (M - n)\sigma\ P_n = (M - \bar{N})\sigma \tag{5.7}$$

For the equilibrium values of channel backlog size and throughput rate given by the condition $S_{out}(n,\sigma) = (M - n)\sigma$ to correctly predict the stationary average values $\bar{N}$ and S_{out}, a necessary condition is

$$S_{out}(\bar{N},\sigma) \cong \sum_{n=0}^{M} S_{out}(n,\sigma)\ P_n = (M - \bar{N})\sigma \tag{5.8}$$

For $p \ll 1$ and $\sigma \ll 1$, the above approximation is very accurate. For example, consider $K = 60$ and $M = 200$ in Fig. 5-8 below. The stationary channel throughput rate (computed by the value-determination operation in the next chapter) is found to be 0.344. The equilibrium value $S_0 = 0.346$.

Both the above equilibrium contours and the equilibrium contours shown in Figs. 3-3 and 3-4 in Chapter 3 are obtained under the condition that the channel input rate is equal to the channel throughput rate. Thus, a point specified by K and S in Fig. 5-3 must give rise to the same values of G and D in Figs. 3-3 and 3-4. Any discrepancy is due to the different approximations made in the two models (the first order approximation model and the linear feedback model).

The above claim can be verified by checking corresponding points on the contours. As an example, consider the point $K = 40$, $S = 0.275$ and $n = 54.5$ in Fig. 5-3. By Little's result [LITT 61], the average backlog time is

$$D_b = \frac{\overline{N}}{S_{out}}$$

Applying Eq. (5.4), we get $D = \frac{54.5}{0.275} + 13 = 211$ slots. Now if we check the corresponding point in Fig. 3-4 for $K = 40$ and $S = 0.275$, we find that $D = 212$ slots. In general, D values given by the linear feedback model are slightly less than those given by the first order approximation in Chapter 3. This is especially true when K is small such that the approximation in Eq. (5.8) becomes less accurate.

Channel state trajectories on the (n,S) plane

Given an equilibrium contour on the (n,S) phase plane, we consider here qualitatively the dynamic behavior of the channel subject to time-varying inputs. The following example serves to clarify similar fluid approximation results in Chapter 4.

Consider the case in which σ is constant while $M = M(t)$ is a function of time as shown in Fig. 5-4. We use the fluid approximation for the trajectory of the channel state vector (N^t, S^t) on the (n,S) plane as sketched in Fig. 5-5. Recall that $S^t = (M - N^t)\sigma$. The arrows indicate the "fluid" flow direction which depends on the relative magnitudes of $S_{out}(n,S)$ and S. Two possible cases are shown corresponding to different values of M_3 in Fig. 5-4. The

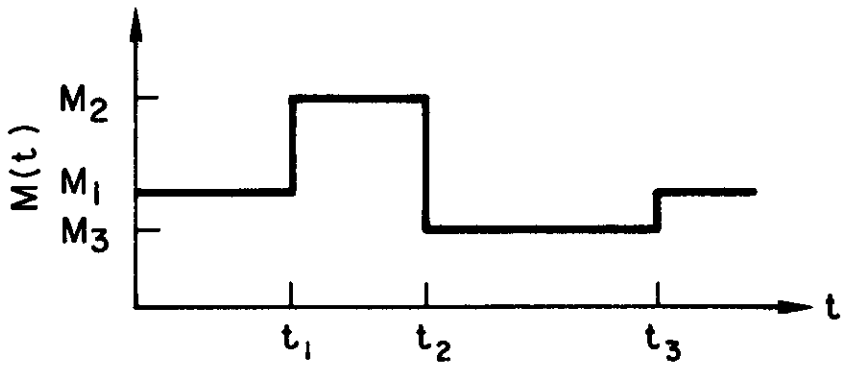

Figure 5-4. M(t)

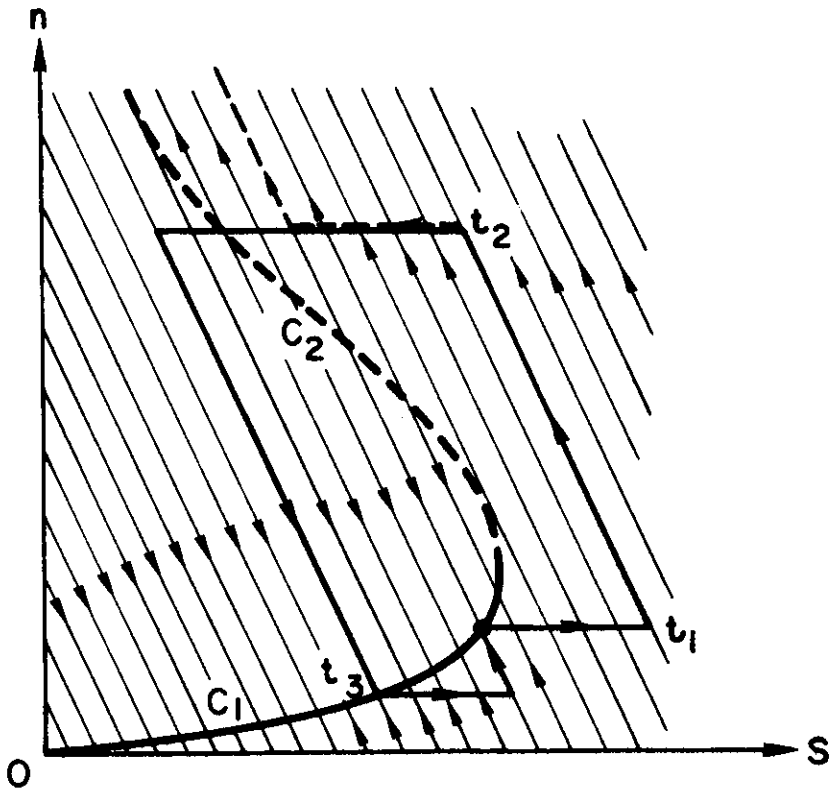

Figure 5-5. Fluid Approximation Trajectories.

solid line (case 1) represents a trajectory which returns to the original equilibrium point on contour C_1 despite the input pulse. The dashed line (case 2) represents a less fortunate situation in which the decrease in the channel input rate at time t_2 is not sufficient to bring the trajectory back into the "safe" region (in which $S < S_{out}(n,S)$). Eventually, the channel "collapses" as a result of an increasing backlog and a vanishing channel throughput rate. Compare these two cases with similar results in Figs. 4-1 and 4-2.

We have demonstrated channel saturation caused by a time-varying input. Next we study the conditions under which the slotted ALOHA channel with a stationary input (constant M and σ) can go into saturation as a result of statistical fluctuations.

5.2 Stability Considerations

We first define what we mean by stable and unstable channels and characterize their behavior. A stability measure is then given to quantify the relative instability of unstable channels.

5.2.1 Stable and Unstable Channels

Given M and σ, we define the channel load line in the (n,S) plane as the line $S = (M - n)\sigma$, which intercepts the n-axis at $n = M$ and has a slope equal to $-\frac{1}{\sigma}$.

The stability definition

> The channel is said to be stable if its load line intersects (nontangentially) the equilibrium contour in exactly one place. Otherwise, the channel is said to be unstable.

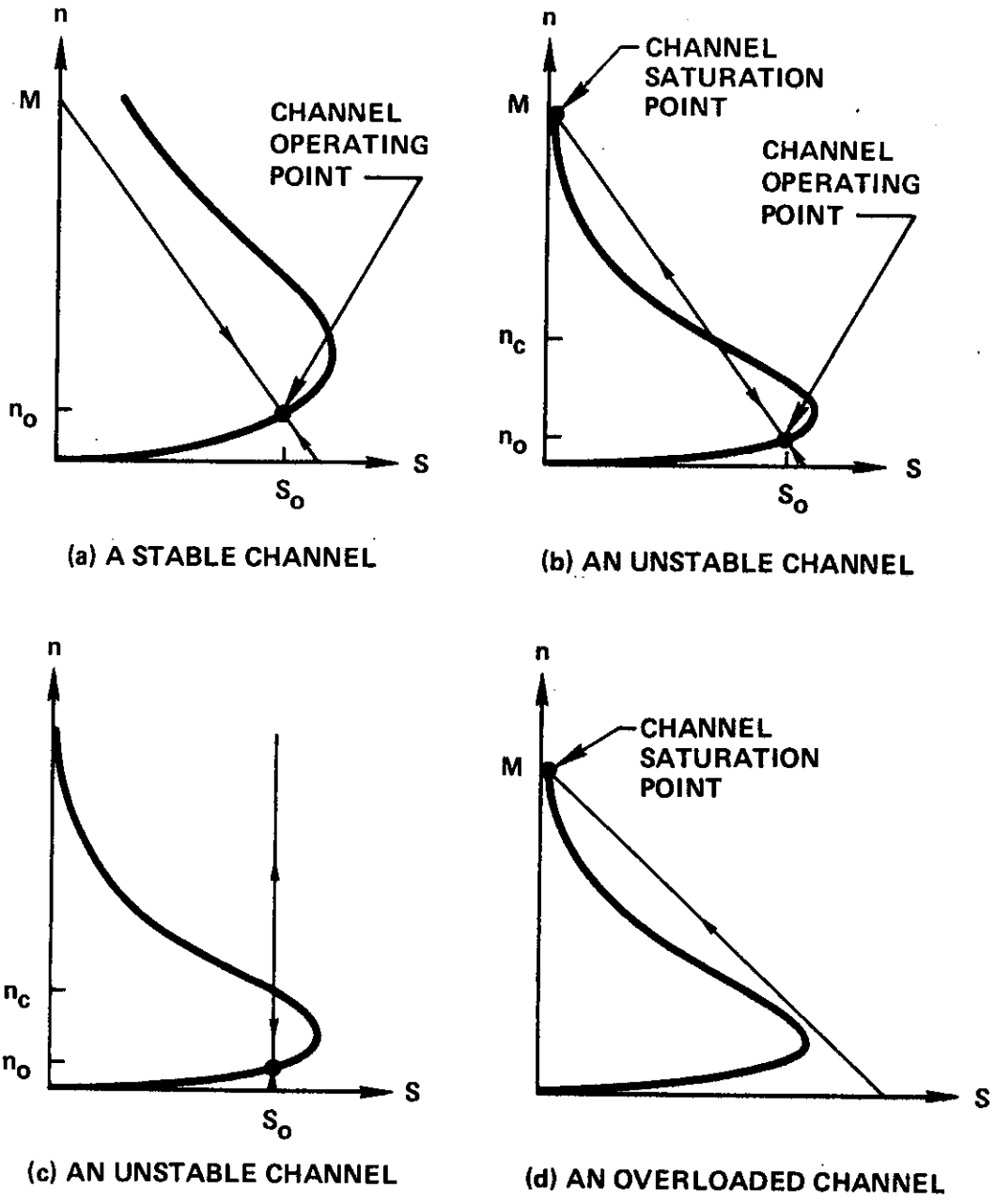

Figure 5-6. Stable and Unstable Channels.

Examples of stable and unstable channels are shown in Figs. 5-6. Arrows on the channel load lines indicate directions of fluid flow given by the fluid approximation. In other words, the arrow points in the direction of increasing backlog size if $S > S_{out}(n,S)$ and in the direction of decreasing backlog size if $S_{out}(n,S) > S$.

Each channel load line may have one or more equilibrium points. A point on the load line is said to be a stable equilibrium point if it acts as a "sink" with respect to fluid flow. It is a globally stable equilibrium point if it is the only stable equilibrium point on the channel load line. Otherwise, it is a locally stable equilibrium point. (Each stable equilibrium point is identified by a dot on channel load lines in Figs. 5-6 except in Fig. 5-6(c), where one of the stable equilibrium points is at $n = \infty$.) An equilibrium point is said to be an unstable equilibrium point if fluid flow emanates from it. Thus, the channel state N^t sitting on such a point will drift away from it given the slightest perturbation.

The stability definition given above is equivalent to defining a stable channel to be one whose channel load line has a globally stable equilibrium point.

In Fig. 5-6(a), we show the channel load line of a stable channel. Since N^t has a finite state space and is irreducible (assuming $p, \sigma > 0$), a stationary probability distribution always exists [PARZ 62]. Since (n_o, S_o) is the only equilibrium point on the load line, it gives the steady-state throughput-delay performance over an infinite time horizon under the approximation in Eq. (5.8). (n_o, S_o) will be referred to as the channel operating point. If M is finite, a stable channel can always be achieved

by using a sufficiently large K (see Fig. 5-3). Of course, a large K implies that the equilibrium backlog size n_o is large. As a result, the average packet delay may be too large to be acceptable.

In Fig. 5-6(b), we show the channel load line of an unstable channel. The point (n_o, S_o) is again the desired channel operating point since it yields the larger channel throughput and smaller average packet delay between the two locally stable equilibrium points on the load line. In fact, the other locally stable equilibrium point, having a huge backlog and virtually zero throughput, corresponds to channel saturation. It will be referred to as the channel saturation point. Since M is finite, and assuming $p, \sigma > 0$, a stationary probability distribution exists for N^t. However, N^t will "flip-flop" between the two locally stable equilibrium points in the following manner. Starting from an empty channel ($N^t = 0$ at time zero) quasi-stationary conditions will prevail at the operating point (n_o, S_o). The channel, however, cannot maintain equilibrium at this point indefinitely since N^t is a random process; that is, with probability one, the channel backlog N^t crosses the unstable equilibrium point n_c in a finite time and as soon as it does, the channel input rate S exceeds $S_{out}(n,S)$. Under this condition, N^t will drift toward the saturation point. (Although there is a positive probability that N^t may return below n_c, all our simulations showed that the channel state N^t accelerated up the channel load line producing an increasing backlog and a vanishing throughput rate.) Since the saturation point is a locally stable equilibrium point, quasi-stationary conditions will prevail there for some finite

(but probably very long) time period. In this state, the communication channel can be regarded as having failed. (In a practical system, external control should be applied at this point to restore proper channel operation.) The two locally stable equilibrium points on the load line of an unstable channel correspond to the channel being "up" or "down." An unstable channel may be acceptable if the average channel up time is large and external control is available to bring the channel up whenever it goes down.

In Figs. 5-7 and 5-8, we see how as the number of channel users M increases, an originally stable channel becomes unstable although the channel input rate S_o at the operating point remains constant (by reducing σ). For $S_o = 0.36$ and $K = 10$, we see that as M exceeds 80, the channel throughput decreases and the average packet delay increases very rapidly. (These results are obtained by solving for the stationary probability distribution of N^t using Algorithm 6.5 in the next chapter. No external control is assumed.) Using the stability definition and Fig. 5-3, the maximum value of M that is possible without rendering the load line unstable is $M_{max} = 79$, which exactly gives the knees of the curves in Fig. 5-7. In Fig. 5-8, by using a larger value of K $(= 60)$, a larger M_{max} is possible. Note, however, that the average packet delay $(\tilde{=} 56$ slots) for $K = 60$ is much larger than the average packet delay $(\tilde{=} 36$ slots) for $K = 10$. Given K and S_o, M_{max} can be obtained graphically from the equilibrium contours such as shown in Fig. 5-3. In Fig. 5-9, we show M_{max} as a function of K with S_o fixed at the maximum

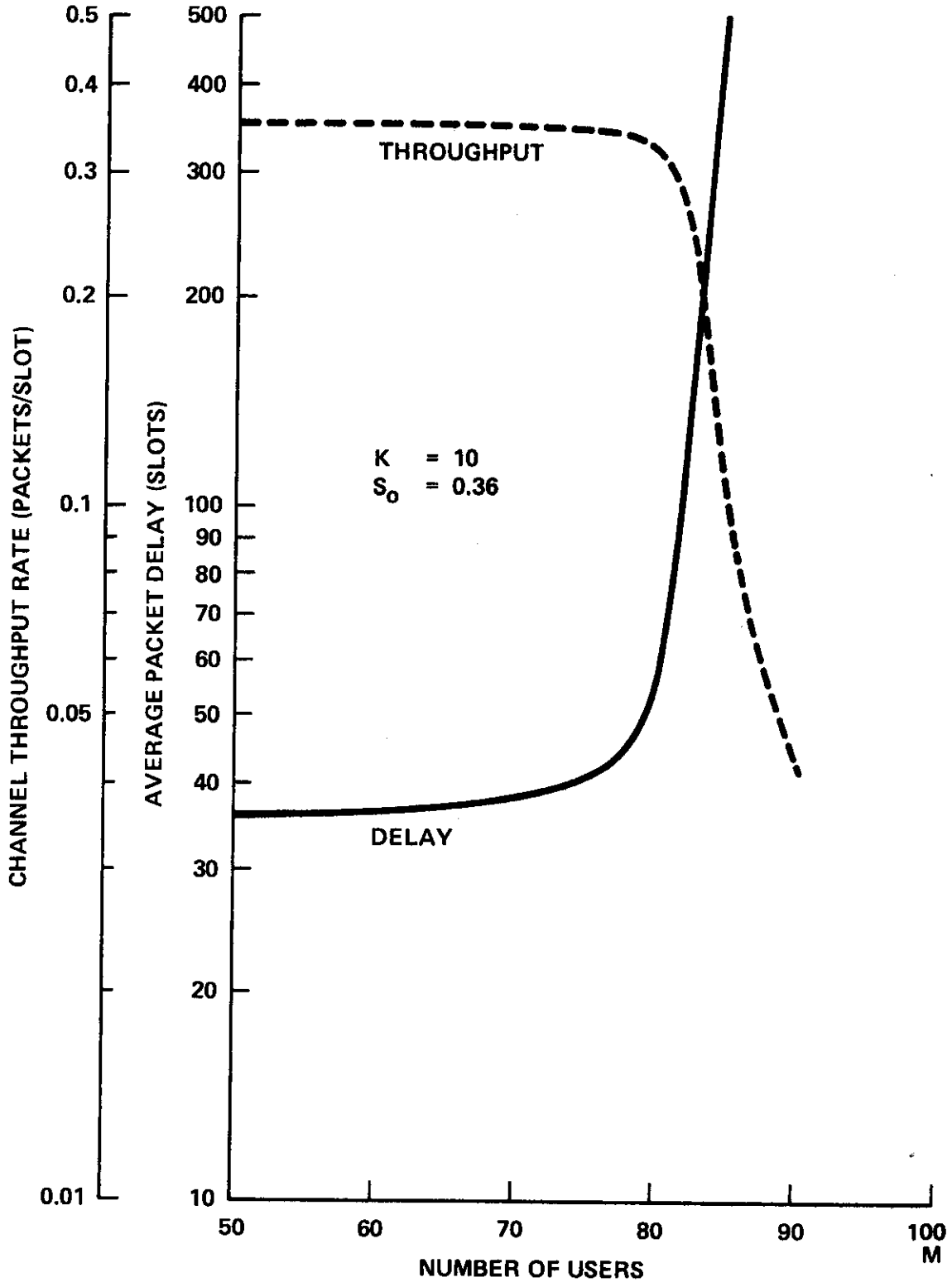

Figure 5-7. Channel Performance Versus M at K = 10 and S_o = 0.36

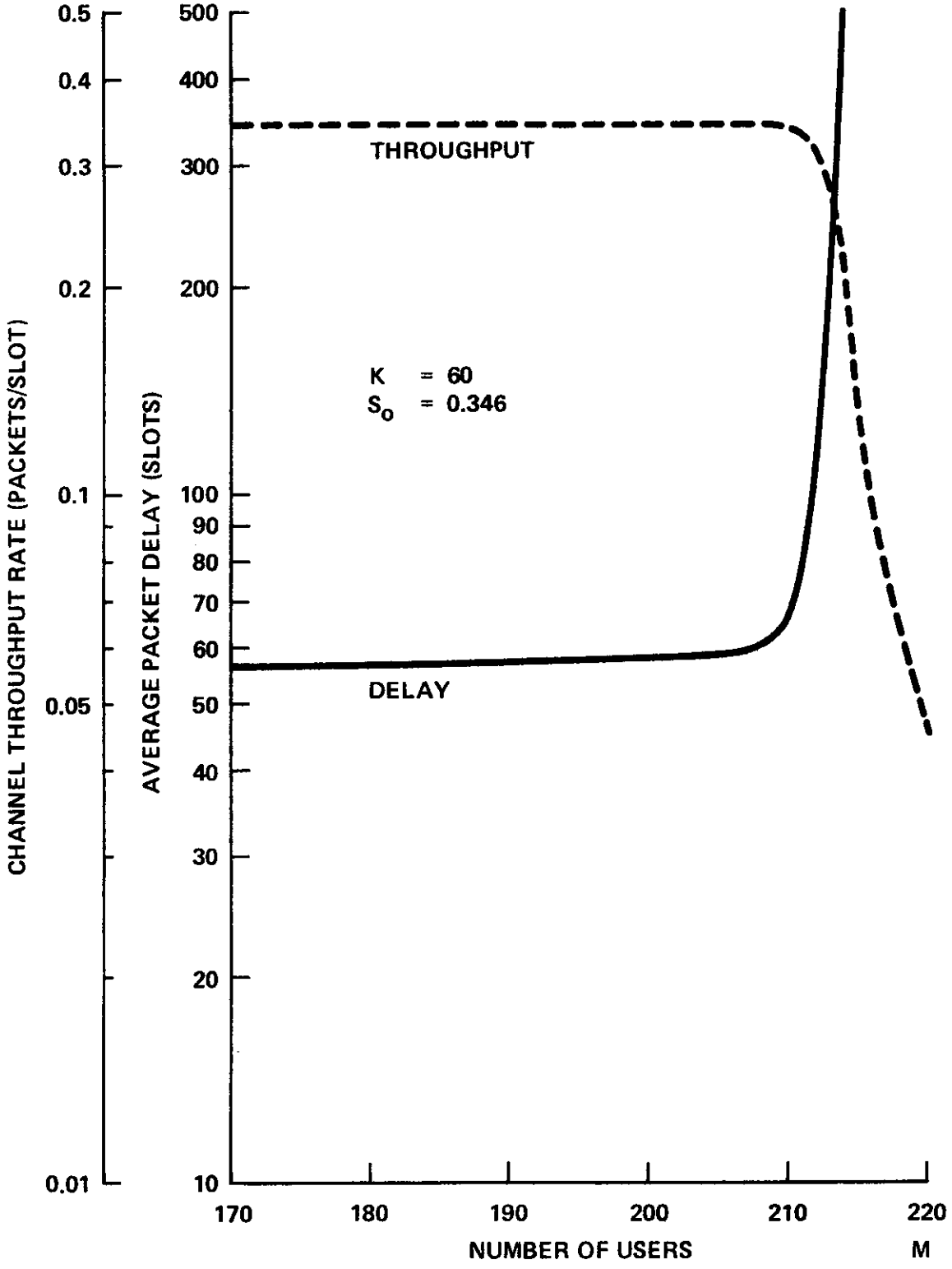

Figure 5-8. Channel Performance Versus M at K = 60 and S_o = 0.346

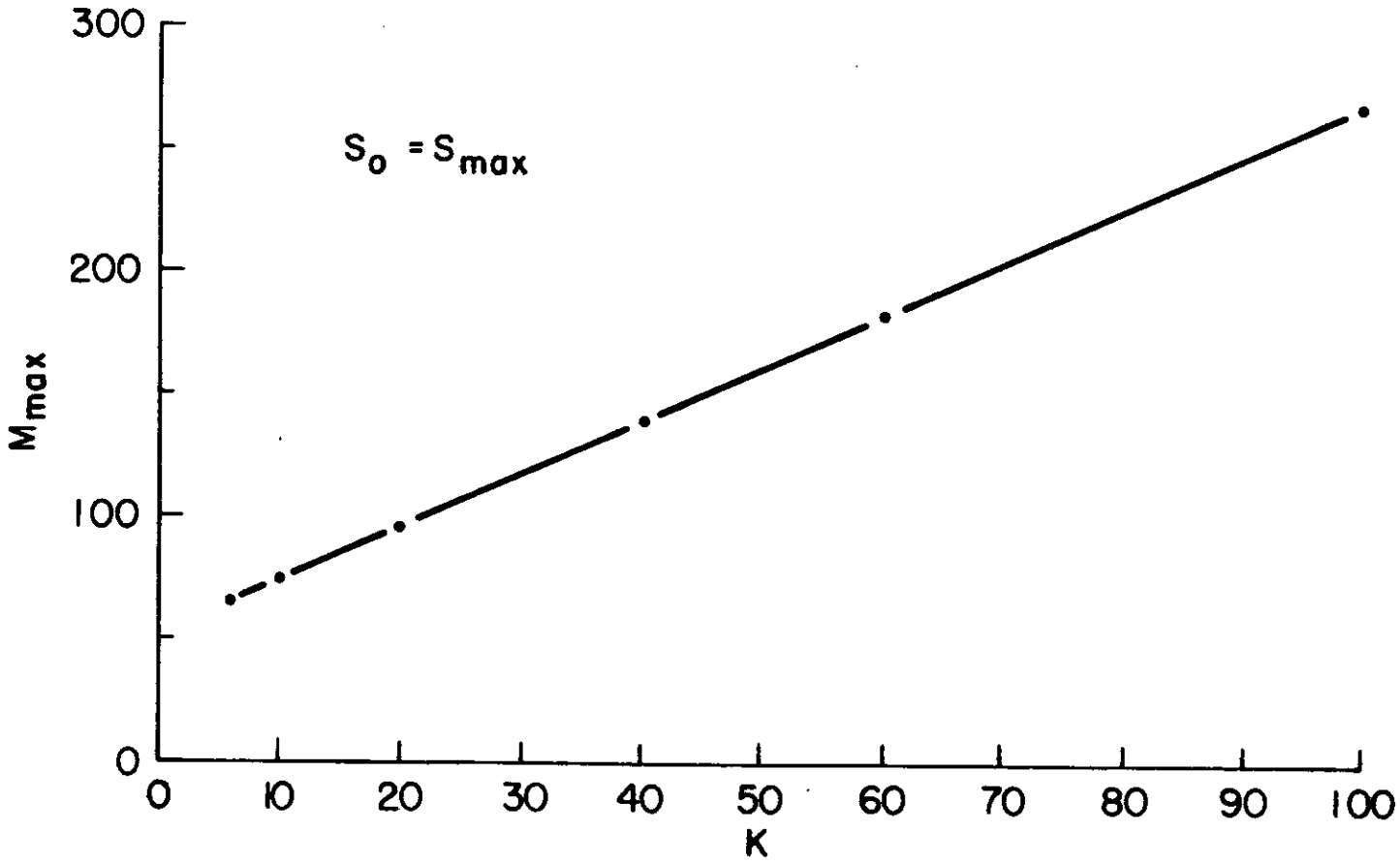

Figure 5-9. M_{max} Versus K.

value given K. Note the linear relationship between M_{max} and K for the values shown. In Fig. 5-10, we show that an originally unstable channel can be rendered stable by using a sufficiently large K.

The channel load line of an infinite population model is depicted in Fig. 5-6(c) as a vertical line. This is an unstable channel according to the stability definition. (Note that $n = \infty$ is a stable equilibrium point.) In fact, since N^t has an infinite state space and $S > S_{out}(n,S)$ for $n > n_c$, a stationary probability distribution does not exist for N^t. (See, for example, Cohen [COHE 69] pp. 543-546 for such proof.)

The channel load line shown in Fig. 5-6(d) is stable according to the stability definition. However, the globally stable equilibrium point in this case is the channel saturation point! Thus, this represents an "overloaded" channel as a result of bad system design. To correct this situation, the number of active users M supported by the channel should be reduced. Note that such an action is distinct from the dynamic control procedures in the next chapter, which are concerned with controlling temporary statistical fluctuations given that the channel is not overloaded in the above sense. From now on, a stable channel will always refer to the load line depicted in Fig. 5-6(a) instead of Fig. 5-6(d).

Let us summarize the major conclusions in the above discussions:

- The steady-state throughput-delay performance of a stable channel is given by its globally stable

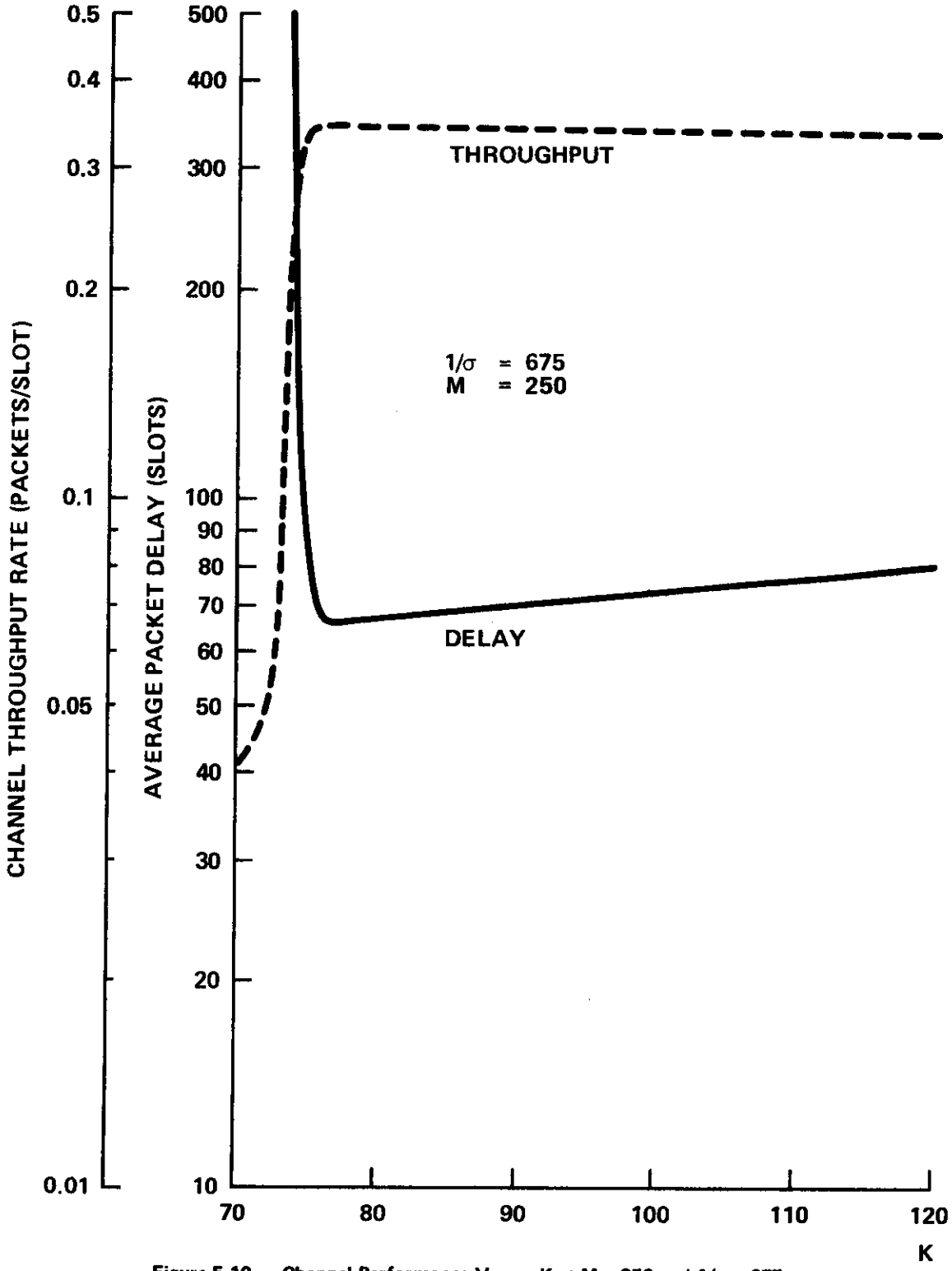

Figure 5-10. Channel Performance Versus K at M = 250 and $1/\sigma$ = 675.

equilibrium point and approximated by the equilibrium throughput-delay results in Chapter 3.

- In an unstable channel, the throughput-delay performance given by a locally stable equilibrium point can be achieved only for some finite time period.

5.2.2 A Stability Measure

From the above discussion and referring to Fig. 5-6(b), the load line of an unstable channel can be partitioned into two regions: the safe region consisting of the channel states $\{0, 1, 2, \ldots, n_c\}$ and the unsafe region consisting of the channel states $\{n_c + 1, \ldots, M\}$. A good stability measure (for these unstable channels!) is the average time to exit into the unsafe region starting from a safe channel state. To be exact, we define FET to be the average first exit time into the unsafe region starting from an initially empty channel ($N^t = 0$ at time zero) . Thus, FET gives an approximate measure of the average up time of an unstable channel. Below we derive the probability distributions and expected values of such first exit times. The derivations are based upon well-known results on first entrance times in Markov chains with stationary transition probabilities [HOWA 71, PARZ 62].

Consider the mathematical model in Section 5.1 with constant M and σ, where M may be infinite. N^t is a Markov process (chain) with stationary transition probabilities $\{p_{ij}\}$ given by Eq. (5.1) or Eq. (5.2). Define the random variable T_{ij} to be the number of transitions which N^t goes through until it enters state j for the first time starting from state i. The probability

distribution of T_{ij} (called the first entrance probabilities from state i to state j) may be defined as

$$f_{ij}(m) = \text{Prob}[T_{ij} = m]$$

$$= \begin{cases} 0 & m = 0 \\ p_{ij} & m = 1 \\ \text{Prob}[N^{t+m} = j,\ N^{t+h} \neq j,\ h = 1, \ldots, m-1 \mid N^t = i] & m \geq 2 \end{cases} \tag{5.9}$$

The state space S for N^t consists of the set of non-negative integers $\{0, 1, 2, \ldots, n_c, n_c + 1, \ldots, M\}$ which is partitioned into the safe region $\{0, 1, 2, \ldots, n_c\}$ and the unsafe region $\{n_c + 1, \ldots, M\}$. Now consider the modified state space $S' = \{0, 1, 2, \ldots, n_c, n_u\}$ where n_u is an absorbing state such that N^t is now characterized by the transition probabilities

$$p'_{ij} = \begin{cases} p_{ij} & i, j = 0, 1, \ldots, n_c \\ \sum_{\ell=n_c+1}^{M} p_{i\ell} & i = 0, 1, \ldots, n_c\ ;\ j = n_u \\ 0 & i = n_u\ ;\ j = 0, 1, \ldots, n_c \\ 1 & i, j = n_u \end{cases} \tag{5.10}$$

Define the random variable T_i to be the number of transitions which N^t goes through before it enters the unsafe region for the

first time starting from state i in the safe region. T_i is called the first exit time from state i. The probability distribution of T_i is defined to be $\{f_i(m)\}_{m=1}^{\infty}$ which are called the first exit probabilities. It is trivial to show that starting from state i $(0 \le i \le n_c)$, the first entrance probabilities into the absorbing state n_u in the modified state space S' are the same as the first exit probabilities into the unsafe region of S. Using Eq. (5.9), such probabilities are given by the following recursive equation [HOWA 71],

$$f_{in_u}(m) = p'_{in_u}\,\delta(m-1) + \sum_{j=0}^{n_c} p'_{ij}\, f_{jn_u}(m-1) \qquad \begin{matrix} m \ge 1 \\ i \ne n_u \end{matrix}$$

where

$$\delta(m) = \begin{cases} 1 & m = 1 \\ 0 & \text{otherwise} \end{cases}$$

The above equation can be rewritten in terms of the first exit probabilities as

$$f_i(m) = \sum_{j=n_c+1}^{M} p_{ij}\,\delta(m-1) + \sum_{j=0}^{n_c} p_{ij}\, f_j(m-1) \qquad \begin{matrix} m \ge 1 \\ 0 \le i \le n_c \end{matrix} \tag{5.11}$$

where $f_i(m)$ can be solved recursively for $m \ge 1$ starting with $f_i(0) = 0$ for all i.

The probability distribution $\{f_i(m)\}_{m=1}^{\infty}$ for the random variable T_i typically has a very long tail and cannot be easily computed. We had defined earlier FET as a stability measure for an unstable channel. By our definition, FET is the same as the expected value of the random variable T_0. Let $\overline{T_i}$ be the expected value and $\overline{T_i^2}$ be the second moment of T_i. These moments can be obtained by solving a set of linear simultaneous equations. It can easily be shown [HOWA 71] that

$$T_i = \begin{cases} 1 & \text{with probability } p_{in_u} \\ 1 + T_j & \text{with probability } p_{ij} \end{cases}$$

from which we obtain [HOWA 71, PARZ 62]

$$\overline{T_i} = 1 + \sum_{j=0}^{n_c} p_{ij}\, \overline{T_j} \qquad i = 0, 1, \ldots, n_c \qquad (5.12)$$

$$\overline{T_i^2} = 2\,\overline{T_i} - 1 + \sum_{j=0}^{n_c} p_{ij}\, \overline{T_j^2} \qquad i = 0, 1, \ldots, n_c \qquad (5.13)$$

Eqs. (5.12) form a set of $n_c + 1$ linear simultaneous equations from which $\{\overline{T_i}\}_{i=0}^{n_c}$ can be solved and FET ($= \overline{T_0}$) determined. After $\{\overline{T_i}\}_{i=0}^{n_c}$ have been found, Eqs. (5.13) can then be solved in a similar manner for $\{\overline{T_i^2}\}_{i=0}^{n_c}$.

5.3 Numerical Results

With the stability measure defined above, we are now in a position to examine quantitatively the tradeoff among channel stability,

throughput and delay for unstable channels. Below we first give a computational procedure to solve for $\overline{T_i}$ and hence, FET. They are then computed for various values of K, S_o and M (corresponding to different load lines). The stability-throughput-delay tradeoff is then shown.

5.3.1 An Efficient Computational Algorithm

The solution of the set of simultaneous equations in either Eq. (5.12) or Eq. (5.13) involves inverting the $(n_c + 1)$ by $(n_c + 1)$ matrix in p_{ij} for $i, j = 0, 1, \ldots, n_c$. When n_c is large, this becomes a nontrivial task because of the large number of computational steps and large computer storage requirement for the $[p_{ij}]$ matrix. The fact that $p_{ij} = 0$ for $j \leq i - 2$ in Eqs. (5.1) and (5.2) enables us to use the following algorithm which is very efficient in terms of both the computer time and space requirements mentioned above when n_c is large.

Algorithm 5.1

This algorithm solves for the variables $\{t_i\}_{i=0}^{I}$ in the following set of $(I + 1)$ linear simultaneous equations,

$$t_0 = h_0 + \sum_{j=0}^{I} p_{0j}\, t_j$$

$$t_i = h_i + \sum_{j=i-1}^{I} p_{ij}\, t_j \qquad i = 1, 2, \ldots, I$$

(1) Define

$$e_I = 1$$

$$f_I = 0$$

$$e_{I-1} = \frac{1 - p_{II}}{p_{I,I-1}}$$

$$f_{I-1} = - \frac{h_I}{p_{I,I-1}}$$

(2) For $i = I - 1, I - 2, \ldots, 1$ solve recursively

$$e_{i-1} = \frac{1}{p_{i,i-1}} \left[e_i - \sum_{j=i}^{I} p_{ij} \, e_j \right]$$

$$f_{i-1} = \frac{1}{p_{i,i-1}} \left[f_i - h_i - \sum_{j=i}^{I} p_{ij} \, f_j \right]$$

(3) Let

$$t_I = \frac{f_0 - h_0 - \sum_{j=0}^{I} p_{0j} \, f_j}{\sum_{j=0}^{I} p_{0j} \, e_j - e_0}$$

$$t_i = e_i \, t_I + f_i \qquad i = 0, 1, 2, \ldots, I-1$$

A derivation of the above algorithm is given in Appendix D. This algorithm is superior to conventional methods such as the Gauss elimination method [CRAI 64] for solving linear simultaneous equations in two respects. First, each p_{ij} is used exactly once and can be computed using Eq. (5.1) or Eq. (5.2) only when it is used in the

algorithm. This eliminates the need for storing the $[p_{ij}]$ matrix and practically eliminates any computer storage constraint on the dimensionality of the problem. Second, the number of arithmetic operations $(+ - x \div)$ required by the above algorithm is in the order of $2I^2$ which is less than that of conventional methods such as Gauss elimination.

5.3.2 Average First Exit Times (FET)

In Fig. 5-11, we have shown FET as a function of K for the infinite population model and for fixed values of the channel throughput rate S_o (at the channel operating point). We see that FET can be improved by either decreasing the channel throughput rate S_o or by increasing K (which in turn increases the average packet delay). The infinite population model results give us the worst case estimates for channel stability as demonstrated in Fig. 5-12 in which we show FET as a function of M for $K = 10$ and four values of S_o. Note that FET increases as M decreases and there is a critical value of M below which the channel is always stable in the sense of Fig. 5-6(a). As M increases to infinity, FET reaches a limiting value corresponding to the infinite population model with a Poisson channel input. Fig. 5-13 is similar to Fig. 5-11 except now the number of users M is 150. Recall that if M is finite, the channel will become stable when K is sufficiently large.

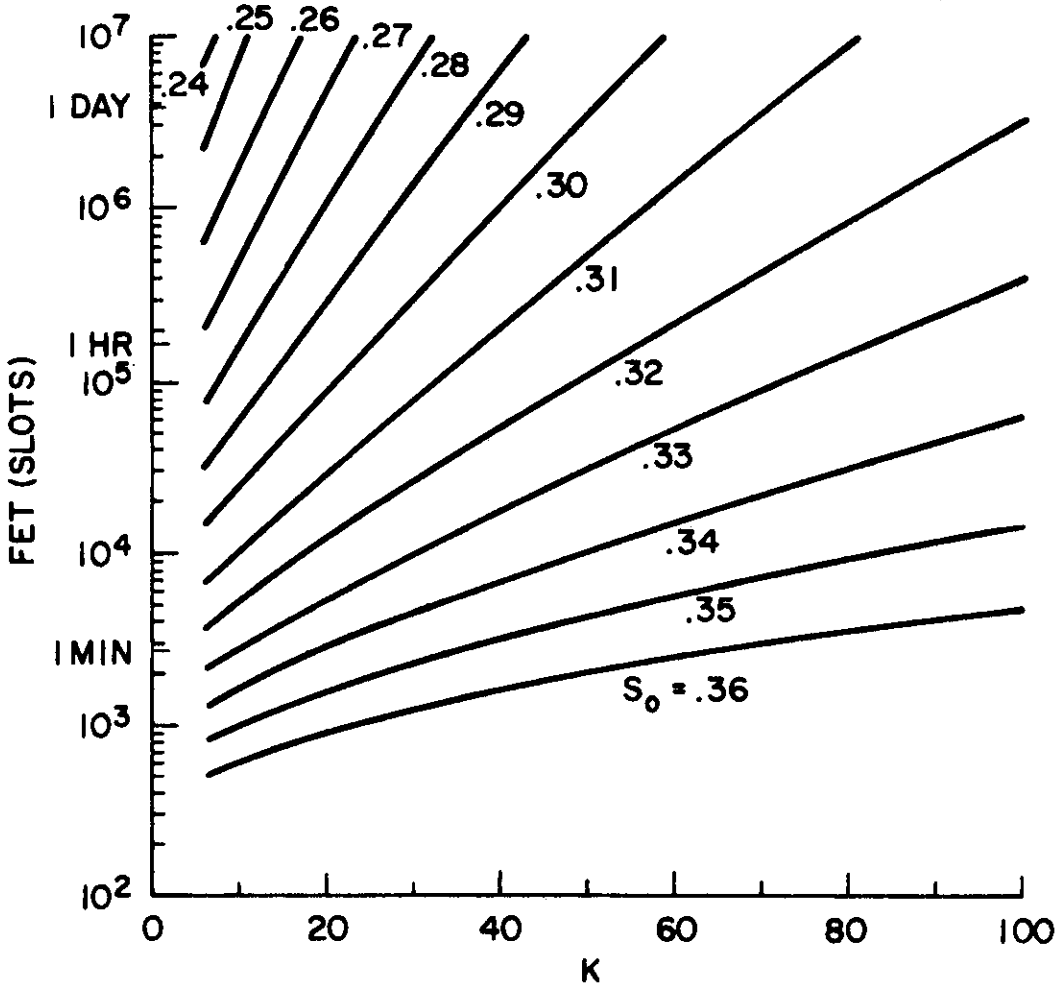

Figure 5-11. FET Values for the Infinite Population Model.

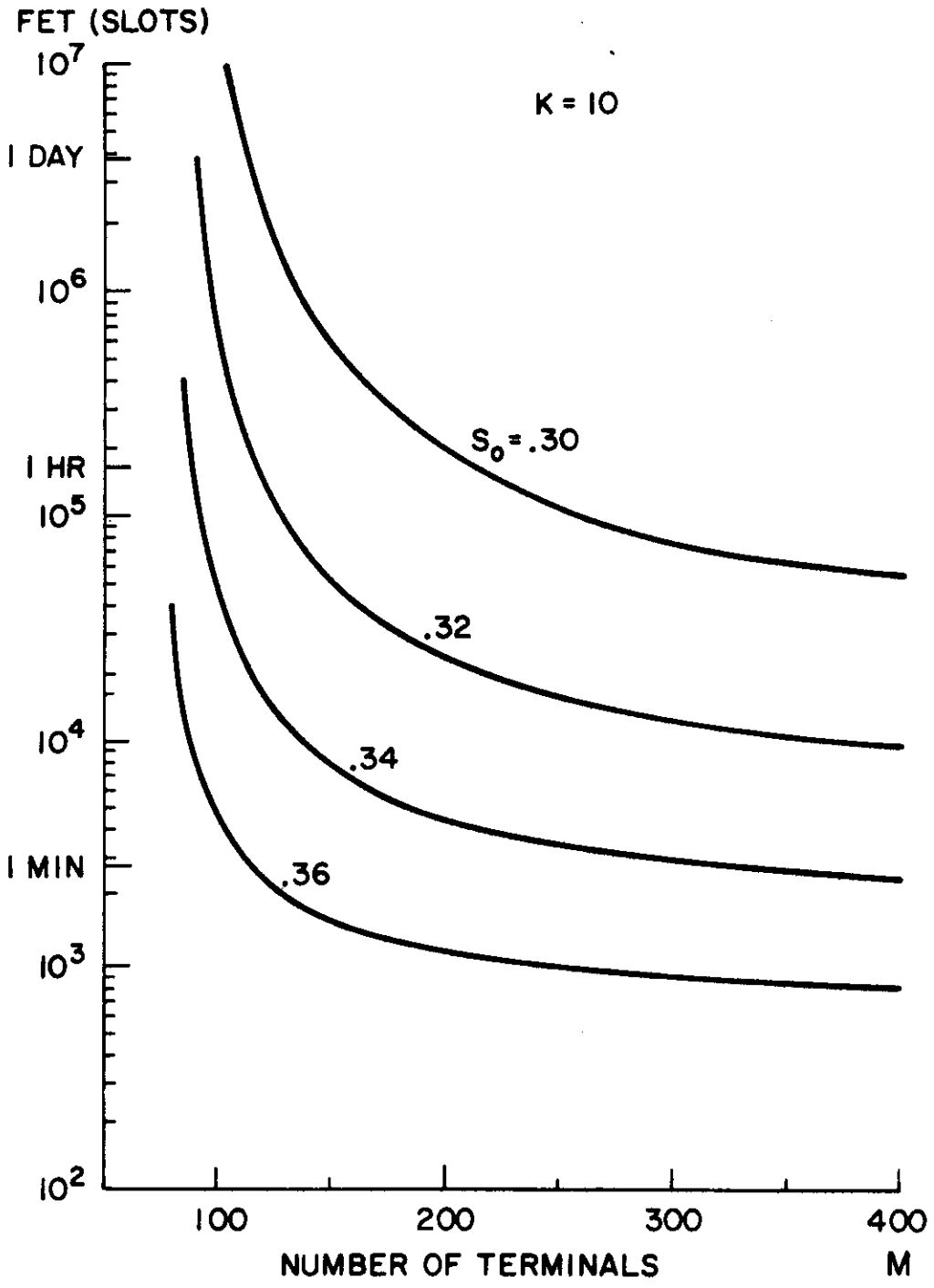

Figure 5-12. FET Versus M.

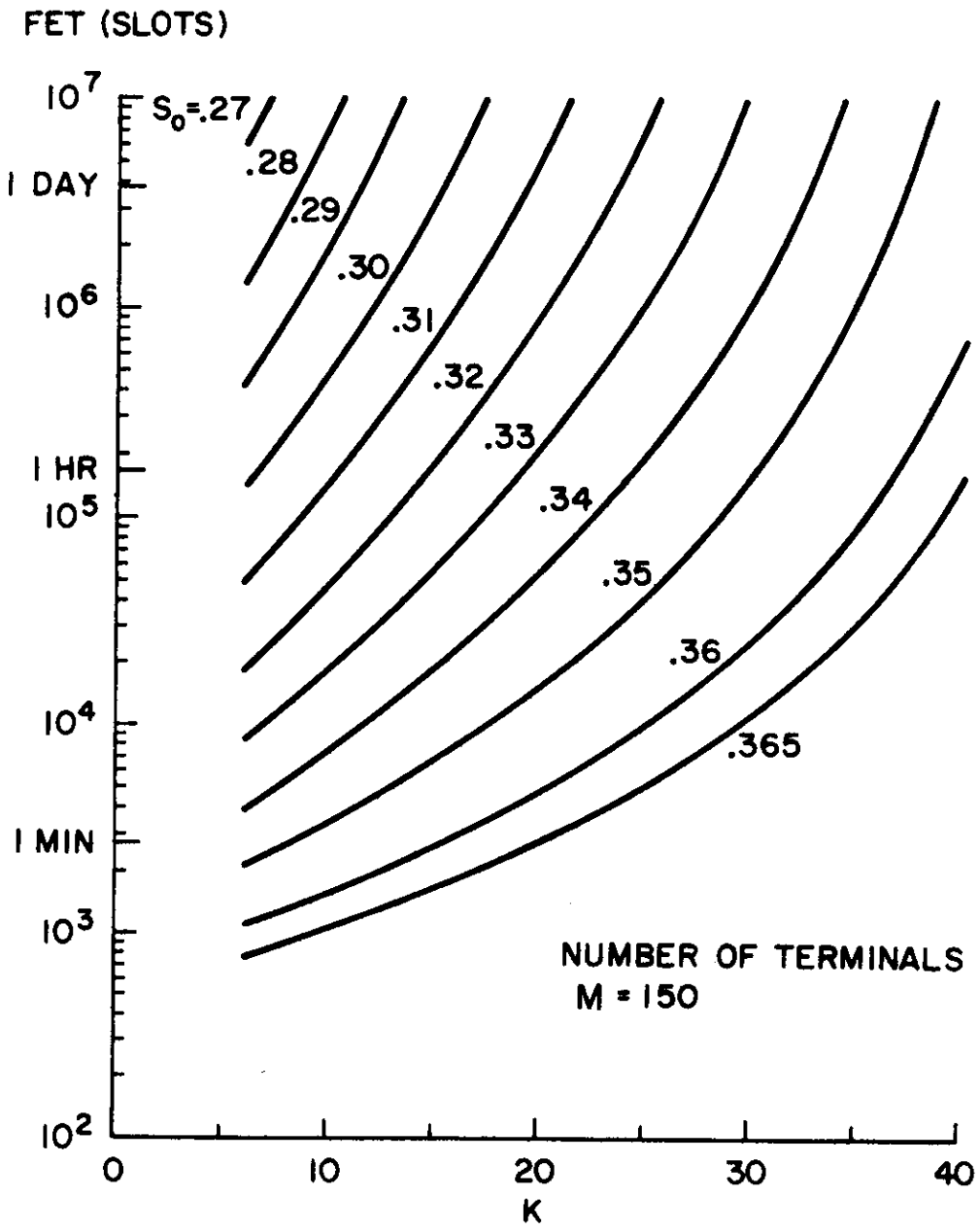

Figure 5-13. FET Values for a Finite User Population (M = 150)

As an example, we see that in Fig. 5-13 for $M = 150$, if the channel throughput rate S_o is kept at approximately 0.28 and $K = 10$ is used, the channel is estimated to fail once every two days on the average. If this is an acceptable level of channel reliability, then no other channel control procedure is necessary except to restart the channel whenever it goes into saturation. However, if absolute channel reliability is required at the same throughput-delay performance, dynamic channel control strategies should be adopted. Channel control schemes will be investigated in the next chapter.

5.3.3 The Stability-Throughput-Delay Tradeoff

In Fig. 5-14, we show as a lower bound the optimum performance envelope in Fig. 3-4 for the throughput-delay tradeoff of the infinite population model. This corresponds to the channel performance at the channel operating point indicated in Figs. 5-6. From these same figures, we see that the channel operating point (n_o, S_o) provides no information on the stability behavior of the channel. The equilibrium performance given by (n_o, S_o) is achievable in the long run if M is small enough such that the channel is stable; else, it is achievable only for some random time period estimated by our stability measure FET.

A design example

The designer of a slotted ALOHA channel is thus faced with the problem of deciding whether he wants a stable channel by using it for a small number of users and sacrifices channel utilization or uses the channel to support a large number of users if he is willing to accept

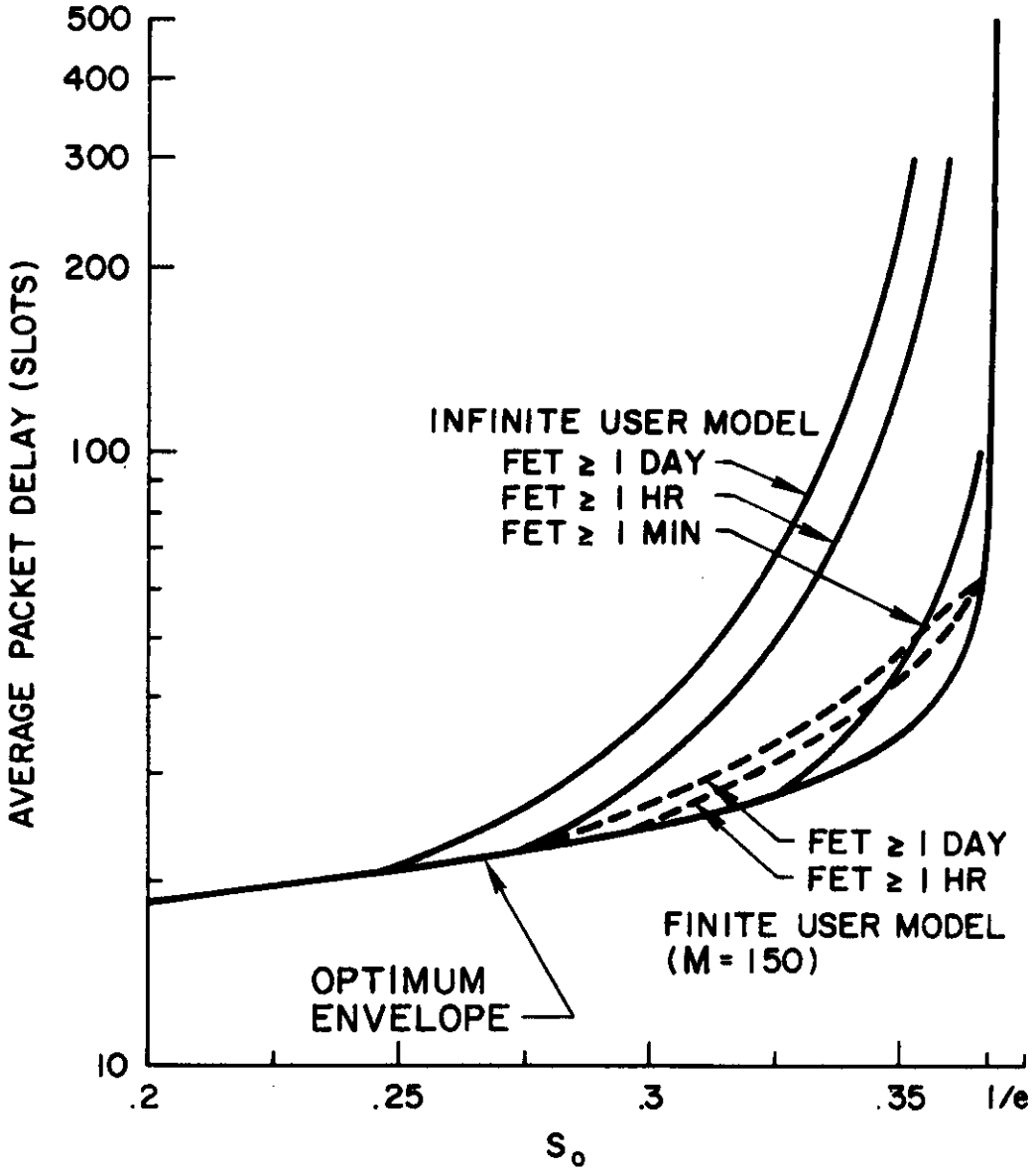

Figure 5-14. Stability-Throughput-Delay Tradeoff.

a certain level of channel reliability (some value of FET). For example, suppose K is chosen to be 10. (Note from Figs. 3-4 and 3-5 that $K = 10$ gives close to optimum equilibrium throughput-delay performance over a wide range of channel throughput rate.) Also, suppose that the channel users have an average think time of 20 seconds which, for our usual channel numerical constants, correspond to 888 time slots. Now if we draw channel load lines on Fig. 5-3 with a slope equal to -888, the channel is stable up to approximately 110 channel users. For $M = 110$, the channel throughput rate S_o is about 0.125 packet/slot. From Fig. 3-4, the average packet delay is roughly 16.5 time slots (= 0.37 second). The same channel can be used to support 220 users at a channel throughput rate of $S_o = 0.25$ packet/slot. The average packet delay is 21 time slots (= 0.47 second). But now the channel is unstable! From Fig. 5-11, for $K = 10$ and $S_o = 0.25$, the average up time (FET) of the channel is approximately two days for an infinite population model. Note that this value represents a lower bound for the FET of $M = 220$. Thus, we see that if a channel failure rate of once every two days on the average is an acceptable level of reliability, the second channel design is much more attractive than the first since the number of channel users is more than doubled at a modest increase in delay.

In addition to the infinite population model optimum envelope, we also show in Fig. 5-14 two sets of equilibrium throughput-delay performance curves with guaranteed FET values. The first set consists of three solid curves corresponding to an infinite population model

with channel FET ≥ 1 day, 1 hour and 1 minute. Again, these results represent worst case estimates when M is finite. The second set consists of two dashed curves corresponding to $M = 150$ with channel FET ≥ 1 day and 1 hour. These results were obtained by looking up the values of K and S_o in Fig. 5-11 or Fig. 5-13 corresponding to a fixed FET. The average packet delay was then obtained from Fig. 3-4. This figure displays the fundamental tradeoff among channel stability, throughput and delay. In the next chapter, we devise strategies to dynamically control the channel to achieve truly stable throughput-delay performance close to the optimum performance envelope.