A QUEUEING MODEL OF INFORMATION FLOW IN A COMMAND AND CONTROL SYSTEM

A. D. MURPHY

A QUEUEING MODEL OF INFORMATION FLOW IN A COMMAND AND CONTROL SYSTEM

by

A. D. Murphy

A QUEUEING MODEL OF INFORMATION FLOW IN A COMMAND AND CONTROL SYSTEM

by

A. D. Murphy II

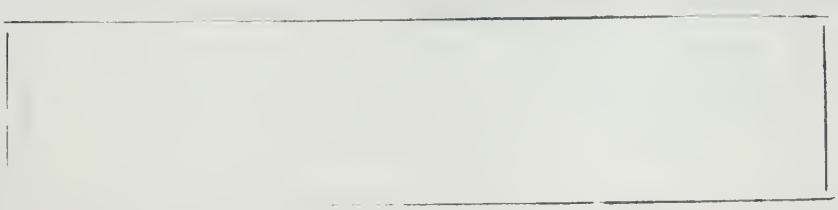
Lieutenant, United States Navy

Submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE

United States Naval Postgraduate School
Monterey, California

1962



A QUEUEING MODEL OF INFORMATION FLOW IN A COMMAND AND CONTROL SYSTEM

by

A. D. Murphy

This work is accepted as fulfilling
the thesis requirements for the degree of

MASTER OF SCIENCE

from the

United States Naval Postgraduate School

ABSTRACT

An important phase in the evolution of a command and control system is an analysis of the information required for human decision at each command level within the system.

To provide a methodology yielding quantitative results which may assist a commander and his staff in this analysis, it is proposed that the problem of the volume of information flow be treated by an application of queueing theory; each command level within the system is considered to behave as a service counter, and the incoming volume of information is related to the concept of customers arriving for service.

The information is the type which is considered to require positive human attention and decision; it may be grouped into classes (depending on content), and may carry designations of priority (depending on urgency).

Standard queueing parameters, results, and measures of effectiveness are re-defined in terms of the analogy proposed. Three queueing situations are presented which lend themselves to the analogy.

The measures of effectiveness may be used by a military commander as performance standards for each command level within the system. The relation is shown between performance standards and the amount of information which may be handled at each command level in the system.

Major conclusions of the paper are that (1) efficient performance at each command level is dependent primarily on the system commander's policy regarding the generation of information, and (2) that training may at times degrade system performance.

Recommendations are made for implementing the results and conclusions.

PREFACE

With the advent of command and control systems, several studies have been made which present guidelines for the proper evolution of such systems. One of the problem areas involves a determination of exactly what information is needed at various command levels for efficient operation of the system as a whole. Some of this information involves positive action-taking (that is, decision-making) on the part of human beings in the system. The determination of what information requires explicit human attention is a responsibility of various high-level commanders and their staffs.

The author has become familiar with some basic command and control concepts during a tour of duty in the Operations Analysis Section (Code 2850), U. S. Navy Electronics Laboratory, San Diego, California; a two-year course of instruction in the Operations Analysis Curriculum at the U. S. Naval Postgraduate School, Monterey, California; and a summer field trip (between the years of postgraduate instruction) to the System Development Corporation, Santa Monica, California. In particular, a ten-week seminar in the Operations Analysis Curriculum was devoted to group discussion of the report of the Summer Study Group, Institute of Naval Studies, on Naval Command and Control.

Despite this exposure to some basic principles of command and control, and specifically to the need for conducting an analysis of the information required for decision, the author feels that there is a need for a methodology yielding quantitative answers to questions of the following type:

(a) How much information can be received by a military commander,

and by his subordinate commanders?

(b) Is there an upper limit to the volume of information received, beyond which a military commander and his subordinate commanders will be unable to function effectively?

(c) Can a military commander prescribe quantitative performance standards for the efficient functioning of all levels of command within his force?

(d) Is there some information, currently in the sphere of positive, human decision-making, which may have to be eliminated (by incorporation into standard operating procedure, for example) in order to preserve positive human decision on "serious" matters (such as the employment of high-yield weapons)?

Without a method which provides quantitative guidelines with which to attempt answers to these and related questions, the most thorough and dedicated analyses of information requirements are felt to be deficient.

In an attempt to assist the military commander and his staff in the determination of what information is required for decision-making at each command level, this paper suggests that queueing theory may provide some quantitative guidelines.

The basic hypothesis in this paper is that each command level in a command and control system (that is, each level of authority in the system) may be considered to be related to the queueing theory concept of a "service stand" which services incoming information, and that this incoming information is not unlike a "queue" of customers awaiting service. Section 1 introduces the analogy. Section 2 considers control systems in general, and Section 3 is a justification of this analogy.

Sections 4, 5, and 6 consider three specific queueing situations which are pertinent to the analogy proposed. In each of these sections, the particular situation under consideration is described in standard queueing terminology; the mathematical results of authors who have examined these situations are presented and interpreted in terms of the analogy; a few standard queueing measures of effectiveness are re-defined in the framework of the analogy, and are discussed from the standpoint of a military commander; specific conclusions for each situation are contained in the Section devoted to that situation, while general conclusions drawn from all situations appear in Section 7. Recommendations are contained in Section 8.

Qualitatively speaking, the standards of performance which may be demanded of a subordinate level of authority by the commander of a command and control system are seen to depend largely on the commander's own information-generating requirements. Quantitatively speaking, queueing theory provides an explicit determination of some upper bounds on these information-generating requirements, and thereby the purpose of this paper is fulfilled.

The author wishes to express his appreciation for guidance in the preparation of this paper to Thomas E. Oberbeck, Chairman of the Department of Operations Research, U. S. Naval Postgraduate School, and to Jack R. Borsting, Associate Professor of Mathematics, Department of Mathematics and Mechanics, U. S. Naval Postgraduate School; for encouragement, to Harold F. Erickson, Chief Technical Editor, Stanford Research Institute Research Office, Fort Ord, California; and for clerical assistance, to Mrs. Norma Stevens of the Naval Engineering Curriculum Office, U. S. Naval Postgraduate School.

TABLE OF CONTENTS

SECTION 1

INTRODUCTION

Throughout recorded history, military commanders have needed information on which to make decisions; they have needed information on an enemy's forces, movements, and readiness; and they have needed similar information about their own forces. In this regard, Alexander the Great and, say, CINCPAC are no different. (No disrespect is intended, -- to either gentleman.)

It is supposed that, even in the time of Alexander, information could be grouped into three gneral types: (1) the "need-to-know" type -- information that was absolutely essential to operations and administration; (2) the "nice-to-know" type -- information that provided a sort of background, setting, or mood; (3) the "unnecessary" type -- information in which the military commander had no interest.

It is supposed further that, even in those days, the definitions of these types of information were completely qualitative and subjective in nature, subject to the changing objectives, missions, and temperament of the commander (if such definitions existed at all).

The passing of centuries has brought about not only an increasing number of problems for the military commander (in turn requiring more information by which he may make decisions), but also an increase in the means of communication by which he receives information.

In an effort to insure that information is provided on an orderly basis, today's military commander has codified, to an extent, his ideas of his information requirements (and the requirements of his

subordinate commanders within his force) in what the author refers to as "information-generating requirements." These information-generating requirements provide for such things as minute-by-minute POSITS, hourly SITREPS, daily OPSUMS, weekly DIGESTS, monthly SUMMARIES, semi-annual REPORTS, and so forth. The information content of this large volume of information so generated is of an operational nature, or an administrative nature, or both.

It is noted that this volume of information tends to increase with an increase in the tempo of the commander's operations, such as, for example, during a short-term world "crisis" of sorts. It also increases with the mere passage of time, which has brought a greater integration of the commander's military strategy with overall national objectives, a greater diversity of forces employed by the commander, a larger "bag" of complex weapons at his disposal, and more channels of communication.

Of course there are periodic reviews of information-generating requirements, in an attempt to eliminate and/or consolidate some of this information. On the other hand, the author's considered opinion is that it is very hard to eliminate a report, a form, or a format, once one has been established; people get used to it, and feel that they need it; it is just human nature.

On several occasions, the author has observed commanding officers and unit commanders staring forlornly at an overflowing action-basket or a bulging message board, and saying in effect: "I suppose it's necessary. But can't 'something' be done about it?" (It is assumed that, on occasion, more senior military commanders react similarly.)

The plaintive cry quoted above includes an assumption which may be occasionally unwarranted; namely, that "it" -- this ever-growing mountain

of information -- is actually "necessary." Does a military commander actually need all the information he receives? Can he afford to spend more and more time absorbing all the subtleties of increasingly complex matters, matters on which he must make positive decisions, and decisions for which he alone is responsible? Can he afford NOT to? There is no clear-cut answer; about the best approximation to an answer is that "it all depends" on the situation, as indeed it does.

To the question "Can't 'something' be done?" there is an affirmative reply. That "something" is the exhaustive study and planning attendant upon the development of a command and control system to assist the modern military commander in the accomplishment of his various missions. This is because the study and planning is meant to define precisely, among other things, the commander's critical problem areas and those of his subordinate levels of command within his force; some of these critical problem areas are those in which the commander, and his subordinate commanders, must take positive action, make positive decisions. Now a natural result of any definition of these particular critical problem areas is a determination of exactly <u>what</u> information is required for decision. It is emphasized that the information so determined is that which is deemed absolutely necessary for decision; the implication is that all "nice-to-know" information has been determined to be either "need-to-know" (and retained) or "unnecessary" (and eliminated).

The author feels that <u>quantitative</u> guidelines are necessary to assist in the determination of information requirements, and that these quantitative guidelines may be available through an application of mathematical queueing theory.

Basically, queueing theory is concerned with "customers" arriving at a "service counter" where they may have to "wait" for service because of the presence of other customers ahead of them in a line (or "queue"). Certain measures of effectiveness have been established for evaluating the efficiency of a queue; for example, the average number of customers waiting in line for service is often used as a measure of effectiveness; another one is the average time that a customer waits for service; a third one is the probability that there are no customers waiting for service and none in service (in other words, the probability that the service counter is idle).

If the results of abstract queueing theory are interpreted in terms of the problem of volume of information flow, the author maintains that quantitative guidelines will be available to assist a military commander and his staff in the determination of information requirements for each command level within his command and control system.

Section 2 of this paper discusses what are considered to be fundamental concepts of control systems in general, and military command and control systems in particular.

Section 3 is devoted to establishing the analogy proposed herein, and to formulating, in standard queueing terminology, the problem of the volume of information flow.

Sections 4, 5, and 6 consider three queueing situations which are interpreted in terms of the basic analogy. Section 4 is more detailed than the other two, in order to give the reader a feel for the queueing situation as applied to the problem of volume of information flow.

Sections 5 and 6 treat queueing situations which are more sophisticated than the one of Section 4, but these sections are less detailed

and assume a familiarity with Section 4.

Section 7 contains general conclusions from the results of Sections 4, 5, and 6, and assumes a familiarity with some basic queueing terminology to be found in Section 4. In non-mathematical terms, the two major conclusions of this paper are:

(1) the efficiency of a command and control system (from the standpoint of timely, positive decision-making on the part of the military commander of the system and of his subordinate commanders) is determined primarily by the commander's own policy regarding the generation of information within the system;

(2) intensive training and exercising of decision-making personnel at all levels within the system is, alone, not the way to achieve higher system performance, since in general, such training (alone) may actually degrade system performance under certain circumstances.

Section 4 illustrates these points at some length, and supports these major conclusions.

Section 8 contains some specific recommendations for implementing the results and conclusions of this application of queueing theory to the problem of volume of information flow.

Appendix A presents the development of a basic equation of queueing theory along the axiomatic lines of modern probability theory. An understanding of the main body of the paper does not depend upon Appendix A.

In summary, the author feels that the analysis of information requirements (which is but one phase in the evolution of a command and control system) demands a methodology which yields quantitative results; it is the purpose of this paper to suggest that queueing theory may provide the quantitative guidelines for this analysis, as well as some performance standards for the command and control system.

If the basic analogy proposed is not accepted, then of course all results and conclusions are meaningless. Nevertheless, it is offered as an approach, on the strength of still another general principle: it is better to light one match than to curse the darkness.

SECTION 2

CONTROL SYSTEMS

In any large organization, much attention is given to controlling its activities in anticipation of improving the extent to which the organization as a whole achieves its purpose and its goals. In some instances there is concern by one component of the organization with the efficiency of another component engaged in the execution of certain plans or operations which have been developed for that component. A consideration of such problems has led to requirements for so-called control systems.

In civilian environments, these control systems are referred to as "management control systems", in military environments they are known as "command and control systems."

For the purpose of this paper, a control system may be defined coarsely as an organization of personnel and equipment, established to perform and to supervise certain operations, tasks, missions, etc.

Control systems differ in several aspects; for example, they differ in function, personnel complement, degree of automation, flexibility, degree of decentralization, and operating environment.

Regardless of such differences, there are certain features which are similar in all control systems. For one thing, there are a finite number of "control points," "check points," or stages, which correspond to the various levels of authority and responsibility defined by the agency which has established the control system. Authority and responsibility at each stage are vested in the human being who personifies each stage of the system. (Thus, the military commander of a command and

control system personifies the supreme, or highest, stage of the system; lower stages in the system are personified by subordinate commanders.)

Furthermore, for the purposes of this paper, all control systems may be characterized by the concepts of information flow between stages of the system, and action-taking on the part of the human being who personifies each stage of the system.[1] Since action-taking is made possible by the flow of information between stages, the concept of information flow is considered the more fundamental.

In the development of this paper, however, it seemed desirable to particularize, and to limit the scope to an area which is neither so confining as to inhibit an extension to general cases, nor so obtuse as to defy reader identification. Therefore, the general area selected for discussion in this paper is that of military command and control systems. (For brevity, the words "system" and "control system" will be used henceforth.) The specific area selected for discussion is the volume of information flow into a typical stage of a command and control system.

---

[1]The terms "action-taker", "action-taking," etc., will be employed henceforth in order to emphasize the fact that each stage of the system acts upon information; avoided are the terms "decision-maker" and "decision-making" which are popularly ascribed to the person and the function, respectively, of the highest stage in the system.

SECTION 3

THE APPLICATION OF QUEUEING THEORY TO INFORMATION FLOW

The results of modern queueing theory are considered adaptable to the problem of information flow into a typical stage of a system. In order to support this hypothesis, the following points are to be noted:

(a) Information flows into a typical stage of the system from various sources.

(b) The content of this information will vary. Some examples are: inventories of supplies, contact reports, material readiness, etc. Based on content, certain classes of information may be established. Also, the information may be in the form of a book, sheet of paper, film, etc. Furthermore, the information may be transmitted to the stage by radio, mail, hand-delivery, etc. Regardless of the class, form, or means of transmission of the information, the information is considered to arrive at the stage in discrete units.

(c) The arriving information units under discussion here are those which require positive action/reaction on the part of the human being who personifies the stage.

(d) Action-taking on the part of the human being requires time; the action-taker may act upon only one unit at a time.

(e) Information units have associated with them various degrees of urgency, or priority;[2] the priorities are assigned arbitrarily,

---

[2] It might be thought that unit "class" and "priority" constitute a distinction without a difference: the distinction is justifiable, however, from the fact that, depending upon the real or simulated military operation, a particular class of information may assume various priorities.

but once assigned, they determine the relative order in which action will be taken upon information units.

(f) While action is being taken on an information unit, other units may be accumulating at the stage, forming a queue.

(g) An action-taker at a stage will generate outgoing information units for transmission to other stages; while this aspect is not to be considered, the point made is that such activity takes time, and may delay further action-taking upon arriving units.

(h) Information units may be of a recurring nature (up-dated at regular intervals) or they may be non-recurring affairs, or both. Among recurring units, action-taking may be required on each unit regardless of timeliness, or else only the most recent unit may be acted upon.

(i) Units are assumed to arrive at a stage free of error or garbles.

In the literature available on queueing theory, various abstract models have been developed, depending upon the number of information unit classes, the number of priorities, and the queue discipline (that is, the order in which incoming units are acted upon at the stage). This paper considers three situations which are pertinent to the application of queueing theory to the problem of volume of information flow into a typical stage of the system:

Situation A: one class of units; one priority; units acted upon on a first-come, first-served basis;

Situation B: several classes of units; one priority; units acted upon on a first-come, first-served basis;

Situation C: several classes of units; several priorities; units acted upon in order of relative priority, but within a priority units are acted upon on a first-come, first-served basis.

Further explanations of these situations appear in the sections devoted to each.

The format to be employed in the sections devoted to Situations A, B, and C is as follows: results of standard queueing theory are stated; an interpretation is made of results to the situation being considered; a few measures of effectiveness are proposed and discussed.

SECTION 4

SITUATION A

The situation to be investigated here is one in which all arriving information units have the same priority and are of the same class. Since timely action upon units is desirable, it seems reasonable to act upon the units in the order received; more formally, the queue discipline is first-come, first-served. (Incidentally, this situation represents the basic queueing model.)

As noted earlier, the format of this section (and the following two sections) is as follows: results of standard queueing theory are stated; an interpretation is made of results to the situation under consideration; a few measures of effectiveness are proposed and discussed.

Negative exponential arrival time and action-time distributions are assumed. Consult Appendix A for the analytic form of these distributions.

Some basic notation is necessary at this point. Let

λ = the mean arrival rate of information units at a system stage; $\lambda > 0$;

μ = the mean action rate upon units at a system stage; $\mu > 0$;

$P_n(t)$ = the probability that, at a time t, the random variable representing the total number of information units at the stage (including the unit being acted upon and those in queue, if any), takes on the value n; n = 0, 1, 2, ...

The reader is referred to Appendix A for a probabilistic development[3] of the following basic differential-difference equation which describes the time rate of change of the probability that, at a time t, there are n information units at the stage:

$$(4\text{-}1) \qquad \frac{dP_n(t)}{dt} = \lambda P_{n-1}(t) + \mu P_{n+1}(t) - (\lambda+\mu)P_n(t)$$

The solution of (4-1) is composed of two parts. Mathematically speaking, the two parts are the homogeneous and non-homogeneous solutions; operationally speaking, these represent, respectively, the time-independent (or steady-state) solution and the time-dependent (or transient) solution.

If $P_n(t)$ does not change with time, then $dP_n(t)/dt = 0$ for all t, and the right side of (4-1) represents a set of simultaneous equations, the solutions of which yield the steady-state probabilities that n units are at the stage. These steady-state probabilities will be denoted by p_n. For this case, the solution of (4-1) is given in standard queueing texts (such as Morse [1]) by

$$(4\text{-}2) \qquad p_n = \left[\frac{\lambda}{\mu}\right]^n p_o \qquad [n = 0,1,2,\ldots]$$

---

[3] Appendix A employs modern probability notation in the development of some basic queueing concepts. The random variables involved are clearly defined. In the main body of this paper, however, standard queueing notation is employed for ease in reading, despite the fact that it often obscures the appropriate random variable. In an effort to compensate, in part, for this notational deficiency, the particular random variable under discussion will always be designated explicitly in the text.

An analysis will now be made of (4-2) in terms of its application to the situation under discussion.

Note first that, for non-trivial applications, there must be some positive probability of the stage being idle (that is, no units in queue and none being acted upon): $p_o > 0$. An explicit determination will be given later.

Also, let $\rho \equiv (\lambda/\mu)$. The dimensions of ρ are arrivals/actions, and clearly $\rho > 0$. Now if $\rho > 1$, this means that there are (on the average) more units arriving at the stage than there are units being acted upon. This situation will cause a backlog of units waiting in queue, and (4-2) indicates that higher probabilities will exist for greater numbers of units at the stage. This is obviously undesirable, so at this point, ρ will be restricted to values ≤ 1.

Observe now the situation if $\rho = 1$. In this case, there is (on the average) one unit arriving for each unit being acted upon; in other words, the action-taker can (on the average) "just about" cope with arriving units. This in itself is not bad, until it is realized that (4-2) indicates that equal probabilities exist for any number of units at the stage; such a situation, in which there is equal probability of the stage being idle and of, say, one hundred units being at the stage, is also unsatisfactory. To obviate this difficulty, ρ is now restricted to values less than unity. In other words, $\lambda < \mu$.

Consider the interpretation of $\rho < 1$. It means that (on the average) information units will arrive at a lower rate than the rate at which they are being acted upon. Now (4-2) indicates that, for this case, the highest

probability exists for p_o: the stage will be idle part of the time. While this situation can be inefficient in a commerical operation, it is not as inefficient in a command and control system as it might appear to be when the following facts are considered: the human action-taker will have various monitoring/supervisory duties; he will also be generating information units for transmission to other stages of the system; and he must assimilate other information on which he is not required to take the positive action envisioned in this model, but on which he depends in order to take action. Thus, it can be said that the idle periods predicted by the classical queueing model do not necessarily mean that the action-taker will be free of any duties.

With ρ now restricted to the open interval (0,1), p_o may now be determined explicitly. Since $\sum_n p_n = 1$, it is evident that

(4-3) $$p_o = 1 - \rho \text{, and}$$

(4-4) $$p_n = \rho^n (1 - \rho) \qquad [\, n = 0,1,\ldots \,],$$

and $p_i < p_j$ for $j < i \leq n$.

Since probabilities may now be calculated for any number of units being at the stage, it is of interest to determine the expected value of the number of units at the stage. This expected value, denoted by $\bar{N}$, is determined by standard procedures:

(4-5) $$E[N] \equiv \bar{N} \equiv \sum_{n=0}^{\infty} n p_n = \frac{\rho}{1 - \rho} .$$

Similarly, if there are n units at the stage, (n - 1) of them will be in queue, and the expected value of the number of units in queue, denoted by $\bar{N}_q$, is given by

(4-6) $$E[N-1] \equiv \bar{N}_q \equiv \sum_{n=1}^{\infty} (n-1) p_n = \frac{\rho^2}{1 - \rho} .$$

At this point, another random phenomenon must be considered: the phenomenon of arriving units, formed in queue, "waiting" for action to be taken and completed upon them. The random variable describing this phenomenon is denoted by W and represents the "waiting time" of units at the stage. That W is indeed a random variable describing a random phenomenon is evident from the following argument: if there are n units at the stage, including the unit on which action is currently being taken, then the total waiting time of the n^{th} unit (that is, the time which the n^{th} unit must wait in queue plus the action time when it finally is acted upon) is the sum of the action times of all units ahead of the n^{th} unit in the queue plus the action time on the n^{th} unit itself, since a unit is acted upon as soon as it reaches the action-taker. Because the "presence" of units at the stage is a random phenomenon described by the random variable N, and because the "action" upon units is a random phenomenon described by the variable T_r as defined in Appendix A, the random phenomenon of "waiting" described by the random variable W is established.

The notion of expectation allows a determination of the expected value of waiting time at the stage, denoted by $\bar{W}$. The summation of action times may be equated to a product of the expected number of units at the stage and the expected value of action times, provided that such expected values exist and that the random variables involved are independent. Independence is a basic assumption in this model (see Appendix A), and the expected values exist; see (4-5) above and (A-8) in Appendix A. Therefore,

$$E[W] \equiv \bar{W} = E[N \cdot T_r] = E[N] \cdot E[T_r] = \frac{\bar{N}}{\mu} \quad . \tag{4-7}$$

Since $\bar{N}_q = \rho \bar{N}$ from (4-5) and (4-6), substitution into the above relation shows that $\bar{W} = \bar{N}_q/\lambda$. Because of the similar form of the two expressions for $\bar{W}$, $\bar{W}$ is considered equal to $\bar{W}_q$, where $\bar{W}_q$ is the expected value of the random variable W_q representing the "waiting time, before action," of units in queue. (Of course, we could have determined $\bar{W}_q$ by a method similar to that employed above in establishing $\bar{W}$, leading to the same result.)

Having determined various probabilities and expected values, we now employ them as measures of effectiveness.

For example, the probability that the stage is busy (i.e., not idle) is apparent from (4-3) as equaling ρ. Thus, ρ may be used as a measure of effectiveness, and in this context it is often called the utilization factor. Thus, a military commander of a command and control system could specify, as a performance standard, a lower bound on the probability of a particular stage being busy. (Alternatively, he could specify an upper bound on the probability that the stage is idle.)

Also, the expected value of the number of units in queue awaiting action, $\bar{N}_q$, determined from (4-6), may be used as a measure of effectiveness. A military commander might specify, as a performance standard, an upper bound on the expected number of units awaiting action. Denote this upper bound by $\bar{N}_q^*$. Then an upper bound on ρ is automatically determined by this specification. Call this upper bound ρ_N. Equation (4-6) indicates that ρ_N is given by

$$\rho_N = \frac{-\bar{N}_q^* \pm \left[\left(\bar{N}_q^*\right)^2 + 4\bar{N}_q^*\right]^{1/2}}{2} \tag{4-8}$$

where only the positive root is applicable.

On the other hand, the expected value of the waiting time of units at the stage, denoted by $\bar{W}$ and determined by (4-7), may be used as a measure of effectiveness. A military commander might specify, as a performance standard, an upper bound on $\bar{W}$. Denote this upper bound by $\bar{W}^*$. Then an upper bound on ρ is automatically specified; denote it by ρ_W. Equations (4-5) and (4-7) indicate that ρ_W is given by

$$\rho_W = \frac{\bar{W}^*\mu}{1 + \bar{W}^*\mu} \quad . \tag{4-9}$$

Finally, p_n itself may be used as a measure of effectiveness. A military commander might specify, as a performance standard, an upper bound on a particular p_n; call this upper bound p_n^*. Again, an upper bound on ρ is automatically specified. Denote this upper bound on ρ by ρ_p. Equation (4-4) indicates that ρ_p is given by the solution of

$$\rho^{n+1} - \rho^n + p_n^* = 0 \tag{4-10}$$

provided, of course, that zeros of this function exist in the open interval (0,1) to which ρ is restricted.

Other measures of effectivnesss may be established, and Rawdin [2] enumerates several more. Those indicated above are considered most appropriate to the application of queueing theory in this paper.

Observe that all measures of effectiveness noted herein involve the factor ρ. Recall that $\rho = \lambda/\mu$. Let us analyze ρ a bit further at this point. It is reasonable to suppose that psychological testing and human factors research may produce good estimates of the factor μ., the mean rate at which a human being reacts under certain conditions

of stress and employment. If this is the case, then ρ is really a function of λ, the mean rate at which information units arrive at a stage of the system. Values of λ may be obtained by an analysis of the information-generating orders, instructions, and procedures of the military commander of the command and control system. Thus, the point is made that, given good estimates of μ, any specified values of ρ, p_o, p_n, $\bar{N}_q$, or $\bar{W}$ imposed by a military commander as performance standards must be consistent with λ, a parameter which is largely under the control of the commander himself!

Assuming now that μ is reasonably well-determined, let us examine the influence on the various measures of effectiveness of variations in ρ (and, through ρ, variations in λ). As ρ increases, the probability of the stage being idle decreases; but the expected number of units at the stage, the expected number of units in queue, and the expected waiting time will all increase; also the p_n will tend to become equal for all n. Thus, the intuitively appealing effort to obtain greater utilization of manpower (through a reduction in the probability that the manpower is idle) forces an acceptance of greater numbers of units at the stage and in queue, as well as of greater expected delay in the action-taking upon information units. Conversely, shorter expected waiting times, and lower expected numbers of units at the stage and in queue mean higher probabilities that the stage is idle. With a known value of μ, all specified values (or acceptable values) of measures of effectiveness which may be imposed by a military commander force him to maintain a consistent λ by means of his own information-generating requirements.

Let us examine an alternate situation. Suppose that a military commander has performed an analysis of his information-generating requirements, and has determined a value of λ. Suppose further that, in his

opinion, he cannot reduce this value of λ for a particular stage in his system. In other words, he feels that there must be positive human action-taking on an amount of information at a particular stage in his system, and he is unwilling to sacrifice any of this information to automation; nor will he eliminate the requirement for any part of it. Finally, suppose that he feels that intensive training will improve performance at the stage; this is another intuitively appealing course of action which, on occasion, will actually effect improved performance. Specifically, if training can increase the value of μ, with λ constant (in other words, if training can actually increase the average number of action-takings per interval of time, with the average arrival rate of information units unchanged), then ρ will decrease, and thus the expected numbers of units at the stage and in queue will decrease, as well as expected waiting time (although the probability of the stage being idle will increase).

With the situation described above (where μ is increased through training while λ is constant, and lower values of certain measures of effectiveness are obtained), there might be a temptation to increase λ; in other words, with higher system performance, it might seem desirable to increase the amount of information arriving for action at the stage per interval of time. Such a course of action may be considered unwise when the effect of fatigue is considered. Although the action-taker may possess a high capacity for acting upon units, the length of time in which he can <u>sustain</u> this capacity is an important factor. If fatigue should set in, the value of μ will certainly decrease, possibly by a large amount. Differentiation of ρ with respect to μ indicates the sensitivity of ρ to small changes in μ. The point is, that if a

military commander increases λ , on the basis of a high value of μ achieved at the stage through training, he must accept the risk of extremely poor performance at the stage should fatigue overcome the action-taker (thereby lowering the value of μ , increasing the value of ρ , and increasing expected numbers of units at the stage and in queue as well as expected waiting time); the effect of such degraded performance at the stage must be considered, since such degradation might take place at a critical time in actual operations.

Then too, the intensive training conducted at the stage might actually result in "over-training", with a resulting value of μ lower than that which prevailed earlier. With a constant value of λ , a decrease in μ through "over-training" will again degrade performance at the stage.

Thus, it is evident that the downward fluctuations in μ (which may be quite sudden and of large magnitude under certain conditions) dictate that the value of λ be rather conservative in order to prevent sudden degradation of performance at the stage; specifically, the value of λ should be related to a value of μ which is not the "peak" value obtained through proper training, but rather one which reflects the possibility of fatigue and "over-training."

Hence, the conclusions are drawn that training <u>alone</u> is not the answer to improved performance at a stage of the system, but rather a thorough analysis of information-generating requirements is necessary.

The conclusions drawn from this analysis of the steady-state queueing model for Situation A are summarized in Section 7, which contains the conclusions of this entire paper. (It may come as no surprise that the conclusions for Situations A, B, and C are essentially the same.)

This section will be concluded with a comment and an explanation why the time-dependent (i.e., transient) solution to (4-1) will not be presented here. The justification for not pursuing an application to the problem of this paper of the transient solution of (4-1) lies in the argument to follow, based on a comment by Saaty.[4]

The reasoning proceeds as follows: a system stage may be expected to vary in the tempo of its operations, depending on the particular real-world or simulated situation; for sustained periods, the stage may be very active, moderately active, or relatively idle; no attempt will be made to define these evanescent terms; instead, the point to be made is that the operating time of the stage may be partitioned into various phases, each phase representing a certain sustained tempo of operations; to each phase, the steady-state solution may be applied (with appropriate values of λ and μ employed.)

This argument is considered to have merit, and, therefore, only steady-state solutions will be presented for the various situations discussed throughout this paper.

---

[4] Saaty [3], page 357.

SECTION 5

SITUATION B

The situation to be investigated here is the logical generalization of Situation A. Several classes of information units are conceived as arriving at a stage of the system at different rates; and corresponding to these several classes are different action rates.

As before, negative exponential arrival time and action time distributions are assumed, and only a steady-state solution will be analyzed.

The results of Ancker and Gafarian [5] are to be presented and interpreted for Situation B. Their notation has been modified, in part, to parallel that of Situation A.

Consider m classes of information units, each with positive mean arrival rate λ_j (j = 1, 2, ...,m), and each with positive mean action rate μ_j (j = 1, 2,....., m). The random phenomenon in this case is the "presence" of information units at the stage, and the random variable describing this random phenomenon is N, the "number of units of all classes in queue". (Note that N is the number of units in queue, in contrast to N of situation A.)

The basic notation is as follows:

$_jP_n$ = the steady-state probability that there are n units in queue at the stage, and that a unit of class j is being acted upon;

$P_n = \sum_{j=1}^{m} {}_jP_n$ = the steady-state probability that there are units in queue at the stage, and that a unit of some class is being acted upon;

$_oP_o$ = the steady-state probability that there are no units in queue, and that no unit is being acted upon (that is, the probability that the stage is idle);

$$ {}_jP = \sum_{n=0}^{\infty} {}_jP_n = $$ the steady-state probability that a unit of class j is being acted upon;

$$ \rho_j \equiv \lambda_j/\mu_j $$ (a generalization of ρ in Situation A).

The basic differential-difference equations are not reproduced; the resultant solutions follow:

(5-1) $$ {}_oP_o = 1 - \beta \text{ , where } \beta = \sum_{j=1}^{m} \rho_j \text{ ;} $$

(5-2) $$ {}_jP_o = \frac{\lambda_j}{(\lambda+\mu_j)(1-\alpha_1)} ({}_oP_o) $$

where $$ \lambda = \sum_j \lambda_j $$

and $$ \alpha_n = \sum_j \frac{\lambda_j}{(\lambda+\mu_j)^n}; $$

(5-3) $$ {}_jP_n = \frac{\lambda}{\lambda+\mu_j} {}_jP_{n-1} + \frac{\lambda_j}{\lambda+\mu_j} P_n \qquad [n \geq 1] , $$

where P_n is given by (5-5);

(5-4) $$ P_o = \sum_j {}_jP_o = \frac{\alpha_1}{1-\alpha_1} {}_oP_o \text{ ;} $$

(5-5) $$ P_n = \sum_j {}_jP_n = \begin{cases} \frac{1}{1-\alpha_1}\left[\frac{\lambda^n\alpha_{n+1}}{1-\alpha_1} {}_oP_o + \sum_{i=1}^{n-1} \lambda^i\alpha_{i+1} P_{n-i}\right] & [n \geq 2] \\ \frac{\lambda\alpha_2}{(1-\alpha_1)^2} {}_oP_o & [n = 1] \end{cases} $$

(5-6) $$ {}_jP = \sum_{n=0}^{\infty} {}_jP_n = \rho_j \text{ .} $$

The expected number of units in queue, denoted as before by $\bar{N}_q$, is given by

$$E\left[N\right] \equiv \bar{N}_q = \frac{\lambda \sum_j \rho_j/\mu_j}{1 - \beta} \quad , \tag{5-7}$$

while the expected number of units at the stage, denoted as before by $\bar{N}$, is, by the definition of N in this section,

$$E\left[N+1\right] \equiv \bar{N} = \bar{N}_q + \beta \quad . \tag{5-8}$$

Consider now another random phenomenon: the "waiting" of units at the stage; and define W as the random variable describing this random phenomenon, in units of "waiting time." The expected value of the waiting time, denoted as before by $\bar{W}$, is given by

$$E\left[W\right] \equiv \bar{W} = \frac{\sum_j \rho_j/\mu_j}{1 - \beta} = \frac{\bar{N}_q}{\lambda} = \frac{\bar{N} - \beta}{\lambda} \quad . \tag{5-9}$$

From an inspection of the various recursion relations given above which involve ${}_oP_o$, it is evident that, for non-trivial interpretations, there must be some positive probability that the stage is idle: ${}_oP_o > 0$. In addition, a probability of unity that the stage is idle makes any further interpretation unnecessary, as well as impossible[5]. Hence,

$$0 < \beta = \sum_j \rho_j < 1 \quad , \tag{5-10}$$

---

[5] Mathematically, such an occurrence is impossible, since ${}_oP_o = 1$ implies $\beta = 0$; but $\beta = \sum_j \rho_j$, and $\rho_j > 0$.

which in turn demands $\lambda_j < \mu_j$ for all j.

Since ${}_oP_o$ must be greater than zero, a further inspection of the above relations indicates that ${}_jP_n > 0$; and since $P_n = \sum_j {}_jP_n$, $P_n > {}_jP_n$ for all n and j.

The next point to be examined is the relation of ${}_oP_o$ to the other probabilities. Although ${}_oP_o$ must be greater than zero, there is for example no reason why this probability should be greater than all others (as was the case in Situation A). Specifically, let us require that P_o exceed ${}_oP_o$; in words, we are requiring that the probability that a unit of some class is being acted upon and that there are no units in queue awaiting action exceed the probability that the stage is idle; this is considered an efficient mode of operation for the stage. Equations (5-1) and (5-4) show that this constraint implies that $\alpha_1 > 1/2$; recalling the definition of α_n given in (5-2), we see that $0 < \alpha_1 < 1$. Combining these inequalities, we find

$$(5\text{-}11) \qquad \frac{1}{2} < \alpha_1 < 1 .$$

Similarly, we introduce the further requirement that ${}_jP_o > {}_oP_o$; in words, we are demanding that the probability that a unit of class j is being acted upon and that there are no units in queue awaiting action exceed the probability that the stage is idle. Using (5-1) and (5-2) and summing over j, we find that, for this constraint,

$$(5\text{-}12) \qquad \alpha_1 > \frac{m}{m+1}$$

where m is the number of distinct information unit classes. The lower bound on α_1 from this relation is seen to increase as m increases. If

(5-2) and (5-4) are differentiated with respect to α_1, we find that ${}_jp_o$ and p_o increase with α_1, hence with m. Also, (5-1) shows that the probability of the stage being idle decreases as m increases, since β increases with m.

Also, let us require that the probabilities of greater numbers of units in queue awaiting action shall decrease as n increases; in symbols, $p_n < p_{n-1}$ for $n \geqslant 1$. For example, to insure $p_1 < p_o$, employ (5-4), (5-5) and (5-11) and recall that $\lambda = \sum_j \lambda_j$; from these equations, we solve for α_2 and find that

$$\alpha_2 < \frac{\alpha_1[1-\alpha_1]}{\lambda} \tag{5-13}$$

Similarly, repeated use of (5-5) for $n \geqslant 2$ will produce upper bounds for the related α_n.

Note further that, as m increases, (5-7), (5-8), and (5-9) indicate greater expected numbers of units in queue, units at the stage, and greater expected waiting time at the stage, respectively. Thus a larger number of distinct information unit classes tends to delay action-taking at the stage.

Utilizing analogous measures of effectiveness as before, namely β (instead of the single ρ of Situation A), ${}_op_o$ (instead of p_o as in Situation A), $\bar{N}_q$, $\bar{W}$, and p_n, it is again apparent that a specification of some upper and/or lower bounds on these measures of effectiveness as performance standards will automatically determine a critical value of β; and, recalling that $\beta = \sum \rho_j$, it is again seen that, if action rates μ_j are reasonably well-known (or capable of being determined), the information-generating rates λ_j must be consistent with the specified

values of the measures of effectiveness.

Thus, the conclusions drawn from a queueing analysis of Situation B are essentially those of Situation A, with one important addition: namely, that the introduction of several classes of units (each with its own mean arrival rate λ and mean action rate μ) tends to increase the expected numbers of units at the stage and in queue, as well as to increase the expected waiting time.

Section 7 contains a summary of conclusions.

SECTION 6

SITUATION C

In the situation considered here, information units arrive at the stage with priority assignments which determine the relative order in which they will be acted upon by the human action-taker. For this reason, Situation C most nearly represents an actual mode of operation of a stage of a command and control system.

Priorities of information units are designated herein by alphabetical letters; a unit of priority A has the highest priority, and units of priority B, C, D,...are of successively lower priority. Within a priority, all units are acted upon on a first-come, first-served basis; but no unit of a lower priority will be acted upon until all units of higher priority have been acted upon.

At this point, we must establish the procedure to be followed if, when a unit of some priority is being acted upon, a unit of higher priority arrives at the stage. In order to reflect standard military procedures, we establish the rule that the higher priority unit will displace the lower priority unit; in other words, the higher priority unit will preempt the action-time of the lower priority unit.

Assuming that a lower priority unit has been preempted, and that all higher priority units which have arrived since the preemption have been acted upon, we further establish that the action-taker will resume his (uncompleted) action upon this unit.

Our preliminary discussion of Situation C concludes with a comment on the classes of information units. Since information units are acted

upon according to relative priority, but within a priority all units are handled on a first-come, first-served basis, it is evident that the results of Situation B apply when considering a particular priority. As we are considering the effect of priority assignment in this Section, we assume that, within each priority, the total number of units have an overall "effective" arrival rate λ and an overall "effective" action rate μ .

White and Christie [6] , Stephan [7] and Jaiswal [8] have considered the two priority cases; some of their findings will be presented here, with notation modified to parallel that employed earlier in this paper. The R-priority case (where $R > 2$) has been examined in part by Heathcote [9] , with more complex results. The two-priority case is sufficient to illustrate the principles.

Consider two priorities of information units, with positive arrival rates λ_A and λ_B, and positive action rates μ_A and μ_B. (Note that, under the assumption made earlier, these rates are really "effective" overall rates which consolidate the effect of unit classes within a priority assignment.) The subscript letter A refers to the highest priority unit. As in previous situations, we assume negative exponential distributions for arrival times and action times, and only the steady-state results are to be presented. The random phenomena here are the "presence" of units of priority A and B; the associated random variables are the"numbers" of each priority present at the stage.

The basic notation scheme is as follows:

p_{n_A} = the steady-state probability that the random variable N_A (representing the number of units of priority A at the stage) takes on the value n_A; $n_A = 0, 1, 2, \ldots$

p_{n_B} = the steady-state probability that the random variable N_B (representing the number of units of priority B at the stage) takes on the value n_B; $n_B = 0, 1, 2, \ldots$

p_o = the steady-stage probability that the stage is idle (that is, the probability that $N_A = N_B \equiv 0$).

$$\rho_A \equiv \lambda_A / \mu_A$$

$$\rho_B \equiv \lambda_B / \mu_B.$$

The steady-state probability that the stage is idle is given by

(6-1) $$p_o = 1 - \gamma \quad ,$$

where $\gamma = \rho_A + \rho_B$;

compare γ with β of Situation B and ρ of Situation A.

Since priority A units preempt priority B units immediately upon arrival, and since priority A units are handled on a first-come, first-served basis, the results of Section 4 for Situation A apply directly to priority A units. With notation altered slightly to direct attention to the priority letter, some results given in the earlier case are repeated here:

(6-2) $$p_{n_A} = \left[\rho_A\right]^n (1 - \rho_A) \quad ;$$

(6-3) $$E\left[N_A\right] \equiv \bar{N}_A = \frac{\rho_A}{1 - \rho_A} \quad ;$$

(6-4) $$E\left[W_A\right] \equiv \bar{W}_A = \frac{\rho_A}{\mu_A\left[1 - \rho_A\right]} \quad ,$$

where W_A is the random variable, "waiting time" for priority A units, associated with the random phenomenon, the "waiting" for action of priority A units.

For priority B units,

(6-5) $$E\left[N_B\right] \equiv \bar{N}_B = \frac{\rho_B}{1 - \gamma}\left[1 + \frac{\mu_B}{\mu_A}\bar{N}_A\right]$$

$$(6\text{-}6) \qquad E\left[W_B\right] \equiv \bar{W}_B = \frac{1}{1-\gamma}\left[\frac{\gamma}{\mu_B} + \bar{W}_A\right] \quad ,$$

where W_B is similarly defined. Note that W_B includes the effect of any preemptions by priority A units, as well as that of other priority B units which have arrived earlier.

The expressions for p_{n_B} are to be found in Stephan [7]; they are not germane to the discussion here.

From the results just presented, note first that reasoning similar to that employed in Situation A and Situation B indicates that $p_o > 0$, and that γ is thus restricted to the open interval (0,1); hence $\lambda_A < \mu_A$ and $\lambda_B < \mu_B$.

Observe also that the assignment of two priorities decreases the probability that the stage is idle (assuming that priority A units have a ρ_A which is identical with ρ of Situation A). Heathcote [9] has shown that in the general case, for R priorities, p_o decreases as the number of priorities increases.

As noted earlier, the assignment of priorities has no effect on the expected number of highest priority units at the stage, or on the expected waiting time of these units.

On the other hand, an assignment of priorities implies that, for example, the expected waiting time of the lower priority units is greater than that of the higher priority units; Equations (6-4) and (6-6) indicate that this will always be so.

If there is a specification on the maximum acceptable expected waiting time of priority A units, an upper bound is automatically determined for ρ_A; see (4-9) of Section 4, and the derivation thereof.

Equation (6-6), above, indicates that, given a specified $\bar{W}_A$, a particular $\bar{W}_B$ is automatically determined. Thus, if such a value of $\bar{W}_B$ is considered excessive, an adjustment must be made in γ, and through γ, an adjustment in ρ_B (assuming ρ_A to be held constant.)

The expected waiting time of various priority units at the stage is considered a major measure of effectiveness; other measures may be investigated of course. Regardless of which one is examined, we always end up with critical values for the various ρ; and if values of μ are reasonably well-known, we ultimately have critical values for the various λ which must be consistent with the information-generating instructions for the stage of the system.

Therefore, the conclusions for Situation C are essentially the same as for the previous situations; they are summarized in Section 7. For this particular situation, we conclude that the assignment of priorities tends to produce greater expected numbers of lower priority units at the stage, and greater expected waiting times for these units, the sole exception being units of the highest priority (which conform to the results of Situation A).

SECTION 7

CONCLUSIONS

The conclusions drawn from the application of queueing theory to the volume of information flow into a typical stage of a command and control system are summarized below. It should be recalled that: the parameter λ represents the mean arrival rate of information units (of a particular class and/or priority) at a stage; μ represents the mean action rate upon these units; ρ is the ratio of these quantities where $\rho = \lambda/\mu$.

The major conclusions are these:

(a) Psychological testing and other human factors research can produce a value for each μ.

(b) An analysis of a military commander's information-generating requirements -- information on which the military commander demands the positive human action-taking considered in this paper -- will provide a value for each λ.

(c) Any specifications on the performance of the human being at a system stage (such specifications being certain values of the various measures of effectiveness which may be used as performance standards) must be consistent with the various λ and μ values.

(d) Should there be incompatibility between these performance standards and the known values of λ and μ, remedial action must be taken.

(e) Remedial action must not be restricted to conducting more intensive training of personnel, for this course of action, alone, may

under certain circumstances lead to reduced values of μ ; and from the results of queueing theory, it has been demonstrated in Section 4 that lower values for μ (with constant values for λ) will tend to increase such values as expected numbers of units at the stage and in queue, and expected waiting time.

(f) The real solution to an incompatibility that may exist between various performance standards and known values of λ and μ , lies in a thorough analysis of a military commander's information-generating requirements; such an analysis should be made with a view to automating, consolidating, or even eliminating some of the information generated; the effect of this analysis would be to reduce values of λ , which in turn would tend to achieve higher efficiency at the stage with respect to those information units which must receive timely, positive human action-taking.

SECTION 8

RECOMMENDATIONS

The results and conclusions of an application of queueing theory to the problem of volume of information flow into a stage of a command and control system are contained in Sections 4 through 7. Specific recommendations follow for implementing these results and conclusions, in order to achieve quantitative guidelines for an analysis of the information requirements for each stage of the system.

(a) On-going operations analyses of the commander's information-generating requirements should be conducted in order to determine mean arrival rates of various classes and/or priorities of information for various real-world and/or simulated environments in which the system is to function, at each command level of the system.

(b) Psychological testing and human factors research should be conducted to determine, for the various operating environments and for the various classes/priorities of information, mean reaction times of the human beings who have the decision-making authority and responsibility at each stage of the system.

(c) Performance standards for each stage of the system should be established by the commander; these standards are specified values of the measures of effectiveness proposed.

(d) Numerical values of the measures of effectiveness proposed in Sections 4 through 6 should be computed, using the data obtained from (a) and (b), above.

(e) The computed values (from (d), above) should be compared

with the performance standards (from (c), above), and incompatibilities should be noted.

(f) Any incompatibility between performance standards and numerical values of measures of effectiveness should be resolved by a further analysis of the information-generating requirements in an attempt to consolidate, automate, or even eliminate some information.

(g) System training alone should not be considered a remedy for a failure to meet performance standards, since it has been shown that training (alone) may under certain circumstances degrade performance even further.

BIBLIOGRAPHY

1. Philip M. Morse, "Queues, Inventories and Maintenance," Wiley, New York, 1958.

2. E. Rawdin, "A New Measure of Effectiveness for Queueing Problems," Opns. Res. 8, 278-280 (1960).

3. Thomas L. Saaty, "Mathematical Methods of Operations Research," McGraw-Hill, New York, 1959.

4. Thomas L. Saaty, "Elements of Queueing Theory," McGraw-Hill, New York, 1961.

5. C. J. Ancker, Jr., and A. V. Gafarian, "Queueing With Multiple Poisson Inputs and Exponential Service Times," Opns. Res. 9, 321-327 (1961).

6. Harrison White and Lee S. Christie, "Queueing With Preemptive Priorities or With Breakdown," Opns. Res. 6, 79-95 (1958).

7. Frederick F. Stephan, "Two Queues Under Preemptive Priority With Poisson Arrival and Service Rates," Opns. Res. 6, 399-418 (1958)

8. N. K. Jaiswal, "Preemptive Resume Priority Queue," Opns. Res. 9, 732-742 (1961).

9. C. R. Heathcote, "A Simple Queue With Several Preemptive Priority Classes," Opns. Res. 8, 630-638 (1960).

10. Emanuel Parzen, "Modern Probability Theory and Its Applications," Wiley, New York, 1960.

APPENDIX A

AN AXIOMATIC DEVELOPMENT OF THE CLASSIC QUEUEING EQUATION

The material in this Appendix describes certain aspects of basic queueing theory in language which is familiar to those who have an understanding of probability based on the modern axiomatic treatment of probability theory.

The general purpose of this Appendix is to provide a simple guide by which the reader may pass from the usual notation of modern probability theory (as found, for example, in Parzen [10]) to the usual notation of queueing theory. The specific purpose of this Appendix is to develop, in an axiomatic way, the basic differential-difference equation (4-1) of Section 4.

Queues are relatively complicated random phenomena. Accordingly, the description of these phenomena in the symbology of modern probability theory is somewhat cumbersome and awkward. The notation which has evolved for theoretical descriptions of queue phenomena does not always make clear the particular random variable(s) under discussion.

To begin with, consider the arrival of an information unit at a stage of the command and control system. The "arrival" is assumed to be a random phenomenon. The "time elapsed between the arrivals of two successive units at the stage" is the random variable associated with the phenomenon.

Let T_a be this random variable; obviously, T_a is non-negative. Also, let A_{T_a} be the cumulative probability distribution function

associated with T_a. That is,

(A-1) $$A_{T_a}(t_a) \equiv P\left[T_a \leqslant t_a\right]$$

where $P\left[T_a \leqslant t_a\right]$ is read: the probability that the random variable T_a takes on a value equal to or less than a specified value t_a. The designation of this function by the capital letter A serves to associate it with the random phenomenon, "arrivals," for which the associated random variable is T_a (representing the "time elapsed between the arrivals of two successive units at a stage of the system").

A further basic assumption to be made at this point is that the random variable T_a has a negative exponential distribution.

Standard queueing treatises, such as Saaty [4] et al, contain a development of the relations to follow.

The cumulative probability distribution function for the negative exponential distribution is given by

(A-2) $$A_{T_a}(t_a) = \begin{cases} 1 - e^{-\lambda t_a} & \left[t_a \geqslant 0\right] \\ 0 & \text{elsewhere,} \end{cases}$$

where λ is a positive constant representing the mean arrival rate of units at a stage; compare (A-4). A dimensional analysis of the exponential term verifies that λ must have the dimensions of $(\text{time})^{-1}$, i.e., a rate.

The probability density function associated with the random variable T_a, denoted by a_{T_a}, is obtained by differentiation of (A-2):

(A-3) $$a_{T_a}(t_a) \equiv \frac{dA_{T_a}(t_a)}{dt_a} = \begin{cases} \lambda e^{-\lambda t_a} & \left[t_a \geqslant 0\right] \\ 0 & \text{elsewhere.} \end{cases}$$

The expected value of the time elapsed between the arrival of two successive units at a stage, denoted by $\bar{T}_a$, is given by usual probability methods:

$$(A\text{-}4) \qquad E\left[T_a\right] \equiv \bar{T}_a \equiv \int_{-\infty}^{+\infty} t_a \cdot a_{T_a}(t_a)\,dt_a = \frac{1}{\lambda} ,$$

and $\bar{T}_a$ is occasionally referred to as the mean time between arrivals.

Consider next the "actions" taken upon information units at a stage. The "actions" are assumed to be random phenomena, and the "time elapsed between the inception of action upon two successive units" is a random variable associated with the random phenomena," "actions". Let t_r denote the value of this random variable; t_r is clearly non-negative. Let R_{T_r}, be the distribution function of the random variable T_r; that is,

$$(A\text{-}5) \qquad R_{T_r}(t_r) \equiv P\left[T_r \leqslant t_r\right]$$

where the designation of the cumulative probability distribution function by the capital letter R serves to associate it with the random phenomena, "actions" (or "reactions," if you will), for which the associated random variable is T_r (representing the "time elapsed between the inception of action upon two successive units at the stage").

An assumption that T_r has a negative exponential distribution leads to an analytic expression for $R_{T_r}(t_r)$:

$$(A\text{-}6) \qquad R_{T_r}(t_r) = \begin{cases} 1 - e^{-\mu t_r} & \left[t_r \geqslant 0\right] \\ 0 & \text{elsewhere} , \end{cases}$$

where μ is a positive constant representing the mean rate at which units are acted upon at the stage; compare (A-8). A dimensional analysis of the exponential term verifies that μ must have the dimensions of $(\text{time})^{-1}$; i.e., a rate.

The probability density function associated with the random variable T_r, denoted by r_{T_r}, is obtained by differentiation of (A-6):

$$(A\text{-}7) \quad r_{T_r}(t_r) \equiv \frac{dR_{T_r}(t_r)}{dt_r} = \begin{cases} \mu e^{-\mu t_r} & [t_r \geqslant 0] \\ 0 & \text{elsewhere} \end{cases}$$

The expected value of the random variable T_r, denoted by $\bar{T}_r$, is obtained in the usual manner:

$$(A\text{-}8) \quad E[T_r] \equiv \bar{T}_r \equiv \int_{-\infty}^{+\infty} t_r \cdot r_{T_r}(t_r) dt_r = \frac{1}{\mu} ,$$

and $\bar{T}_r$ may be referred to as the mean time between actions.

We introduce next another random phenomenon which is illustrated by performing (mentally) the following experiment. Observe a stage of the system as it operates. Immediately after a unit arrives at the stage being observed, start counting the number of units which arrive after that starting time. After a length of time t (where t is NOT a random variable, but rather some experimental time), cease counting and determine the total number of units which have arrived in that period of duration t; the total number will be either zero or a positive integer. The "arrivals" in the time t are random phenomena, and the "number of arrivals" in a time t, denoted by N, is an integral-valued random variable describing this random phenomena; N = 0, 1, 2,... .

Standard treatises on modern probability theory show that, under certain assumptions, the probability mass function of the random variable N, for an observation time of duration t (measured from the time of arrival of an arbitrary unit), is given by

$$a_N[n;t] = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \qquad [n = 0,1,2,\ldots] \tag{A-9}$$

which is recognized as the Poisson distribution.

Let us investigate still another random phenomenon; to illustrate it, perform another experiment. Observe a stage of the system as it operates in time. Select an arbitrary point in time (a "clock" time). Any one of the three following (exhaustive) conditions may be found to exist: (1) the action-taker is idle; (2) the action-taker is acting upon a unit, and there are no units in queue; or (3) the action-taker is acting upon a unit, and there is a queue of units awaiting action. Whatever condition is found, count and record the number of units at the stage, including the unit being acted upon, and the units in queue, if any.

Now the arrivals of units at the stage, and the actions taken upon them, are two independent random phenomena; therefore, the interaction of these two random phenomena produces another random phenomenon, which is the "state"of the stage at a time of observation t. Corresponding to this random phenomenon, "state", we establish a random variable N, representing the "number of units" at the stage at a time of observation t. (Note that t is <u>not</u> a random variable; it is a time of observation. Note also that the random variable N is not the same as the random variable in (A-9).)

In order to become more analytic (and less experimental), we wish to obtain some relations which will enable us to determine the probability that an observer will find, at some arbitrary observation time t, a total of N units at the stage. More formally, define

$P_N\left[n;t\right] \equiv$ the probability that, at an arbitrary time t, the random variable N (representing the number of units at the stage, including the unit being acted upon as well as those in queue awaiting action, if any) takes on the value n, where n = 0, 1, 2,...

$E_N(t) \equiv$ the state of the stage in which, at a time of observation t, there are N units at the stage, including the unit being acted upon and those units in queue awaiting action, if any.

Thus, $P_N\left[n;t\right] \equiv P\left[E_N(t)\right]$.

While the exact form of $P_N\left[n;t\right]$ is as yet unknown, we may employ it to obtain the basic differential-difference equation (4-1) of Section 4. To do this, we must examine how the state of the stage may change as time passes. Formally, we seek some relations describing the state of the stage at a time (t + dt), given the state at some arbitrary time t. As is usual in proceedings of this sort, the magnitude of the differential element of time, (dt), is never precisely defined; it is "large" enough to allow "some" things to happen, but "small" enough to prevent "too much" from occurring; a slippery notion, indeed! Nevertheless, it is quite useful.

To begin with, we seek the conditional probability that the stage is in state $E_N(t + dt)$, given that it was in state $E_{N-1}(t)$; in symbols, we want

$$P\left[E_N(t + dt)\middle|E_{N-1}(t)\right] \quad .$$

Now <u>one</u> way in which the stage can be in state $E_N(t + dt)$, given that it was in state $E_{N-1}(t)$, is the simultaneous occurrence of two independent

random phenomena: one unit arrives in (dt), and action is not completed on a unit in this interval of time. By the assumption of independence, the conditional probability sought is given by the product of the probability that there were (N-1) units at the stage at time t, the probability that one unit arrived in the interval (dt), and the probability that action was not completed upon a unit in the interval (dt):

$$(A\text{-}10)\quad P\left[E_N(t+dt)\middle|E_{N-1}(t)\right] = P_{N-1}\left[(n-1);t\right]\cdot\left(a_N\left[1;(dt)\right]\right)\cdot\left(e^{-\mu(dt)}\right);$$

in this relation, (A-9) is used to determine the probability of one arrival in (dt), Also (A-6) has been used to determine the probability that action has not been completed on a unit in (dt); an explanation of this is in order. The probability that the random variable T_r takes on values equal to or less than a specified value t_r is given by (A-6). In the situation under discussion here, identify the element (dt) with the specified value t_r; then

$$P\left[T_r \leq (dt)\right] = 1 - e^{-\mu(dt)},$$

and so the probability that action has <u>not</u> been completed on a unit in the interval (dt) is equivalent to the probability that the random variable T_r exceeds (dt); in symbols,

$$P\left[T_r > (dt)\right] = e^{-\mu(dt)};$$

hence the last term on the right side of (A-10).

Inserting the probability of the arrival of one unit in the interval (dt) from (A-9), into (A-10), we find that

$$P\left[E_N(t+dt)\middle|E_{N-1}(t)\right] = P_{N-1}\left[(n-1);t\right]\cdot\left(\lambda(dt)e^{-\lambda(dt)}\right)\left(e^{-\mu(dt)}\right).$$

At this point, we establish as a general rule in the development to follow the convention that exponential terms in (dt) will be expanded, and that

terms involving orders of (dt) greater than the first will be discarded. Rationalization for this procedure lies in the fact that, since (dt) is "small", terms involving $(dt)^2$, $(dt)^3$, etc., are even "smaller", and do not contribute to, or detract from, any expression "substantially". Employing this convention, we find, at last, that

$$P\left[E_N(t + dt) \middle| E_{N-1}(t)\right] \approx \left[\lambda(dt)\right] \cdot P_{N-1}\left[(n-1);t\right] \quad . \tag{A-11}$$

Now another way in which the state can be in state $E_N(t + dt)$, given that it was in state $E_{N-1}(t)$, is for k arrivals to have occurred in (dt) and also for (k - 1) units to have had action completed on them; k = 2,3,... . But (A-9) indicates that the probability of k arrivals in (dt) involves higher orders of (dt), and by our convention such probabilities are considered insignificant. Thus, the change in state from $E_{N-1}(t)$ to $E_N(t + dt)$ is reckoned primarily on the arrival of one unit in (dt).

Proceeding further, we next seek the conditional probability that the stage is in state $E_N(t + dt)$, given that it was in state $E_{N+1}(t)$. One way in which this can occur is the simultaneous occurrence of two independent random phenomena: action is completed on a unit in (dt), and no unit arrives in (dt). By the assumption of independence, the conditional probability of this occurrence is given by the product of the probability that there were (N + 1) units at the stage at time t, the probability that no unit arrived in (dt), and the probability that action was completed on a unit in (dt):

$$P\left[E_N(t + dt) \middle| E_{N+1}(t)\right] = P_{N+1}\left[(n+1);t\right] \cdot \left(e^{-\lambda(dt)}\right)\left(1 - e^{-\mu(dt)}\right)$$

which becomes on expansion,

$$P\left[E_N(t + dt) \middle| E_{N+1}(t)\right] \approx \left[\mu(dt)\right] \cdot P_{N+1}\left[(n+1);t\right] \quad . \tag{A-12}$$

By reasoning similar to that employed above, it is clear that the change in state from $E_{N+1}(t)$ to $E_N(t + dt)$ is dependent primarily on the completion of action on one unit.

Finally, one way in which the stage can be in state $E_N(t + dt)$, given that it was in state $E_N(t)$, is the simultaneous occurrence of two independent random phenomena: no unit arrives in (dt), and action is not completed on a unit in (dt). Thus,

$$(A\text{-}13) \qquad P\left[E_N(t + dt) \middle| E_N(t)\right] \approx \left[1 - (\lambda+\mu)(dt)\right] \cdot P_N\left[n;t\right] \quad ,$$

and, again, this is the only significant probability.

In (A-11), (A-12), and (A-13), we have probability expressions for three mutually exclusive ways in which the stage can be in state $E_N(t + dt)$. Now because of the convention regarding orders of (dt), these three ways are the only ways for the stage to be in state $E_N(t + dt)$; hence, the enumeration of these three ways constitutes an exhaustive listing of mutually exclusive events, none of which permits the number of units at the stage to change by more than unity. (Note that the restriction on orders of (dt) has us saying, in effect, that (dt) is "large" enought to permit one unit to arrive, or to permit the completion of action upon one unit, but not both; and also that (dt) is "small" enough to prevent more than one occurrence of one of these independent random phenomena.)

Since we now have an exhaustive listing of all possible ways in which the stage can be in state $E_N(t + dt)$, the total unconditional probability that the stage is in state $E_N(t + dt)$ is given by the sum of (A-11), (A-12), and (A-13):

$$
\begin{aligned}
P_N\left[n;\ t+dt\right] &= \lambda(dt) \cdot P_{N-1}\left[(n-1);\ t\right] \\
&+ \mu(dt) \cdot P_{N+1}\left[(n+1);\ t\right] \\
&+ \left[1 - (\lambda+\mu)(dt)\right] \cdot P_N\left[n;t\right] \\
&+ \text{terms involving } (dt)^2,\ (dt)^3, \text{ etc.}
\end{aligned}
\tag{A-14}
$$

If (A-14) is rearranged slightly, we have

$$
\begin{aligned}
P_N\left[n;\ t+dt\right] - P_N\left[n;t\right] = &\ \lambda(dt) \cdot P_{N-1}\left[(n-1);t\right] \\
&+ \mu(dt) \cdot P_{N+1}\left[(n+1);t\right] \\
&- \left[(\lambda+\mu)\ (dt)\right] \cdot P_N\left[n;t\right] \\
&+ \text{terms involving } (dt)^2,\ (dt)^3, \text{ etc.}
\end{aligned}
$$

Dividing both sides by (dt), we obtain

$$
\begin{aligned}
\frac{P_N\left[n;\ t+dt\right] - P_N\left[n;t\right]}{dt} = &\ \lambda P_{N-1}\left[(n-1);t\right] \\
&+ \mu P_{N+1}\left[(n+1);t\right] \\
&- \left[(\lambda+\mu)\right] P_N\left[n;t\right] \\
&+ \text{terms involving } (dt),\ (dt)^2, \text{ etc.}
\end{aligned}
$$

If we let (dt) approach zero in the expression above, we have in the limit

$$
\frac{dP_N\left[n;t\right]}{dt} = \lambda P_{N-1}\left[(n-1);t\right] + \mu P_{N+1}\left[(n+1);t\right] - (\lambda+\mu) P_N\left[n;t\right] . \tag{A-15}
$$

Equation (A-15) has been developed along the axiomatic lines of modern probability theory. Standard queueing notation employs a different notation. As long as it is clearly understood that (1) the presence of information units at the stage at a time t, that is, the "state" of the stage at a time t, is a random phenomenon which is itself the effect of the interaction of two other independent random phenomena (namely, "arrivals"

and "actions"); (2) N is the non-negative, integral-valued random variable describing this random phenomenon in terms of the number of units at the stage at a time t; and (3) time t is NOT a random variable, but merely a time of observation and/or prediction of the state of the stage, we may simplify the notation of (A-15) into the standard queueing theory form:

$$\text{(A-16)} \qquad \frac{dP_n(t)}{dt} = \lambda P_{n-1}(t) + \mu P_{n+1}(t) - (\lambda+\mu)P_n(t) .$$

Equation (A-16) is (4-1) of Section 4.

Equation (A-16) has been developed with the assumptions that the random variables T_a and T_r obey negative exponential distribution laws (with parameters λ and μ, respectively), and that the random variable N representing the number of arrivals in an observation time of duration t obeys a Poisson distribution law (with parameter λ). It should be noted that Parzen [10] derives (A-16) as a special case of a more general situation (the "pure birth and death process") with the aid of fewer basic assumptions; specifically, no underlying distribution laws are assumed.