SOME RECENT PROGRESS IN ANALYTIC STUDIES OF SYSTEM PERFORMANCE

Hisashi Kobayashi

(IBM T. J. Watson Research Center, Yorktown Heights, New York)

ABSTRACT

In recent years, the evolution of studies on system performance evaluation has been characterized by an increasing use of stochastic process representations of programs and of statistical techniques in the measurement and analysis. The present paper attempts to place some recent progress of this field in perspective. First, characterization of computer workload is discussed in terms of time-series analysis of the job traffic, and a random walk model of the program behavior in a virtual memory system, etc. Secondly, applications of statistical analysis and inference techniques to measurement and evaluation are discussed in an expository form, including recent empirical results. Thirdly, queueing network models are reviewed. In particular, we demonstrate how J.R. Jackson's results on network of queues can be applied to multiprogrammed/multiprocessing systems. The last section focuses on the diffusion approximation approach of queueing processes and program behavior.

I. INTRODUCTION

Construction of a general quantitative methodology for system performance evaluation is still in embryo. Yet, we observe an increasing interest and progress in this subject. The present paper is intended to review the state-of-the-art of some important aspects of analytical studies of system performance evaluation. We hope the present paper will serve as an introduction and stimulate further evolution of this important field of computer science.

The problem of performance evaluation is essentially identification of a large complex system driven by complex inputs. Thus, the first required step is to characterize job streams and service demands, which is the subject of Sect. 2.1. Study on the program behavior in a paging environment is drawing an increasing attention. In Sect. 2.2, we review several models of page reference patterns.

A barely explored aspect of system performance study is the systematic procedure to identify key system variables which can capture essential information of the "state" of the system and relate themselves to a chosen performance criterion. Sect. III discusses some of the statistical analyses and experimental design techniques which serve for that purpose.

A recent development in network of queue models promotes a renewed interest in queueing theoretic models, and give rise to some hope that more realistic models of multiprogrammed/multiprocessing systems will be obtained. Sect. IV interprets J.R. Jackson's result [30] in this context, emphasizing its generality and more flexibility than closed network models.

A technique which may overcome limitations of queueing models is an approximation method via the diffusion process representation. Section V reviews recent work on this subject and presents some new results as well. Applications of similar treatment to program behavior studies are also touched upon.

II. CHARACTERIZATION OF WORKLOADS

2.1 Message Traffic and Service Time Requirements

One important aspect in modelling interactive systems or on-line data base systems is characterization of the job arrival process. Some recent studies [1, 2, 3] give empirical evidence that an arrival process can be approximated by a Poisson process, thus a large body of queueing theory based on the Poisson arrival may be justified for its use. Lewis and Yue [2] analyzed the measurement data of a teleprocessing information retrieval system, and Anderson and Sargent [3] conducted a similar study on an experimental APL/360 system. The analysis proceeds as follows: first, the question of stationality is examined. Here, a graphical analysis is of prime importance. For the analysis of trends, there exist several formal statistical tests [4]. After an appropriate sample interval is chosen to make the traffic rate trend-free over each interval, the sequence of arrival moments $\{t_1, t_2, t_3, \ldots\}$ is examined in two steps: first, we must show that the arrival process is a renewal process, i.e. the independence of the interarrival time sequence $\{x_i\}$ must be tested, where $x_i = t_i - t_{i-1}$. If this test is passed, then apply uniform conditional tests [4] where the null hypothesis is that the data arises from a Poisson process. If the series has been observed for a fixed time interval T, and n events occur in $(0,T)$ then we define

$$u_i = \frac{t_i}{T}, \quad i = 1,2,\ldots,n. \tag{2.1}$$

Under the null hypothesis, the random variables $\{u_i\}$ are clearly order statistics whose probability distribution is uniform over the unit interval, i.e., its distribution is

$$F_0(u) = u, \quad 0 \le u \le 1. \tag{2.2}$$

Denote by $F_n(u)$ the empirical distribution of the observation $\{u_i\}$. Then the problem has been reduced to the canonical form of distribution-free tests of fit [4]. We accept the null hypothesis, if the "distance" between $F_0(u)$ and $F_n(u)$ is not large. As for measures of distance, the Kolmogorov-Smironov Statistics are chosen. Other distance measures investigated are the Anderson-Darling statistics, Durbin's modification of these statistics and the Moran statistics [4]. In both studies [2, 3], the null hypothesis that the input traffic is Poisson was not rejected at the significance level of 5% in all of the tests discussed above.

In queueing models an exponential distribution of service time is often assumed for analytical tractability. Although some measurement studies report validity of such an assumption, several authors [3, 5] observed that computer processing demand tends to have a fairly large coefficient of variation (c.v). Anderson and Sargent [3] report that in the APL system they measured the coefficient of variation (c.v) was typically as large as 9 to 10. A large c.v. is often due to a long tail at the upper end of the probability density function. A graphical tool to examine the tail end is the completion rate (the hazard function or the failure rate). We denote by s the service time demand variable and by $F(s)$, $R(s) = 1-F(s)$ and $f(s)$ the distribution function, survivor function and probability density function. Then the completion rate $\phi(s)$ is defined by

$$\phi(s) = \frac{f(s)}{R(s)} = -\frac{d}{ds} \log R(s) \quad . \tag{2.3}$$

Thus

$$R(s) = \exp\{-\int_0^s \phi(u)du\}. \tag{2.4}$$

If $\phi(s)$ is some power of s, i.e.,

$$\phi(s) = \alpha\beta s^{\beta-1}, \ \alpha>0, \ \beta>0 \quad , \tag{2.5}$$

then

$$F(s) = 1-R(s) = 1-e^{-\alpha s^{\beta}}. \tag{2.6}$$

This distribution is called the Weibull distribution with parameters α, β [4]. The completion rate is monotone decreasing if β is less than one. The computer processing distribution of the experimental APL/360 system is well approximated by the Weibull distribution with $\beta = 0.25 \sim 0.35$ [3].

The hyperexponential (or mixed-exponential) distribution [4] is sometimes a better model than the Weibull ($0<\beta<1$) or Gamma distributions. The initial value of the probability density function is not infinite as with the two other distributions, and this allows for greater flexibility. Another advantage of the hyperexponential distribution over other models is that the model can be incorporated with the notion of parallel stages in queueing theory [6, 7].

2.2' Representation of Program Behavior

In a virtual memory system with paging, the behavior of a program is represented by the page reference string:

$$\underline{r} = r_1 r_2 \cdots r_i \cdots \tag{2.7}$$

Let us transform $\underline{r}$ into a numerical sequence $\underline{d} = d_1 d_2 d_3 \cdots d_i, \ldots$, where d_i is the total number of distinct pages referenced since the last reference to page name r_i. If r_i has never been referenced in the past, we assign ∞ to d_i. The sequence $\underline{d}$ is called the LRU stack distance string. The notion of stack distance was originally introduced by Mattson et al. [8] in the stack processing technique which is an efficient evaluation technique for storage hierarchies. Lewis and Yue [9] and Shedler and Tung [10] considered the LRU stack distance representation of a program in order to quantitatively characterize the dynamical behavior of a program. The empirical study by Lewis and Yue [9] indicates that $\underline{d}$ would very well be stationary even though $\underline{r}$ is clearly nonstationary. They also observed that the power spectrum of sequence $\underline{d}$ reveals individual characteristics of programs. When a program is moving toward a new locality, values of d_i's tend to be large, since new pages or those which have not been referenced for a long period are likely to appear. The d_i's retain small values while the program remains within a locality. Therefore, the rate of transition from one locality to another has a strong effect on the spectral shape. Shedler and Tung [10] have proposed to model distance strings by a certain class of finite state Markov chains. Values for parameters of the model can be chosen to make the page-exception characteristics of the generated sequences of page references consistent with those of actual program traces. Empirical studies to validate these models and development of "synthetic" programs for actual use are yet to be explored.

Some probabilistic characteristics of the working set and working set size [11] have been investigated by several authors [12, 13]. The working set $W_i(T)$ at time i is the set of distinct pages references in the T most recent references $r_{i-T+1}, r_{i-T+2}, \ldots, r_i$, and the working set size $w_i(T)$ is simply the size of $W_i(T)$. Properties of the working set are discussed by Denning and Schwartz [12] under the stationarity assumption for $\underline{r}$. Most of their results seem extendable to a more general case where we require only the weak stationarity of $\underline{w}(T)$. Let

$$E[w_i(T)] = s(T) \tag{2.8}$$

and let $m_i(T)$ be the missing page rate: the probability that r_{i+1} is not found in $W_i(T)$. The stationarity of $w_i(T)$ implies that $m_i(T)$ is independent of i and

$$s(T+1) - s(T) = m(T) \tag{2.9}$$

and

$$s(T+1) - s(T) = 1-F_x(T) \quad , \qquad (2.10)$$

where $F_x(.)$ is the distribution of the inter-reference interval variable x. An inter-reference interval x_i is defined as the total number of pages referenced since the last reference to r_i. Clearly, $x_i \geq d_i$. Equation (2.9) is obtained from the following observation: $E[w_i(T+1) - w_i(T)] = s(T+1) - s(T)$ is the probability that r_{i-T}, which has just left the window $[i-T+1,i]$, does not find its copy in the current working set $W_i(T)$. This quantity should be equal to $m(T)$, the probability that a new incoming page r_{i+1} does not find itself in $W_i(T)$, because of the equilibrium of the working set size. Similarly, the probability that r_{i-T} does not find its copy in $W_i(T)$ is equal to the probability that the inter-reference interval is greater than T.

The working set size sequence is representable as

$$w_i(T) = w_{i-1}(T) + z_i, \quad i = 1,2,... \qquad (2.11)$$

where the steps z_i can only take the values 1, 0 or -1. So if we can assume that $Prob\{z_i=1\} = p$, and $Prob\{z_i = -1\} = q$, then the process w_i is a simple random walk with reflecting barriers [14] at w=1 and w=T. Then the limiting equilibrium distribution of the state occupation probabilities is given by the truncated geometric distribution:

$$Prob\{w=\ell\} = \frac{1 - \frac{p}{q}}{1 - (\frac{p}{q})^T} (\frac{p}{q})^{\ell-1}, \quad \ell = 1,2,...,T \qquad (2.12)$$

The above assumption may be acceptable when T is relatively small but is clearly unrealistic for a large value of T. The probability p that z_i takes on 1 should be dependent on the state w_{i-1}. Similarly, the probability q should be a monotone increasing function of w_{i-1}. The process $\{w_i\}$ may be described as a random walk with some kind of central restoring tendency. If the probability of moving one unit towards the center (E[w] = s) is greater than the probability of moving one unit away from the center by an amount proportional to the distance from the center, i.e., $|w-s|$, then such a random walk has an interpretation in terms of the Ehrenfest model of diffusion [14, 15]. The limiting equilibrium distribution is given by a symmetric binomial distribution. It can be shown by use of the diffusion approximation that for a large T the Ehrenfest model goes over into the Ornstein-Uhlenback (O-U) process. Furthermore, the equilibrium distribution for the O-U process between two reflecting barriers is a truncated Gaussian distribution [14, 15]. Coffman and Ryan [13] applied the Gaussian distribution approximation to a study of storage partitioning in multiprogram environments. Specifically, the fixed partioning and dynamic partitioning schemes are compared in terms of the probability that no regions are in saturation.

III. APPLICATION OF STATISTICAL DESIGN AND ANALYSIS TECHNIQUES TO PERFORMANCE EVALUATION

The need for more thorough statistical studies of empirical data has been recently recognized. See, for example, Grenander and Tsao [16]. During the past several years we have observed a significant amount of efforts made on development of both hardware and software measurement tools. Although some need still exists for improvement and economization of these techniques, a more important and difficult question to be answered is "what parameters should be measured?" rather than how to measure these parameters. A branch in statistics called "design of experiments" [17] has something to contribute to this problem. An essential feature of modern experimentation is the "randomization out" of the experiment of the effects of factors outside the experimental structure. Consider, for instance, a time-sharing system which uses a round-robin scheduling. We want to examine whether different values of the quantum size of a time slice q affects the average response time per job x. So we choose different values of the quantum size: q_i, $i = 1,2,...,m$ and perform some experiment. However, the workload changes according to the time of day and this unrecognized causal factor may vary in such a way as to overdominate the effect due to the quantum size. One obvious method which can avoid such an effect will be to assign different sizes of time quantum randomly to different observation intervals and hence randomize out unrecognized nuisance factors. Such an arrangement is called a completely randomized design.

Statistical tests of significance are often useful in interpreting the results of experiments. Let μ_i be the (unknown) expected value of x when the quantum size is q_i. Then the hypothesis we postulate is

$$H: \ \mu_1 = \mu_2 = \ldots = \mu_m. \qquad (3.1)$$

$\underline{X}$ stands for the set of observations and $p(\underline{X})$, the probability density function. We define the likelihood ratio statistic λ by

$$\lambda = \frac{\max_{H\cap\Omega} p(\underline{X})}{\max_{\Omega} p(\underline{X})} \quad , \qquad (3.2)$$

where Ω is the set of assumptions made on the probability density function $p(\underline{X})$. Let x_{ij} stand for the j^{th} sample under the i^{th} treatment, $j = 1,2,...,n_i$, $i = 1,2,...,m$. Let us assume that x_{ij} is normally distributed:

$$p(\underline{X}) = (\frac{1}{2\pi\sigma^2})^{\frac{N}{2}} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij}-\mu_i)^2\},$$

$$N = \sum_{i=1}^{m} n_i. \qquad (3.3)$$

Then some manipulation leads to

$$\lambda = (1 + \frac{Q_{Tr}}{Q_a})^{-\frac{N}{2}} \quad , \qquad (3.4)$$

where

$$Q_a = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2,$$

$$Q_{Tr} = \sum_{i=1}^{m} = n_i(\bar{x}_{i.} - \bar{x}_{..})^2, \qquad (3.5)$$

and

$$\bar{x}_{..} = \frac{1}{N} \sum_{i=1}^{m} n_i \bar{x}_{i.} \qquad (3.6)$$

The quantity Q_a represents the sum of dispersion within groups and Q_{Tr} represents dispersion among different treatments. Instead of λ we may take any monotone function of λ, say,

$$F_{m-1,N-m} \triangleq \frac{N-m}{m-1} (\lambda^{-\frac{2}{N}} - 1) = \frac{Q_{Tr}/(m-1)}{Q_a/(N-m)} . \qquad (3.7)$$

Thus, if we define $F_0 = F(\lambda_0)$, then $\lambda < \lambda_0$ if and only if $F > F_0$. The statistic $F_{m-1,N-m}$ has the F-distribution with m-1 and n-m degrees of freedom. The F-test plays a central role in the analysis of variance. Both the likelihood ratio test and F-test assume normality of the observation. Investigation has shown, that failure to satisfy this condition has little effect upon the F-test for equal means [18].

In the completely randomized design, we avoid the appearance of bias due to unattended causal factors. It is often possible to make more precise comparisons of the treatments by grouping the experimental units into blocks of m units such that units within a block resemble each other more than units in different blocks. This layout is called a randomized block design, if within each block the M treatments are assigned at random to the m experimental units. The analysis of variance can be performed by assuming that the block and treatment effects are additive; the total sum of squares of deviation from the mean value is then

$$\sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \bar{x}_{..})^2 = m \sum_{j=1}^{n} (\bar{x}_{.j} - \bar{x}_{..})^2 + n \sum_{i=1}^{m} (\bar{x}_{i.} - \bar{x}_{..})^2 + \sum_{i=1}^{m} \sum_{j=1}^{n} \{x_{ij} - \bar{x}_{.j} - (\bar{x}_{i.} - \bar{x}_{..})\}^2 \qquad (3.8)$$

or

$$Q_T = Q_B + Q_{Tr} + Q_E , \qquad (3.9)$$

where Q_T, Q_B, Q_{Tr}, and Q_E stands for sums of squares for total, blocks, treatments and for errors, respectively. What was called Q_a is now divided into two parts: $Q_a = Q_B + Q_E$.

Analogous to (3.7), we can show that

$$F_{m-1,(n-1)(m-1)} = \frac{Q_{Tr}/(m-1)}{Q_E/(n-1)(m-1)} . \qquad (3.10)$$

is the statistic to test the hypothesis of equal treatment effects.

In the design methods discussed above, we were concerned with only one factor in the treatments. If the treatments in an experiment consist of all possible combinations of a set of underlying factors, it is called a factorial experiment. We may use a factorial arrangement with any of the designs discussed above. Consider for example, two-factor completely randomized designs where the two factors are denoted by A and B. Thus, analogously to those equations we derived above, we can show that the total sum of squares is partitioned as

$$Q_T = Q_A + Q_B + Q_{AB} + Q_E \qquad (3.11)$$

where Q_{AB} corresponds to a measure of interaction between factors A and B. These quantities are then used in F-tests of various significance tests including the test of examining whether factors A and B interact. An application of factorial experiments to the problem of computer performance is found in Tsao et al. [19]. Anderson and Sargent [3] applied a two-factor completely randomized experiment to their study where they attempted to characterize the degradation of response time in terms of two factors: the number of active users and the traffic rate per user. They showed through the analysis of variance that these two factors were statistically significant but that the total variation of the response time contained a large amount of unexplained terms, Q_E. This result led them to consider the internal queue size as an additional system variable for inclusion in the empirical model of an experimental APL/360 system. Silverman and Yue [21] applied the analysis of variance technique to software measurement data of a teleprocessing information retrieval system, and drew a conclusion that the following four parameters are most significant factors in characterizing the response time of a given message task: (1) the priority position in the CPU dispatching queue, (2) the number of concurrent tasks, (3) the number of other tasks trying to gain access to the same file and (4) the number of I/O operations of the given task. The first three factors are considered major system congestion factors and influence the wait times for CPU, channels, and file accesses, respectively.

The analysis of variance discussed above, in relation to different types of experimental design, treats factors qualitatively in the sense that the method refers to the presence or absence of the effects of some factors under the condition in which the observations are taken. In regression analysis all factors are quantitative and treated quantitatively. The theory of regression is concerned with prediction or estimation of one or more variables on the basis of information provided by other measurements or concomitant variables. For details of regression analysis (the theory of least squares, in general), the readers are referred to [17, 20] or any books on statistical estimation theory. An application of regression analysis to perform evaluations is reported by Bard [22] in his study of CP-67/

CMS, in which he attempted to relate CPU supervisor state time to a set of system variables in a linear regression model. Anderson and Sargent [3] applied linear regression models to quantify the inter-relationship between the response time and a set of key system state variables.

As mentioned in the Introductory section, identification of key system variables is a most important and difficult task in analysis of computer systems. There is no unique statistical procedure for doing this. Several procedures have been proposed to select the "best regression equation", given a response variable (i.e., performance measure), a set of candidate predictor variables (i.e., system factors) and a series of observations on all of them. They are (1) all possible regressions, (2) backward elimination, (3) forward selection, (4) stepwise regression, and (5) stagewise regression. Draper and Smith [20] discuss details and relative merits of these different methods. Applications of these methods to measurement data of computer systems are yet to be seen.

IV. QUEUEING NETWORK MODELS OF COMPUTER SYSTEM

Most of the queueing theoretic models studied in the past treat a computer system as a single entity; namely, it is tacitly assumed that a central processor unit (CPU) is the only "bottleneck" resource. These "single server" models of computer systems are well documented in the forthcoming books by Kleinrock [7], and Coffman and Denning [23]. A class of analytic models which has been drawing attention during the past few years is the queueing network type [24-29]. These models are able to incorporate parallel processing capabilities of a typical multiprogrammed system.

Let us assume that the arrival process of a given system is Poisson with rate λ (Figure 1). Once a program enters the system, the behavior of the program is representable as alternations of CPU and I/O executions. Let the execution intervals of processors be exponentially distributed with parameters $\mu_i(n_i)$, $0 \le i \le N$. Here n_i is the number of programs residing in the i^{th} processor of its queue. Let λ_i be the arriving rate to the i^{th} processor, i.e.,

$$\lambda_i = \lambda\delta_{i,0} + \sum_{j=0}^{N} \lambda_j p_{ji}, \quad i = 0,1,\ldots,N \qquad (4.1)$$

where p_{ji} is the probability of transition from the j^{th} processor to the i^{th} processor. Then we can apply the result due to J.R. Jackson [30]: under these Markovian assumptions, the equilibrium distribution of the queue size of individual processors can be solved by pretending as though they were separate and independent queueing systems having Poisson arrivals with rates λ_i's given by (4.1). Therefore, the probability that the i^{th} processor finds n_i programs is

$$p_i(n_i) = p_i(0) \prod_{n=1}^{n_i} \frac{\lambda_i}{\mu_i(n)}, \quad n_i = 0,1,2.... \qquad (4.2)$$

So, letting $\underline{n} = [n_0 n_1 \ldots n_N]$ be the state vector which indicates that there are n_i programs at the i^{th} processor, we have

$$p(\underline{n}) = \prod_{i=0}^{N} p_i(n_i). \qquad (4.3)$$

For simplicity, we assume hereafter that $\mu_i(n) = \mu_i$. Then equations (4.2) and (4.3) are reduced to

$$p(\underline{n}) = \prod_{i=0}^{N} \{(1-\rho_i)\,\rho_i^{n_i}\} \qquad (4.4)$$

where

$$\rho_i = \frac{\lambda_i}{\mu_i}. \qquad (4.5)$$

From (4.4), we can derive easily the utilization factor of each processor, the average queue size, the average stay time in each of the queues, etc. Furthermore, the response time and throughput are also readily computable.

In the above model the degree of multiprogramming $M = \sum_{i=1}^{N} n_i$ is a random variable and is unbounded. If one is interested in the behavior of the system at its maximum degree of multiprogramming, one should modify the model as follows: let M_0 programs reside in the system initially and set $\lambda=0$. Connect the departure path from the I/O processors to the input processor and put $\mu_0(n_0) = \infty$. This means that as soon as one of the M_0 programs leaves the system, a new program enters the central processor immediately. Then we obtain the solution of (4.1):

$$\lambda_i = \lambda_1 p_{1i}, \quad i = 2,3,\ldots,N \qquad (4.6)$$

where λ_1 is arbitrary. Let us define S by

$$S = \{\underline{n}: \sum_{i=1}^{N} n_i = M_0\}. \qquad (4.7)$$

Then the probability $p^*(\underline{n})$ that $\underline{n}$ is the state of the system is

$$p^*(\underline{n}) = \begin{cases} 0 & \underline{n} \notin S \\ p(\underline{n}|S) & \underline{n} \in S \end{cases} \qquad (4.8)$$

where $p(\underline{n}|S)$ is the probability of $\underline{n}$ conditioned on $\sum_{i=1}^{N} n_i = M_0$:

$$p(\underline{n}|S) = \frac{p(\underline{n})}{p(S)} = \frac{\prod_{i=1}^{N} \rho_i^{n_i}}{\sum_{\underline{n}' \in S} \prod_{i=1}^{N} \rho_i^{n_i'}} \qquad (4.9)$$

where

$$\rho_1 = \frac{\lambda_1}{\mu_1}, \quad \rho_i = \frac{\lambda_1 p_{1i}}{\mu_i}, \quad i = 2,\ldots,N. \qquad (4.10)$$

Putting $\lambda_1 = \mu_1$, we obtain the equilibrium state probability of the closed queue network as follows:

$$p^*(\underline{n}) = \begin{cases} 0 & , \text{ for } \underline{n} \notin S \\ \dfrac{\prod_{i=2}^{N} \left(\frac{\mu_1 p_{1i}}{\mu_i}\right)^{n_i}}{\sum_{\underline{n}' \in S} \prod_{i=2}^{N} \left(\frac{1 p_{1i}}{\mu_i}\right)^{n_i'}} & , \text{ for } \underline{n} \in S \end{cases} \qquad (4.11)$$

Gordon and Newell [31] developed a solution technique which, for closed systems, involves essentially the same steps as Jackson's technique and which could be used to derive (4.11).

Jackson's work is in a more general framework than those cases discussed above. For example, he considers the case where the arrival rate λ is an arbitrary function of M. Thus, we shall be able to impose constraint on the maximum degree of multiprogramming in a more realistic fashion than we do in the closed queue model. Furthermore, his results included the case where an immediate injection of a new job is triggered when M falls below a specified lower bound M*. This enables us to treat the situation where we keep background jobs to maintain high utilization of resources. It will not be difficult to see that the closed queue network can be regarded as a special case where $M^* = M_0$ and $\lambda(M) = 0$ for $M \geq M_0$.

Pozner and Bernholtz [32] have extended the work of Gordon and Newell [31] to the treatment of systems having several classes of jobs (programs) in a closed finite queue network. Baskett [33] and Chandy [34] have recently shown that in a closed network of queues the queue-length distributions at equilibrium state are solely dependent on the mean values of service time distributions, if the processors (service stations) associated with non-exponential service time distributions are "processor shared". Processor sharing [7] is an asymptotic form of round-robin scheduling with its time quantum approaching zero, and its properties in a single server system with various queue disciplines are discussed by Kleinrock, Muntz and Hsu [35].

Sekino [36] discusses a queueing model for a multiprogrammed/multiprocessor time-sharing system based on the machine repairman model [6], which is in fact a special case of a closed queue network; thus the result of Gordon and Newell [31] could be applied. Sekino has derived a closed form expression for the distribution of response time under the assumption that the processing time and user's think time are both exponentially distributed.

V. APPROXIMATION METHODS FOR PROBABILISTIC MODELS

The practical value of queueing theoretic analysis has been severely limited by the lack of approximation methods which would enable one to attach more realistic models and by the lack of sensitivity analysis which would give estimates of errors introduced by simplification in the model.

A few attempts, however, have been made in recent years to break away from the vogue in queueing theory. Kingman [37] has shown in his treatment of "heavy traffic theory" that properties of nearly saturated queues are rather insensitive to the detailed form of the arrival or service distributions. The heavy traffic approximation relies on the central limit theorem and so does the approximation approach based on the diffusion process. Although the idea of approximating a discrete state process by a diffusion process with continuous path is not new, applications to congestion theory are apparently rather recent [38, 39]. Gaver [39] applied this method to the waiting time in an M/G/1 queue and Gaver and Shedler [40] proposed a method to evaluate CPU utilization of a multiprogrammed system represented by a cyclic queue model. The diffusion approximation method can be useful because mathematical methods associated with the continuum (e.g. differential equations, integration) very often lend themselves more easily to analytical treatment. A recent book by Newell [41] serves as an excellent introduction to this subject.

In this section we use the cyclic queue model as an illustrative example. If we set N=2 in the closed queueing model discussed in Sect. IV, we obtain a cyclic queue system (Figure 2). Let us denote the service distribution functions of two processors of the cyclic queue by A(s) and B(s), respectively. If both A(s) and B(s) are general distributions, no known techniques exist to solve this model. Thus, we are motivated to overcome this mathematical difficulty by using some approximation technique. Let M(t) denote the number of programs at the central processor or its queue at time t. A typical realization of M(t) is shown in Figure 3. Suppose that $0<M(t)<M_0$ at a given instant t. The mean and variance of the incremental change of M(t) per unit time are approximately given by [14, 40, 41]

$$\alpha = \frac{1}{\Delta} E[M(t+\Delta)-M(t)] \tilde{=} \frac{1}{\mu_b} - \frac{1}{\mu_a} \qquad (5.1)$$

and

$$\beta = \frac{1}{\Delta} \mathrm{Var}[M(t+\Delta)-M(t)] \tilde{=} \frac{\sigma_b^2}{\mu_b^3} + \frac{\sigma_a^2}{\mu_a^3} \qquad (5.2)$$

This suggests taking the following process X(t) as an approximation to M(t)

$$dx(t) = \alpha \cdot dt + \sqrt{\beta}\ (t)\ \sqrt{dt} \qquad (5.3)$$

where x(t) is bounded by

$$0 \leq x(t) \leq M_0. \qquad (5.4)$$

The process z(t) of Equation (5.3) is a white Gaussian process with zero mean and unit variance. If the boundary condition (5.4) were not imposed, then the stochastic differential equation (5.3) defines a Wiener-Levy process with drift: if the initial value of x(t) is x_0, then the unrestricted process x(t) has the following conditional probability density function at time t:

$$p(x_0,x;t) = \frac{1}{\sqrt{2\pi\beta t}} \exp\{ - \frac{(x-x_0-\alpha t)^2}{2\beta t} \}. \qquad (5.5)$$

which satisfies the following partial differential equation:

$$\frac{\partial}{\partial t} p(x_0,x;t) = -\alpha \frac{\partial}{\partial x} p(x_0,x;t) + \frac{\beta}{2} \frac{\partial^2}{\partial x^2} p(x_0,x;t) \tag{5.6}$$

Equation (5.6) is known as the forward or Fokker-Planck equation [14].

In what follows, we are interested in solving Equation (5.6) with the boundary condition (5.4). A natural way to handle this condition is to treat $x=0$ and $x=M_0$ as reflecting barriers. The boundary conditions for reflecting barriers at $x=0$ and $x=M_0$ are [14]:

$$\frac{\beta}{2} \frac{\partial}{\partial x} p(x_0,x;t) - \alpha\, p(x_0,x;t) = 0 \quad \text{at } x=0 \text{ and } x=M_0. \tag{5.7}$$

The statistical equilibrium distribution is readily obtainable as

$$p(x) = \lim p(x_0,x;t) = \frac{2\gamma e^{2\gamma x}}{e^{2\gamma M_0}-1}, \quad 0 \le x \le M_0 \tag{5.8}$$

where

$$\gamma = \frac{\alpha}{\beta} = \frac{\mu_a^3\mu_b^2 - \mu_b^3\mu_a^2}{\mu_a^3\sigma_b^2 + \mu_b^3\sigma_a^2} \tag{5.9}$$

In order to compare this solution with some known result, consider the exponential distribution case, i.e., $\sigma_a = \mu_a$, $\sigma_b = \mu_b$. The distribution of the queue size is easily obtainable as

$$p_m = \frac{(1-\rho)\rho^m}{1-\rho^{M_0+1}}, \quad m = 0,1,2,\ldots,M_0 \tag{5.10}$$

where

$$\rho = \frac{\mu_b}{\mu_a} \tag{5.11}$$

whereas the solution via the diffusion approximation is given by (5.8) where

$$\gamma = \frac{1-\rho}{1+\rho}. \tag{5.12}$$

Note that the exact solution (5.10) is a truncated geometric distribution, and its approximation is a truncated exponential function. In order to illuminate the analogy further, let us set $\rho=1$. Then we obtain

$$p_m = \frac{1}{M_0+1}, \quad m = 0,1,2,\ldots,M_0 \tag{5.13}$$

$$p(x) = \frac{1}{M_0}, \quad 0 \le x \le M_0 \tag{5.14}$$

which are both uniform distributions.

Thus far we were concerned with the steady state distribution. Equally important to practical applications is the rate of approach to the equilibrium. The diffusion approximation is often advantageous in this respect since the transient behavior can in many cases be obtainable in relatively simple closed form. The solution for the diffusion equation (5.6) with the conditions (5.7) and $x(0) = x_0$ is obtained in [42]:

$$p(x_0,x;t) = p(x) + e^{\gamma(x-x_0-\frac{\alpha}{2}t)} \cdot \sum_{n=1}^{\infty} \phi_n(x)\phi_n(x_0) e^{\frac{\alpha\lambda_n^2}{2}t} \tag{5.15}$$

where $p(x)$ is given by (5.8) and

$$\lambda_n = \frac{n\pi}{M_0}, \tag{5.16}$$

and

$$\phi_n(x) = \frac{2\lambda_n^2}{M_0(\lambda_n^2+\gamma^2)} \{\cos \lambda_n x + \frac{\alpha}{\lambda_n} \sin \lambda_n x\} \tag{5.17}$$

for $n = 1,2,3,\ldots$. A more detailed study of the transient behavior and a more comprehensive treatment of the diffusion approximation as applied to a general queueing network model are discussed in [42].

VI. CONCLUDING REMARKS

In this concluding section we shall refer to some subjects which are left unmentioned.

An automatic or adaptive algorithm for resource allocation is among the least explored subject. Doherty [43] applies the concept of working set size to a TSS/360 scheduler in which the length of a time slice is set approximately inversely proportional to its current working set size. This policy allows programs with good locality to progress rapidly. Several authors have pointed out that performance of programs in virtual memory systems can be significantly improved by rearranging program sectors that make up the program layout in virtual memory. Hatfield and Gerald [44] report semi-automated procedures to improve program structure.

The "conservation-law" which Kleinrock [7,35] and Wolff [45] discuss for a single-server system should be extendable to a general network of queue. Therefore the queue size distribution (hence the throughput also) obtained in Section V is independent of the queue disciplines of a network, as far as they are work-conserving priorities. However, the wait time distribution for non-FCFS disciplines is not easy to obtain. Kobayashi and Silverman [46] showed that the CPU dispatching rule adopted in IBM 360/OS MVT is equivalent, under a special environment, to a preemptive-resume scheme based on seniority, and analyzed its effect on the response time distribution. An earlier work by Gaver [47] on priority queues seems extendable to a general network of queues.

Despite the widespread practice of simulation models for performance evaluation, the present paper did not address itself to this subject due to the limitation of space. Statistical techniques discussed

in Sect. III are equally applicable to simulation plans and analyses. What characterizes most of the simulation studies conducted in the past are enormous amounts of costs associated with the development and operations, their extreme specificity and poor capability to yield general insights. This remark, however, is not intended to underrate the value of simulation studies. An ingenuity often brings about a significant amount of savings in simulation costs. For example, the stack processing techniques discussed earlier [8] provided a means of avoiding the tremendous amount of direct simulation which would otherwise be necessary to obtain "hit ratios" for a range of page-sizes and memory capacities. As for Monte Carlo simulations, the cost to run a simulator can be reduced substantially by proper use of variance reduction techniques [48, 49].

ACKNOWLEDGEMENTS

I have profited from discussions with many colleagues at the IBM T. J. Watson Research Center: in particular, Drs. I. Adiri, H.A. Anderson, H.F. Silverman and P.C. Yue. I would also like to thank Prof. S. Kimbleton of the University of Michigan, Dr. Y. Bard of the IBM Cambridge Scientific Center and anonymous referees for their comments and suggestions given to the original manuscript of the present paper.

REFERENCES

1. E. Fuchs and P.E. Jackson, "Estimates of Distributions of Random Variables for Certain Computer Communications Traffic Models", CACM, Vol. 13, No. 12, pp. 752-757. (1970).
2. P.A.W. Lewis and P.C. Yue, "Statistical Analysis of Series of Events in Computer Systems", in W. Freiberger (ed), Statistical Computer Performance Evaluation, pp. 265-280, Academic Press (1972)
3. H.A. Anderson and R. Sargent, "The Statistical Evaluation of the Performance of an Experimental APL/360 System". ibid. pp. 73-98. Also, H.A. Anderson, "An Empirical Investigation into Foreground-Background Scheduling for an Interactive Computing System" Ph.D Thesis, Syracuse Univ. (1972)
4. D.R. Cox and P.A.W. Lewis, The Statistical Analysis of Series of Events, Methuen, (1966)
5. E.S. Walter and V.L. Wallace, "Further Analysis of a Computing Center Environment", CACM, Vol. 10, No. 5, pp. 266-272, (1967)
6. D.R. Cox and W.L. Smith, Queues, Methuen, (1961)
7. L. Kleinrock, Queueing Systems: Theory and Applications, Wiley, (1972)
8. R.L. Mattson, J. Gecsei, D.R. Slutz and I.L. Traiger, "Evaluation Techniques for Storage Hierarchies", IBM Sys. Jour., Vol. 9, No. 2, pp. 78-117, (1970)
9. P.A.W. Lewis and P.C. Yue, "Statistical Analysis of Program Reference Patterns in a Paging Environment", IEEE Conf. Computer Soc., Boston, P. 133, (1971).
10. G.S. Shedler and C. Tung, "Locality in Page Reference Strings", IBM Res. Rpt. RJ-932, (1971)
11. P.J. Denning, "The Working Set Model for Program Behavior", CACM, Vol. 11, No. 5, pp. 323-333, (1968)
12. P.J. Denning and S.C. Schwartz, "Properties of the Working Set Model", CACM, Vol. 15, No. 3, pp. 191-198, (1972)
13. E.G. Coffman and T.A. Ryan, "A Study of Storage Partitioning Using a Mathematical Model of Locality ", CACM, Vol. 15, No. 3, pp. 185-190. (1972)
14. D.R. Cox and H.D. Miller, The Theory of Stochastic Processes, John Wiley and Sons, (1965)
15. G.E. Uhlenbeck and L.S. Ornstein, "On the Theory of the Brownian Motion", Physical Review, Vol. 36, No. 3, pp. 823-841. (1930) Also in N. Wax (ed.) Selected Papers on Noise and Stochastic Processes, Dover Publications (1954)
16. U. Grenander and R.F. Tsao, "Quantitative Methods for Evaluating Computer System Performance: A Review and Proposals", in Statistical Computer Performance Evaluation, Academic Press, pp. 3-24 (1972)
17. M.G. Kendall and A. Stuart, The Advanced Theory of Statistics, Vol. 3: Design and Analysis and Time Series, Charles Griffin, (1966)
18. W.C. Guenther, Analysis of Variance, Prentice-Hall, (1964)
19. R.F. Tsao, L.W. Comeau and B.H. Margolin, "A Multifactor Paging Experiment" in Statistical Computer Performance Evaluation, Academic Press, pp. 103-134, (1972)
20. N.R. Draper and H. Smith, Applied Regression Analysis, John Wiley & Sons. (1966)
21. H.F. Silverman and P.C. Yue, "The Response Time Characterization of an Information Retrieval System' IBM Res. Rpt. RC-3796 (1972)
22. Y. Bard, "Performance Criteria and Measurement for a Time-sharing System", IBM Sys. Jour., Vol. 10, No. 3, pp. 193-216. (1971)
23. E.G. Coffman and P.J. Denning, Operating Systems Theory, Prentice-Hall, (1972)
24. J.P. Buzen, "Queueing Network Models of Multiprogramming", Ph.D. Thesis Harvard Univ. (1971)
25. C.G. Moore, "Network Models for Large-Scale Time-sharing Systems" Ph.D. Thesis, Univ. of Michigan, (1971)
26. S.R. Arora and A. Gallo, "The Optimal Organization of Multiprogrammed Multi-level Memory", Proc. of ACM-SIGOPS Workshop on System Performance Evaluation, pp. 104-141. (1971)
27. H. Tanaka, "An Analysis of On-Line Systems Using Parallel Cyclic Queues", Trans. of Inst. of Elec. and Comm. Engrs. of Japan, Vol. 53-C, No. 10, pp. 756-764, (1970)
28. I. Adiri, "Queueing Models for Multiprogrammed Computers", IBM Res. Rpt. RC-3802, (1972)
29. A. Chang and S.S. Lavenberg, "Work-Rates in Closed Queueing Networks with General Independent Servers", IBM Res. Rpt. RJ-989, (1972)
30. J.R. Jackson, "Jobshop-Like Queueing Systems", Management Science, Vol. 10, No. 1, pp. 131-142. (1963)
31. W.J. Gordon and G.F. Newell, "Closed Queueing Systems With Exponential Servers", Oper. Res. Vol. 15, No. 2, pp. 254-265. (1967)
32. M. Posner and B. Bernholtz, "Closed Finite Queuing Networks with Time Lags and with Several Classes of Units", Oper. Res. Vol. 16, pp. 977-985. (1968)
33. F. Baskett, "The Dependence of Computer System Queues upon Processing Time Distribution and Central Processing Scheduling". Proc. of ACM SIGOPS Third Symp. pp. 109-113. (1971)

34. K.M. Chandy, "The Analysis and Solutions for General Queuing Networks", presented at the Sixth Annual Princeton Conference on Information Sciences and Systems, Princeton University, New Jersey, March 1972
35. L. Kleinrock, R.R. Muntz and J. Hsu, "Tight Bounds on the Average Response Time for Time-shared Computer Systems", Presented at IFIP Congress 71. Ljubljana, August 1971, Booklet TA-2, pp. 50-58
36. A. Sekino, "Response Time Distribution of Multiprogrammed Time-Shared Computer Systems" presented at the Sixth Annual Princeton Conference on Information Sciences and Systems, Princeton University, New Jersey, March 1972
37. J.F.C. Kingman, "The Heavy Traffix Approximation in the Theory of Queues", Chapter 6 of Proceedings of the Symposium on Congestion Theory, edited by W.L. Smith and W.E. Wilkinson, The University of North Carolina Press, 1965, pp. 137-169
38. G.F. Newell, "Approximation Methods For Queues With Application to the Fixed-Cycle Traffic Light", SIAM Review, Vol. 7, No. 2, April 1965
39. D.P. Gaver, "Diffusion Approximations and Models for Certain Congestion Problems", Journal of Applied Probability, Vol. 5, 1968, pp. 607-623
40. D.P. Gaver and G.S. Shedler, "Multiprogramming System Performance Via Diffusion Approximations", IBM Research Report, RJ-938, November 1971
41. G.F. Newell, Applications of Queueing Theory, Chapman and Hall Ltd, London, 1971
42. H. Kobayashi, "Applications of the Diffusion Approximation to Queuing Networks", to appear as IBM Research Report, July 1972
43. W.J. Doherty, "Scheduling TSS/360 for Responsiveness", Proceedings of 1970 Fall Joint Computer Conference, Vol. 37, pp. 97-111
44. D.J. Hatfield and J. Gerald, "Program Restructuring for Virtual Memory" IBM Systems Journal, Vol. 10, No. 3, 1971, pp. 168-192
45. R.W. Wolff, "Work-Conserving Priorities", Journal of Applied Probabilities, Vol. 7, 1970, pp. 327-337
46. H. Kobayashi and H.F. Silverman, "Some Dispatching Priority Schemes and Their Effects on Response Time Distribution, Part I", IBM Research Report RC-3584, T.J. Watson Research Center, Oct. 1971
47. D.P. Gaver, "A Waiting Line with Interrupted Service, including Priorities" J. Roy. Statis. Soc. {B} Vol. 24, No. 1, 1962, pp. 73-90
48. J.M. Hammersley and D.C. Handscomb, Monte Carlo Methods, Methuen and Co., Ltd., London, 1964, Chapter 5
49. D.P. Gaver and G.S. Shedler, "Control Variable Methods in the Simulation of a Model of Multiprogrammed Computer System", IBM Research Report RJ-818, Feb. 1971

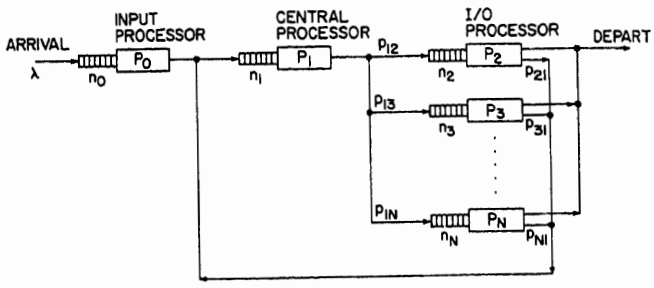

Figure 1. Queuing Network Model

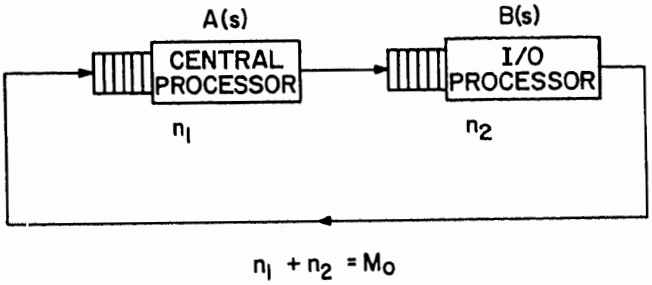

Figure 2. Cyclic Queue Model

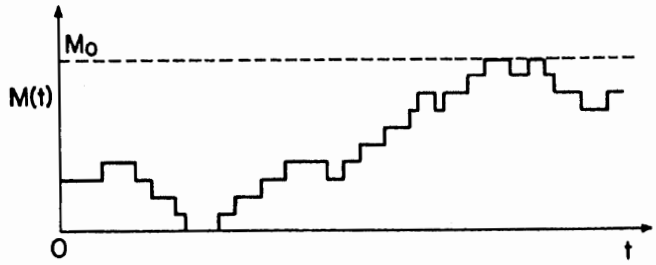

Figure 3. A Typical Behavior of M(t)