HOPKINS COMPUTER RESEARCH REPORTS

REPORT #20

AUGUST 1972

MULTI-SERVER QUEUING MODELS OF
MULTIPROCESSING SYSTEMS

BY

RENAUD REGIS

RESEARCH PROGRAM IN COMPUTER SYSTEMS ARCHITECTURE

COMPUTER SCIENCE PROGRAM

THE JOHNS HOPKINS UNIVERSITY

BALTIMORE, MARYLAND

MASTER

RECEIVED BY TIC NOV 27 1972

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Multi-server Queueing Models of Multiprocessing
Systems*

by Renaud Regis

Mailing Address:

Dr. Renaud Regis
c/o Mrs. Carol Rogers
Computer Science Dept.
The Johns Hopkins University
Baltimore, Maryland 21218

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Atomic Energy Commission, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

*This work was supported in part by the U. S. Atomic Energy Commission
under contract AT (11-1 3288

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

ABSTRACT

Conventional time-sharing and multiprogramming systems have been extensively modeled as single-server queues. In contrast, multiprocessing systems must be modeled as multi-server queueing systems.

In this paper, a general stochastic approach is proposed as a uniform basis for modeling multiprocessing systems and assessing their performance, under broad assumptions.

The performance is evaluated in terms of global measures; the utilization factor of the servers, as a measure of internal efficiency, and the mean response time, as a measure of external (user oriented) performance.

Of special interest is the influence of the queue discipline on these measures - a controversial issue even in single-server systems - under very general assumptions: arbitrary arrival process and arbitrary service times distribution. We distinguish two classes of queue disciplines: preemptive and non-preemptive; and two types of arrival processes: dependent (on the state of the queue) and independent.

It is shown that, non-preemptive disciplines leave both measures invariant, for given, but arbitrary (dependent or not) arrival process and arbitrary service distribution. For preemptive disciplines, the utilization factor is invariant only if the arrival process is independent, but the mean response time depends on the discipline, whatever the arrival process. It is also shown that the exponential distribution is the only distribution for which the queue discipline does not affect the performance of the system, thus underlining the very peculiar nature of this standard assumption and preventing us from safely generalizing the conclusions reached under this assumption.

The so-called processor-sharing system is investigated in conjunction with arbitrary service time distributions and it is shown that limiting queue length is the same as in a FIFO system with exponentially distributed service times - a result which has been conjectured by several authors for various distributions.

I. Introduction

A great variety of queueing theoretic models of computer systems have been proposed in the last few years. Yet, due to the restrictiveness of the basic assumptions and the diversity of the operating conditions, much confusion still exists about the validity of these results. Such is the case for the influence of the scheduling strategy on the system's performance. In this paper we try to clarify this issue and characterize the effect of the queue discipline, considered as a parameter, under the least restrictive (section 3) assumption. Most of the proposed models are variations of a single server queue subjected to Poisson arrival process and exponential service distribution*. Here, we consider multi-server queues and arbitrary distributions. The exponential assumption is not very realistic: measurements have shown that "real" service distributions tend to be either more or less skewed than the exponential, which has a coefficient of variation of 1.

Yet much more critical is the memoryless property of that distribution. (It should be emphasized that the negative exponential distribution is the only continuous distribution to have this property, while there are "many" distributions with a coefficient of variation of 1). This property is heavily used in the analytical derivations of the transient as well as the limiting distributions of the basic stochastic processes, and prevents us from applying this modeling technique to more general situations.

*Although some consideration has been recently given to networks of queues[1] and rational distributions[2], these models essentially resort of the same techniques.

Further, as we shall see, the exponential assumption induces into the system certain behavioral properties which cannot be generalized.

Thus, more powerful techniques are needed. Such a technique is presented in section 4 and its practicality and generality are demonstrated. In particular it enables us to formally establish, under very general assumptions, a result proved in isolated cases[3],[4], namely that the so-called processor-sharing discipline essentially recreates the memoryless property of the exponential distribution (section 5). Finally, in section 6 we consider the influence of the service distribution on the performance of the system.

II. The General Model

The general model is defined by the following assumptions:

- Requests join a single queue. Their inter-arrival times constitute a stationary process, although they are not necessarily independent, nor identically distributed. Their distribution will be specified whenever necessary. Unless otherwise indicated, we assume a "no-loss" system, in which every request joins the queue. We shall use the abbreviation i.i.d. for independent, identically distributed random variables.
- The queue is served by s identical servers. A request can be served by at most one server at a time. When a server is freed, it immediately starts serving the next request in the queue, if any.
- The service requirements of the requests are arbitrarily distributed and independent of the interarrival times.

We assume further that during the period of observation there is no external interruption of the servers (due, for instance, to a hardware failure). Such events would only degrade the overall performance of the system but would not alter the essential results.

Queue disciplines.

The queue discipline specifies the order in which the requests are considered for service and the amount of continuous service an individual server gives to a particular request before switching to the next request. We distinguish two types of queue disciplines:

- non-preemptive: the server continuously serves a request up to completion (natural interruption).
- preemptive: service of the current request may be interrupted before it is completed (artificial interruption). In that case, the interrupted request is recycled (feedback) for later processing.

With a preemptive-resume discipline, processing will resume at the point of interruption, whereas with a preemptive-repeat discipline service will be given from the beginning.

Preemptive-repeat disciplines are seldom used, however, because of the work loss involved. We will therefore consider only preemptive-resume disciplines and, in the following, the term preemptive will be used in lieu and place of preemptive-resume.

We assume that the overhead involved in switching from request to request is either negligible or already taken into account in the service distribution.

The quantum size, i.e. the amount of time allocated to a request before an artificial interrupt occurs, may be a constant or a random variable (possibly depending on the state of the system).

The queue discipline is an important system parameter because it is that over which we have probably the most direct control. The problem of selecting a proper strategy (sometimes referred to as the job scheduling problem) tends to be the central issue in operating system design. It has been studied by many in the context of uniprocessor multiprogramming and

time-sharing systems under the exponential assumption ([3], [4], etc.). In the more general context of multi-server systems and in conjunction with arbitrary service distributions, we will see that this issue takes on an accrued significance.

Performance measures.

In models of time-sharing systems, the performance was traditionally measured by the response times given to particular classes of requests (i.e. conditioned on some variable). Yet, the (conditional) response time does not completely characterize the dynamic behavior of a system. In this paper, we consider more global measures: the utilization factor of the servers, as a measure of internal efficiency, and the mean response time, averaged over the entire population of requests, as a measure of external performance. This choice is motivated by the following consideration. The utilization factor of the resources is directly related to the cost of a system and is therefore particularly critical in multi-server organizations. Further, giving preferential treatment to certain requests may in fact degrade the global performance of the system. It is therefore important to be able to evaluate the cost of such a strategy.

The utilization factor, d , is defined as the average proportion of busy servers, at statistical equilibrium; that is, the proportion of time spent in actual processing.

The mean response time, w , is the average total time spent in the system by a request, from arrival to completion, at statistical equilibrium. It is sometimes referred to as the mean total waiting time.

III. Effect of the Queue Discipline

The impact of the queue discipline depends essentially on whether it is preemptive or not. But it also depends on the arrival process. We

distinguish two types of arrivals:

- independent arrival processes which are independent of the state of the queue. Note, however, that we do not place any restriction on the stochastic relation between successive inter-arrival times: they could well be stochastically dependent.
- dependent arrival processes which depend in some way on the state of the queue. These occur, in particular, in the study of closed and mixed networks of multi-server queues. They also model the arrival to finite waiting room queues.

Theorem 1

For a given but arbitrary arrival process and given arbitrarily but independently, identically distributed service times, the stationary queue length process, when it exists, is invariant with respect to the queue discipline, as long as it is non-preemptive.

Proof

From the point of view of the servers, all requests are identical, since their service times are independent identically distributed, and service is given up to completion. Therefore the departure process $D(t)$ is independent of the order in which the requests are serviced. For an independent arrival process $A(t)$, it is then obvious that the queue length process $Q(t) = \max(0, A(t) - D(t))$ is independent of the queue discipline. For a dependent arrival, the proof will be given in Section 4 when we develop adequate analytical tools.

This result was recognized by Takacs[5] in the case of a single server queue subjected to an independent Poisson arrival process and an arbitrary service distribution (M/G/1). It was proved here in the more general case of a multi-server system with both arbitrary arrival and service distributions. In particular, the inter-arrival times need not be i.i.d.

Corollary 1.

Under the conditions of Theorem 1, the utilization factor and the mean response time are also independent of the queue discipline.

Proof.

Since the limiting queue length distribution is independent of the (non-preemptive) discipline, so is the number of busy servers and therefore, the utilization factor. Further, the expected queue length is a fortiori invariant and, by Little's formula[6], $E(Q) = \lambda w$ (where $\frac{1}{\lambda}$ is the average inter-arrival time) so is w , the mean response time.

With Kleinrock, who proved the invariance of the mean response time in the M/G/1 system[7], we can say that, although some (non-preemptive) disciplines may give preferential treatment to certain requests, the average waiting time is conserved. We shall see that this is not so for preemptive disciplines or if the service requirements are not i.i.d. When we consider preemptive disciplines, Theorem 1 does not hold anymore. Yet, we can establish that:

Theorem 2.

For a given but arbitrary independent arrival process and arbitrarily (not necessarily identically) distributed service times, the utilization factor is invariant with respect to the queue discipline, although the limiting distribution of the queue length, and a fortiori, the mean response time, will in general depend on the (preemptive) discipline.

Proof of Theorem 2.

Consider first the uni-server case. Without loss of generality, suppose $t=0$ is the beginning of a busy period and let $L(t)$ be remaining processing load at time t , i.e. the amount of service still to be given as of time t . $L(t)$, a random variable, decreases at a unit rate as long as it is

Figure 1: A Realization of $L(t)$.

positive, remains at 0 until the next arrival and jumps by an amount equal to the service requirement at every instant of arrival of a new request. A typical realization of $L(t)$ is depicted in figure 1. $L(t)$ is clearly independent of the queue discipline, since the times between successive jumps (inter-arrival times), the amounts of the discontinuities (service requirements) and the (constant) rate of decrease are all independent of the strategy. So are the lengths of the busy period T_1 and the busy cycle Z_1 , since by definition.

$$T_1 = \inf\{t | L(t) \leq 0\} \text{ and } Z_1 = \inf\{t > T_1 | L(t) > 0\},$$

and therefore the utilization factor.

In the multi-server case, the proof is slightly more involved. We shall prove that the limiting distribution of the number of busy servers is independent of the strategy, although the limiting distribution of the queue length may not be.

Let t_n be the successive instants of "full-employment" i.e.

$$t_n = \min\{t > v_{n-1} | Q(t) \geq s\}$$

v_n the successive instants of "partial-employment", i.e.

$$v_n = \min\{t > t_n | Q(t) \leq s-1\}$$

$V_n = v_n - t_n$ the n^{th} full-employment period, and

$T_n = t_{n+1} - v_n$ the consecutive partial-employment period.

An argument similar to the uni-server case shows that the length of the n^{th} full-employment period is independent of the discipline, although it depends on the state of the system (more precisely, on the remaining processing load) at the end of the previous partial-employment period (t_n^-). During a full-employment period, the rate of decrease of $L(t)$ is constant (equal to s); further the instants of full-employment are also instants of arrival of new requests, thus V_n depends only on $L(t_n^-)$.

Figure 2: A Realization of $L(t)$ for $s-2$.

Figure 3: The corresponding realization of $N(t)$.

During a partial-employment period, if a request is preempted, it will be taken over immediately by a free server (there is at least one free server in such a period). Since the servers are identical, we can consider that the request is served continuously by the same server; thus service is given in the order of arrival and up to completion until the end of the period. Therefore, neither the number of busy servers, nor the length of the partial-employment period, nor the remaining processing load at the end of the period, $L(t_n^-)$ depends on the strategy, since, besides the arrival process is itself independent of the latter.

Juxtaposing both periods, it is then clear that the number of busy servers process is independent of the queue disciplines, implying the independence of d . QED.

Figure 2 shows a typical realization of the load process in the case $s=2$ while figure 3 shows the corresponding realization of the busy servers process $N(t)$.

It should be emphasized that the steady-state distribution of the queue length (and a fortiori of the waiting time) will in general be affected by the queue discipline. Only the utilization factor is invariant. We shall see in section 6 that even in the simple M/G/1 system the queue process does depend on the queue discipline.

However, if the service distribution is exponential, the queue length process is independent of the queue discipline, even if it is preemptive. This is due to the memoryless property of this distribution. We show in Appendix A that, under the conditions of theorem 2, the memory less property is a necessary and sufficient condition for theorem 1 to extend to preemptive disciplines. This already underlines the special role played by the exponential assumption in the behavior of multi-server systems. We will see in the next section other implications of this assumption.

In the proof of theorem 2, we did not require that the service times be identically distributed. In particular, theorem 2 applies to priority queueing systems in which each priority class has its own arbitrary service requirement distribution, G_i , $i = 1, \dots, p$. Since it is true of preemptive disciplines, it is a fortiori true of non-preemptive disciplines. However, theorem 1 cannot be so generalized: the queue length process will in general depend on the queue discipline if the service requirements are not i.i.d., even for non-preemptive disciplines and independent arrival processes. Therefore, the mean response time will in general depend on the discipline in this case.

On the other hand, theorem 2 does not apply to dependent arrivals even if the service times are i.i.d. Thus we have:

Theorem 3.

For dependent arrivals and preemptive disciplines, the utilization factor (and a fortiori, the queue length process and the mean response time) will depend in general on the discipline.

We shall see in section 6 that this is indeed the case for the $M_d/G/1$ system, where M_d denotes a state-dependent birth arrival process.

From a practical point of view, the implication is that, if the system at hand is subjected to an independent arrival, the optimization of secondary objectives (e.g., response-time, overhead etc.) will not degrade the efficiency of the system; but, if the arrival is dependent - and usually this is the case - the optimization will be done at the expense of the system's efficiency. It is then important to evaluate the "cost", in terms of the performance measures of enforcing particular objectives, so as to determine a meaningful trade-off.

Note, also, that, for dependent arrivals, it is necessary that the service requirements be independent and identically distributed for d

to be invariant, even with a non-preemptive discipline. In particular, in a priority system with different service requirements and a restricted arrival, the order in which requests are given service does affect the overall performance of the system.

There is, however, a noticeable exception to Theorem 3, namely when the service requirements are exponentially (and identically) distributed. Due to the memoryless property, whenever a request is preempted, its remaining service requirement has the same exponential distribution, regardless of the number of times it was recycled. Thus from the point of view of the servers, all requests in the queue are identical (whether recycled or new) and, whatever the queue discipline, the queue length process and, a fortiori, the arrival process, will not be affected. It can be shown that this is the only exception to theorem 3.

Once again, we see the impact of the exponential assumption. But here it is even more dramatic than previously, since it affects all of the performance measures of restricted models, while it only affected the mean response time in the unrestricted case.

Thus the exponential assumption is not a "harmless", mere simplification of the models, but in fact "hides" from us some of the most fundamental aspects of the dynamic behavior of a system. The properties of exponential models appear to be direct consequences of the exponential assumption, rather than structural properties of the model, and it is therefore dangerous to extrapolate these properties to other systems.

To summarize briefly, the queue discipline is an extremely important optimization parameter of systems subjected to dependent arrivals, while, for independent arrivals, it affects only partially the performance. Further, non-preemptive disciplines leave invariant both performance measures, whereas preemptive strategies in general affect both.

Thus, we need now to evaluate quantitatively this effect. Yet, because of the particular properties of the exponential distribution, the techniques developed under this assumption are no longer adequate and more powerful techniques must be devised. In the following section, such a technique is presented.

IV. Analytical Solution of the General Model

We will assume in the following that the arrival process is a state dependent birth process[8], i.e. the probability of an arrival between t and $t+\zeta$, given that the system is in state k at time t , is $\lambda_k \zeta + o(\zeta)$. The state of the system will in general be described by the queue length and/or the number of busy servers.

In particular, we can model and evaluate the effect of constraints on the capacity of the system: suppose that the queue can accomodate a maximum of K requests (waiting or in service). In general, when the queue reaches its maximum capacity, the source(s) will be prevented from issuing a new request. The actual arrival process to the queue is then described by:

$$\lambda_0 = \lambda_1 = \dots = \lambda_{K-1} = \lambda, \quad \lambda_K = 0$$

Another important measure in such a situation is the probability of overflow. It is shown in Appendix D that the well-known Erlang's loss formula for exponential service generalizes to arbitrary service, and differs noticeably from that derived in the unconstrained case.

Another situation of importance, previously discussed in the case of exponential service, is one in which requests are initiated by a finite number n of sources; in this case $\lambda_k = (n-k)\lambda$.

Because of the arbitrary service distribution, the basic stochastic processes of the models studied here (restricted or unrestricted) are not markovian.

One possible approach to study non-markovian processes is to try and isolate an embedded Markov chain (see Kendall[16]). However this approach presents two difficulties:

First, the determination of regeneration points (at which the process is markovian) may not be a trivial problem.

Secondly, and more important, the embedded process usually does not have the same limiting distribution than the "true" process; in particular the distribution of the number of busy servers is usually different (see [9]) and can no longer be directly related to the busy period analysis.

An alternate approach, which does not present these problems, is to study an "augmented process" which is then markovian. This technique is, of course, not limited to the models of interest here and could be successfully applied to the modeling of a great variety of situations.

As an illustrative example, we apply it now to the restricted models as defined above.* For the sake of simplicity, we assume that the arrival rates depend on the total queue length. A more involved case would be treated in a similar manner. The service times are i.i.d. according to the arbitrary distribution G.

We first consider a non-preemptive discipline.

Let $\xi_i(t)$ $i=1, \dots, s$ denote the "expended" service time of the request being served by the i^{th} server at time t , i.e. the total amount of service received by that request up to time t , and $Q(t)$ the queue length at time t .

If the i^{th} server is idle, $\xi_i(t) = 0$

Consider the augmented process

$$X(t) = \{Q(t); \xi_1(t), \dots, \xi_s(t)\}$$

*Unrestricted models are a special case, namely $\lambda_k = \lambda, \forall k$

It is clear that the augmented process is markovian, since the arrival process is an homogenous birth process.

Let $P_k(t; x_1, \dots, x_k)$ for $0 \leq k \leq s-1$ and $P_k(t; x_1, \dots, x_s)$ $k \geq s$ be the joint cumulative distribution function of $X(t)$ and $P_k(t_1, x_1, \dots, x_s)$ its density function.

I.e. $P_k(t; x_1, \dots, x_k) = \Pr\{Q(t)=k, \xi_{i_1}(t) \leq x_1, \dots, \xi_{i_k}(t) \leq x_k\}$ where $\{i_1, \dots, i_k\}$ is particular combination of the s servers.

Although this is not essential to the derivation of the steady-state probabilities, it is convenient to consider the system as being symmetrical, i.e. when $k < s$, an incoming request has equal probability to be served by any of the $(s-k)$ free servers.

With this convention and the notation:

$P\{(t; k; x_1 \dots x_k) | (\zeta; j; x_1 \dots x_j)\}$ standing for the conditional probability that $X(t) = \{k; x_1 \dots x_k\}$ given that $X(\zeta) = \{j; x_1 \dots x_j\}$ the transition probabilities are:

$$P\{(t+\Delta t; 0) | (t; 0)\} = 1 - \lambda_0 \Delta t + o(\Delta t) \quad (\text{no arrival})$$

$$\begin{aligned} P\{(t+\Delta t; k-1; x_1+\Delta t, \dots, x_{k-1}+\Delta t) | (t; k; x_1 \dots x_k)\} \\ = \frac{G(x_k+\Delta t) - G(x_k)}{1 - G(x_k)} \prod_{i=1}^{k-1} \frac{1 - G(x_i+\Delta t)}{1 - G(x_i)} + o(\Delta t) \quad 0 < k \\ (\text{one departure, no arrival}) \end{aligned}$$

$$\begin{aligned} P\{(t+\Delta t; k+1; x_1+\Delta t, \dots, x_k+\Delta t, 0) | (t; k; x_1 \dots x_k)\} \\ = \prod_{i=1}^k \frac{1 - G(x_i+\Delta t)}{1 - G(x_i)} \frac{\lambda_k}{s-k} \Delta t + o(\Delta t) \quad 0 < k \leq s-1 \\ (\text{one arrival, no departure}) \end{aligned}$$

$$\begin{aligned} P\{(t+\Delta t; k+1; x_1+\Delta t, \dots, x_s+\Delta t) | (t; k; x_1 \dots x_s)\} \\ = \prod_{i=1}^k \frac{1 - G(x_i+\Delta t)}{1 - G(x_i)} \lambda_k \Delta t + o(\Delta t) \quad k \geq s \end{aligned}$$

$$P\{(t+\Delta t; k; x_1+\Delta t, \dots, x_{\min(k,s)}+\Delta t) | (t; k; x_1 \dots x_{\min(k,s)})\}$$

$$= [1 - \lambda_k \Delta t + o(\Delta t)] \prod_{i=1}^k \frac{1-G(x_i+\Delta t)}{1-G(x_i)} \quad 0 < k$$

(no arrival, no departure)

$$P\{(t+\Delta t; k-1; x_1+\Delta t, \dots, x_{s-1}+\Delta t, 0) | (t; k; x_1 \dots x_k)\}$$

$$= \frac{G(x_s+\Delta t) - G(x_s)}{1-G(x_s)} \prod_{i=1}^{s-1} \frac{1-G(x_i+\Delta t)}{1-G(x_i)} + o(\Delta t) \quad k > s$$

The other transitions either are impossible or have a probability of $O(\Delta t)$.

Since the process is markovian, it must satisfy the Chapman-Kolmogorov equations, i.e. the density functions p_k must satisfy the equations derived from these (see Feller[10] Chapter X), which are the counterpart, in continuous time, of the Kolmogorov equations for a Markov chain.

(1.a)

$$k=0 \quad p_0(t+\Delta t) = p_0(t)(1 - \lambda_0 \Delta t) + s \int_0^\infty p_1(t; x) \frac{G(x_1+\Delta t) - G(x_1)}{1-G(x_1)} dx_1 + o(\Delta t)$$

$$0 < k < s \quad p_k(t+\Delta t; x_1+\Delta t, \dots, x_k+\Delta t) = p_k(t; x_1 \dots x_k) (1 - \lambda_k \Delta t) \prod_{i=1}^k \frac{1-G(x_i+\Delta t)}{1-G(x_i)} \\ + (s-k) \int_0^\infty p_{k+1}(t; x_1 \dots x_k, x_{k+1}) \frac{G(x_{k+1}+\Delta t) - G(x_{k+1})}{1-G(x_{k+1})} \prod_{i=1}^k \frac{1-G(x_i+\Delta t)}{1-G(x_i)} dx_{k+1} + o(\Delta t)$$

$$k=s \quad p_s(t+\Delta t; x_1+\Delta t, \dots, x_s+\Delta t) = p_s(t; x_1 \dots x_s) \prod_{i=1}^s \frac{1-G(x_i+\Delta t)}{1-G(x_i)} (1 - \lambda_s \Delta t) + o(\Delta t)$$

$$k > s \quad p_k(t+\Delta t; x_1+\Delta t, \dots, x_s+\Delta t) = p_k(t; x_1 \dots x_s) \prod_{i=1}^s \frac{1-G(x_i+\Delta t)}{1-G(x_i)} (1 - \lambda_k \Delta t) \\ + p_{k-1}(t; x_1 \dots x_s) \prod_{i=1}^s \frac{1-G(x_i+\Delta t)}{1-G(x_i)} \lambda_{k-1} \Delta t + o(\Delta t)$$

and

(1.b)

$$0 < k \leq s \quad p_k(t+\Delta t, x_1+\Delta t, \dots, x_{k-1}+\Delta t, 0) dx_k$$

$$= p_{k-1}(t; x_1 \dots x_{k-1}) \prod_{i=1}^{k-1} \frac{1-G(x_i+\Delta t)}{1-G(x_i)} \frac{\lambda_{k-1}}{s-k+1} \Delta t + o(\Delta t)$$

$$k=s \quad p_s(t+\Delta t; x_1+\Delta t, \dots, x_{s-1}+\Delta t, 0) dx_s$$

$$= p_{s-1}(t; x_1 \dots x_{s-1}) \prod_{i=1}^{s-1} \frac{1-G(x_i+\Delta t)}{1-G(x_i)} \lambda_{s-1} \Delta t + o(\Delta t)$$

$$+ \int_0^\infty p_{s+1}(t; x_1 \dots x_s) \prod_{i=1}^{s-1} \frac{1-G(x_i+\Delta t)}{1-G(x_i)} \frac{G(x_s+\Delta t)-G(x_s)}{1-G(x_s)} dx_s + o(\Delta t)$$

$$k > s \quad p_k(t+\Delta t; x_1+\Delta t, \dots, x_{s-1}+\Delta t, 0) dx_s$$

$$= \int_0^\infty p_{k+1}(t; x_1 \dots x_s) \prod_{i=1}^{s-1} \frac{1-G(x_i+\Delta t)}{1-G(x_i)} \frac{G(x_s+\Delta t)-G(x_s)}{1-G(x_s)} dx_s + o(\Delta t)$$

$$\text{Let } \Pi_k(t; x_1 \dots x_k) = \frac{p_k(t; x_1 \dots x_k)}{\prod_{i=1}^k [1-G(x_i)]} \quad \text{with } \Pi_0(t) = p_0(t)$$

(1a) and (1b) can be rewritten in terms of the Π 's and G only. Taking the limit as $\Delta t \rightarrow 0$ and observing that

$$\lim_{\Delta t \rightarrow 0} \int_0^\infty \Pi_k(t; x_1 \dots x_k) \frac{G(x_k+\Delta t)-G(x_k)}{\Delta t} dx_k = \int_0^\infty \Pi_k(t; x_1 \dots x_k) dG(x_k)$$

since Π_k is a continuous, bounded function of x_k and G is a distribution, we obtain

2.a

$$\frac{\delta \pi_0}{\delta t} + \lambda_0 \pi_0 = s \int_0^\infty \pi_1(t; x_1) dG(x_1) \quad k=0$$

$$\frac{\delta \pi_k}{\delta t} + \sum_{i=1}^k \frac{\delta \pi_k}{\delta x_i} + \lambda_k \pi_k = (s-k) \int_0^\infty \pi_{k+1}(t; x_1 \dots x_k, x_{k+1}) dG(x_{k+1}) \quad 0 < k < s$$

$$\frac{\delta \pi_s}{\delta t} + \sum_{i=1}^s \frac{\delta \pi_s}{\delta x_i} + \lambda_s \pi_s = 0 \quad k=s$$

$$\frac{\delta \pi_k}{\delta t} + \sum_{i=1}^s \frac{\delta \pi_k}{\delta x_i} + \lambda_k \pi_k = \lambda_{k-1} \pi_{k-1} \quad k > s$$

2.b

$$\pi_{k+1}(t; x_1 \dots x_{i-1}, 0, x_{i+1}, \dots, x_{k+1}) = \frac{\lambda_k}{s-k} \pi_k(t; x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_{k+1}) \quad 0 \leq k \leq s-2$$

$$\pi_s(t; x_1 \dots x_{i-1}, 0, x_{i+1}, \dots, x_s) = \lambda_{s-1} \pi_{s-1}(t; x_1 \dots x_{i-1}, x_{i+1}, \dots, x_s) + \int_0^\infty \pi_{s+1}(t; x_1 \dots x_s) dG(x_1)$$

$$\pi_k(t; x_1 \dots x_{i-1}, 0, x_{i+1}, \dots, x_s) = \int_0^\infty \pi_{k+1}(t; x_1 \dots x_s) dG(x_i) \quad k > s+1$$

The functions $\pi_k(t; x_1 \dots x_k)$ that we are seeking are those solutions of the integro-differential system (2a) which satisfy the boundary conditions (2b). From these, the transient behavior of the process follows.

To obtain the limiting distribution of the process, we solve (2) for functions independent of t . Let $\pi_k^*(x_1 \dots x_k)$ be the solution.

The limiting joint density is then given by

$$p^*(x_1 \dots x_k) = \pi_k^*(x_1 \dots x_k) \prod_{i=1}^k [1 - G(x_i)]$$

from which we derive the limiting distribution of the queue length:

3.

$$p_0^* = \pi_0^*$$

$$p_k^* = \binom{s}{k}^+ \int_0^\infty \dots \int_0^\infty \pi_k^*(x_1 \dots x_k) \prod_{i=1}^k [1-G(x_i)] dx_1 dx_2 \dots dx_k$$

$$\text{and } \binom{s}{k}^+ = \begin{cases} \binom{s}{k} & 0 \leq k \leq s \\ 1 & k > s \end{cases}$$

The system of equations (2) has not yet been solved in the most general case. But, the simple form of equations (2) suggests that a numerical solution will be fairly straightforward.

It should be noted that at no point in the derivation of equations (2) did we have to be concerned with the order of service of the requests. Thus, equations (2) are valid whatever the non-preemptive queue discipline. Clearly, this result extends to arbitrary dependent arrival processes. The derivation of the steady-state equations is essentially similar: it suffices to augment the process $X(t)$ with a variable representing the time elapsed since the last arrival. This completes the proof of theorem 1.

On the other hand, equations (2) are still valid for preemptive disciplines for $0 \leq k \leq s$, if we assume, without loss of generality, that an interrupted request is immediately recycled to the same server (which is possible since $k \leq s$). However, when $k > s$, we know from the proof of theorem 2 that, during a full employment period, the queue length depends on the strategy. Since now the arrival process itself depends on the queue length, the length of the full employment period will depend on the strategy and therefore theorem 1 does not apply to dependent arrival models.

A significant property of the method presented in this section is that it constitutes a uniform and powerful basis for the analytical treatment of a great variety of models. Thus, the equations of a large

number of models which, if at all considered, have been solved through very different and ad-hoc techniques are contained in equations (2). In particular, the limiting distribution of the queue length of the M/G/1 system, first solved by Kendall [11] in the unrestricted case and by Takacs[12] in the restricted case, is derived from equations (2) in Appendices B and C respectively. The M/G/1 system is the basis of most studies of preemptive discipline in the Computer Science literature, in conjunction with exponential service requirement the foregoing results show that the preemptive discipline can be replaced by a FIFO non-preemptive strategy. Also in Appendix C, we show that the solution of the dependent arrival model studied by Baht[9] is a special case of equations (2) and we derive an analytical expression for the limiting distribution of the queue length, where Baht obtained only an iterative scheme for numerical solution. In that respect, equations (2) yield directly the limiting distribution of the process, whereas other techniques usually solve for its Laplace-Stieljes transform (LST). It is then a non-trivial problem to invert that transform and in general requires a numerical (as opposed to analytical) inversion.

Other less popular models can be straightforwardly solved by equations (2). For example, in Appendix D, we solve equations (12) for a capacity constrained multi-server system in which no actual waiting space is reserved: when all servers are busy, sources are prevented from issuing a new request (as it is lost). It is shown that Erlang's loss formula, established under the exponential assumption generalizes to arbitrary service requirements; this is an interesting result because the determination of the limiting distribution requires only the knowledge of the mean service requirements. In practical situations, it is often the case that the real service distribution is unknown, whereas its mean value can be evaluated fairly accurately by statistical estimations methods.

In the following sections, we discuss other models of special interest, whose equations are a special form of equations (2).

V. The Capacity-Sharing System.

A system, which has been considered by several authors under various particular assumptions is the so-called capacity-sharing system. The service station provides a constant amount of servicing power which is equally shared among the requests present in the system. Thus, if k requests are present, rather than waiting and being serviced one at a time, they are all taken into service simultaneously but each receive only a fraction of the total capacity: in a real time interval Δt , each request will receive $c_k \Delta t (c_k \leq 1)$ of processing time. There is of course a lower bound on the service time of a request, namely its service time when it is alone in the system, which we will take as the service requirement. We can then define the capacity of the service mechanism as the maximum number of requests which can be served simultaneously at their maximum service speed. Let s be the capacity. Then, obviously, the c_k 's must satisfy

(4.)

$$0 < c_k \leq 1, \quad k c_k \leq s$$

This hypothetical model has a substantial theoretical interest: it is the limiting case of a Round Robin discipline when the quantum size tends to zero and therefore constitutes a good approximation of its behavior when the quantum size is small compared to the mean service requirement. But, in addition to this theoretical interest, it also has a practical interest in the modeling of some generalized parallel processing architectures being currently investigated[13].

The essential difference between the multi-server system and the capacity-sharing system can be best understood from the point of view of

(a) Non-preemptive Discipline

(b) Preemptive Discipline

Figure 4: Realizations of the service rate in the multiserver system.

an individual request.

In the multi-server system, the service received by a request is in fact a binary-valued function of time: during some periods of time (service periods), a request will be serviced at the maximum rate (1) while during consecutive periods (waiting periods) it will not be serviced at all. The instantaneous rate of service is represented in figure 4 for a non-preemptive discipline (a) and for a preemptive discipline (b).

In the capacity sharing system, a request continuously receives service, although the instantaneous rate may not be constant: as long as a request needs service the rate is strictly positive. Figure 5 shows a typical realization of the service rate in this case.

The cumulated amount of service received by an individual request is depicted in figure 6. Note that for the capacity-sharing system it is strictly increasing.

Further, while in the multi-server system only one request at a time can receive service, with the capacity-sharing service mechanism, all requests present in the system receive service at the same rate.

Before deriving the limiting distribution of the queue length process in the general case, let us consider various combinations of the c_k 's.

The capacity-sharing system has been investigated essentially in the case $s=1$, $c_k = \frac{1}{k}$ ([14],[15] e.g.). However, other cases are interesting.

For $c_k = \begin{cases} 1 & k \leq s \\ \frac{s}{k} & k > s \end{cases}$, up to s request can be served simultaneously at their maximum speed (as if they were alone), while for $k > s$ the server's capacity is then equally shared by the requests.

We can also model a probably more realistic situation in which, beyond a certain point, interferences may develop that prevent the total capacity

Figure 5: Capacity Sharing System.

Figure 6: Cumulated Service

from being effectively used. In such a case $k c_k < s$ (say for $k > n$).

We now derive the limiting distribution of the queue length under the following assumptions:

- Birth arrival process with parameters λ_k
- Arbitrary service distribution with mean $(1/\mu)$
- Capacity-shared server with parameters c_k satisfying (4).

The solution of the capacity-sharing model is essentially similar to the previous technique. However, since individual requests each receive only a fraction $c_k \Delta t$ of service time Δt , Δt must now be replaced by $c_k \Delta t$ in the derivation of the transition probabilities.

Thus the steady-state equations are given by:

$$(5a.) \quad \lambda_0 \Pi_0 = c_1 \int_0^\infty \Pi_1(u) dG(u)$$

$$c_k \sum_{i=1}^k \frac{\partial \Pi_k}{\partial x_i} + \lambda_k \Pi_k = (k+1) c_{k+1} \int_0^\infty \Pi_{k+1}(x_1, \dots, x_k, u) dG(u)$$

$$(5b) \quad c_{k+1} \Pi_{k+1}(x_1, \dots, x_k, 0) = \frac{\lambda_k}{k+1} \Pi_k(x_1, \dots, x_k)$$

The solution is then:

$$\Pi_k^* = \Pi_0 \cdot \frac{1}{k!} \prod_{i=1}^k \left(\frac{\lambda_{i-1}}{c_i} \right)$$

and therefore :

$$P_k = P_0 \cdot \frac{1}{k!} \left(\frac{1}{\mu} \right)^k \prod_{i=1}^k \frac{\lambda_{i-1}}{c_i}$$

$$P_0 = \left[1 + \sum_{k=1}^{\infty} \left(\prod_{i=1}^k \frac{\lambda_{i-1}}{c_i} \right) \frac{1}{k!} \left(\frac{1}{\mu} \right)^k \right]^{-1}$$

*We do not have to specify the queue discipline since we know from theorem 1 that it is irrelevant for exponential distribution.

The limiting distribution has a remarkably simple form: it is identical to that of a birth and death process with parameters λ_k and $\mu k c_k$ (see, for instance, [8]).

In particular, if $c_k = \begin{cases} 1 & k \leq s \\ s/k & k > s \end{cases}$, the capacity-sharing system with an arbitrary service distribution is equivalent, from the point of view of the stationary queue length, to an s-server system subjected to the same (dependent) arrival and exponentially distributed service times with the same mean $(1/\mu)^*$. In fact, we conjecture that this equivalence is true whatever the arrival process.

This property of the capacity-sharing system has been independently reported in the case of (independent) Poisson arrival, rational service distributions and $s=1$ $c_k=1/k$ [14]. We see that it generalizes to many less restrictive models.

By Little's formula we know that the mean response time is also the same as in $M_d/M/s$. It is conjectured that the equivalence carries over to the limiting distribution of the total waiting time (not only its mean)

Thus, the capacity-sharing discipline essentially recreates the memoryless property of the exponential distribution. We shall now take a closer look at the implications of this property.

VI. Performances of Preemptive and Non-Preemptive Strategies.

Consider two uni-server queueing systems subjected to the same dependent arrival process (for simplicity, we take $\lambda_k = (n-k)$), the same arbitrary service requirement distribution G and differing only by the service discipline.

System #1 implements a non-preemptive service discipline. According to theorem 1, the queue length process is independent of the queue discipline and we can assume, without loss of generality, a First-in First-out

discipline. The limiting distribution of the queue length is derived in Appendix C.

System #2 implements a capacity-sharing service discipline which is assumed to be 100% effective, i.e. $c_k = \frac{1}{k} \psi_k$. The limiting distribution of the queue length directly follows from the results of the previous section.

Ample motivation for these systems has been given elsewhere, but it is interesting to note that they can also be interpreted as the two extreme cases of a Round Robin strategy when the quantum size is respectively infinitely large and infinitely small.

Various simulations[15] have indicated that the performance of the Round Robin strategy is a monotone function of the quantum size and lies between those of the two limiting situations.

Denoting the first system by NP and the second by CS, the queue length distributions are given by (see Appendix C):

$$P_{np}(v) = P_{np}(0) \sum_{k=1}^v (-1)^{v-k} \binom{n-k}{v-k} \frac{n}{n-k} \sum_{i=0}^{k-1} \binom{n-1}{i} \prod_{j=1}^{k-1} \frac{1-\epsilon_{n-1-j}}{\epsilon_{n-1-j}} \quad v=1, \dots, n-1$$

$$\text{where: } P_{np}(0) = [1 - n\lambda\phi'(0) \sum_{i=0}^{n-1} \binom{n-1}{i} \prod_{j=1}^{n-1} \frac{1-\epsilon_{n-1-j}}{\epsilon_{n-1-j}}]^{-1}$$

$\phi(\theta) = \int_0^{\infty} e^{-\theta x} dG(x)$ is the Laplace-Stieljes transform of the service requirement distribution,

$-\phi'(0)$ is the mean service requirement

$$\epsilon_1 = \phi(1\lambda) \quad \text{and} \quad \frac{1-\epsilon_0}{\epsilon_0} = 1$$

The probability generating function has a simpler expression:

$$\psi_{np}(z) = \sum_{v=0}^n P_{np}(v) z^v = \frac{1 + n\lambda \sum_{k=1}^n \alpha_k z^k (1-z)^{n-k}}{1 + n\lambda \alpha_n}$$

$$\text{where } \alpha_k = \frac{1}{\lambda(n-k)} \sum_{i=0}^{k-1} \binom{n-1}{i} \prod_{j=i}^{k-1} \frac{1 - \epsilon_{n-1-j}}{\epsilon_{n-1-j}} \quad k=1, \dots, n-1$$

$$\text{and } \alpha_n = -\phi'(0) \sum_{i=0}^{n-1} \binom{n-1}{i} \prod_{j=i}^{n-1} \frac{1 - \epsilon_{n-1-j}}{\epsilon_{n-1-j}}$$

For the capacity-sharing system

(6)

$$P_{cs}(v) = P_{cs}(0) \frac{n!}{(n-v)!} \rho^v \quad v = 0, \dots, n$$

$$\text{where } P_{cs}(0) = \left[1 + \sum_{v=1}^n \frac{n!}{(n-v)!} \rho^v \right]^{-1}$$

and $\rho = -\lambda\phi'(0)$ is the traffic intensity.

We obtain for the utilizations factor:

(7.)

$$d_{np} = 1 - P_{np}(0) = 1 - \frac{1}{1 + n \sum_{i=0}^{n-1} \binom{n-1}{i} \prod_{j=0}^i \frac{1 - \epsilon_j}{\epsilon_j}}$$

$$d_{cs} = 1 - P_{cs}(0) = 1 - \frac{1}{1 + n \sum_{i=0}^{n-1} \frac{(n-1)!}{(n-1-i)!} \rho^i}$$

Further, a trivial but tedious algebraic manipulation shows that, in both cases, the average queue length, $E(Q)$, is related to the utilization factor by

(8.)

$$E(Q) = n - \frac{d}{\rho} \quad \text{while the average throughput rate is}$$

(9.)

$$r = -d\phi'(0)$$

From the form of equations (7) it is clear that

$$d_{cs} > d_{np} \quad \text{if} \quad -\phi'(0) > \frac{1-\phi(\theta)}{\theta\phi(\theta)} \quad \forall \theta > 0 \quad \text{and}$$

$$d_{cs} < d_{np} \quad \text{if} \quad -\phi'(0) < \frac{1-\phi(\theta)}{\theta\phi(\theta)} \quad \forall \theta > 0$$

Definitions:

A distribution whose Laplace-Stieljes Transform satisfies

$$(a) \quad -\phi'(0) > \frac{1-\phi(\theta)}{\theta\phi(\theta)} \quad \forall \theta > 0 \quad \text{is called}$$

an hyperexponential distribution. It is hypoexponential if it satisfies

$$(b) \quad -\phi'(0) < \frac{1-\phi(\theta)}{\theta\phi(\theta)} \quad \forall \theta > 0$$

Note that the only distribution for which $-\phi'(0) = \frac{1-\phi(\theta)}{\theta\phi(\theta)} \quad \forall \theta > 0$ is the exponential distribution since, in that case $\phi(\theta) = \frac{1}{1-\theta\phi'(0)}$

which is precisely the LST of the exponential distribution with mean $-\phi'(0)$, and the inverse transform is unique. In this case, both systems will have identical performance, illustrating the point made previously about the exponential assumption.

However, for hyperexponential distributions the capacity-sharing strategy will yield a better performance than a non-preemptive strategy, from the point of view of resource utilization and throughput but also from the user's standpoint, since, by equation (8), the average queue

length (and therefore the mean response time) will be lowered. This result holds whatever ρ and n . This point should be emphatically stressed, for it is a common assertion in the computer literature (see [D1] for instance) that system's efficiency and user's satisfaction are conflicting objectives.

On the other hand, for hypoeponential distributions, the capacity-sharing arrangement will be less performant, in terms of resource usage, throughput and delays, than a non-preemptive discipline. Thus, for such service distributions, a Round Robin strategy will give a better response time to some privileged classes of requests, at the expense of global system performances, including users' satisfaction.

In particular, this result implies that, in some cases, a simple FIFO non-preemptive discipline will be far more efficient in all respects than a sophisticated strategy (especially if overhead is taken into account). Again this holds irrespective of n and the traffic intensity ρ .

The common distributions can be easily classified as either hyperexponential or hypoeponential.

It can be easily shown by convexity arguments that any rational distribution* whose poles are real is hyperexponential. This class includes in particular any convex combination of exponentials.

The Gamma distribution

$$dG_{\alpha}(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} e^{-x/\beta} \quad x, \alpha, \beta > 0$$

is hyperexponential for $0 < \alpha < 1$ and hypoeponential for $\alpha > 1$.

The generalized Erlang distribution (LST: $\gamma(\theta) = \prod_i \frac{\lambda_i}{\lambda_i + \theta}$), the deterministic, the uniform are all hypoeponential.

*whose LST is a rational function and, thus, can be decomposed as

$$\gamma(\theta) = w_0 + \sum_{i=1}^k \frac{w_i \mu_i}{\mu_i + \theta}, \quad w_0 + \sum_{i=1}^k w_i = 1.$$

Incidentally, it can be shown from the properties of Laplace-Stieljes transforms (in particular, their complete monotonicity; see Feller[10] Chapter XIII) that the coefficient of variation $\frac{\sigma}{\mu}$ (σ = standard deviation) is greater than one for hyperexponential distributions and less than one for hypoexponential distributions. However, this is not a sufficient condition.

VII. Conclusions.

In this paper, we have characterized the effect of the queue discipline in queueing models of computer systems. We have attempted to clarify this controversial issue under extremely general assumptions and demonstrated the need for such assumptions since the standard exponential assumption induces special properties which cannot be generalized.

A more general technique has been presented to solve these models and illustrated in the $M_d/G/s$ system and the $M_d/G/$ capacity-sharing system. The latter discipline has been shown to essentially recreate memoryless property.

Finally the influence of the distribution of the service times on the performance of the system has been investigated, leading to a classification between hyperexponential and hypoexponential distributions.

Acknowledgements.

The author gratefully acknowledges the invaluable support and encouragement he received throughout this study from Professor M. J. Flynn.

APPENDIX A

Theorem.

For independent arrivals and preemptive queue disciplines the steady state queue length process is invariant with respect to the queue discipline if and only if the service requirements are independent identically distributed according to an exponential distribution.

Proof.

The sufficiency of the condition is obvious in light of the memoryless property of the exponential distribution.

On the other hand, if the steady-state queue length distribution is invariant with respect to the preemptive queue discipline, it must be the same in particular for the M/G/1 system under the two extreme cases of the Round Robin discipline: FCFS ($q=\infty$) and P.S ($q=0$).

Let $P(s)$ be the p.g.f of the queue length. With the FCFS discipline (or for that matter, any non-preemptive discipline),

$$P_{NP}(s) = (1-\rho) \frac{(1-s)\Psi[\lambda(1-s)]}{\Psi[\lambda(1-s)]-s}$$

Under the PS discipline

$$P_{PS}(s) = \frac{1-\rho}{1-\rho s} \quad \text{where} \quad \rho = \frac{\lambda}{\mu} \quad \text{and} \quad \frac{1}{\mu} = -\Psi'(0)$$

Then $P_{NP}(s) \equiv P_{PS}(s)$ if and only if

$$\frac{(1-s)\Psi[\lambda(1-s)]}{\Psi[\lambda(1-s)]-s} \equiv \frac{1}{1-\rho s}$$

which is equivalent to

$$\Psi[\lambda(1-s)] \equiv \frac{1}{1+\frac{\lambda}{\mu}(1-s)}$$

i.e.,

$$\Psi(x) \equiv \frac{\mu}{\mu+x}$$

which is precisely the LST of the exponential distribution with parameter μ QED.

APPENDIX B

The M/G/1 Queueing System

For $s=1$, equations (12a) and (12b) yield for the limiting values of $\Pi_k(t;x)$, as $t \rightarrow \infty$

$$\begin{aligned}\lambda \Pi_0(0) &= \int_0^\infty \Pi_1(x) dG(x) \\ \frac{d\Pi_1(x)}{dx} + \lambda \Pi_1(x) &= 0\end{aligned}\tag{B.1a}$$

$$\frac{d\Pi_k(x)}{dx} + \lambda \Pi_k(x) = \lambda \Pi_{k-1}(x) \quad k \geq 1$$

and

$$\begin{aligned}\Pi_1(0) &= \lambda \Pi_0(0) + \int_0^\infty \Pi_2(x) dG(x) \\ \Pi_k(0) &= \int_0^\infty \Pi_{k+1}(x) dG(x) \quad k \geq 1\end{aligned}\tag{B.1b}$$

Let $\Pi(s,x) = \sum_{k=1}^{\infty} \Pi_k(x) s^k$ be the generating function associated with $\Pi_k(x)$.

Multiplying each equation of (B.1a) by s^k , $k \geq 1$ and summing them, we obtain

$$\frac{\partial \Pi(s,x)}{\partial x} + \lambda \Pi(s,x) = \lambda s \Pi(s,x)\tag{B.2}$$

The general solution of (B.2) is of the form

$$\Pi(s,x) = f(s) e^{-\lambda(1-s)x}$$

where $f(s)$ is an arbitrary function of s .

The particular solution must satisfy the boundary conditions (B.1b)

Multiplying each equation of (B.1b) by s^k , $k \geq 1$ and summing them, we obtain:

$$\Pi(s, 0) = \lambda s \Pi_0(0) + \frac{1}{s} \int_0^\infty \left[\sum_{k=2}^\infty \Pi_k(x) s^k \right] dG(x) \quad (B.3)$$

$$\text{Since } \sum_{k=2}^\infty \Pi_k(x) s^k = \Pi(s, x) - s \Pi_1(x)$$

$$\text{and } \int_0^\infty \Pi_1(x) dG(x) = \lambda \Pi_0(0) \text{ by the first equation of (B.1a),}$$

(B.3) becomes:

$$\Pi(s, 0) = \lambda(s-1) \Pi_0(0) + \frac{1}{s} \int_0^\infty \Pi(s, x) dG(x)$$

The solution sought must then satisfy

$$f(s) = \lambda(s-1) \Pi_0(0) + \frac{1}{s} f(s) \psi[\lambda(1-s)]$$

where $\psi(\theta)$ is the L.S.T. of $G(x)$

Therefore the generating function of the Π_k 's satisfying both (B.1a) and (B.1b) is:

$$\Pi(s, x) = \frac{\lambda s(1-s) \Pi_0(0)}{\psi[\lambda(1-s)] - s} e^{-\lambda(1-s)x}$$

Coming back to the original probabilities $P_k(x) = \Pi_k(x) [1-G(x)]$

Let $P(s)$ be the probability generating function of the steady state probabilities, i.e.:

$$P(s) = \sum_{k=0}^\infty s^k \int_0^\infty \Pi_k(x) [1-G(x)] dx$$

We obtain:

$$P(s) - P(0) = \int_0^\infty \Pi(s, x) [1-G(x)] dx = \frac{\lambda s(1-s) \Pi_0(0)}{\psi[\lambda(1-s)] - s} \int_0^\infty e^{-\lambda(1-s)x} [1-G(x)] dx$$

Integrating the last expression by parts:

$$P(s) - P(0) = \frac{\lambda s \Pi_0(0) [1 - \psi[\lambda(1-s)]]}{\psi[\lambda(1-s)] - s}$$

But since $\Pi_0(0) = P(0)$, then

$$P(s) = P(0) \frac{(1-s)\psi[\lambda(1-s)]}{\psi[\lambda(1-s)] - s} \quad (\text{B.4})$$

which is exactly Kendall's result [P1]

$P(0)$ is determined by the requirement that $P(1) = 1$

$$\text{i.e.} \quad P(0) = 1 - \lambda \psi'(0) = 1 - \rho \quad (\text{B.5})$$

APPENDIX C

The system M /G/1

In this case, $\lambda_n = (N-n)\lambda$ and $s=1$ Equations (2a) and (2b) become:

$$\begin{cases} N\lambda\pi_0(0) = \int_0^\infty \pi_1(x) dG(x) & (C.1a) \\ \frac{d\pi_1(x)}{dx} + (N-1)\lambda\pi_1(x) = 0 \\ \frac{d\pi_n(x)}{dx} + (N-n)\lambda\pi_n(x) = (N-n+1)\lambda\pi_{n-1}(x) & 2 \leq n \leq N \end{cases}$$

$$\begin{cases} \pi_1(0) = N\lambda\pi_0(0) + \int_0^\infty \pi_2(x) dG(x) & (C.1b) \\ \pi_n(0) = \int_0^\infty \pi_{n+1}(x) dG(x) & 2 \leq n \leq N-1 \\ \pi_N(0) = 0 \end{cases}$$

The form of these equations suggests that we look for a solution of the form

$$\pi_n(x) = \sum_{k=1}^n \alpha_{k,n} e^{-(N-k)\lambda x}$$

Let $P(s)$ be the probability generating function of the steady state queue length. Then

$$P(s) = P(0) + \sum_{n=1}^N s^n \int_0^\infty \sum_{k=1}^n \alpha_{k,n} e^{-(N-k)\lambda x} [1-G(x)] dx$$

$$\text{i.e. } P(s) = P(0) + \sum_{n=1}^N s^n \sum_{k=1}^n \alpha_{k,n} \frac{1 - \Psi[\lambda(N-k)]}{\lambda(N-k)}$$

the coefficient of $\alpha_{N,N}$ being $-\Psi'(0)$. Note also that:

$$P(0) = \pi_0(0)$$

Let $\varepsilon_k = \Psi[\lambda(N-k)]$

We now solve for $\alpha_{k,n}$, $k \leq n$

The first equation (C.1a) yields $\alpha_{11} = \frac{N\lambda P(0)}{\varepsilon_{N-1}}$

The second is identically satisfied.

The third yields $\alpha_{12} = -(N-1)\alpha_{11}$

etc.

The $(n+1)$ st equation of (C.1a) yields by identification:

$$\alpha_{k,n} = -\frac{N-n+1}{n-k} \alpha_{k,n-1} \quad 1 \leq k \leq n-1$$

which implies, by induction on n

$$\alpha_{k,n} = (-1)^{n-k} \binom{N-k}{n-k} \alpha_{k,k} \quad k \leq n$$

Thus, the non-diagonal elements are determined in terms of the diagonal elements by equations (C.1a). The latter are determined by equations (C.1b):

$$\sum_{k=1}^n \alpha_{k,n} = \sum_{k=1}^n \alpha_{k,n+1} \varepsilon_{N-k} + \alpha_{n+1,n+1} \varepsilon_{N-n-1} \quad (N \geq n \geq 1) \quad (C.2)$$

Since both summations can be expressed as homogenous linear combination of the $\alpha_{k,k}$'s $k=1, \dots, n$, (C.2) determines $\alpha_{n+1,n+1}$ as an homogenous linear combination of $\alpha_{k,k}$ $k=1, \dots, n$.

Since α_{11} has the factor $N\lambda P(0)$, all $\alpha_{k,k}$ ($k > 1$) will have this same factor and $N\lambda P(0)$ will be factored out and, from now on, $\alpha_{k,k}$ refers to the old α_{kk} divided by $N\lambda P(0)$.

$P(s)$ can now be written as:

$$P(s) = P(0) \left[1 + N\lambda \sum_{n=1}^N s^n \sum_{k=1}^n \alpha_{k,n} \frac{1-\epsilon_{n-k}}{\lambda(N-k)} \right] \quad \text{i.e.}$$

$$P(s) = P(0) \left[1 + N\lambda \sum_{k=1}^N \frac{1-\epsilon_{N-k}}{\lambda(N-k)} \alpha_{k,k} s^k (1-s)^{N-k} \right] \quad (C.3)$$

In the last expression the factor of $\alpha_{N,N} s^N$ is $-\Psi'(0)$.

Since $P(s)$ is a probability generating function, $P(1)=1$

and $P(0)$ can then be determined as:

$$P(0) = [1 - N\lambda \Psi'(0) \alpha_{N,N}]^{-1} \quad \rho = -\lambda \Psi'(0) \quad (C.4)$$

Further $\left. \frac{dP(s)}{ds} \right|_{s=1}$ is the average queue size and is given by

$$E(Q) = NP(0) [N\rho \alpha_{N,N}^{-(1-\epsilon_1)} \alpha_{N-1,N-1}] \quad (C.5)$$

The resolution by induction of equations (C.2) yield:

$$\begin{cases} \alpha_{k,k} = \frac{1}{1-\epsilon_{N-k}} \sum_{i=0}^{k-1} \binom{N-1}{i} \left(\frac{1-\epsilon_{N-k}}{\epsilon_{N-k}} \right) \left(\frac{1-\epsilon_{N-k+1}}{\epsilon_{N-k+1}} \right) \dots \left(\frac{1-\epsilon_{N-1-i}}{\epsilon_{N-1-i}} \right) & k=1, N-1 \\ \alpha_{N,N} = \sum_{i=0}^{N-1} \binom{N-1}{i} \left(\frac{1-\epsilon_1}{\epsilon_1} \right) \left(\frac{1-\epsilon_2}{\epsilon_2} \right) \dots \left(\frac{1-\epsilon_{N-1-i}}{\epsilon_{N-1-i}} \right) & \text{where } \frac{1-\epsilon_0}{\epsilon_0} = 1 \end{cases} \quad (C.6)$$

Formula (C.6) completes the resolution of the system and, under this form, it is particularly easy to verify that it coincides with Takacs solution [T3]

In particular, the expressions (C.3), (C.4) and (C.6) give a closed form solution to the model studied by Baht in [B1] where only an iterative computational scheme was given.

The system studied by Baht is a single server, restricted arrival system with $\lambda_k = (N-k)\lambda$. Service is given, in a Round Robin fashion with a constant quantum q , but the service requirements are assumed to be exponentially distributed and we showed in section 5.4 that the queue

length was exactly the same as under a non-preemptive FIFO discipline. There is however a constant overhead swap time δ to be added every time a request is taken in service.

Let X^* be the total service time spent on behalf of a request. If X denotes the (exponential) service time

$$X^* = X + C\delta \quad \text{where} \quad C-1 < X/(q-\delta) < C$$

Therefore, if g denotes the density function of X and h that of X^*

$$h(x^*) = \begin{cases} g(x^* - c\delta) & (c-1)q + \delta < x^* \leq cq \\ 0 & cq \leq x^* < cq + \delta \end{cases}$$

The LST of X^* is given by

$$\Psi^*(\theta) = E[e^{-\theta X^*}] = \sum_{c=1}^{\infty} \int_{(c-1)q+\delta}^{cq} e^{-\theta x^*} e^{-\mu(x^*-c\delta)} dx^* \quad (C.7)$$

i.e

$$\Psi^*(\theta) = \frac{\mu}{\mu+\theta} \left[1 - \frac{1 - e^{-\theta\delta}}{1 - e^{-[(\theta+\mu)q-\mu\delta]}} \right] \quad (C.8)$$

Replacing Ψ by Ψ^* in (C.4), (C.5) and (C.6) then solves Baht's model in closed-form.

APPENDIX D

Equations (12a) and (12b) of Chapter V can also solve a multiple server system submitted to a general service requirement and a "loss" arrival: if a request is made when all servers are busy, the request immediately departs from the system. There is no actual waiting line. Erlang studied such a system in connection with telephone trunklines, under the exponential assumption, and established the so-called Erlang's loss formula, i.e. the probability that all servers are busy. We show now that Erlang's formula generalizes to arbitrary service requirements.

As usual, $G(x)$ denotes the service distribution and $1/\mu$ its mean. Equations (12a) and (12b) yield for the limiting Π_k 's:

$$\begin{aligned} \lambda \Pi_0(0) &= s \int_0^\infty \Pi_1(x) dG(x) \\ \sum_{i=1}^k \frac{\partial \Pi_k(x_1 \dots x_k)}{\partial x_i} + \lambda \Pi_k(x_1 \dots x_k) &= (s-k) \int_0^\infty \Pi_{k+1}(x_1, \dots, x_{k+1}) dG(x_{k+1}) \quad 0 \leq k < s \\ \sum_{i=1}^s \frac{\partial \Pi_s(x_1 \dots x_s)}{\partial x_i} &= 0 \end{aligned} \quad (D.1a)$$

The last equation probably requires some explanation. Since in state s there cannot be a new arrival (it would be lost), λ_s is effectively zero.

The boundary conditions are given by (D.1b) for $0 \leq k \leq s-1$

$$\Pi_{k+1}(x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_{k+1}) = \frac{\lambda}{s-k} \Pi_k(x_1 \dots x_{i-1}, x_{i+1} \dots x_{k+1})$$

It is then clear that the solution of (D.1a) and (D.1b) is:

$$\Pi_k(x_1 \dots x_k) = \Pi_0(0) \frac{\lambda^k (s-k)!}{s!} \quad 0 \leq k \leq s$$

which implies by equation (13)

$$p_k = p_0 \frac{(\lambda/\mu)^k}{k!} \quad 0 \leq k \leq s$$

and p_0 is such that $p_k = 1$

$$\text{Finally} \quad p_k = \frac{\frac{(\lambda/\mu)^k}{k!}}{\sum_{k=0}^s \frac{(\lambda/\mu)^k}{k!}} \quad 0 \leq k \leq s \quad (D.2)$$

which is precisely Erlang's result for the exponential case [P1 p 62].

We also obtain the solution to the restricted arrival loss system by replacing λ with $(N-k)\lambda$ in the k^{th} equation.

The solution in that case is:

$$p_k = \frac{\binom{N}{k} (\lambda/\mu)^k}{\sum_{k=0}^s \binom{N}{k} (\lambda/\mu)^k} \quad 0 \leq k \leq s \quad (D.3)$$

This is the Engset distribution.

This generalization is particularly interesting since it does not require the complete knowledge of the service requirement distribution but only its mean $1/\mu$.

REFERENCES

- [1] Chandy, K. M., "The analysis and solutions for general queueing networks", Proc. of the 6th Annual Princeton Conference on Information Sciences and Systems, Princeton, New Jersey (March 1972).
- [2] Baskett, F. & Gomez, F. P., "Processor sharing in a central server queueing model of multiprogramming with applications", ibid.
- [3] Coffman, E. G. & Kleinrock, L., "Feedback queueing models for time-shared computer systems", JACM, vol 15, no 4 (October 1968) pp. 549-576.
- [4] Rasch, P. J., "A queueing theory study of Round Robin scheduling of time-shared computer systems", JACM, vol 17, no 1 (January 1970) pp. 131-145
- [5] Takacs, L., "Delay distributions for one line with Poisson input, general holding times and various orders of service", Bell System Technical Journal, March 1963.
- [6] Little, J. D. C., "A proof of the queueing formula $L = \lambda w$ ", ORSA, vol 9, (1961) pp. 383-387
- [7] Kleinrock, L., "A conservation law for a wide class of queueing disciplines", Naval Research Logistics Quarterly, vol 12 (1965) pp. 181-191.
- [8] Feller, W., "An introduction to probability theory and its applications", vol 1, 3rd edition (1968) J. Wiley Pub.
- [9] Bhat, U. N. & Nance, R. E., "Busy period analysis of a time-sharing system modeled as a semi-Markov process", JACM, Vol 18, no 2 (April 1971) pp. 221-238.
- [10] Feller, W., "Introduction to probability theory and its applications", vol 2, Second Edition (1971) J. Wiley Pub.
- [11] Kendall, D. G., "Some problems in the theory of queues", Journal of the Royal Statistical Society, vol 13 - 13 (1951) pp. 151-185
- [12] Takacs, L., "On a stochastic process concerning some waiting time problems", Theory of probability and its applications, vol 11, no 1 (1951) pp. 90-103.
- [13] Flynn, M. J., "Shared internal resources in a multiprocessor", Johns Hopkins Computer Research Report no 2 (April 1971)
- [14] Coffman, E. G., Muntz, R. R. & Trotter, H., "Waiting time distributions for processor-sharing systems", JACM, vol 17, no 1 (January 1970) pp. 123-130.
- [15] Baskett, F., "Mathematical models of multiprogrammed computer systems", Ph.D. Dissertation, University of Texas (January 1971).
- [16] Kendall, D. G., "Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov Chain", Annals of Mathematical Statistics, vol 24 (1953) pp. 338-358.