

NAVAL POSTGRADUATE SCHOOL

Monterey, California



Servicing Impatient Tasks That Have Uncertain Outcomes

by

Donald P. Gaver
Patricia A. Jacobs

October 1999

Approved for public release; distribution is unlimited.

Prepared for: Space-C2-Information Warfare, Strategic Planning Office,
N6C3, Washington, DC 20350-2000
Institute for Joint Warfare Analysis,
NPS, Monterey, CA 93943

DTIC QUALITY INSPECTED 4

19991122 085

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA 93943-5000

Rear Admiral R. C. Chaplin
Superintendent

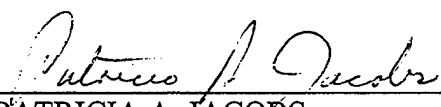
Richard Elster
Provost

This report was prepared for and funded by Space-C2-Information Warfare, Strategic Planning Office, N6C3, Washington, DC 20350-2000; and the Institute for Joint Warfare Analysis, NPS, Monterey, CA 93943.

Reproduction of all or part of this report is authorized.

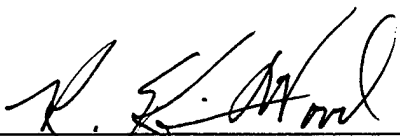
This report was prepared by:

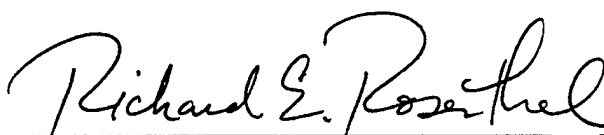

DONALD P. GAVER
Professor of Operations Research

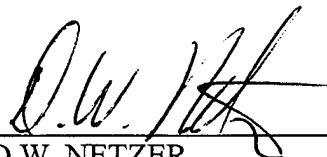

PATRICIA A. JACOBS
Professor of Operations Research

Reviewed by:

Released by:


R. KEVIN WOOD
Associate Chairman for Research
Department of Operations Research


RICHARD E. ROSENTHAL
Chairman
Department of Operations Research

 11/9/99
DAVID W. NETZER
Associate Provost and Dean of Research

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE October 1999		3. REPORT TYPE AND DATES COVERED Technical
4. TITLE AND SUBTITLE Servicing Impatient Tasks That Have Uncertain Outcomes			5. FUNDING NUMBERS N0001499WR30108	
6. AUTHOR(S) Donald P. Gaver and Patricia A. Jacobs				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943			8. PERFORMING ORGANIZATION REPORT NUMBER NPS-OR-00-001	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Space-C2-Information Warfare, Strategic Planning Office, N6C3 2000 Navy Pentagon, Washington, DC 20350-2000 Institute for Joint Warfare Analysis, NPS, Monterey, CA 93943			10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) Many service systems confront tasks of uncertain identity, with limited time for service; that is, the uncertain tasks have deadlines, or are behaviorally impatient. Examples occur in medical care (especially emergencies), telephone help systems, and in military operations. This paper presents modifications of the M/G/1 system to illustrate the impact of the above features. Imperfect task classification is modeled, as is imperfect service and error-afflicted assessment: tasks can be processed, and reprocessed, either correctly or incorrectly depending upon classification, performance, and performance assessment skills. The impact of exponential deadlines, either behavioral or server-controlled, is represented using both a modification of the Takaçs-Beneš integro-differential equation, and a simple and accurate fixed-point approximation.				
14. SUBJECT TERMS service systems, deadlines, information warfare, battle damage assessment (BDA), emergency relief			15. NUMBER OF PAGES 44	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL	

Servicing Impatient Tasks that have Uncertain Outcomes

D. P. Gaver

P. A. Jacobs

Abstract

Many service systems confront tasks of uncertain identity, with limited time for service; that is, the uncertain tasks have deadlines, or are behaviorally impatient. Examples occur in medical care (especially emergencies), telephone help systems, and in military operations. This paper presents modifications of the M/G/1 system to illustrate the impact of the above features. Imperfect task classification is modeled, as is imperfect service and error-afflicted assessment: tasks can be processed, and reprocessed, either correctly or incorrectly depending upon classification, performance, and performance assessment skills. The impact of exponential deadlines, either behavioral or server-controlled, is represented using both a modification of the Takaçs-Beneš integro-differential equation, and a simple and accurate fixed-point approximation.

1. Background

Many service situations are characterized by currently unmodeled uncertainties at least as influential as those identified with the usual stochastic arrival and service processes. Additional and important uncertainty sources or components include the true identity or nature of the task, hence the "optimal", or at least satisfactory, mode of its service (in the light of resources available and the other tasks on hand or anticipated). Further complications occur if the task has a deadline: is impatient or perishable, i.e. has an unknown life or time available for useful service: e.g. telephone callers placed on hold

(prone to hang up or abandon an attempt), medical emergency patients (who may die), mobile military targets (that tend to move, or possibly fire first), and many others.

These questions naturally arise: once a service is performed, is it “complete”, or should additional service work, possibly of a different kind, be performed? And what fraction of time-sensitive perishable or deadline-afflicted tasks actually finish service? Such issues arise because the initial classification of tasks is realistically uncertain, as is the degree of task accomplishment. Post-service assessment is, also realistically, error-prone, which can sometimes lead to premature release of an incompletely served task, with costly legacy, *or* to the needless expenditure of extra time and resources to “complete” a task already completed. We call the above generic *uncertain service situations*, and study several of their features. Important practical issues are to discover system performance sensitivities, such as which additional uncertainty-reduction capabilities are likely to be most cost-effective.

As emphasized above, uncertain service situations arise in many contexts. Examples are medical emergencies, e.g. those that arise from terrorist attacks or natural hazards, such as hurricanes, floods, or earthquakes; forest-fire fighting; engineering and operational problems with new computer software and hardware; and in military operations, e.g. on the battlefield, where “service” means actual or potential destruction or deterrence of opponent assets or troops, but may also involve inadvertent server deception, fratricide, collateral damage, and injury to non-combatants. Avoidance of these latter limits service options.

In this report the deadlines, or impatience, that characterize our tasks are described probabilistically: the server only knows the distribution of deadline elapse and task disappearance, not individual durations. There is a large literature on task scheduling in the face of hard, *known*, deadlines; cf. Liu and Layland (1973), Jiang, Lewis, and Colin (1996), and others. More recently, stochastic scheduling of queued tasks with deadlines

that are subjected to particular queue disciplines are studied; cf. Lehoczky (1996, 1997a, 1997b), and Doytchinov, Lehoczky and Shreve (1998). The latter problems come from situations for which it is natural to know the deadline; our examples are otherwise. Imperfect service (repair) has also been studied by Brown and Proschan (1983), but not in a congested setting. Quite possibly there is related work by others that is unknown to us. A specialized and more intricate *deterministic* version of our problem is described in Gaver and Jacobs (1999). Software illustrating that model is available from the authors.

2. Analytical Setting

The purpose of this paper is to explore and expose operating characteristics of a simplified *totally stochastic* version of the above setup. The mathematical-probability theory of a single-server queue, cf. Cox and Smith (1961), Kleinrock (1976), and many others, is adapted to study an *uncertain service* situation in which *deterioration of task value* also occurs (deadlines are missed, patients die, military targets move).

Our basic model utilizes a first-in, first-out (FIFO) basic queue discipline for several reasons. Simplicity is first among these: understanding how to establish priorities in the face of the other uncertainties faced is difficult, and is postponed. Real-time maintenance of priority or other control is also time-consuming, so the time cost of control should be included; see the model of Appendix C for a start. Alternatively, one can view the present model as being that for a service system that experiences triaged traffic from an initial screener. Disciplines other than FIFO are ignored at this level. We do plan to attack the uncertain service problem in a more general control environment in future.

3. Model Formulation

A task of type j is *classified* as being of type k with probability c_{jk} ; to be interesting, $c_{jj} < 1$, meaning that task misclassification may occur, perhaps with appreciable probability. A model input is the classification matrix \underline{c} . The (possibly misclassified) task

is then *prosecuted* with probability of success m_{jk} ; this probability depends on treatment choice; let $\bar{m}_{jk} = 1 - m_{jk}$ be the probability of failure. A model input is success probability matrix, \underline{m} . *Assessment* (service outcome inspection) is next conducted; if the task has been completed/carried out successfully this fact is ascertained correctly (verified) with probability b_{jk} ; if the task has been completed successfully it is incorrectly reported as not complete, hence is a candidate for needless repetition with probability $1 - b_{jk} = \bar{b}_{jk}$. If the task is *not* completed successfully it is correctly reported as non-successful with probability b_{jk}^* , and incorrectly reported as successful with probability $\bar{b}_{jk}^* = 1 - b_{jk}^*$. All of these parameters can be made functions of other variables: for example, patient condition such as age or auxiliary measures and symptoms in medical diagnosis, or such as range, atmospheric conditions, terrain, deception tactics in military command and control. We treat them as constants for the present. Thus further model inputs are the post-service assessment matrices \underline{b} and \underline{b}^* . The actual time spent servicing the task, which depends upon the classified state, is subject to the decision maker's influence (e.g. it may be truncated); it too is a decision variable.

In general, there may be a number of attempts made to complete a task, and to confirm that completeness. Depending on the classification (\underline{c}), the probability of success and task completion (\underline{m}), and assessment (\underline{b} , \underline{b}^*) skills of the service facility, the system will either provide good and timely service, or not live up to its promise. In various cases, the reason for performance degradation may well depend dominantly upon timely *information* available concerning the task and its accomplishment, and less on the actual or true probability of task prosecution success. In all real situations the required information and capabilities are only available at a cost. It is the purpose of this study to illustrate the nature of the various possible cost tradeoffs. The models proposed are a beginning, but ultimately a means to that end.

4. Stochastic Model

Arrivals appear at a service facility according to a Poisson (λ) process. The probability that an arrival is of type j is p_j ($j = 1, 2, \dots, J$) independently from task to task. Let S_k denote the processing time initially allotted to a task *classified* as of type k , an assigned service time. Note that this is not necessarily (or even frequently, in the present context) the time to successfully service an item of true type j , especially one different from type j . We use S_k to represent the time to carry out a particular process that it has been selected to apply, distinguishing this from the probability of process success, m_{jk} , or $m_{jk}(S_k)$ if desired. As suggested, S_k may have a decision component, i.e. be subject to a decision maker's choice.

4.1 Individual Server Occupancy Times (ISOT)

Suppose a task of type j presents itself to a servicing facility. In what follows we characterize its continuous occupancy of the servicing facility, e.g. a diagnostic and treatment sojourn with a medical facility, or as the current target of a generic shooter in a military context. The task may actually complete long before the generalized server recognizes that fact; on the other hand, the generalized server (server plus reassessment asset) may act, and prematurely and incorrectly decide that the task is complete. Some "completed" tasks are thus released in misdiagnosed and dangerous condition, either to themselves (medicine) or others (military). Our models allow understanding of system tradeoffs that control the probability of such happenings.

4.2 Random Reclassification after Each Assessment

Suppose the system is arranged so that reclassification occurs independently and "with replacement" immediately after each assessment that *declares* an unsuccessful service attempt. If the assessment declares success a new task begins. This is *just one simple option*; see Appendix C, which proposes a decision rule that may reclassify if the

task service is perceived to be incomplete; more sophisticated options are available at a price in time. Here is the corresponding model for the individual server occupancy time (ISOT) of a task of true/actual type j under random reclassification:

for $j, k = 1, 2, \dots, J$

$$C_j = \begin{cases} S_k & \text{with probability } c_{jk} [m_{jk}(S_k)b_{jk} + \bar{m}_{jk}(S_k)\bar{b}_{jk}^*]; \\ S_k + C'_j & \text{with probability } c_{jk} \bar{m}_{jk}(S_k)b_{jk}^*; \\ S_k + K_j & \text{with probability } c_{jk} m_{jk}(S_k)\bar{b}_{jk}; \end{cases} \quad (4.1)$$

K_j represents the random time until a completed task is so identified, the *recognition time*:

$$K_j = \begin{cases} S_k & \text{with probability } c_{jk} \cdot 1 \cdot b_{jk} \\ S_k + K'_j & \text{with probability } c_{jk} \cdot 1 \cdot \bar{b}_{jk} \end{cases} \quad (4.2)$$

Notice that the task accomplishment probability is allowed to depend explicitly on the allocated service time, S_k , and that the ISOT can terminate with unrecognized incomplete task service, i.e. the task may be terminated although incompletely served. The random variables C'_j and K'_j above are independent stochastic replicas of C_j and K_j .

4.3 Expectations

Taking conditional expectations we obtain these expressions for mean ISOT:

$$E[C_j] = \sum_k c_{jk} E[S_k] + \sum_k c_{jk} E[\bar{m}_{jk}(S_k)] b_{jk}^* E[C_j] + \sum_k c_{jk} E[m_{jk}(S_k)] \bar{b}_{jk} E[K_j], \quad (4.3)$$

where

$$E[K_j] = \sum_k c_{jk} E[S_k] + \sum_k c_{jk} \bar{b}_{jk} E[K_j], \quad (4.4)$$

which leads to the formula

$$E[C_j] = \sum_k c_{jk} E[S_k] \left\{ \frac{1 + \sum_k c_{jk} E[m_{jk}(S_k)] \bar{b}_{jk} / \left(1 - \sum_k c_{jk} \bar{b}_{jk}\right)}{1 - \sum_k c_{jk} E[\bar{m}_{jk}(S_k)] b_{jk}^*} \right\}. \quad (4.5)$$

An expression for the second moment appears in Appendix A. It is very clear from (4.5) that degradation of performance can be associated with misclassification and faulty assessment.

4.4 Task Queue

Tasks appear at the service facility at Poisson rate λ . Any task is, independently, of type j with probability p_j ($j = 1, 2, \dots, J$), but the type is only known with uncertainty, c_{jk} . If tasks are treated according to the first-come, first-served discipline by a single server then the system is M/G/1 with effective service times

$$E[C] = \sum_j p_j E[C_j] \quad (4.6)$$

hence, traffic intensity

$$\rho = \lambda E[C]. \quad (4.7)$$

One measure of the system congestion is then the long-run expected number of enqueued tasks (if $\rho < 1$):

$$E[N(\infty)] = \frac{\rho^2 E[C^2] / (E[C])^2}{2(1 - \rho)}. \quad (4.8)$$

This measure does not reflect the number of tasks that are terminated before service is complete. If a goal is to minimize weighted expected waiting time then prioritization in accordance with the index $w_j/E[C_j]$ is optimal. Here w_j is the desirability weight associated with completing class j ; task groups are served in order of increasing index. This, however, takes no account of the influence of deadlines. When deadlines are an important feature of the problem then different measures of system effectiveness are needed.

5. Successful Accomplishment Probability for Tasks with Deadlines

In a variety of contexts the value of task service, or the probability of successful completion, decreases with the delay experienced. This is often true of medical diagnosis and treatment, particularly in emergencies, and in military targeting (the object targeted may move). Using adaptations of the M/G/1 queuing model, cf. Cox and Smith (1961), we study tradeoffs among service capabilities in such situations. Queuing models in which an arriving customer is lost when it waits more than a fixed time in queue have been studied by Boots and Tijms (1999), and Whitt (1999). Note that *server-imposed* deadlines are a control device that may improve certain measures of system performance. These are in effect a *refusal* to provide service; *balking* has traditionally been a task-initiated refusal; see Whitt (1999). Our models address refusals in general.

5.1 The M/G/1 Service System with Ignored Exponential Deadlines

Express the delay sensitivity or deadline for tasks of type j arriving at the M/G/1 system above as an exponential random variable with rate θ_j . A well-known queuing theory result is that the long-run waiting time, W , in an M/G/1 system has Laplace-Stieltjes transform

$$E[e^{-sW}] = \frac{1-\rho}{1-\rho\left(\frac{1-E[e^{-sC}]}{sE[C]}\right)} \equiv \frac{1-\rho}{1-\rho\delta(s)} \quad (5.1)$$

where the role of service time is played here by the ISOT, C . For an arriving task of type j her probability of surviving the wait in queue without deadline elapse is $e^{-\theta_j W}$, conditional on W , so, unconditionally, the probability of initial wait-survival is

$$E[e^{-\theta_j W}] = \frac{1-\rho}{1-\rho\left(\frac{1-E[e^{-\theta_j C}]}{\theta_j E[C]}\right)}. \quad (5.2)$$

See Appendices B and C for expressions for $E[e^{-sC}]$. In the present model all tasks are served to completion, regardless of deadline elapse. This may be more reasonable in some situations than others; it is changed in a subsequent model.

By the memoryless/Markov exponential property the type j task survives a subsequent completion time, duration D_j , that terminates with successful service with probability $E[e^{-\theta_j D_j}]$; see Appendix B for the transform of the improper/dishonest random variable D_j .

It follows that the long-run marginal probability that a random task completes service satisfactorily is

$$\begin{aligned}
 P(\text{parameters}) &= \sum_{j=1}^J p_j E[e^{-\theta_j W}] \cdot E[e^{-\theta_j D_j}] \\
 &= \sum_{j=1}^J p_j \left\{ \frac{1-\rho}{1-\rho \left(\frac{1-E[e^{-\theta_j C}]}{\theta_j E[C]} \right)} \right\} \cdot E[e^{-\theta_j D_j}].
 \end{aligned} \tag{5.3}$$

The above analytical expression may be evaluated numerically, and explored for parameter dependencies. The results of such investigations appear later.

Note that the above model does not assume the capability of detecting “dead” or deadline-elapsed tasks in the queue or upon entering service. Under many conditions such could be refused or purged, thus increasing the chance of successful service for others. The above model results thus tend to be pessimistic or conservative from the server perspective, but not necessarily unrealistically so: additional capabilities may be needed, but unavailable, to monitor enqueued tasks for real-time viability. The next model addresses such capability by the server.

5.2 The M/G/1 System with Deadline-Sensitive Delay

Consider the arrival of tasks with an exponential (θ) deadline, and suppose that when the task reaches the end of the queue it (the probability) can be determined that the deadline will not elapse before reaching the server, given the virtual waiting time, $W(t)$. With that probability, the task is accepted into the queue. We first propose the following heuristic analyses, but follow up with a more formal treatment in Appendix D.

Approximation I

Given the waiting time encountered on generic arrival, W , the *effective* service (ISOT) time is

$$C^\# = \begin{cases} 0 & \text{with probability } 1 - e^{-\theta W} \quad (\text{refused admission, or balks}) \\ C & \text{with probability } e^{-\theta W} \quad (\text{admitted}). \end{cases} \quad (5.4)$$

So, marginally,

$$E[C^\#] = E[C]E[e^{-\theta W}] = E[C]\psi(\theta), \quad (5.5,a)$$

$$E[e^{-\theta C^\#}] = (1 - E[e^{-\theta W}]) + E[e^{-\theta C}] \cdot E[e^{-\theta W}]. \quad (5.5,b)$$

Now model the above system as M/G/1 with state-dependent (thinned Poisson) arrivals as follows:

$$\begin{aligned} \psi(\theta) \equiv E[e^{-\theta W}] &= \frac{1 - \lambda E[C^\#]}{1 - \lambda E[C^\#] \left\{ \frac{1 - E[e^{-\theta C^\#}]}{\theta E[C^\#]} \right\}} \\ &= \frac{1 - \rho \psi(\theta)}{1 - \rho \psi(\theta) \left\{ \frac{1 - [1 - \psi(\theta) + E[e^{-\theta C}]\psi(\theta)]}{\theta E[C]\psi(\theta)} \right\}} \\ &= \frac{1 - [\psi(\theta)\lambda]E[C]}{1 - [\psi(\theta)\lambda]E[C] \left\{ \frac{1 - E[e^{-\theta C}]}{\theta E[C]} \right\}}. \end{aligned} \quad (5.6)$$

This expression asserts that the probability of successful (deadline unviolated) task arrival for service, $\psi(\theta)$, is that of an M/G/1 system whose arrivals are filtered *by the same probability* in the long run. In effect, each arrival flips a biased coin with success probability $E[e^{-\theta W}]$ to be permitted to *join* the queue. The result differs somewhat from the solution of the forward Kolmogorov (Takačs-Beneš) equation for the same assumed arrival-queue interaction; see Appendix D.

The expression (5.6) is a quadratic in the desired probability, the solution of which is

$$\psi(\theta) = \frac{2}{1 + \rho + \sqrt{(1 + \rho)^2 - 4\rho\delta(\theta)}} \quad (5.7)$$

where

$$\delta(\theta) = \frac{1 - E[e^{-\theta C}]}{\theta E[C]} \quad (5.8)$$

the transform of the service/completion time *tail* or survivor distribution. Of course the probability of successful transit to the server is unity when $\theta \rightarrow 0$ (no degradation, or infinite deadline), regardless of the value of $\rho < 1$; if $\theta \rightarrow \infty$ then, since deadlines are now stringent, the only hope of initiating service is to arrive when there is no server activity, i.e. with probability $1/(1 + \rho)$, irrespective of (positive) ρ -value. Likewise, there are no restrictions on ρ in (5.7): a long queue generates many rejections, and thus does not remain long, or grow indefinitely. Numerically, the above simple expressions, (5.7) and (5.16), supply a lower bound that has been shown numerically to be a good approximation to the exact solution of a refusal model proposed in Appendix D. Note that the present approximation gives for the transform of virtual waiting time of non-refused tasks, W , the formula

$$\psi(\xi, \theta) = \frac{1 - \lambda\psi(\theta)E[C]}{1 - \lambda\psi(\theta)E[C] \left\{ \frac{1 - E[e^{-\xi C}]}{\xi E[C]} \right\}} \quad (5.9)$$

where $\psi(\theta)$ is given by (5.7). This can be modified to represent refusal during service, and to provide approximations to the long-run mean waiting time. No numerical discussion or comparisons are available at present, although comparison to results from (D.12) and (D.13) is of interest. These models can be compared to those of Whitt (1999).

Approximation II

A refined version of the above accounts for the different experience of a new task that arrives to find the server busy ($W > 0$), as contrasted to one that arrives to find it idle ($W = 0$). Put

$$\psi_+(\theta) = E[e^{-\theta W} | W > 0] \quad (5.10)$$

the marginal long-run rate of task acceptance given that the server is busy. From (5.1)

$$\begin{aligned} E[e^{-sW} | W > 0] &= \left[\frac{1-\rho}{1-\rho\delta(s)} \right] \frac{1}{\rho} \\ &= \frac{(1-\rho)\delta(s)}{1-\rho\delta(s)}. \end{aligned} \quad (5.11)$$

Approximate as follows ($s = \theta = 1/\text{mean deadline}$)

$$\psi_+(\theta) = (1 - \rho\psi_+(\theta)) \frac{\delta(\theta)}{1 - \rho\psi_+(\theta)\delta(\theta)}; \quad (5.12)$$

this asserts that the probability that an arriving task that encounters a busy period and reaches the service stage is that of an M/G/1 system whose busy-period arrivals are filtered by the same probability. In other words, an auxiliary randomization (biased coin flip) adjusts for the imposition of the deadline, as before in Approximation I, but in a somewhat more refined manner. The solution of (5.11) is

$$\psi_+(\theta) = \frac{2\delta(\theta)}{(1 + \rho\delta(\theta)) + \sqrt{(1 + \rho\delta(\theta))^2 - 4\rho\delta(\theta)^2}}. \quad (5.13)$$

For such a ψ_+ -filtered system the expected duration of a busy period, $E[B]$, satisfies

$$\begin{aligned}
E[B] &= E[C] + \rho\psi_+(\theta)E[B] \\
&= E[C]/(1 - \rho\psi_+(\theta)).
\end{aligned}
\tag{5.14}$$

Consequently in the long run (by alternating renewal process results),

$$P\{W = 0\} = \frac{\lambda^{-1}}{\lambda^{-1} + E[B]} = \frac{1 - \rho\psi_+(\theta)}{1 + \rho[1 - \psi_+(\theta)]}. \tag{5.15}$$

Now the probability that an arriving task is admitted (not refused, and eventually served) is

$$\begin{aligned}
\psi_2(\theta) &= P\{W = 0\} + (1 - P\{W = 0\})\psi_+(\theta) \\
&= \frac{1}{1 + \rho[1 - \psi_+(\theta)]},
\end{aligned}
\tag{5.16}$$

which differs from (5.7) owing to the more refined conditioning imposed.

It will be seen that Approximation II improves on Approximation I in all cases considered.

6. Numerical Exploration

In this section we display graphs of the effects of the various capability parameters on long-run probability of successful service completion.

For reference, the following parameters are varied

- λ ; the Poisson arrival rate of tasks;
- c_{jk} ; ($j, k \in [1, 2, \dots, J]$): probability of *initially classifying a task of type j as being type k* ;
- m_{jk} , or generally $m_{jk}(S_k)$: *probability of success of a service of task type j when prosecuted/served as type k* ; $\bar{m}_{jk} = 1 - m_{jk}$ is the probability of an unsuccessful outcome;
- b_{jk} : probability of *successful/correct assessment as successful treatment of task of type j , given it has been successfully prosecuted/served as type k* ; $1 - b_{jk} = \bar{b}_{jk}$ is

probability of classifying successful treatment as unsuccessful. In the former case the task is discharged correctly as completed; in the latter case it is incorrectly re-served.

- b_{jk}^* : *probability of successful/correct assessment as unsuccessful treatment of a task of type j , given unsuccessfully prosecuted/served as task type k* . In this case the task is reclassified and served again.
- There are two customer types in our present examples; their basic service times are constant (delta-function distributed) with means denoted s_1, s_2 : $S_1 = s_1, S_2 = s_2$ with probability one.

6.1 Discussion of the Figures

In each case investigated, we display the probability of successfully completing service: reaching the server through the initial queue *and* subsequently being successfully served before deadline elapse. Three models are compared: β_0 = the probability of successful completion of service with no task refusals (5.3); β_1 = the probability of successful completion of service with task refusal depending on the virtual waiting time (D.11); and β_2 = the probability of successful completion of service with task refusal depending on the virtual waiting time and *allowed service time*, (D.15) and (D.11). We also illustrate the numerical quality of Approximation II in the graphs that follow, and compare Approximation I and II in the tables.

Probability of successful task completion
 Probability of correct task completion assessment: $b=0.5$ $b^*=0.5$
 $c_{ii}=0.7$ $m_{11}=0.7$ $m_{22}=0.7$ $\theta=0.1$ $s_1=.5$ $s_2=1$

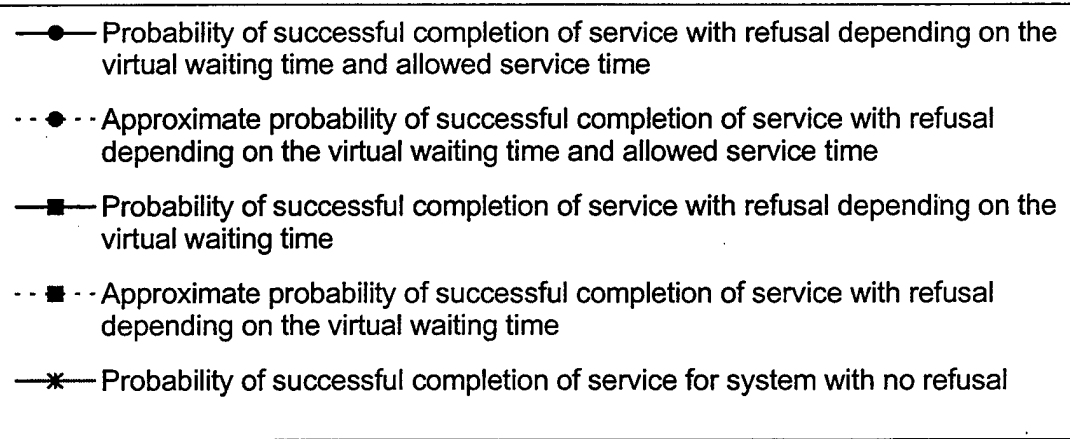
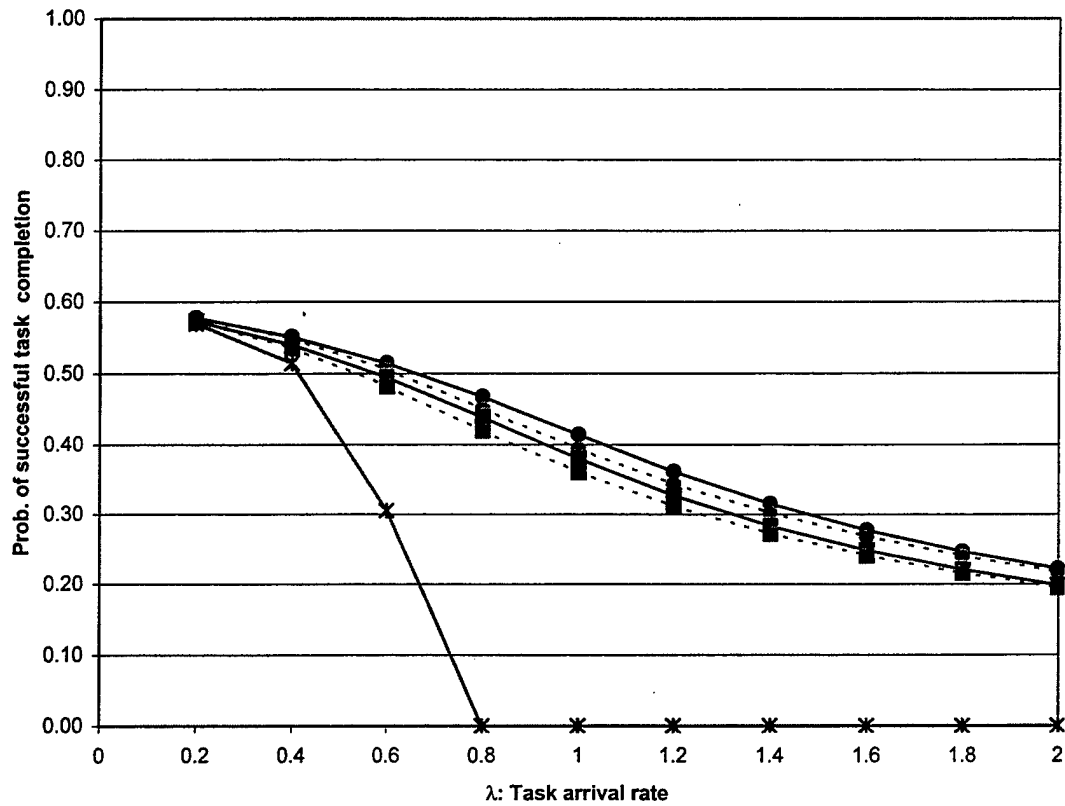


Figure 1

Table for Figure 1

task arrival rate	prob. of receiving service with refusal depending on virtual waiting time	approx. prob. of receiving service with refusal depending on virtual waiting time Approx I/II	prob. of receiving service with refusal depending on virtual waiting time and allowed service time	approx. prob. of receiving service with refusal depending on virtual waiting time and allowed service time Approx I/II	prob. of success. comple. of service with refusal depending on virtual waiting time	approx. prob. of success. comple. of service with refusal depending on virtual waiting time Approx I/II	prob. of success. comple. of service with refusal depending on virtual waiting time and allowed service time	approx. prob. of success. comple. of service with refusal depending on virtual waiting time and allowed service time Approx I/II	prob. of success. comple. of service for system with no refusals
0.2	0.96	0.96 / 0.96	0.97	0.97 / 0.97	0.57	0.57 / 0.57	0.58	0.58 / 0.58	0.57
0.4	0.91	0.89 / 0.90	0.93	0.92 / 0.92	0.54	0.53 / 0.54	0.55	0.55 / 0.55	0.51
0.6	0.83	0.79 / 0.81	0.86	0.84 / 0.85	0.49	0.47 / 0.48	0.51	0.50 / 0.51	0.31
0.8	0.74	0.68 / 0.70	0.79	0.74 / 0.76	0.44	0.41 / 0.42	0.47	0.44 / 0.45	0.00
1.0	0.64	0.58 / 0.60	0.70	0.64 / 0.66	0.38	0.35 / 0.36	0.41	0.38 / 0.39	0.00
1.2	0.55	0.50 / 0.52	0.61	0.56 / 0.58	0.33	0.30 / 0.31	0.36	0.33 / 0.34	0.00
1.4	0.48	0.44 / 0.46	0.53	0.49 / 0.51	0.28	0.26 / 0.27	0.32	0.29 / 0.30	0.00
1.6	0.42	0.39 / 0.40	0.47	0.43 / 0.45	0.25	0.23 / 0.24	0.28	0.26 / 0.27	0.00
1.8	0.37	0.35 / 0.36	0.41	0.39 / 0.40	0.22	0.21 / 0.22	0.25	0.23 / 0.24	0.00
2.0	0.33	0.32 / 0.33	0.37	0.35 / 0.36	0.20	0.19 / 0.20	0.22	0.21 / 0.22	0.00

Parameters for Figure 1

arrival rate of tasks, $\lambda = 0.2$ (0.2) 2

prob. of correct task class: $c_{11} = c_{22} = 0.7$

prob. of correctly classifying a complete task as complete: $b_{11} = b_{12} = b_{21} = b_{22} = 0.5$

prob. of correctly classifying an incomplete task as incomplete: $b_{11}^* = b_{12}^* = b_{21}^* = b_{22}^* = 0.5$

prob. complete task of type j that is correctly classified as type j : $m_{11} = m_{22} = 0.7$

prob. complete task of type j that is incorrectly classified as type k : $m_{12} = m_{21} = 0$

service time for task classified as type 1: $s_1 = 0.5$

service time for task classified as type 2: $s_2 = 1$

prob. an arriving task is of type i : $p_1 = p_2 = 0.5$

mean of the exponential deadline: $(\theta)^{-1} = 10.00$

Discussion of Figure 1

This demonstrates the anticipated decrease in the probability of receiving service (transiting queue before deadline elapse), and the probability of ultimate correct task completion as the arrival rate, λ , increases. Notice that Approximation II, (5.16), relatively closely, but conservatively, tracks the exact solution of (D.11), bounding the latter from below. Approximation I does nearly as well. The payoff from being able to recognize deadline elapse in service (and task ejection) is evident, but the dramatic effect is caused by the existence of a deadline-recognized queue admission capability: if deadlines are ignored then the queue quickly saturates and the success probability plummets. This happens despite the fact that the mean deadline, 10.0, is much greater than the mean ISOT ("service time"), 1.5, in this example.

Next we study the effect of varying the classification parameters.

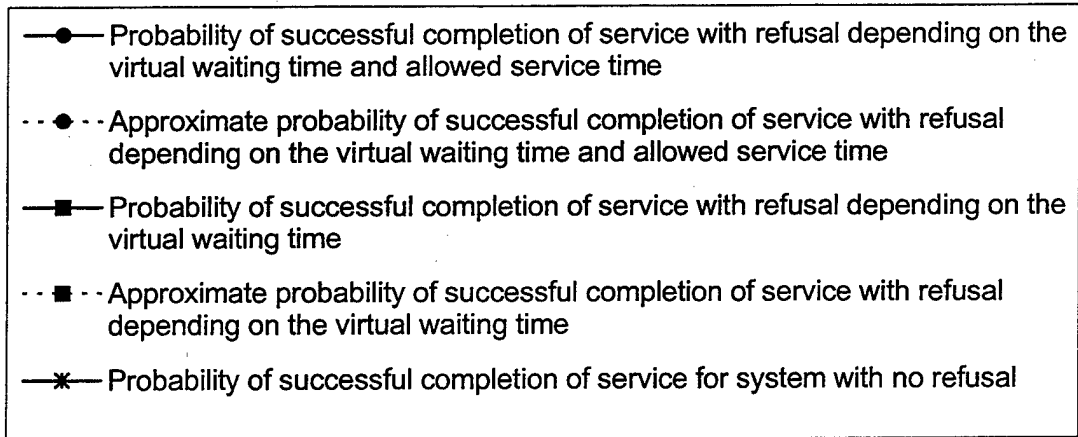
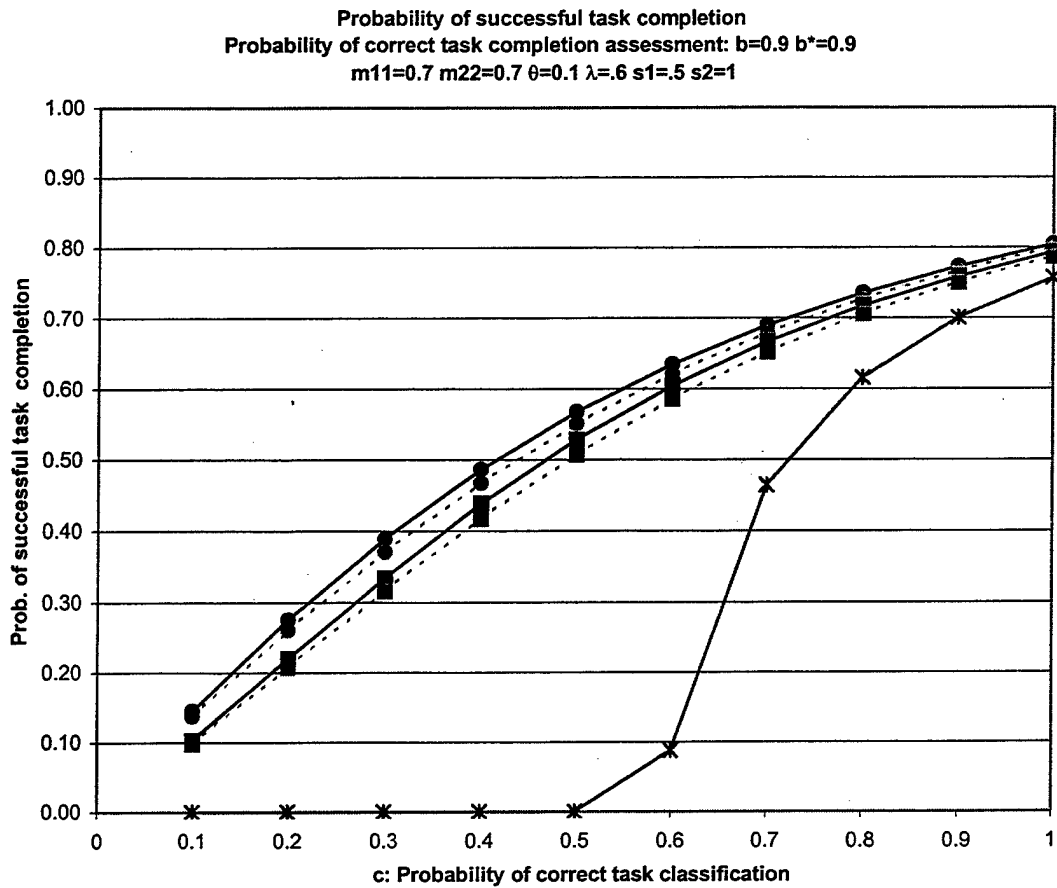


Figure 2

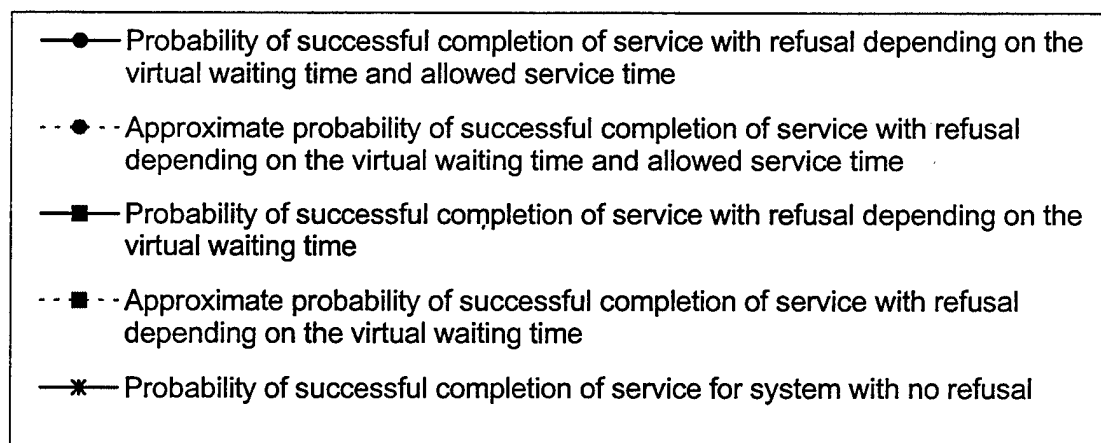
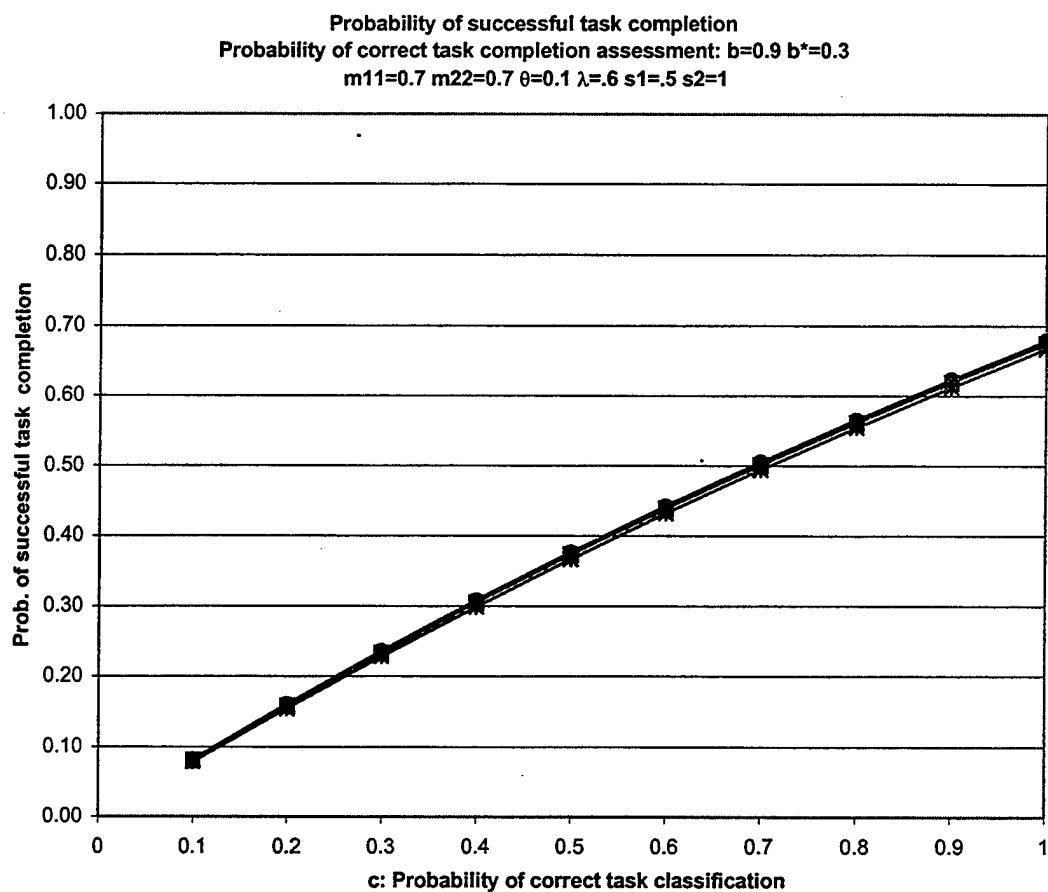


Figure 3

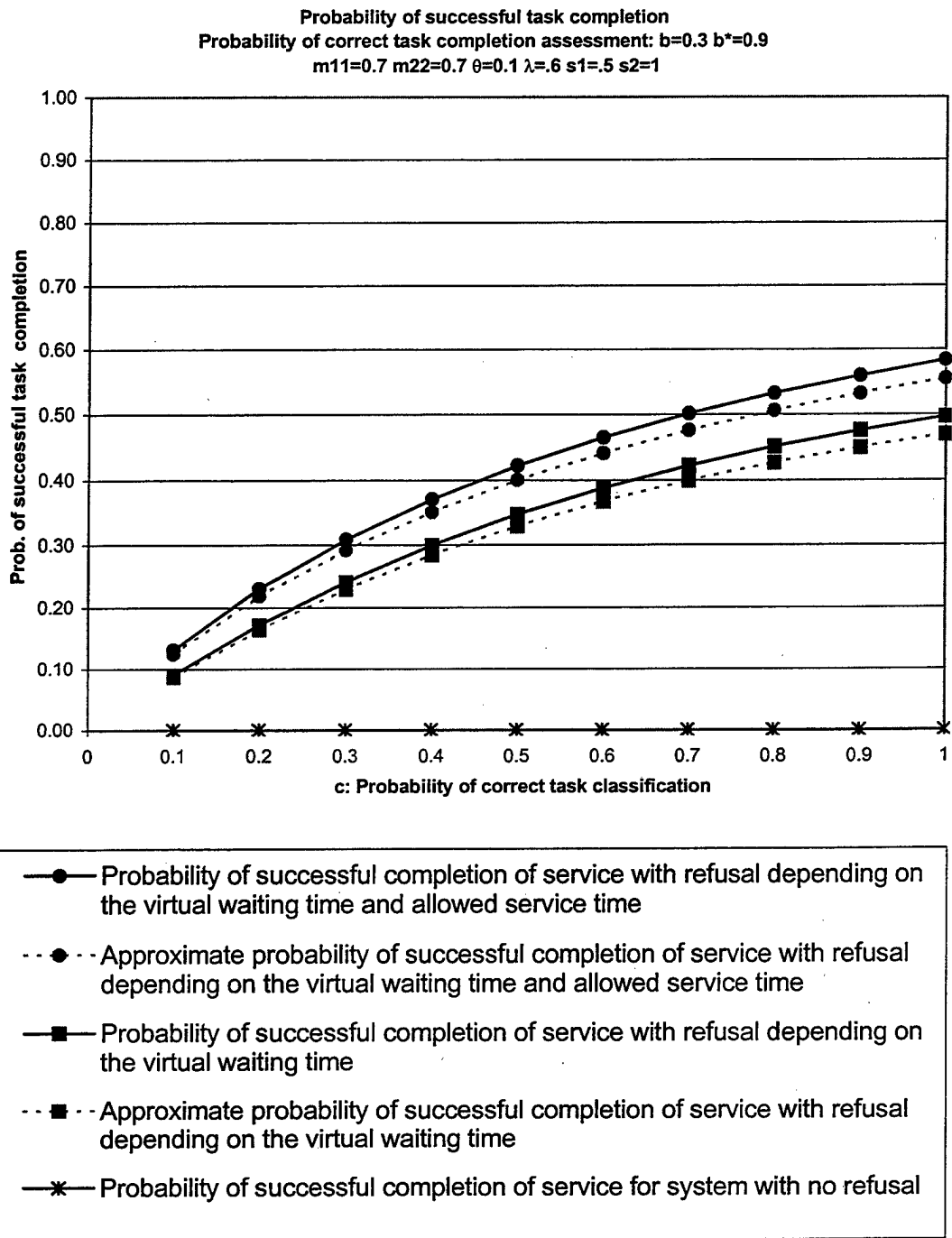


Figure 4

Parameters for Figures 2-4

prob. of correct task class: $c_{11} = c_{22} = 0.1$ (0.1) 1

prob. of correctly classifying a complete task as complete: $b_{11} = b_{12} = b_{21} = b_{22} = 0.9$
(Figures 2 and 3), 0.3 (Figure 4)

prob. of correctly classifying an incomplete task as incomplete: $b_{11}^* = b_{12}^* = b_{21}^* = b_{22}^* = 0.3$
(Figure 3), 0.9 (Figures 2 and 4)

prob. complete task of type j that is correctly classified as type j : $m_{11} = m_{22} = 0.7$

prob. complete task of type j that is incorrectly classified as type k : $m_{12} = m_{21} = 0$

service time for task classified as type 1: $s_1 = 0.5$

service time for task classified as type 2: $s_2 = 1$

arrival rate of tasks, $\lambda = 0.6$

prob. an arriving task is of type i : $p_1 = p_2 = 0.5$

mean of the exponential deadline: $(\theta)^{-1} = 10$

Table for Figure 2

prob. of correct task class.	prob. of receiving service with refusal depending on virtual waiting time	approx. prob. of receiving service with refusal depending on virtual waiting time Approx I/II	prob. of receiving service with refusal depending on virtual waiting time and allowed service time	approx. prob. of receiving service with refusal depending on virtual waiting time and allowed service time Approx I/II	prob. of success. comple. of service with refusal depending on virtual waiting time	approx. prob. of success. comple. of service with refusal depending on virtual waiting time Approx I/II	prob. of success. comple. of service with refusal depending on virtual waiting time and allowed service time	approx. prob. of success. comple. of service with refusal depending on virtual waiting time and allowed service time Approx I/II	prob. of success. comple. of service for system with no refusals
0.1	0.35	0.31 / 0.34	0.50	0.44 / 0.47	0.10	0.09 / 0.10	0.15	0.13 / 0.14	0.00
0.2	0.48	0.42 / 0.45	0.60	0.53 / 0.56	0.22	0.19 / 0.21	0.28	0.25 / 0.26	0.00
0.3	0.58	0.52 / 0.55	0.68	0.62 / 0.65	0.33	0.30 / 0.32	0.39	0.35 / 0.37	0.00
0.4	0.67	0.61 / 0.64	0.75	0.69 / 0.72	0.44	0.40 / 0.42	0.49	0.45 / 0.47	0.00
0.5	0.74	0.69 / 0.71	0.80	0.75 / 0.77	0.53	0.49 / 0.51	0.57	0.54 / 0.55	0.00
0.6	0.80	0.75 / 0.77	0.84	0.80 / 0.82	0.60	0.57 / 0.59	0.63	0.61 / 0.62	0.09
0.7	0.84	0.81 / 0.82	0.87	0.85 / 0.86	0.67	0.64 / 0.65	0.69	0.67 / 0.68	0.46
0.8	0.87	0.85 / 0.86	0.89	0.88 / 0.89	0.72	0.70 / 0.71	0.74	0.72 / 0.73	0.62
0.9	0.90	0.88 / 0.89	0.91	0.90 / 0.91	0.76	0.74 / 0.75	0.77	0.76 / 0.77	0.70
1.0	0.91	0.90 / 0.91	0.93	0.92 / 0.92	0.79	0.78 / 0.79	0.79	0.80 / 0.80	0.76

Table for Figure 3

prob. of correct task class.	prob. of receiving service with refusal depending on virtual waiting time	approx. prob. of receiving service with refusal depending on virtual waiting time Approx I/II	prob. of receiving service with refusal depending on virtual waiting time and allowed service time	approx. prob. of receiving service with refusal depending on virtual waiting time and allowed service time Approx I/II	prob. of success. comple. of service with refusal depending on virtual waiting time	approx. prob. of success. comple. of service with refusal depending on virtual waiting time Approx I/II	prob. of success. comple. of service with refusal depending on virtual waiting time and allowed service time	approx. prob. of success. comple. of service with refusal depending on virtual waiting time and allowed service time Approx I/II	prob. of success. comple. of service for system with no refusals
0.1	0.93	0.91 / 0.92	0.94	0.93 / 0.93	0.08	0.08 / 0.08	0.08	0.08 / 0.08	0.08
0.2	0.93	0.92 / 0.92	0.94	0.93 / 0.94	0.16	0.16 / 0.16	0.16	0.16 / 0.16	0.15
0.3	0.93	0.93 / 0.93	0.94	0.94 / 0.94	0.23	0.23 / 0.23	0.24	0.23 / 0.23	0.23
0.4	0.94	0.93 / 0.93	0.95	0.94 / 0.94	0.30	0.30 / 0.30	0.31	0.31 / 0.31	0.30
0.5	0.94	0.93 / 0.94	0.95	0.94 / 0.95	0.37	0.37 / 0.37	0.38	0.37 / 0.37	0.37
0.6	0.94	0.94 / 0.94	0.95	0.95 / 0.95	0.44	0.44 / 0.44	0.44	0.44 / 0.44	0.43
0.7	0.95	0.94 / 0.94	0.95	0.95 / 0.95	0.50	0.50 / 0.50	0.50	0.50 / 0.50	0.49
0.8	0.95	0.94 / 0.95	0.95	0.95 / 0.95	0.56	0.56 / 0.56	0.57	0.56 / 0.56	0.55
0.9	0.95	0.95 / 0.95	0.96	0.95 / 0.95	0.62	0.62 / 0.62	0.62	0.62 / 0.62	0.61
1.0	0.95	0.95 / 0.95	0.96	0.95 / 0.96	0.68	0.67 / 0.67	0.68	0.68 / 0.68	0.67

Discussion of Figures 2-4

These figures illustrate the possible dependence of successfully joining the queue, and successful task completion on the value of $c_{11} = c_{22} = c$: the probability of correctly classifying the task type (for fixed arrival rate, λ , and other parameters). Apparently there is strong dependence: advantage accrues to systems with task refusals. Accepting all tasks can send the system into saturation for small values of $c_{11} = c_{22} = c$; it only approaches the task refusal systems if c closely approaches unity (current task type classification is nearly perfect). The size of the advantage depends on the ability to correctly assess task completion. If the probability of assessing an incomplete task as incomplete is small ($b^* = 0.3$), then many tasks are thrown out of service before they are complete. Thus, the traffic intensity is less than 1 for the system with no refusal and the probability of successful task completion is about the same as for a system with task refusal. If the probability of

assessing a complete job as complete is small ($b = 0.3$), the additional (non-productive) service on already complete tasks, can saturate the system with no refusals; it also increases the probability a task will be refused in a system with refusal. Thus, increasing the probability of correct task classification can have less effect if b^* is small.

In the next case we explore the effect of quality of post-service assessment.

Probability of successful task completion
 Probability of correct task completion assessment: $b^*=0.9$
 $c_{ii}=0.7$ $m_{11}=0.7$ $m_{22}=0.7$ $\theta=0.1$ $\lambda=.6$ $s_1=.5$ $s_2=1$

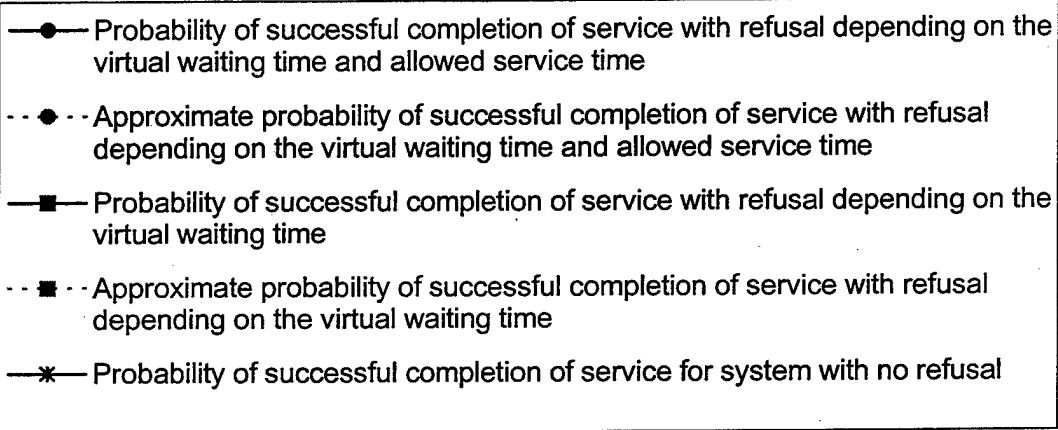
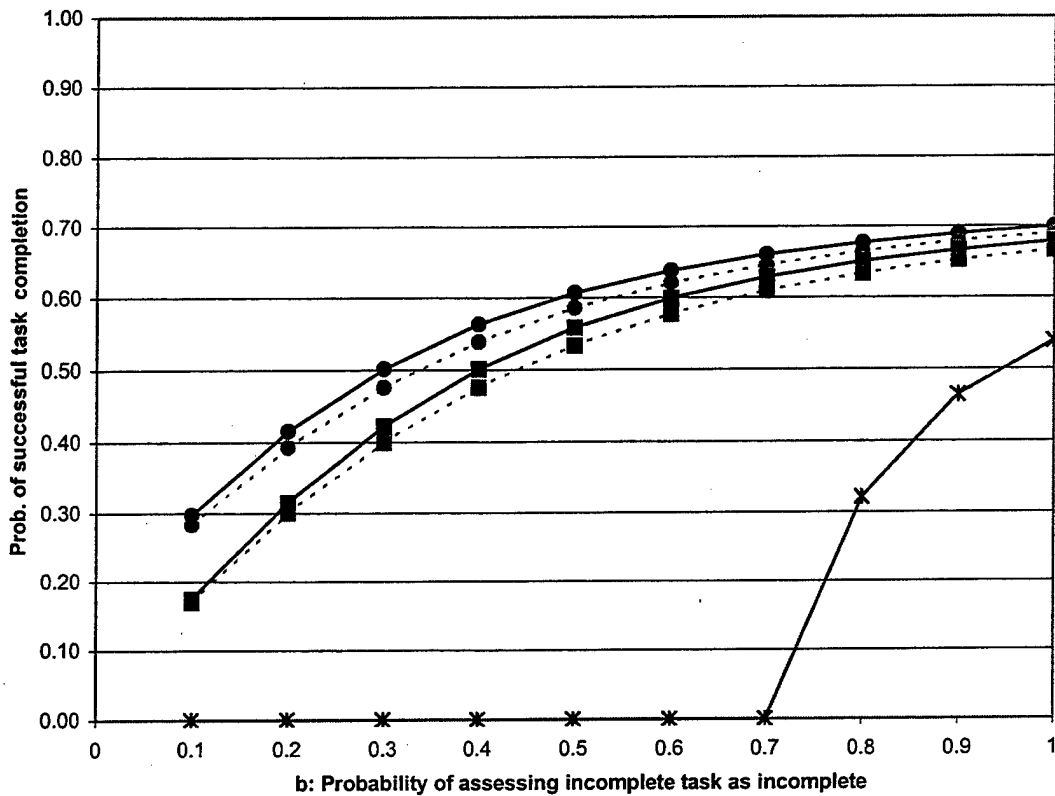


Figure 5

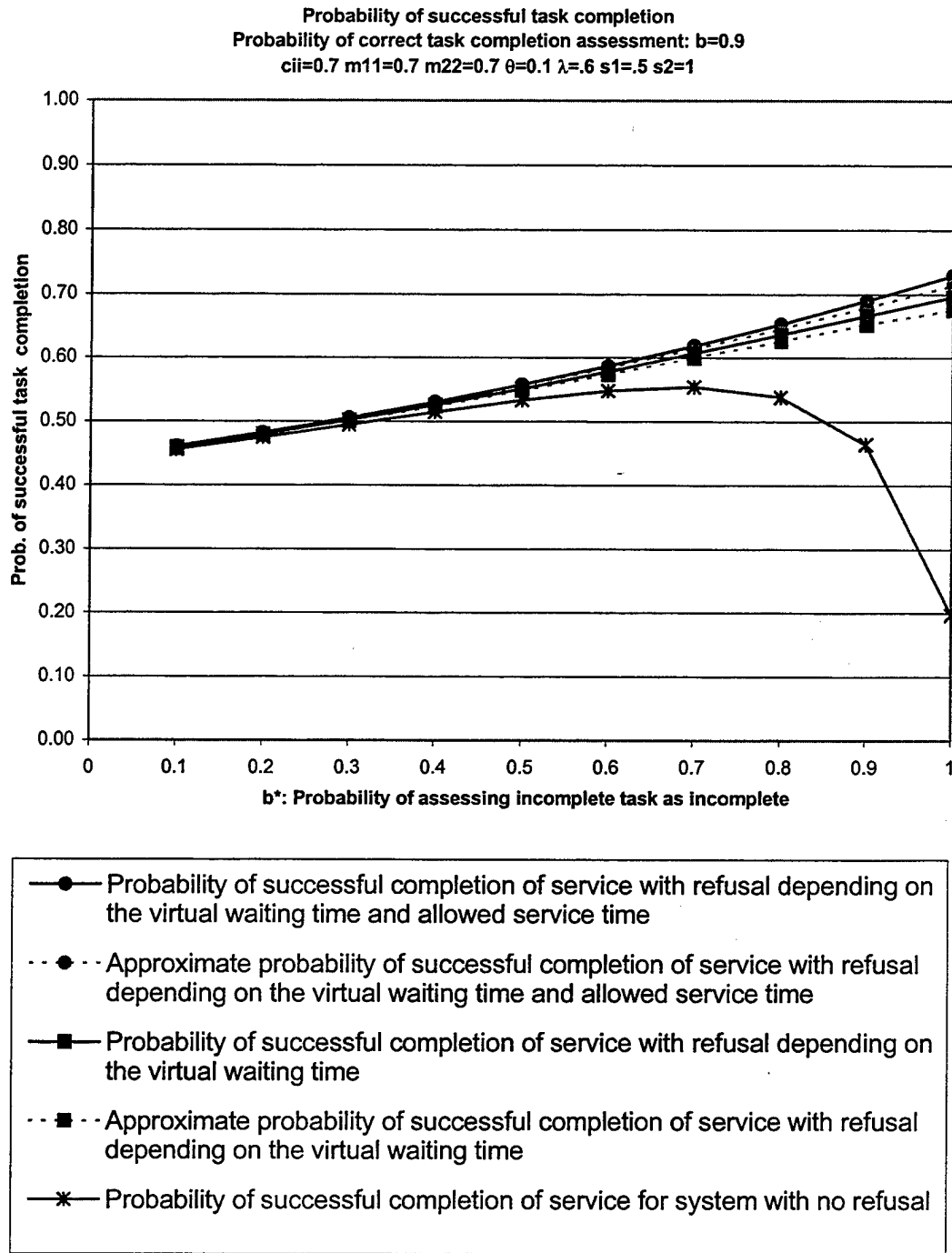


Figure 6

Parameters for Figures 5-7

prob. of correctly assessing a complete task as complete: $b_{11} = b_{12} = b_{21} = b_{22} = 0.1$ (0.1) 1
(Figure 5), $b_{11} = b_{12} = b_{21} = b_{22} = 0.9$ (Figure 6), $b_{11} = b_{12} = b_{21} = b_{22} = 0.1$ (Figure 7)

prob. of correct task class: $c_{11} = c_{22} = 0.7$

prob. of correctly assessing an incomplete task as incomplete: $b_{11}^* = b_{12}^* = b_{21}^* = b_{22}^* = 0.9$
(Figure 5), $b_{11}^* = b_{12}^* = b_{21}^* = b_{22}^* = 0.1$ (0.1) 1 (Figure 6), $b_{11}^* = b_{12}^* = b_{21}^* = b_{22}^* = 0.7$
(Figure 7)

prob. complete task of type j that is correctly classified as type j : $m_{11} = m_{22} = 0.7$

prob. complete task of type j that is incorrectly classified as type k : $m_{12} = m_{21} = 0$

service time for task classified as type 1: $s_1 = 0.5$

service time for task classified as type 2: $s_2 = 1$

arrival rate of tasks, $\lambda = 0.6$ (Figures 5-6), $\lambda = 0.8$ (Figure 7)

prob. an arriving task is of type i : $p_1 = p_2 = 0.5$

mean of the exponential deadline: $(\theta)^{-1} = 10$ (Figures 5-6), $\theta = (0.1 \text{ (0.2) (1.9)})$ (Figure 7)

Discussion of Figures 5-6

Increasing the probability of correctly assessing an incomplete task as incomplete $b_{ij}^* = b^*$ results in increases in the probability of correct task completion for those systems with refusals. However, for the system with no refusals, increasing $b_{ij}^* = b^*$ results in larger service times and thus decreases the probability of successful task completion.

Increasing the probability of correctly assessing a complete task as complete, $b_{ij} = b$ results in increases in the probability of correct task completion for those systems with and without customer refusals. For $b \leq 0.7$, the system with no refusals is saturated and the probability of correct task completion is 0.

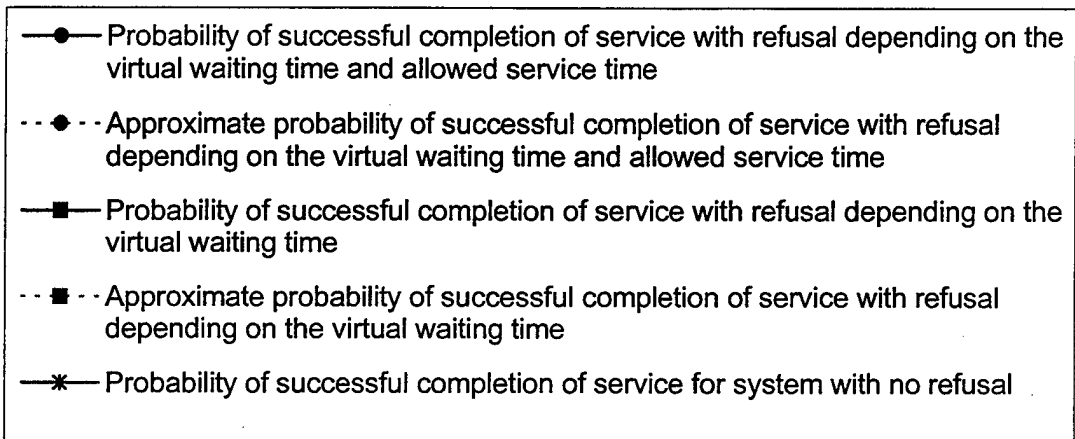
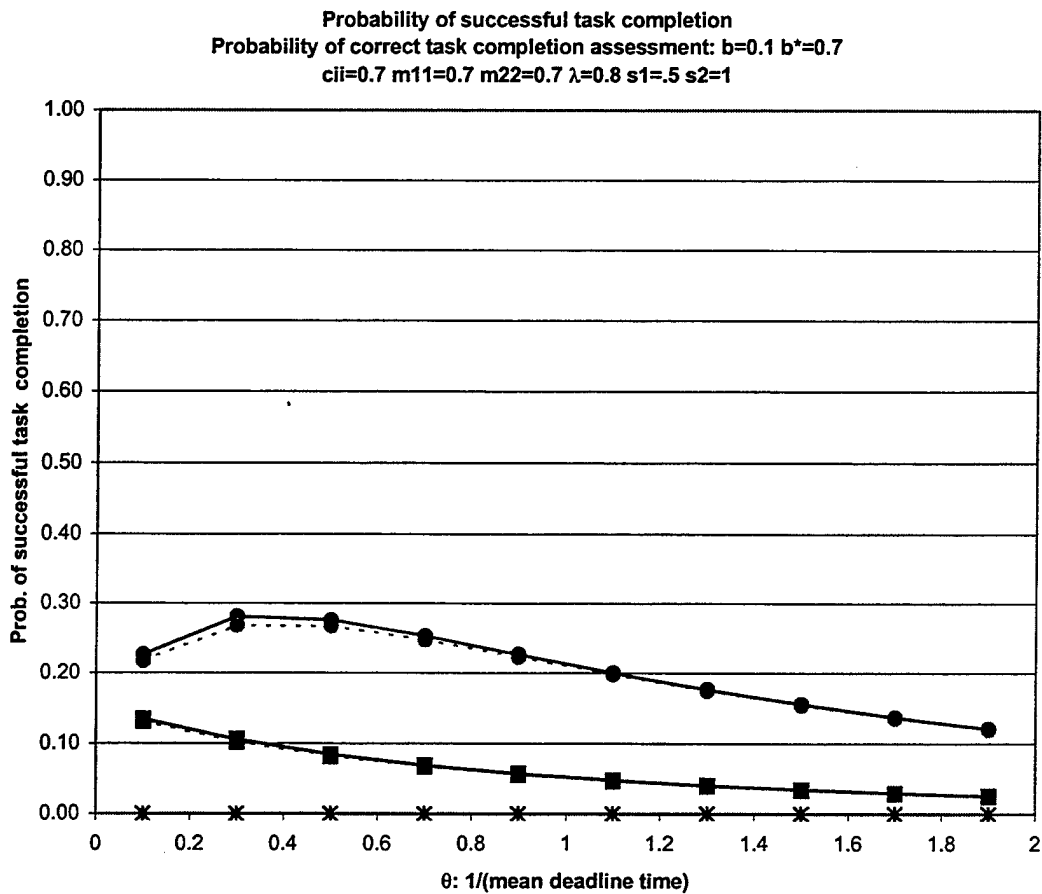


Figure 7

Discussion of Figure 7

Figure 7 displays the probability of successful service completion as a function of the deadline rate, θ . In Figure 7 the probability of correctly assessing a complete job as complete is small; $b = b_g = 0.1$. Note that for the case of task refusal depending on both the virtual waiting time and the allowed time in service, there is an optimum θ . Reason: the small probability of assessing a complete task as being complete is creating much unproductive work for the server. A θ that is too small can result in a task remaining in service when it is complete. A θ that is too large results in too many tasks being turned away.

7. Summary

A stochastic model has been introduced that allows initial discussion of the general problem of uncertain impatient service. The influence of the various processes that may affect such service has been numerically explored in special cases; it can be seen that the ability to adapt to long queues (by refusing admission) as in Section 5.2 and Appendix D, can improve overall performance. In fact, if deadlines did not exist they might well be imposed in order to improve overall long-run system performance. Likewise, improvement of performance by improving either pre-service classification, and/or post-service assessment can be quantitatively traced and choices made. Degradation of either of the latter capabilities can substantially degrade overall system performance, as measured by the probability of successful complete task processing (before deadline elapse).

The present paper scratches the surface of an important and largely neglected service system design and control problem. It is planned to pursue other ramifications, such as non-stationary phenomena (via fluid approximations) and adaptive control (dynamic priorities) in subsequent work.

8. Acknowledgements

The writers wish to acknowledge useful comments by Gideon Weiss, Lee Schruben, and especially by John Lehoczky, who suggested Approximation II.

References

- Boots, N.K. and H. Tijms (1999) "A multi-server queueing system with impatient customers," *Management Science*, **45**(3), 444-448.
- Boots, N.K. and H. Tijms (1998) "An M/M/C queue with impatient customers," presented at the First International Workshop on Retrial Queues, Universidad Complutense de Madrid, Madrid, September 22-24, 1998.
- Brown, M. and F. Proschan (1983) "Imperfect repair," *J. Appl. Prob.* **20**, 851-859.
- Cox, D.R. and W.L. Smith (1961) *Queues*, Chapman & Hall, London.
- Doytchinov, B., J. Lehoczky, and S. Shreve (1998) "Real-time queues in heavy traffic with earliest-deadline-first queue discipline," Technical Report, Depts. of Statistics and Mathematical Sciences, Carnegie-Mellon University, Pittsburgh, PA.
- Gaver, D.P. and P.A. Jacobs (1999) "A model for analyzing Blue force response to region invasion by multi-type Red forces," Naval Postgraduate School Technical Report, forthcoming.
- Jiang, Z., T.G. Lewis, and J-Y. Colin (1996) "Scheduling hard real-time constrained periodic tasks on multiple processors," *J. Systems & Software*, **19**(11), 102-118.
- Kleinrock, L. (1976) *Queueing Systems, Vol. I: Theory*, Wiley (Interscience), New York.
- Lehoczky, J.P. (1996) "Real-time queueing theory," *Proceedings of the IEEE Real-Time Systems Symposium*, December 1996, 186-195.
- Lehoczky, J.P. (1997a) "Using real-time queueing theory to control lateness in real-time systems," *Performance Evaluation Review*, **25**(1), 158-168.
- Lehoczky, J.P. (1997b) "Real-time queueing network theory," *Proceedings of the IEEE Real-Time Systems Symposium*, December 1997, 58-67.
- Liu, C.L. and J.W. Layland (1973) "Scheduling algorithms for multiprogramming in a hard real-time environment," *J. Automatic Computing Machinery*, **20**(1), 40-61.
- Whitt, Ward (1999) "Improving service by informing customers about anticipated delays," *Management Science*, **45**(2), 192-207.

APPENDIX A

Second Moment of ISOT

Squaring and taking conditional expectations leads to an expression for the second moment of completion time (needed for calculating long-run expected system occupancy when tasks are queued). We express the formula in terms of the expressions for the mean $E[C_j]$, (4.4), and the mean and second moment of recognition time K_j :

$$\begin{aligned}
 E[C_j^2] = & \left\{ \sum_k c_{jk} E[S_k^2] + 2 \sum_k c_{jk} E[S_j \bar{m}_{jk}(S_k)] b_{jk}^* E[C_j] \right. \\
 & + 2 \sum_k c_{jk} E[S_k m_{jk}(S_k)] \bar{b}_{jk} E[K_j] + \sum_k c_{jk} E[m_{jk}(S_k)] \bar{b}_{jk} E[K_j^2] \left. \right\} \\
 & \div \left(1 - \sum_k c_{jk} E[\bar{m}_{jk}(S_k)] b_{jk}^* \right)
 \end{aligned} \tag{A.1}$$

where

$$E[K_j] = \sum_k c_{jk} E[S_k] \div \left(1 - \sum_k c_{jk} \bar{b}_{jk} \right) \tag{A.2}$$

and

$$E[K_j^2] = \left(\sum_k c_{jk} E[S_k^2] + 2 \sum_k c_{jk} E[S_k] \bar{b}_{jk} E[K_j] \right) \div \left(1 - \sum_k c_{jk} \bar{b}_{jk} \right). \tag{A.3}$$

APPENDIX B

Model with Unobservable Exponential Deadlines

The Laplace-Stieltjes transform (LST) of the "completion" time can be calculated by taking conditional expectations in (4.1). The result is

$$E[e^{-\alpha_j}] = \left\{ \sum_k c_{jk} E[e^{-\alpha_k} [m_{jk}(S_k)b_{jk} + \bar{m}_{jk}(S_k)\bar{b}_{jk}^*]] + \sum_k c_{jk} E[e^{-\alpha_k} m_{jk}(S_k)\bar{b}_{jk}] E[e^{-\alpha_j}] \right\} \div \left(1 - \sum_k c_{jk} E[e^{-\alpha_k} \bar{m}_{jk}(S_k)] b_{jk}^* \right) \quad (B.1)$$

with

$$E[e^{-\alpha_j}] = \left\{ \sum_k c_{jk} E[e^{-\alpha_k}] b_{jk} \right\} \div \left(1 - \sum_k c_{jk} E[e^{-\alpha_k}] \bar{b}_{jk} \right) \quad (B.2)$$

Note that the above can be interpreted as the probability that a sojourn at the server, including repeats, ends before the termination of an independent exponential random variable with mean $1/\theta$.

To obtain the probability that a task service has been satisfactorily completed at sojourn completion and that sojourn ends before the termination of the independent exponential random variable with mean $1/\theta$, define

$$D_j = \begin{cases} S_k & \text{with probability } c_{jk} m_{jk}(S_k); \\ S_k + D_j' & \text{with probability } c_{jk} \bar{m}_{jk}(S_k) b_{jk}^*. \end{cases} \quad (B.3)$$

This is simply (4.1) with a term omitted. It is seen that

$$E[e^{-\alpha_j}] = \left(\sum_k c_{jk} E[e^{-\alpha_k} m_{jk}(S_k)] \right) \div \left(1 - \sum_k c_{jk} E[e^{-\alpha_k} \bar{m}_{jk}(S_k) b_{jk}^*] \right)$$

APPENDIX C
A Model with the Option to Reclassify
Unobservable Deadlines

In this appendix we present results for a model in which a server may opt to reclassify the task type if it perceives that the task is not complete. There is a time penalty Δ for reclassification. Let α be the probability that the server decides to reclassify the task after a task service which is perceived not to have completed the task.

$$C_{jk} = \begin{cases} S_k & \text{with probability } [m_{jk}(S_k)b_{jk} + \bar{m}_{jk}(S_k)\bar{b}_{jk}^*]; \\ S_k + C'_{jk} & \text{with probability } \bar{m}_{jk}(S_k)b_{jk}^*(1-\alpha); \\ S_k + C'_j + \Delta & \text{with probability } \bar{m}_{jk}(S_k)b_{jk}^*\alpha; \\ S_k + K_{jk} & \text{with probability } m_{jk}(S_k)\bar{b}_{jk}(1-\alpha); \\ S_k + K_j & \text{with probability } m_{jk}(S_k)\bar{b}_{jk}\alpha; \end{cases} \quad (C.1)$$

C_{jk} represents the service time of a task of type j that has been classified as a type k . C_j represents the service time of a task of type j . K_{jk} represents the random time until a completed task of type j that has been classified as a type k is so identified. K_j represents the random time until a completed task of type j is so identified, the recognition time:

$$K_{jk} = \begin{cases} S_k & \text{with probability } b_{jk} \\ S_k + K'_{jk} & \text{with probability } \bar{b}_{jk}(1-\alpha) \\ S_k + K'_j + \Delta & \text{with probability } \bar{b}_{jk}\alpha \end{cases} \quad (C.2)$$

$$C_j = C_{jk} \text{ with probability } c_{jk}, k = 1, \dots, J$$

$$K_j = K_{jk} \text{ with probability } c_{jk}, k = 1, \dots, J$$

Notice that the task accomplishment probability is allowed to depend explicitly on the allocated service time, S_k , and that the "completion" time can terminate with unrecognized incomplete task service, i.e. incompletely. The random variables C'_j , C'_{jk} , K'_j , and K'_{jk} above are independent stochastic replicas of C_j , C_{jk} , K_j and K_{jk} , as usual.

The expectations required are now calculated.

$$E[e^{-\theta_{jk}}] = \frac{E[e^{-\theta_k}]b_{jk} + E[e^{-\theta_k}]E[e^{-\theta_j}]E[e^{-\theta_j}]\bar{b}_{jk}\alpha}{1 - E[e^{-\theta_k}]\bar{b}_{jk}\alpha} \quad (C.3)$$

Thus

$$\begin{aligned} E[e^{-\theta_j}] &= \sum_k c_{jk} E[e^{-\theta_{jk}}] \\ &= \frac{\sum_k c_{jk} f_{jk}(K)}{1 - \sum_k c_{jk} a_{jk}(K)} \end{aligned} \quad (C.4)$$

where

$$\begin{aligned} f_{jk}(K) &= \frac{E[e^{-\theta_k}]b_{jk}}{1 - E[e^{-\theta_k}]\bar{b}_{jk}(1 - \alpha)} \\ a_{jk}(K) &= \frac{E[e^{-\theta_k}]E[e^{-\theta_j}]\bar{b}_{jk}\alpha}{1 - E[e^{-\theta_k}]\bar{b}_{jk}(1 - \alpha)}. \end{aligned} \quad (C.5)$$

Similiarly

$$E[e^{-\theta_{jk}}] = \frac{N_{jk}(C) + g_{jk}(C)E[e^{-\theta_j}]}{D_{jk}(C)} \quad (C.6)$$

where

$$\begin{aligned} N_{jk}(C) &= E[e^{-\theta_k} [m_{jk}(S_k)b_{jk} + \bar{m}_{jk}(S_k)\bar{b}_{jk}^*]] \\ &\quad + E[e^{-\theta_k}]E[e^{-\theta_j}]E[e^{-\theta_j}]\bar{b}_{jk}\alpha \\ &\quad + E[e^{-\theta_k}m_{jk}(S_k)]E[e^{-\theta_{jk}}]m_{jk}\bar{b}_{jk}(1 - \alpha). \end{aligned} \quad (C.7)$$

Further,

$$D_{jk}(C) = 1 - E[e^{-\theta_k}\bar{m}_{jk}(S_k)]\bar{b}_{jk}^*\alpha \quad (C.8)$$

and

$$g_{jk}(C) = E[e^{-\alpha_k} \bar{m}_{jk}(S_k)] E[e^{-\alpha_k}] b_{jk}^* \alpha. \quad (C.9)$$

Thus,

$$E[e^{-\alpha_j}] = \sum_k c_{jk} E[e^{-\alpha_{jk}}] = \frac{\sum_k c_{jk} f_{jk}(C)}{1 - \sum_k c_{jk} a_{jk}(C)} \quad (C.10)$$

where

$$f_{jk}(C) = \frac{N_{jk}(C)}{D_{jk}(C)} \quad (C.11)$$

and

$$a_{jk}(C) = \sum_k c_{jk} \frac{g_{jk}(C)}{D_{jk}(C)}. \quad (C.12)$$

After appropriate weighting by p_j , the probability that an arrival is of type j , the needed moments and transforms can be calculated as was done previously.

APPENDIX D

Forward Kolmogorov (Takaçs-Beneš) Equations with Exponential Balking

Suppose that tasks arrive at a service facility according to a Poisson process with rate λ . Service times are independent and identically distributed. Let $W(t)$ be the total virtual work in the system at time t . Each task has a deadline which is exponentially distributed with mean $1/\theta$: if the waiting time or virtual work present when the task arrives exceeds the deadline the task does not enter the system. This *approximates* the situation in which tasks whose deadlines have elapsed *when they reach the server* are not served. With some modification it addresses the situation in which a deadline elapses *during* service.

D.1 Statistically Identified Deadlines and Service to Completion

Let the distribution function of $W(t)$ be

$$F_W(x; t; \theta) = P\{W(t) \leq x\}$$

and express this as

$$F_W(x; t; \theta) = p_0(t; \theta) + \int_0^x p(z; t; \theta) dz,$$

where

$$p_0(t; \theta) = P\{W(t) = 0\}.$$

Since, given $W(t)$, the task joins the queue with probability $e^{-\theta W(t)}$, the probability its deadline does not expire while in queue, one can write

$$\begin{aligned} p(x; t + \Delta t; \theta) &= p(x + \Delta t; t; \theta) [1 - \lambda e^{-(x + \Delta t)\theta} \Delta t] \\ &\quad + p_0(t; \theta) b(x) \lambda \Delta t + \lambda \Delta t \int_0^x e^{-\theta y} p(y; t; \theta) b(x - y) dy + o(\Delta t) \end{aligned} \tag{D.1a}$$

where b is the density function of the positive service time C . Also,

$$p_0(t + \Delta t; \theta) = p_0(t; \theta) [1 - \lambda \Delta t] + p(0, t; \theta) \Delta t (1 - \lambda \Delta t) + o(\Delta t). \tag{D.1b}$$

Taking limits as $\Delta t \rightarrow 0$ results in

$$\frac{d}{dt} p_0(t; \theta) = -\lambda p_0(t; \theta) + p(0, t; \theta), \quad (\text{D.2})$$

and also

$$\begin{aligned} \frac{\partial}{\partial t} p(x; t; \theta) &= \frac{\partial}{\partial x} p(x; t; \theta) - p(x; t; \theta) \lambda e^{-\alpha} + \lambda p_0(t; \theta) b(x) \\ &\quad + \lambda \int_0^x e^{-\theta y} p(y; t; \theta) b(x-y) dy \end{aligned} \quad (\text{D.3})$$

If $t \rightarrow \infty$, then a steady-state density satisfies

$$0 = \frac{\partial}{\partial x} p(x; \theta) - p(x; \theta) \lambda e^{-\alpha} + p_0 b(x) \lambda + \lambda \int_0^x e^{-\theta y} p(y) b(x-y) dy, \quad (\text{D.4a})$$

$$0 = -\lambda p_0 + p(0). \quad (\text{D.4b})$$

Laplace transform to obtain

$$p^*(s; \theta) = \int_0^\infty e^{-sx} p(x; \theta) dx \text{ and } b^*(s) = \int_0^\infty e^{-sx} b(x) dx.$$

Thus, (D.4a) implies

$$s p^*(s; \theta) = [p(0; \theta) + \lambda p^*(s + \theta; \theta)] [1 - b^*(s)] \quad (\text{D.5})$$

Let

$$\begin{aligned} \psi(s; \theta) &= \int_0^\infty e^{-sx} dF_w(x; \theta) = p_0(\theta) + \int_0^\infty e^{-sx} p(x; \theta) dx \\ &= p_0(\theta) + p^*(s; \theta); \end{aligned} \quad (\text{D.6})$$

then

$$\psi(s; \theta) = p_0(\theta) + \rho \psi(s + \theta; \theta) \delta(s) \quad (\text{D.7})$$

where

$$\delta(s) = \frac{1 - b^*(s)}{s E[C]}. \quad (\text{D.8})$$

Since $\psi(0; \theta) = 1 = p_0 + \rho \psi(\theta; \theta)$,

$$p_0 = 1 - \rho\psi(\theta, \theta) = 1 - \lambda E[e^{-\theta W}] E[C]. \quad (D.9)$$

which motivates the heuristic approximation (actually lower bound) of Section 5.2.

Iterative solution to the equation (D.7)

Since

$$\psi(s; \theta) = [1 - \rho\psi(\theta, \theta)] + \rho\psi(s + \theta, \theta)\delta(s)$$

it follows that, putting $s = \theta$, and defining

$$\psi(\theta) \equiv \psi(\theta, \theta) = [1 - \rho\psi(\theta)] + \rho\psi(2\theta, \theta)\delta(\theta)$$

and

$$\psi(2\theta) \equiv \psi(2\theta, \theta) = [1 - \rho\psi(\theta)] + \rho\psi(3\theta, \theta)\delta(2\theta).$$

Substituting the expression for $\psi(2\theta)$ into that for $\psi(\theta)$ results in

$$\begin{aligned} \psi(\theta) &= [1 - \rho\psi(\theta)] + \rho\{[1 - \rho\psi(\theta)] + \rho\psi(3\theta, \theta)\delta(2\theta, \theta)\}\delta(\theta) \\ &= 1 + \rho\delta(\theta) - \rho\psi(\theta)[1 + \rho\delta(\theta)] + \rho^2\psi(3\theta, \theta)\delta(2\theta)\delta(\theta). \end{aligned}$$

Continuing in this manner results in the equation

$$\begin{aligned} \psi(\theta) &= A(0, \theta) + \dots + A(n; \theta) - \rho\psi(\theta)[A(0, \theta) + \dots + A(n; \theta)] \\ &\quad + \rho\psi((n+1)\theta, \theta)A(n; \theta). \end{aligned} \quad (D.10a)$$

where

$$\begin{aligned} A(0, \theta) &= 1 \\ A(n; \theta) &= \rho^n \delta(n\theta) \delta((n-1)\theta) \times \dots \times \delta(\theta). \end{aligned} \quad (D.10b)$$

For $\theta > 0$, $A(n; \theta) \rightarrow 0$. Thus the probability that an arriving task joins and survives the queue before deadline elapse is

$$\psi(\theta) = \frac{\sum_{k=0}^{\infty} A(k; \theta)}{1 + \rho \sum_{k=0}^{\infty} A(k; \theta)}. \quad (D.11)$$

It is clear that the infinite sums converge exponentially rapidly for $\theta > 0$, and that this is true for any ρ -value.

Further, for $s \neq \theta$

$$\psi(s; \theta) = [1 - \rho\psi(\theta, \theta)] \sum_{k=0}^{\infty} C(k; s) \quad (\text{D.12})$$

where

$$C(k; s) = \rho^k \prod_{i=0}^{k-1} \delta(s + i\theta), \quad k \geq 1 \quad (\text{D.13})$$

and

$$C(0; s) = 1.$$

D.2 Services Subject to Exponential Deadline

If a task deadline's elapse is detectable during service and the task then removed, then the distribution of service time, C , must be replaced by that of $C_T = \min(C, \text{deadline})$, the *allowed service time*. Consequently the service times *that contribute to the virtual waiting time* are, thanks to the exponential deadline assumption, iid with mean

$$E[C_T] = \frac{1 - E[e^{-\theta C}]}{\theta} \quad (\text{D.14})$$

and tail-transform now

$$\delta_T(s; \theta) = \frac{1 - E[e^{-(\theta+s)C}]}{(\theta+s)E[C_T]} = \frac{\delta(\theta+s)}{\delta(\theta)}. \quad (\text{D.15})$$

These replace $E[C]$ in ρ , and $\delta(s)$ in the previous solution, (D.11).

D.3 Class-Specific Deadlines

A natural generalization of the above is to allow independent Poisson arrivals from J task classes, the j^{th} rate being λ_j with service time density b_j and exponential deadline parameter θ_j . Arguments analogous to those in Appendix D enable us to show that

$$\psi(s; \theta) = p_0(\theta) + \sum_j \rho_j \gamma_j(s) \psi(s + \theta_j; \theta) \quad (\text{D.16})$$

where

$$\gamma_j(s) = \frac{1 - b_j^*(s)}{s E[C_j]}, \quad (\text{D.17})$$

where $\rho_j = \lambda_j E[C_j]$, and

$$p_0(\theta) = 1 - \sum_j \rho_j \psi(\theta_j; \theta). \quad (\text{D.18})$$

It can be seen that

$$p_0(\theta) \geq \frac{1}{1 + \sum_j \rho_j} \equiv \frac{1}{1 + \rho}. \quad (\text{D.19})$$

The equation (D.16) can be solved in closed form (a series) by successive substitution/iteration, but this step is omitted.

DISTRIBUTION LIST

1. Research Office (Code 09) 1
 Naval Postgraduate School
 Monterey, CA 93943-5000

2. Dudley Knox Library (Code 013) 2
 Naval Postgraduate School
 Monterey, CA 93943-5002

3. Defense Technical Information Center 2
 8725 John J. Kingman Rd., STE 0944
 Ft. Belvoir, VA 22060-6218

4. Therese Bilodeau (Editorial Assistant) 1
 Dept of Operations Research
 Naval Postgraduate School
 Monterey, CA 93943-5000

5. Prof. Donald P. Gaver (Code OR/Gv) 1
 Dept of Operations Research
 Naval Postgraduate School
 Monterey, CA 93943-5000

6. Prof. Patricia A. Jacobs (Code OR/Jc) 1
 Dept of Operations Research
 Naval Postgraduate School
 Monterey, CA 93943-5000

7. SEW Strategic Planning Office, N6C3 1
 2000 Navy Pentagon, Rm 5C633
 Washington, DC 20350-2000

8. Dr. Alfred G. Brandstein 1
 MCCDC
 Studies and Analysis Division
 3093 Upshur Avenue
 Quantico, VA 22134-5130

9. Prof. Sir David Cox 1
 Nuffield College
 Oxford OX1 1NF
 ENGLAND

10. Dr. D. F. Daley 1
Statistics Dept. (I.A.S.)
Australian National University
Canberra, A.C.T 2606
AUSTRALIA
11. Prof. J. Michael Harrison 1
Graduate School of Business
Stanford University
Stanford, CA 94305-5015
12. Dr. F. P. Kelly 1
Statistics Laboratory
16 Mill Lane
Cambridge
ENGLAND
13. Prof. J. Lehoczky 1
Department of Statistics
Carnegie-Mellon University
Pittsburgh, PA 15213