MANAGEMENT SCIENCE
Vol. 34, No. 9, September 1988
Printed in U.S.A.

$M/G/c$ QUEUEING SYSTEMS WITH MULTIPLE CUSTOMER CLASSES: CHARACTERIZATION AND CONTROL OF ACHIEVABLE PERFORMANCE UNDER NONPREEMPTIVE PRIORITY RULES*

A. FEDERGRUEN AND H. GROENEVELT
Graduate School of Business, Columbia University, New York, New York 10027
Simon School of Business Administration, University of Rochester, Rochester, New York 14627

This paper considers an $M/G/c$ queueing system serving a finite number (J) of distinct customer classes. Performance of the system, as measured by the vector of steady-state expected waiting times of the customer classes (the *performance vector*), may be controlled by adopting an appropriate priority discipline.

We show that the performance space, the set of performance vectors which are achievable under some nonpreemptive work conserving priority rule, is a polyhedron described by $2^J - 1$ inequalities. The special (polymatroidal) structure of this polyhedron, nevertheless, allows for efficient ($O(J^2 \log J)$) procedures to minimize any convex (separable) function of the performance vector.

Linear objectives are shown to be minimized by absolute priority rules, thus generalizing a well-known result for $M/G/1$ systems. We also show that each point in the performance space may be achieved by a *unique*, generalized dynamic priority rule, specified by $J - 1$ parameters, which may be determined by the recursive solution of $J - 1$ single variable quadratic equations. This class of rules contains the absolute priority rules and the (pure) dynamic rules as special cases. Our results are accurate up to one, extremely accurate, approximation and completely exact for $M/G/1$ and $M/M/c$ systems as well as in heavy traffic.
(QUEUEING SYSTEMS; NONPREEMPTIVE PRIORITY RULES; CUSTOMER CLASSES)

1. Introduction and Summary

Queueing models are increasingly used for the analysis and design of complex production and service systems in which different classes of users (or "customers") compete for a limited number of shared resources (or "servers"). It is often possible to classify the customers in a finite number of distinct classes and to apply a specific type of preferential treatment to one class at the expense of others. Such schemes are referred to as priority queueing systems.

Examples include production facilities which manufacture batch orders for a number of distinct products with the same equipment and/or operators. Often, different service level requirements and/or holding cost rates apply to different items, so that significantly different economic consequences result from the delays or sojourn times experienced by the various items. In modern telecommunication systems, heterogeneous data types (e.g., interactive messages, computer outputs, file transfers, facsimile, etc.) compete with voice for the limited availability of shared transmission equipment, e.g., buses in a local area network or frequency bands in a satellite channel. Appropriate priority systems need to be designed to achieve an optimal trade-off between (the economic consequences of) the delays encountered by the different traffic types. In other systems, the objective is to achieve an *equitable* scheduling procedure of the different customer types for access to the shared resource(s).

* Accepted by John P. Lehoczky, former Departmental Editor; received June 29, 1986. This paper has been with the authors 3 months for 1 revision.

0025-1909/88/3409/1121$01.25

When designing such priority systems, it is natural to think in terms of minimizing some cost function with respect to the vector of (average) delays experienced by the different customer classes. Most of the literature on priority queueing systems is concerned with the performance analysis of a specific priority rule in a given queueing model. Surprisingly little attention has been given to the *design* of queueing disciplines which minimize well-stated and realistic cost functions.

The $M/G/c$ queueing system (with Poisson arrivals, independent service times with an arbitrary probability distribution, and a pool of identical servers) is arguably the most commonly used multi-server model. This paper considers an $M/G/c$ queueing system serving a finite number of distinct customer classes $E = \{1, \ldots, J\}$. Performance of the system, as measured by the vector of steady-state expected waiting times of the customer classes (the *performance vector*), may be controlled by adopting an appropriate priority discipline. We consider the class of all nonpre-emptive and strongly *work conserving* rules; see §2 for a precise definition.

Exact evaluation of the performance vector of even a simple priority rule like FIFO is not possible in the general $M/G/c$ model. However, various approximation methods exist some of which are extremely accurate, see Tijms (1985) for a survey. Our results are accurate up to one such approximation (which is exact for $M/G/1$ and $M/M/c$ and asymptotically exact for heavy traffic systems).

The main results are the following: we characterize the *performance space*, the set of performance vectors which are achievable under some (nonpreemptive) work conserving rule. The latter is shown to be a polyhedral set in $\mathbf{R}^J$ described by $2^J - 1$ (in)equalities. Normally this characterization would preclude tractability of any kind of optimization (or trade-off analysis) over the performance space for all but very small values of J. Fortunately the performance space is a polyhedron of a very special structure: up to a simple scaling transformation the polyhedron is the base (of the independence polytope) of a so-called *polymatroid* (cf. e.g., Edmonds 1970, Welsh 1976). This result allows for efficient algorithms to minimize system-wide performance measures expressed as convex functions of the performance vector. Additional structure may be brought to bear to obtain efficient implementations of these algorithms requiring no more than $O(J^2 \log J + J\chi)$ operations where χ is the time needed to solve a certain type of single variable (nonlinear) equation. (For an important class of objective functions $\chi = O(J)$, so the computational bound reduces to $O(J^2 \log J)$.)

In addition, the polymatroidal structure explains the optimality of absolute priority rules for *linear* objectives, a result well-known for the *single* server case (see Gelenbe and Mitrani 1980, Fife 1965, Smith 1956 and Kleinrock 1976). An absolute priority rule ranks the classes in a given sequence and determines priorities on the basis of class ranks only (breaking ties on a FIFO basis).

In addition to characterizing the performance space and reviewing algorithms to optimize performance measures over this space, we address the issue of *synthesis*: for a given achievable performance vector specify a *simple* priority discipline under which this vector may be achieved.

While a randomization of absolute priority rules can easily be constructed to correspond with any given achievable performance vector, such randomizations may be hard to implement and exhibit large variances in the long-run waiting times. Instead we show that the synthesis problem may be resolved using a slight generalization of the *dynamic* (Jackson 1960) or *delay dependent* (Kleinrock 1976) scheduling rules where a customer's priority is proportional to his time spent in queue, the proportionality constant being class dependent.

The above results are only partially extendable to more general models in systems with general (non-Poisson) arrival streams. The performance space remains *contained within* a polyhedral set of the above described type. A counterexample shows however

that the performance space may be a strict (as of yet uncharacterized) subset of this polyhedron.

Federgruen and Groenevelt (1986) discuss the characterization and control of achievable performance in *preemptive* systems. Results similar to ours are obtained for systems with general arrival processes but exponential service times. Gelenbe and Mitrani (1980) characterized the performance space for the single server case, i.e., for nonpreemptive $M/G/1$ systems. Mitrani (1982) achieved the same in $M/G/1$ systems in which the service time of each customer is known upon his arrival and where this information may be used in assigning priorities. (A partial characterization of this case can already be found in Kleinrock et al. 1971.)

The synthesis problem in multiclass $M/M/1$ models with processor sharing was addressed by Fayolle et al. (1978) and Mitrani and Hine (1971). Generalizations of Kleinrock's delay dependent priority rules have been investigated by Kleinrock and Finkelstein (1967), Netterman and Adiri (1979), and Bagchi and Sullivan (1985). Our synthesis algorithm modifies and generalizes a procedure in Wood and Sargent (1984).

In §2 we give notation and some preliminary results. The performance space is characterized in §3; we conclude that section with a brief review of optimization algorithms for system wide performance measures. §4 gives a synthesis algorithm determining a dynamic priority rule for each achievable performance vector. In §5 we discuss possible generalizations of our results.

2. Notation and Preliminaries

We first introduce some notation and assumptions. The customer classes arrive to the system according to independent Poisson processes; λ_j denotes the arrival rate of class j, $j \in E$. The service times of the customers in a given class $j \in E$ are assumed to be independent and identically distributed as a random variable V_j with finite second moment. Let $\rho_j = \lambda_j E V_j$, $j \in E$. When a customer arrives, only his class is known but not his actual service time.

A rule R is called strongly work conserving if

(W_1) no server is free when a customer is in the queue;

(W_2) the discipline does not affect the amount of service time given to a customer or the arrival time of any customer;

(W_3) priorities are assigned on the basis of the history of the process and the time elapsed since the last epoch at which the system became empty.

Let **R** be the class of rules satisfying (W_1)–(W_3). Conditions (W_1) and (W_2) are standard, see e.g. Heyman and Sobel (1982). Condition (W_3) is similar to one stated in Gelenbe and Mitrani (1980) and appears to be the most general, easily describable restriction under which the existence of long-run averages of waiting times can be verified, i.e., under which the performance vector is properly defined. (The statement on p. 432 in Heyman and Sobel (1982) that conditions (W_1) and (W_2) are sufficient, appears incorrect.)

To ensure that the work-in-system process is (stochastically) independent of the priority rule used we need the following restriction:

(C) if $c > 1$, assume all customers have the same service time distribution.

A key tool in the characterization of the performance space is provided by the following work conservation law which is due to Heyman and Sobel (1982), generalizing a proof in Schrage (1970) for $G/G/1$ queues. (This work conservation law applies in fact for systems with far more general arrival processes, see §5.) For a given priority rule in **R**, let

W_{nj} = delay of the nth customer of class j ($j \in E$; $n \geq 1$),

$A(t)$ = work in system at time t ($t > 0$),

$A_{\text{FIFO}}(t)$ = work in system at time t, under FIFO ($t > 0$).

LEMMA 1 (Work conservation law). *Assume* $\sum_{j=1}^{J} \rho_j/c < 1$ *and condition* (C). *Fix a rule* $R \in \mathbf{R}$.

(a) *There are numbers* $W_j^*, j \in E$ *such that*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} W_{nj} = W_j^* \qquad \text{(w.p.1)}, \qquad j = 1, \ldots, J.$$

(b) *The long-run average work in system* A^* *exists and is independent of the priority rule*:

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T A(t)dt = \lim_{T\to\infty} \frac{1}{T} \int_0^T A_{\text{FIFO}}(t)dt = A^* \qquad \text{w.p.1.} \tag{1}$$

(c) $\sum_{j=1}^{J} \rho_j W_j^* = A^* - \frac{1}{2} \sum_{j=1}^{J} \lambda_j E V_j^2$.

PROOF. (a) The system is regenerative under any rule in $\mathbf{R}$, with ends of busy cycles as regeneration points: The condition $\sum_j \rho_j/c < 1$ guarantees that the length of a busy cycle has a finite expectation, see Wolff (1984) or Whitt (1982). Part (a) now follows from a standard application of the renewal reward theorem.

(b) The existence of the long run average may be verified as in part (a); the independence with respect to the adopted priority rule follows from (W_1), (W_2) and the service time assumptions.

(c) See Theorem 11–13 in Heyman and Sobel (1982) the proof of which is based on an application of the $H = \lambda G$ identify, cf. ibid. and Heyman and Stidham (1980). ■

Next define for all $S \subset E$ and $R \in \mathbf{R}$, $A_R^*(S)$ and $W_R^*(S)$ as the long-run average work in system and the long run average waiting time under rule R for customers in the collection of classes S. Also, let

$$A^*(S) = \inf\{A_R^*(S): R \in \mathbf{R}\}, \qquad S \subset E.$$

The following lemma shows that $A^*(S)$ is achieved by *any* rule which assigns absolute priority to S-customers above all other classes. In particular, $A^*(S)$ may be achieved by lumping all S-customers in a single "class" and all other customers in a second "class" giving head-of-the-line priority to the S-customers and breaking ties according to FIFO, otherwise. We refer to this discipline as the S-priority rule. Let $W^*(S)$ denote the long-run average waiting time for S-customers under this rule.

LEMMA 2. *Assume condition* (C) *holds.*

(a) *Let* $V_0 \overset{\text{def}}{=}$ *the initial delay an arbitrary customer experiences (if any) until the first epoch at which a server becomes available for service. The distribution of* V_0 *is independent of the rule* $R \in \mathbf{R}$, *and* $E(V_0) < \infty$.

(b) $A^*(S)$ *is achieved by any rule in* $\mathbf{R}$ *which assigns absolute priority to customers in the collection of classes* S *over all other classes,* $S \subset E$.

(c) *Consider an absolute priority rule* R *and assume the classes are numbered such that under rule* R, *class* i *has priority over class* j *iff* $i > j$. *Then, simultaneously,* $A_R^*(\{l, \ldots, J\}) = A^*(\{l, \ldots, J\})$, $l = 1, \ldots, J$.

PROOF. (a) We distinguish between two cases:

(i) *the single server case,* $c = 1$: Let $(V_0 | j)$ denote the residual service time of the customer in service (at an arbitrary epoch), given a customer of class $j \in E$ is being served. Also, let p_j = steady-state probability of a customer of class j being served, $j \in E$. Clearly, V_0 is a mixture of the $(V_0 | j)$-distributions with $\{p_j, j = 1, \ldots, J\}$ as the mixing probabilities. It thus suffices to show that the distributions of all $(V_0 | j)$ variables as well as $\{p_j, j \in E\}$ are independent of the rule $R \in \mathbf{R}$. The former follows from

Green (1982) who also showed that $E(V_0 \mid j) = \frac{1}{2}(EV_j^2)/EV_j$; $p_j = \rho_j$ (independent of $R \in \mathbf{R}$) follows from an application of Little's law. Note, finally that

$$EV_0 = \tfrac{1}{2} \sum_{j=1}^{J} \lambda_j EV_j^2. \tag{2}$$

(ii) *the multiserver case*, $c > 1$: immediate from condition (C).

(b) Let $A_R(S; t)$ be the work in systems due to S-customers, at time t and under rule R. We distinguish between the same two cases as in part (a).

(i) The $\{A_R(S; t),\ t \geqq 0\}$ process has jumps whenever S-customers arrive. These arrival epochs and the sizes of the jumps (the customers' service times) are independent of the priority rule in view of (W_2). In addition $A_R(S; t)$ decreases at rate 1 whenever an S-customer is served. We conclude that $A_R(S; t)$ is minimized (simultaneously for all $t > 0$ and on each sample path) by giving absolute priority to S-customers whenever possible. Moreover, the distribution of $A_R(S; t)$ is independent of the relative priorities assigned among S-customers. Consider thus a rule R which determines these relative priorities according to FIFO. Let $W_j(R)[N_j(R)]$ denote the expected steady state waiting time [number of customers] in queue for customers in class j and under rule R. As before we obtain using part (a) that

$$\begin{aligned} W_j(R) &= E(V_0) + \sum_{l \in S} N_l(R) E(V_l) = E(V_0) + \sum_{l \in S} \lambda_l W_l(R) E(V_l) \\ &= E(V_0) + \sum_{l \in S} \rho_l W_l(R), \qquad j \in S, \end{aligned}$$

invoking Little's law. Since $\sum_{j \in S} \rho_j < 1$ this system of equations has the unique solution

$$W_j(R) = W^*(S) = E(V_0)/[1 - \sum_{l \in S} \rho_l]. \tag{3}$$

It follows from the proof of Lemma 1(c) that

$$\begin{aligned} A^*_R(S) &= \sum_{j \in S} \rho_j W_j(R) + \tfrac{1}{2} \sum_{j \in S} \lambda_j E(V_j^2) \\ &= (\sum_{j \in S} \rho_j) W^*(S) + \tfrac{1}{2} \sum_{j \in S} \lambda_j E(V_j^2). \end{aligned} \tag{4}$$

Thus, $A^*_R(S)$ is identical for all rules R giving absolute priority to customers in S above all other customers.

(ii) As in the single-server case, one easily verifies that $A_R(S; t)$ is minimized (simultaneously for all t and on every sample path) by assigning absolute priority to S-customers (whenever possible). Note that priorities need only to be determined at service completion epochs at which $c - 1$ servers remain busy. The state of the system at such epochs is described by the queue lengths for all classes $j \in E$ and the elapsed service times of the $c - 1$ busy servers. In view of condition (C), the distribution of the continuation of the $\{A_R(S; t)\}$ process (given the current state of the system) is only dependent on whether a customer in S or in $E \backslash S$ is given priority; it does *not* depend on the *specific* customer in S or in $E \backslash S$ to be granted priority. The distribution of the $\{A_R(S; t)\}$ process is thus independent of the *relative* priorities assigned to customers in S and in $E \backslash S$.

(c) Immediate from part (b). ■

We conclude this section with a number of definitions. A set function $h: 2^E \to \mathbf{R}$ is called *nondecreasing* if $h(T) \leqq h(S)$ whenever $T \subset S$, and *supermodular* (*submodular*) if $h(S \cup \{j\}) - h(S) \geqq (\leqq)\ h(T \cup \{j\}) - h(T)$ for all $T \subset S$ and $j \notin S$.

DEFINITION 1. Let f be a real valued function. Let $\alpha \in R^J$ be a vector of positive

weights. The set function $h: 2^E \to \mathbf{R}$ defined by $h(S) = f(\sum_{j\in S} \alpha_j)$ is called *generalized symmetric*.

One easily verifies that a generalized symmetric set function is nondecreasing if f is nondecreasing, and supermodular (submodular) if f is convex (concave). For a given set function $h: 2^E \to \mathbf{R}$, a polyhedron $X = \{x \in \mathbf{R}^J: \sum_{j\in S} x_j \leqq h(S), S \subset E\}$ is called the (independence polytope) of a *polymatroid* provided $h(\varnothing) = 0$ and $h(\cdot)$ is nondecreasing and submodular (cf., e.g., Edmonds 1970, Welsh 1976).

3. The Performance Space

In this section we characterize the performance space.

THEOREM 1 (Necessary conditions for achievability). *Assume condition* (C) *holds. If a vector W represents an achievable performance vector corresponding with a rule $R \in \mathbf{R}$, then*

$$\sum_{j\in S} \rho_j W_j \geqq (\sum_{j\in S} \rho_j) W^*(S) = A^*(S) - \tfrac{1}{2} \sum_{j\in S} \lambda_j E(V_j^2), \tag{5}$$

$$\sum_{j=1}^{J} \rho_j W_j = (\sum_{j=1}^{J} \rho_j) W^*(E) = A^*(E) - \tfrac{1}{2} \sum_{j=1}^{J} \lambda_j E(V_j^2). \tag{6}$$

Each of the lower bounds in (5) *is tight.*

PROOF. The proof of Theorem 11–13 in Heyman and Sobel (1982) shows for any $S \subset E$, that

$$\sum_{j\in S} \rho_j W_j = A_R^*(S) - \tfrac{1}{2} \sum_{j\in S} \lambda_j E(V_j^2) \geqq A^*(S) - \tfrac{1}{2} \sum_{j\in S} \lambda_j E(V_j^2).$$

Strict equality holds for the S-priority rule (see Lemma 2). Under the latter rule, $W_j = W^*(S)$, $j \in S$. This proves (5) and (together with Lemma 1) (6). ■

Let $\mathbf{W}^* = \{W \in \mathbf{R}^J: W \text{ satisfies (5) and (6)}\}$. Subtracting the inequalities (5) from (6) we obtain the following alternative representation of $\mathbf{W}^*$: Let

$$b^*(S) = A^*(E) - A^*(E \backslash S) - \tfrac{1}{2} \sum_{j\in S} \lambda_j E(V_j^2), \qquad S \subset E. \tag{7}$$

Thus,

$$\mathbf{W}^* = \{W \in \mathbf{R}^J: W \geq 0 \text{ satisfies (6) and the inequalities } \sum_{j\in S} \rho_j W_j \leqq b^*(S), S \subset E\}. \tag{8}$$

We state the following assumption:

Assumption (A). $A^*(\cdot)$ is a supermodular set function.

This assumption clearly holds in the single server model: substitute (3) into (4) to conclude

$$A^*(S) = (\sum_{j\in S} \rho_j) E(V_0)/[1 - \sum_{l\in S} \rho_l] + \tfrac{1}{2} \sum_{j\in S} \lambda_j E(V_j^2) \tag{9}$$

with $E(V_0)$ independent of $S \subset E$, see (2). Note that the first term in (9) is in fact a generalized symmetric, nondecreasing and supermodular set function, with $f(x) = E(V_0)x/(1 - x)$ and $\alpha_j = \rho_j$, $j \in E$. (It follows that $b^*(\cdot)$ is generalized symmetric nondecreasing and submodular.) Likewise in the multiserver case with exponential service times ($M/M/c$) we have (see Gross and Harris 1974, p. 194) $W^*(S) = E(V_0)/[1 - c^{-1} \sum_{l\in S} \rho_l]$, and hence (as in the proof of Lemma 2)

$$A^*(S) = (\sum_{j\in S} \rho_j) E(V_0)/[1 - c^{-1} \sum_{l\in S} \rho_l] + \tfrac{1}{2} \sum_{j\in S} \lambda_j E(V_j^2). \tag{10}$$

The first term in (10) is again a generalized symmetric nondecreasing and supermodular set function, so that $b^*(S)$ is generalized symmetric, nondecreasing and submodular in this case as well. (The derivation of (10) is analogous to that of (9), see the proof of Lemma 2.)

For multiserver models with a deterministic service time distribution, it has been shown in Federgruen and Groenevelt (1987, Theorem 2) that the $A^*(\cdot)$ function is supermodular, i.e., Assumption (A) holds. It is, however, unknown whether the supermodularity property holds for multiserver models with general (nondeterministic and nonexponential) service time distributions. (Federgruen and Groenevelt 1987 provide a counterexample, however with deterministic interarrival times rather than Poisson arrivals.) In fact, for this most general case, no *exact* expressions for $W^*(S)$ are known; even the expected delay under FIFO cannot be exactly evaluated. For the latter an approximation formula does, however, exist, derived independently by a number of authors (Lee and Longton 1957, Krampe et al. 1973, Maaloe 1973, Stoyan 1976, Nozaki and Ross 1978, Hokstad 1978 and Tijms et al. 1981) under different approximation assumptions:

$$W_{\text{FIFO}} = \left(\frac{EV^2}{2EV}\right)^{-1} cB/[1 - c^{-1}\rho] \tag{11}$$

where $\rho = \sum_{j=1}^{J} \rho_j$, V is the service time random variable (common to all classes under condition (C) and B is the probability of delay in the $M/M/c$ system with the same expected service time. The approximation formula is exact in heavy traffic (see Boxma et al. 1979) as well as in the $M/G/1$ and $M/M/c$ cases. Most importantly, empirical studies have shown that the relative approximation errors are very small indeed, see Tijms et al. (1981), Seelen et al. (1985), Seelen and Tijms (1985), Van Hoorn (1984) and Groenevelt et al. (1984). In the remainder we use:

Assumption (A'). When $c > 1$, and when all servers are busy, the intervals between consecutive service completion are independent of the queue size and distributed as V/c.

Assumption (A') is one of several under which approximation formulae (11) for the expected waiting times W_{FIFO} may be derived, see, e.g., Tijms et al. (1981). Moreover, under (A'), Assumption (A) holds. Assumption (A') holds of course *exactly* in single-server systems and in $M/M/c$ queues. On the basis of the same assumption, one easily derives the approximation formula

$$W(S) = \left(\frac{EV^2}{2EV}\right) Bc^{-1}/[1 - c^{-1} \sum_{l \in S} \rho_l],$$

and hence

$$A^*(S) = \left(\sum_{l \in S} \rho_l\right)\left(\frac{EV^2}{2EV}\right) Bc^{-1}/[1 - c^{-1} \sum_{l \in S} \rho_l] + \tfrac{1}{2}\left(\sum_{j \in S} \lambda_j\right) EV^2. \tag{12}$$

Note that the first term to the right of (12) is again generalized symmetric, nondecreasing and supermodular. Employing formula (12) this yields a generalized symmetric, nondecreasing and submodular $b^*(\cdot)$ function.

Theorem 2 below shows that W^* is in fact the performance space (up to approximation (A)).

THEOREM 2. *Let condition* (C) *and Assumption* (A) *hold.*

(a) $X^* \overset{\text{def}}{=} \{x \in \mathbf{R}^J : x \geq 0, \sum_{j \in S} x_j \leqq b^*(S), S \subset E^J \text{ and } \sum_{j=1} x_j = b^*(E)\}$ *is the base of a polymatroid.*

(b) *The performance vector of any absolute priority rule is an extreme point of*

$\mathbf{W}^*$; *conversely, each extreme point of* $\mathbf{W}^*$ *is the performance vector of an absolute priority rule.*

(c) $\mathbf{W}^*$ *is the performance space.*

PROOF. (a) It is easily shown that $A^*(\cdot)$ is nondecreasing. (The proof is analogous to that of Federgruen and Groenevelt 1985, Lemma 1.) In view of (A) and (7), it follows that $b^*(\cdot)$ is nondecreasing and submodular with $b^*(\varnothing) = 0$.

(b) The proof of part (B) is analogous to that of Theorem 2 in Federgruen and Groenevelt (1987b).

(c) We conclude from part (b) that each point in $\mathbf{W}^*$ is the performance vector of an appropriate randomization of absolute priority rules. ■

Optimization of System Performance Measures

The performance space is thus a polyhedron described by $2^J - 1$ inequalities. Normally this would preclude (for any but the smallest values of J), tractability of any algorithm optimizing a system performance measure which is expressed as a linear (let alone a nonlinear) function of the performance vector.

However, since X^* is the base of a polymatroid (see Theorem 2(a)) it follows that *simple* polynomial (or pseudopolynomial) algorithms exist to minimize any convex separable function $f(W) = \sum_j f_j(W_j)$. (Certain nonseparable cases can be handled as well, see Federgruen and Groenevelt 1986.)

As a direct corollary to Theorem 2 (Corollary 1 below) we obtain that the minimum of any *linear* objective $\sum_{j=1}^J c_j W_j$ is achieved by an *absolute* priority rule which gives (absolute) priority to a customer in class i if and only if $c_i/\rho_i \leq c_j/\rho_j$ $(i, j \in E)$. Minimization of a linear cost objective thus reduces to the ranking of the ratios $\{c_j/\rho_j : j \in E\}$ which requires $O(J \log J)$ time only.

COROLLARY 1. *Assume assumption* (A) *and conditions* (C) *hold. Consider the cost objective* $\sum_{j=1}^J c_j W_j$ $(c_j \geq 0, j \in E)$ *and assume that the customer classes are numbered such that* $c_1/\rho_1 \geq c_2/\rho_2 \geq \cdots \geq c_J/\rho_J$. *The absolute priority rule which* (*at each service completion*) *assigns priority to a waiting customer of the lowest indexed class minimizes the cost objective among all work conserving priority rules.*

PROOF. Let $x_j = \rho_j W_j$ $(j \in E)$. With this substitution of variables, our minimization problem may be formulated as $\min \sum_{j \in E} (c_j/\rho_j) x_j$ s.t. $x \in X^*$. Since X^* is the base of a polymatroid (Theorem 2(a)) it follows from Edmonds' (1970) famous result that an optimal extreme point for this linear program may be obtained by the greedy procedure:

Step 0. (Since x_J has the largest coefficient in the objective function) set x_J to its maximum feasible value, i.e., $x_J = b^*(\{J\})$; $l = J - 1$.

Step 1. Given fixed values for $x_{l+1}, \ldots, x_J$, (and since x_l has the next largest coefficient in the objective function) set x_l to its maximum feasible value, i.e.,

$$x_l := b^*(\{l, \ldots, N\}) - \sum_{i=l+1}^{J} x_i = b^*(\{l, \ldots, N\}) - b^*(\{l+1, \ldots, N\}).$$

Each extreme point of X^* is the performance vector of an absolute priority rule, see Theorem 2(b); moreover, the specific extreme point constructed by the greedy procedure, is obtained by a lexicographic *maximization* of $(x_J, x_{J-1}, \ldots, x_1)$ and hence of $(W_J, W_{J-1}, \ldots, W_1)$ and must therefore correspond with the absolute priority rule which assigns priority to a customer in class i over one in class j if $i < j$. ■

We note that optimality of an absolute priority rule for linear objectives (determined by a simple ranking of the ratios $\{c_j/\rho_j : j \in E\}$) has been shown for a number of special cases (see Fife 1965, Smith 1956, Kleinrock 1976, and Gelenbe and Mitrani 1980).

More generally, Theorem 2(b) establishes that an absolute priority rule is optimal for any concave (or even *quasi-concave*) objective $f(W_1, \ldots, W_J)$, since such objectives achieve their maximum in an extreme point of X^*.

For *convex* system performance measures, the simplest polynomial algorithm is the so-called decomposition algorithm, see Groenevelt (1985). Since the right-hand sides of the constraints (3) and (4) are generated by a *generalized symmetric* function, an efficient implementation of this algorithm may be achieved with a running time of $O(J^2 \log J + J\chi)$ where χ is the time needed to solve a certain type of single variable (nonlinear) equation. When all of the terms in the objective function are of the form

$$f_j(W) = \alpha_j h((W - \beta_j)/\alpha_j), \tag{13}$$

for some $\alpha_j, \beta_j > 0$ and a strictly convex function $h(\cdot)$, $\chi = O(J)$ so that the overall running time of the algorithm is $O(J^2 \log J)$. (We call such objectives *homoform*.) It is noteworthy that, in the homoform case, the optimal solution is independent of the specific choice for the function $h(\cdot)$, an observation which goes back to Veinott (1971), see also the discussion below. We refer to Groenevelt (1985) and Federgruen and Groenevelt (1988) for a full specification and analysis of the algorithm.

A simple alternative is provided by a greedy or marginal allocation procedure. This algorithm is, however, only pseudopolynomial, see Federgruen and Groenevelt (1986) for details.

The above procedures thus generate an *exact* optimal solution for the *single*-server case; in the general multi-server case, an approximation error may, however, arise since an approximation assumption (Assumption (A')) is invoked in the determination of the (achievable) performance space. We conclude this section with a discussion of the resulting accuracy with which optimal performance vectors are approximated by the recommended procedures.

Assumption (A') may in fact be viewed as introducing two potential approximation steps:

(i) Assumption (A') implies supermodularity of the $A^*(\cdot)$ function (the weaker Assumption (A)) which is used to establish that *all* points in $\mathbf{W}^*$ are feasible, or equivalently that constraints (8) are sufficient (as well as necessary) conditions for feasibility, see Theorem 2. In other words, should the supermodularity property fail to hold, only *relaxations* of the true optimization problems may be solved by the above described algorithms, see also the discussion in §5.

As discussed above, the supermodularity property holds when the service times are exponential or deterministic but its exact validity is unknown for other types of service time distributions. On the other hand, the $A^*(\cdot)$ function is, even for this most general case, at least approximately supermodular since Assumption (A') results in very accurate approximations of the values of $\{A^*(S), S \subset E\}$ as substantiated above. (Recall, e.g., that the approximations are *always* asymptotically exact in heavy traffic.)

(ii) Assuming henceforth that the performance space is indeed given by $\mathbf{W}^*$ with the $b^*(\cdot)$ set function in (8) submodular, approximation errors may be incurred in the evaluation of the righthand sides $\{b^*(S): S \subset E\}$ of the constraints in (8). More specifically, Assumption (A') results in the approximation formulae (12) for $\{A^*(S): S \subset E\}$ and hence in approximations $\{\hat{b}(S): S \subset E\}$ for the true right-hand sides $\{b^*(S): S \subset E\}$ of the constraints in (8). Again, as substantiated above, these approximations are exact when the service times are exponential as well as in heavy traffic and their accuracy has numerically been established in several above-mentioned studies.

Here we demonstrate that

—in the case of a *linear* cost objective, the *exact optimal* performance vector is obtained in spite of any approximation errors in the computation of the right-hand sides of (8);

—in the case of nonlinear, convex cost objectives, "small" approximation errors in the determination of the $\{b^*(S): S \subset E\}$ numbers result in "small" approximation errors for the computed "optimal" performance vector.

The first conclusion follows directly from the observation that for *linear* system performance measures, an optimal priority rule exists which depends on the ratios $\{c_j/\rho_j: j \in E\}$ only, see Corollary 1!

To substantiate our second conclusion for nonlinear, convex objectives, we first show the following proposition. (A function is C^k if it is k times continuously differentiable, $k \geq 1$. Let $\{b(S): S \subset E\}$ be the parameterized right-hand sides of the constraints in (8).)

PROPOSITION 1. *Consider a system performance measure* $f(W) = \sum_{j \in E} f_j(W_j)$ *with* f_j *strictly convex and* C^2 $(j \in E)$. *The* unique *optimal solution* W^* *in* (8) *is a continuously differentiable function of* $\{b(S): S \subset E\}$.

PROOF. The fact that the optimal performance vector is unique is a standard result in convex programming with strictly convex objectives.

Apply, as in the proof of Corollary 1, the substitution of variables $x_j = \rho_j W_j$ $(j \in E)$. Also, let $g_j(x_j) = Mx_j - f_j(x_j/\rho_j)$ $(j \in E)$ for some large constant M. ($M \geq \max_{j \in E} f_j'(b^*(\{j\}))$.) Note that the functions $g_j(\cdot)$ are increasing, strictly concave, and C^2, $j \in E$. Our optimization problem is thus equivalent to:

$$\max \sum_{j \in E} g_j(x_j) \qquad \text{s.t.} \qquad x \geq 0, \qquad \sum_{j \in S} x_j \leq b(S), \qquad S \subset E. \tag{P}$$

(Note that we have relaxed the equality $\sum_{j \in E} x_j = b^*(E)$ to an inequality; since the objective is increasing, this relaxation does not affect the optimal solution.) Let x^* be an optimal solution of (P) and let $\{S_1, \ldots, S_L\}$ be an enumeration of the sets S for which the constraints are binding. Likewise, let $\lambda^*(S)$ denote the optimal Lagrange multiplier (dual variable) associated with the constraint for set $S \subset E$. (These multipliers are unique since the objective is strictly concave.) Finally, let $e(S) \in \mathbf{R}^J$ denote the indicator vector of set $S \subset E$, i.e., $e(S)_j = 1$ if $j \in S$, and 0 otherwise.

The proposition now follows from Theorem 2.1 in Fiacco (1976) by verifying that

(i) the vectors $\{e(S_1), \ldots, e(S_L)\}$ are linearly independent;

(ii) $\lambda^*(S_l) > 0$, $l = 1, \ldots, L$.

(i) follows from the fact that the sets $\{S_1, \ldots, S_L\}$ are nested, i.e., either $S_i \subseteq S_j$ or $S_j \subseteq S_i$ $(1 \leq i, j \leq L)$. This is a well-known consequence of the strict submodularity of $b^*(\cdot)$, see, e.g., Lemma 2.1 in Lawler and Martel (1982); see also the proof of Lemma 4 below. Thus (after appropriate numbering) $S_1 \subseteq S_2 \subseteq \cdots \subseteq S_L$ and (i) follows.

Moreover, in view of (i) and the fact that $x_j^* > 0$ $(j \in E)$, the Kuhn-Tucker conditions imply that

$$\sum_{j \in S_l} x_j^* = b(S_l), \qquad l = 1, \ldots, L, \tag{14}$$

$$g_j'(x_j^*) = \lambda^*(S_l) + \lambda^*(S_{l+1}) + \cdots \lambda^*(S_L), \qquad l = 1, \ldots, L$$

$$\text{and} \qquad j \in S_l \backslash S_{l+1}. \tag{15}$$

(ii) is now easily verified by complete induction (starting with $l = L$) since $g_j(\cdot)$ is strictly increasing, i.e., $g_j'(x_j^*) > 0, j \in E$. ■

We note that the optimal performance vector W^*, viewed as a function of the right-hand side vector $\{b(S): S \subset E\}$ is in fact *as smooth* as the functions $\{g_j'(\cdot): j \in E\}$, i.e., if $g_j(\cdot) \in C^{k+1}$ for some $k \geq 1$, then $W^*(b) \in C^k$, see Corollary 4.1 in Fiacco (1976).

In view of equations (14) and (15) it is in fact possible to derive bounds for the possible approximation errors in the computed optimal performance vectors. Thus, let

$\hat{A}(\cdot)$ ($\hat{b}(\cdot)$) denote the approximation for $A^*(\cdot)$ ($b^*(\cdot)$), as obtained from Assumption (A'), and let $\epsilon = \max_{S \subset E} \{ | b^*(S) - \hat{b}(S) | \} = \max_{S \subset E} \{ | A^*(S) - \hat{A}(S) | \}$. (As argued above, ϵ is small.) Likewise, let $\hat{W}(\cdot)$ be the computed approximation to the optimal performance vector $W^*(\cdot)$.

Note first from the proof of Theorem 2.1 in Fiacco (1976) that the constraints corresponding with the collection of sets $\{S_1, \ldots, S_L\}$ are the binding constraints for $\hat{W}$ as well as W^* when ϵ is sufficiently small. It thus follows from (14) (provided ϵ is sufficiently small) that

$$| \sum_{j \in S_l} \rho_j \hat{W}_j - \sum_{j \in S_l} \rho_j W_j^* | = | b^*(S_l) - \hat{b}_j(S_l) | \leq \epsilon.$$

Moreover, it is possible to derive an exact expression for the approximation errors in the individual components of the optimal performance vector. This is most easily accomplished when the objective function is homoform, see (13). Let

$$B_l = \sum_{j \in S_{l+1} \setminus S_l} \beta_j \quad \text{and} \quad A_l = \sum_{j \in S_{l+1} \setminus S_l} \alpha_j, \qquad l = 1, \ldots, L.$$

By a simple algebraic manipulation of (15) one verifies for all $l = 1, \cdots L$ and $j \in S_l \setminus S_{l+1}$, that $(\rho_j W_j^* - \beta_j)/\alpha_j = (b^*(S_l) - B_l)/A_l$ while $(\rho_j \hat{W}_j - \beta_j)/\alpha_j = (\hat{b}(S_l) - B_l)/A_l$, for ϵ sufficiently small. Thus, for all $l = 1, \ldots, L$ and $j \in S_l \setminus S_{l+1}$,

$$| W_j^* - \hat{W}_j | = \rho_j^{-1} \frac{\alpha_j}{(\sum_{i \in S_{l+1} \setminus S_l} \alpha_i)} | b^*(S_l) - \hat{b}_j(S_l) | \leq \rho_j^{-1} \epsilon.$$

Similar bounds (in terms of $\sum_{j \in S_l} g_j'^{-1}(\cdot)$, $l = 1, \ldots, L$) may be derived for general nonhomoform objectives.

4. A Synthesis Algorithm

We observed (in the proof of Theorem 2) that each point in the performance space corresponds with an appropriately chosen randomization of absolute priority rules. Since, in view of Carathéodory's theorem (see, e.g., Bazaraa and Shetty 1979), each point in a J-dimensional polyhedron can be written as a convex combination of no more than $J + 1$ extreme points, it follows that each point in $\mathbf{W}^*$ is the performance vector of a randomization of no more than $(J + 1)$ absolute priority rules. (The randomization probabilities may, at least in principle, be determined by a linear program.)

Randomizations of absolute priority rules are, however, difficult to implement; moreover, the *variances* of the steady-state waiting times tend to be large under such rules. In this section we show that each point in W^* corresponds with a slight generalization of the, far more attractive, so-called *dynamic* (Jackson 1960) or *delay dependent* (Kleinrock 1976) scheduling disciplines where a customer's priority is proportional to his time spent in queue, the proportionality constant being class dependent. More specifically, in a delay dependent priority rule, positive weights α_j ($j \in E$) are specified such that a customer of class j who arrives at time τ is given a priority value of $\alpha_j(t - \tau)$ at time $t > \tau$. At a service completion epoch, the customer with the highest priority value (among all queueing customers) is taken into service.

We show below that each *interior* point of $\mathbf{W}^*$ may be achieved by some dynamic priority rule, i.e., by an appropriate choice of the weight vector α. To cover the entire performance space we need a larger class of rules which includes the *dynamic* and *absolute* priority disciplines as special cases:

DEFINITION 2. A mixed dynamic priority rule is characterized by a partition $\{E_1, \cdots E_L\}$ of E. The subsets $\{E_l: l = 1, \cdots L\}$ are referred to as *leagues*. A customer in E_k has *absolute* priority over a customer in E_l if and only if $k > l$. Relative priorities

within a league l ($l = 1, \ldots, L$) are determined according to a dynamic priority rule with weight vector $\alpha^{(l)} = (\alpha_j)_{j \in E_l}$.

Absolute priority rules have J leagues each consisting of a simple class; *pure* dynamic priority rules have a single league consisting of all classes in E. In the following we assume, without loss of generality, that the customer classes are numbered in ascending order of their attributed priorities (and hence expected waiting times). We can thus restrict ourselves to rules with *consecutive* leagues and nondecreasing weight vectors: (A set $S \subset E$ is consecutive if it consists of a collection of consecutive integers.)

We first derive a system of linear equations from which the (approximate) performance vector of any mixed dynamic rule can be obtained. This derivation is based on Assumption (A').

As in Kleinrock (1976, p. 109) consider a customer in class $p \in E$ and define

N_{ip} = number of customers from class i who are in the queue when the tagged customer (in class p) arrives and who receive service before the tagged customer does ($i \in E$).

M_{ip} = number of customers from class i who arrive to the system while the tagged customer (in class p) is in queue and who receive service before he does ($i \in E$).

In view of Assumption (A'), since Poisson arrivals see time averages, and using Lemma 2(a), we obtain:

$$W_p = E(V_0) + c^{-1} \sum_{j=1}^{J} E(N_{jp} + M_{jp}) E(V_j), \qquad p \in E. \tag{16}$$

For any given rule, let $S_l = \bigcup_{k=l}^{L} E_k$, $l = 1, \ldots, L$. Following the analysis in Kleinrock (1976, §3.7) we obtain:

$$EN_{jp} = \begin{cases} 0 & \text{for} \quad j \in \bar{S}_l = E \backslash S_l, \\ \lambda_j W_j \alpha_j / \alpha_p & \text{for} \quad j \in E_l, \quad j < p, \\ \lambda_j W_j & \text{for} \quad j \in S_l, \quad j \geqq p, \end{cases} \tag{17}$$

$$EM_{jp} = \begin{cases} 0 & \text{for} \quad j \leqq p, \\ \lambda_j W_p \left(1 - \dfrac{\alpha_p}{\alpha_j}\right) & \text{for} \quad j \in E_l, \quad p < j, \\ \lambda_j W_p & \text{for} \quad j \in S_{l+1}. \end{cases}$$

Substitution of (16) into (17) results in

$$W_p = E(V_0) + \frac{1}{c} \sum_{\substack{j \in E_l \\ j < p}} \rho_j W_j \frac{\alpha_j}{\alpha_p} + \frac{1}{c} \sum_{\substack{j \in S_l \\ j \geqq p}} \rho_j W_j + \frac{1}{c} \sum_{\substack{j \in E_{l_p} \\ j \geqq p}} \rho_j W_p \left(1 - \frac{\alpha_p}{\alpha_j}\right) + \frac{1}{c} \sum_{j \in S_{l+1}} p_j W_p,$$

$$p \in E_l \quad (l = 1, \ldots, L). \tag{18}$$

LEMMA 3. *Let Assumption* (A') *and condition* (C) *hold. The performance vector of any mixed dynamic rule is the unique solution of the linear system of equations* (18).

PROOF. In view of the conservation law in Lemma 1, we have

$$\sum_{\substack{j \in S_l \\ j \geqq p}} \rho_j W_j = \rho W^*(E) - \sum_{j=1}^{p-1} \rho_j W_j.$$

Substitution into (15) transforms this system into a triangular one. □

We now prove that each point in $\mathbf{W}^*$ is the performance vector of a mixed dynamic rule (the *synthesis proof*). While the proof itself fails to be constructive, it is followed by

a simple algorithm which determines the parameters of a rule corresponding with any given $W \in \mathbf{W}^*$ (the *synthesis algorithm*).

The synthesis proof uses an alternative characterization of mixed dynamic rules through a vector $r \in D = \{(r_1, \ldots, r_{J-1}) \in \mathbf{R}^{J-1}: 0 \leqq r_j \leqq 1, j = 1, \ldots, J-1\}$. A rule with leagues $\{E_1, \ldots, E_L\}$ and weight vectors $\{\alpha^{(l)}: l = 1, \cdots L\}$ corresponds with the vector $r \in D$ defined by:

$$r_{i_l} = 0, \qquad l = 1, \ldots, L-1;$$

$$r_p = \alpha_p^{(l)}/\alpha_{p+1}^{(l)}, \qquad \text{if} \quad p \in E_l, \quad p < i_l \quad (l = 1, \ldots, L), \quad \text{where}$$

$$i_l = \max\{p: p \in E_l\} \qquad (l = 1, \ldots, L).$$

Conversely for any $r \in D$, let $\{i_l: l = 1, \ldots, L-1\}$ be the (possibly empty) collection of zero components. (Note, $1 \leqq L \leqq J-1$ and $L = 1$ if all components of r are positive; assume $i_1 < i_2 < \cdots < i_{L-1}$ and set $i_L = J$.) The vector r corresponds with a rule with L leagues; the lth league $\{i_{l-1}+1, \ldots, i_l\}$ has a weight vector $\alpha^{(l)}$ defined (recursively) by

$$\alpha_{i_l}^{(l)} = 1; \qquad \alpha_j^{(l)} = \alpha_{j+1}^{(l)} r_j, \qquad j = i_{l-1}+1, \ldots, i_l - 1.$$

Describing a rule via its associated r-vector, one easily verifies that (14) simplifies to

$$EN_{jp} = \lambda_j W_j \prod_{k=j}^{p-1} r_k \qquad (j, p \in E),$$

$$EM_{jp} = \lambda_j W_p \Big(1 - \prod_{k=p}^{j-1} r_k\Big) \qquad (j, p \in E),$$

(with the convention that empty products equal one).

Substitution into (16) results in:

$$W_p = E(V_0) + c^{-1} \sum_{j=1}^{J} \rho_j \Big(\prod_{k=j}^{p-1} r_k\Big) W_j + c^{-1} \sum_{j=1}^{J} \rho_j \Big(1 - \prod_{k=p}^{j-1} r_k\Big) W_p, \qquad p \in E. \quad (19)$$

Define the polyhedron $\hat{\mathbf{W}} = \mathbf{W}^* \cap \{W \in \mathbf{R}^J: W_1 \geqq W_2 \geqq \cdots \geqq W_J\}$. Recall from (5) and (10) that $\hat{\mathbf{W}} = \{W \in \mathbf{R}^J: \sum_{l=i+1}^{J} \rho_l W_l \geqq f(\sum_{l=i+1}^{J} \rho_l),\ i = 1, \ldots, J-1;\ \sum_{l=1}^{J} \rho_l W_l = f(\rho);\ W_1 \geqq W_2 \geqq \cdots \geqq W_J\}$ where $f(z) = E(V_0)z/(1 - c^{-1}z)$.

We first need the following lemma.

LEMMA 4. *Let Assumption* (A′) *and condition* (C) *hold. There exists a piecewise linear transformation* $\chi: \hat{\mathbf{W}} \to D$ *with the following properties*:

(a) $W \in \hat{\mathbf{W}}$ *satisfies*

$$\sum_{l=i+1}^{J} \rho_l W_l = f\Big(\sum_{l=i+1}^{J} \rho_l\Big)\} \Rightarrow \chi(W)_i = 0 \qquad (i = 1, \ldots, J-1). \quad (20)$$

(b) $W \in \hat{\mathbf{W}}$ *satisfies*

$$W_i = W_{i+1} \Rightarrow \chi(W)_i = 1 \qquad (i = 1, \ldots, J-1). \quad (21)$$

PROOF. Let $W \in \hat{\mathbf{W}}$. We first show for any $i = 1, \ldots, J-1$:

$$\sum_{l=i+1}^{J} \rho_l W_l = f\Big(\sum_{l=i+1}^{J} \rho_l\Big) \Rightarrow W_i > W_{i+1}, \quad (22)$$

thus showing that properties (20) and (21) are consistent. Subtract the equality to the left of the $\Rightarrow$ sign in (22) from the inequality $\sum_{l=i}^{J} \rho_l W_l \geqq f(\sum_{l=i}^{J} \rho_l)$ to conclude that

$$W_i \geqq \Big[f\Big(\sum_{l=i}^{J} \rho_l\Big) - f\Big(\sum_{l=i+1}^{J} \rho_l\Big)\Big]/\rho_i > \Big[f\Big(\sum_{l=i+1}^{J} \rho_l\Big) - f\Big(\sum_{l=i+2}^{J} \rho_l\Big)\Big]/\rho_{i+1}$$

$$\geqq W_{i+1}.$$

(Empty sums are again assumed to be zero. The second strict inequality follows from the fact that f is strictly convex; the last inequality is derived by subtracting the inequality $\sum_{l=i+2}^{J} \rho_l W_l \geqq f(\sum_{l=i+2}^{J} \rho_l)$ from the equality to the left of the $\Rightarrow$ sign in (22).)

We now define the piecewise linear transformation χ. Each vertex of $\hat{\mathbf{W}}$ is mapped to a vertex of D which satisfies (20) and (21). Let $\{W^{*(1)}, \ldots, W^{*(M)}\}$ be an enumeration of the vertices of $\hat{\mathbf{W}}$. Fix a specific triangulation of $\hat{W}$, i.e., fix a partition of $\hat{\mathbf{W}}$ into a collection of simplices, each containing $J+1$ vertices. (A simplex in $\mathbf{R}^J$ is the convex hull of $J+1$ points.) Such a partition always exists, see e.g. Corollary 1.7 in Hudson (1969). Thus, each point W in $\hat{\mathbf{W}}$ is either in the interior of a unique simplex or on the boundary of two adjacent simplices. In the former case W may be written as a convex combination of the $(J+1)$ vertices of its simplex and in the latter case as a convex combination of the J vertices in the intersection of the two adjacent simplices. For a given triangulation, this procedure *uniquely* specifies for each $W \in \hat{\mathbf{W}}$, a vector $\alpha \in \mathbf{R}^M$ with $\sum_{p=1}^{M} \alpha_p = 1$ and $\alpha_p \geqq 0$ $(p = 1, \ldots, M)$ such that $W = \sum_{p=1}^{M} \alpha_p W^{*(p)}$.

Define $\chi(W) = \sum_p \alpha_p \chi(W^{*(p)}) \in D$. Clearly, χ is piecewise-linear and it is easily verified that χ is continuous as well. (χ is a so-called simplicial map, see e.g. §4 in Hudson 1969.) Moreover, (20) and (21) are easily verified. □

THEOREM 3. *Let Assumption* (A′) *and condition* (C) *hold. Each* $W \in \hat{\mathbf{W}}$ *is the performance vector of a* (*mixed*) *dynamic rule.*

PROOF. In view of Lemma 3, there exists, for any $r \in D$, a unique solution $W(r)$ of (16) and $W(r) \in \hat{\mathbf{W}}$. Note from the implicit function theorem that $W(\cdot)$ is a continuous mapping from D into $\hat{W}$. Assume first (to the contrary) that some *interior* point W^0 of $\hat{\mathbf{W}}$ is *not* contained in the image of $W(\cdot)$. Let ψ denote the central projection with W^0 as its center, projecting each point in the polyhedron $\hat{\mathbf{W}}$ onto its boundary. Note that ψ is continuous on $\hat{\mathbf{W}} \setminus \{W^0\}$. Next, let $\nu: D \to D$ be the point symmetry with center $(\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2})$, i.e., $\nu(r)_i = 1 - r_i$ for $i = 1, \ldots, J-1$. Finally, define $\sigma: D \to D$: $\sigma(r) = \nu \circ \chi \circ \psi \circ W(r)$. Since σ is the composition of continuous mappings, it is a continuous mapping of a convex compact set into itself and has a fixed point r^*, in view of Brouwer's fixed point theorem. Observe that ψ maps $\hat{\mathbf{W}} \setminus \{W^0\}$ onto its boundary; χ maps the boundary of $\hat{\mathbf{W}}$ into the boundary of D, in view of (20) and (21). Likewise, ν maps the boundary of D into itself. Thus σ maps D into its boundary and r^* is a boundary point. Thus, for some $i = 1, \ldots, J-1$ either $r_i^* = 0$ or $r_i^* = 1$. In the former case, the collection $\{i+1, \ldots, J\}$ has absolute priority over all other classes under the rule associated with r^*. In view of Lemma 2 this implies that $W(r^*)$, and hence $\psi \circ W(r^*)$ is a point on the hyperplane to the left of the $\Rightarrow$ sign in (20). In view of (20), $\chi \circ \psi \circ W(r^*)_i = 0$ and hence $\sigma(r^*)_i = 1$, which contradicts $\sigma(r^*) = r^*$. If $r_i^* = 1$, a similar argument leads to a contradiction. We conclude that the interior of $\hat{\mathbf{W}}$ is contained within the image of $W(\cdot)$. Moreover, the image of the compact set D is compact, since $W(\cdot)$ is continuous. Since $\hat{\mathbf{W}}$ is the smallest among all compact sets which include the interior of $\hat{\mathbf{W}}$, it follows that $W(\cdot)$ maps D *onto* $\hat{\mathbf{W}}$. □

A Synthesis Algorithm

We are now ready to specify a simple synthesis algorithm. Thus, fix $W \in \hat{\mathbf{W}}$. If W is the performance vector of a rule described by the vector $r \in D$, then r must satisfy the system of equations (19). In particular, for any $p = 1, \ldots, J-1$, r must satisfy the equation obtained by subtracting r_p times the p-th equation in (19) from the $p+1$st equation in (19):

$$W_{p+1} - r_p W_p = E(V_0)(1 - r_p) + c^{-1} \sum_{j=1}^{J} \rho_j (1 - \prod_{k=p+1}^{j-1} r_k) W_{p+1} - c^{-1} \sum_{j-1}^{J} \rho_j (r_p - r_p^2 \prod_{k=p+1}^{j-1} r_k) W_p.$$

Thus, given values for $(r_{p+1}, \ldots, r_{J-1})$, r_p must be a root of the quadratic equation:

$$[c^{-1}\sum_{j=1}^{J}\rho_j(\prod_{k=p+1}^{j-1}r_k)]W_p x^2 + \{(1-c^{-1}\rho)W_p - E(V^0)\}x$$

$$- W_{p+1}\{1 - c^{-1}\sum_{j=1}^{J}\rho_j(1 - \prod_{k=p+1}^{j-1}r_k)\} + E(V_0) = 0, \qquad p = 1, \ldots, J-1. \quad (23)$$

This suggests that a rule with W as performance vector, may be obtained by recursive solution of the $(J-1)$ quadratic equations (23):

Synthesis Algorithm.

Step 0. Set $p = J - 1$.

Step 1. Find the unique nonnegative root x^* of the quadratic equation in (23); $r_p := x^*$.

Step 2. If $p > 1$, $p := p - 1$ and return to Step 1; otherwise, terminate.

THEOREM 4. *Let Assumption* (A′) *and condition* (C) *hold. For each* $W \in \hat{\mathbf{W}}$, *the synthesis algorithm determines a rule with* W *as performance vector.*

PROOF. Fix $w \in \hat{\mathbf{W}}$. It follows from Theorem 3 that a rule *exists* with W as performance vector. As explained above, the corresponding vector r of any such rule must satisfy the recursive quadratic equations (23). It is sufficient to show that these equations have a unique nonnegative root. Since the coefficient of the quadratic term is positive, the proof is complete if we show that either (i) the constant term is negative or (ii) the constant term is zero and the coefficient of the linear term nonnegative. Note, however, from (19) that

$$W_{p+1} \geqq E(V^0) + c^{-1}\sum_{j=1}^{J}\rho_j(1 - \prod_{k=p+1}^{j-1}r_k)W_{p+1}$$

and that equality holds iff $r_p = 0$. Hence the constant term is nonpositive and if it is zero, we have $r_p = 0$. But this implies, again from (19), that $W_p \geqq E(V_0) + \rho c^{-1} W_p$, so the coefficient of the linear term is indeed non-negative, in this case. □

A special case of the above synthesis algorithm for interior points of the performance space in $M/G/1$ systems was first proposed by Wood and Sargent (1984).

The following corollary strengthens Theorem 3. Its proof is immediate from that of Theorem 4.

COROLLARY 2. *Let Assumption* (A′) *and condition* (C) *hold. For each* $W \in \hat{\mathbf{W}}$, *there exists a* unique *mixed dynamic rule with* W *as its performance vector.*

We conclude this section with a brief discussion of the approximation errors that may arise in determining a mixed dynamic rule which optimizes a given system objective. Such errors may be due to possible errors in the computation of W^* which in turn may result from possible approximations in the evaluation of $\{A^*(S): S \subset E\}$ (in view of Assumption (A′)). As argued at the end of §3, approximation errors in the computation of W^* arise only in certain cases, and are "small" when they arise. As mentioned there, the collection of binding constraints in (8) for the computed approximation $\hat{W}$ of W^* is the same as that of W^*, when the approximation errors for the $A^*(\cdot)$ function are sufficiently small and it follows that the *computed league structure is identical to that of the optimal performance vector* W^*! (Use Lemma 3 and Theorem 4.)

Recall that within a given league, $l = 1, \ldots, L$, the dynamic weight factors $\alpha^{(l)}$ are determined by repeated evaluations of the unique positive root of single variable quadratic equations. Bounds on the possible errors that may arise in the computation of these weights are thus easily derived.

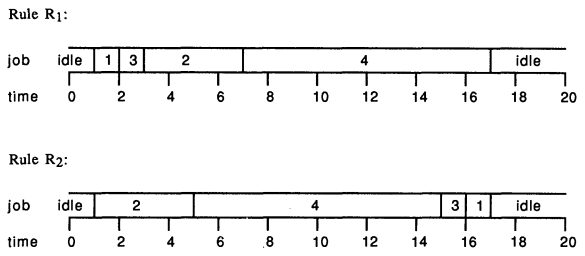

FIGURE 1. Schedules for Example.

5. More General Models

Several elements of the analysis in §§2 and 3 may be extended to more general queueing models. In particular, under the stability condition $\sum_j \rho_j < c$, Lemma 1 (the work conservation law) applies to general arrival processes as long as long run average arrival rates exist and the long run averages of the customers' waiting times converge for each class in E, see Theorem 11–13 in Heyman and Sobel (1982). Lemma 1 thus holds in particular for $GI/G/c$ queues (with renewal arrival processes), see Wolff (1984) and Whitt (1982) for a verification of the required conditions.

Since Lemma 1 holds for general arrival processes, the necessary conditions for achievability of a performance vector in Theorem 1, may be extended to systems with such arrival processes as well: the polyhedron $\mathbf{W}^*$, described by equations (3) and (4), thus *contains* the performance space under rather general conditions.

$A^*(\cdot)$ (viewed as a set function on 2^E) is supermodular in *single*-server systems with general arrival processes, see Corollary 1 in Federgruen and Groenevelt (1985). The supermodularity property may, however, fail to hold in multiserver systems with non-Poisson arrival streams and nondeterministic service times, cf., ibid. Thus Theorem 2 may fail to hold for general arrival processes. $\mathbf{W}^*$ may fail to be the base of a polymatroid.

Most importantly, the following single-server example with deterministic service and interarrival times shows that $\mathbf{W}^*$ may fail to represent the performance space when the arrival processes are more general than Poisson.

EXAMPLE. Let $E = \{1, 2, 3, 4\}$ and $c = 1$. Assume all interarrival times are deterministic with the kth customer of classes 1 and 2 arriving at time $20(k-1)+1$, the kth customer of class 3 at time $20(k-1)+2$ and the kth customer of class 4 at time $20(k-1)+3.5$, $k \geq 1$. Service times are deterministic with $V_1 = V_3 = 1$; $V_2 = 4$ and $V_4 = 10$. Let R_i be the rule which gives absolute priority to class i, $i = 1, 2$. Note that $A^*(\{3\}) = \frac{1}{40}$ and $A^*(\{3\})$ is achieved *only* under rule R_1, see Figures 1 and 2. Note also that $A^*(\{3,4\}) = 3.925$ and $A^*(\{3,4\})$ is achieved *only* under rule R_2, see Figure 3. Thus consider any extreme point of W^* satisfying the equations, (see (3))

$$\rho_3 W_3 = A^*(\{3\}) - \tfrac{1}{2}\lambda_3 EV_3^2 = \tfrac{1}{40} - \tfrac{1}{40} = 0,$$

$$\rho_3 W_3 + \rho_4 W_4 = A^*(\{3,4\}) - \tfrac{1}{2}\lambda_3 EV_3^2 - \tfrac{1}{2}\lambda_4 EV_4^2 = 1.4.$$

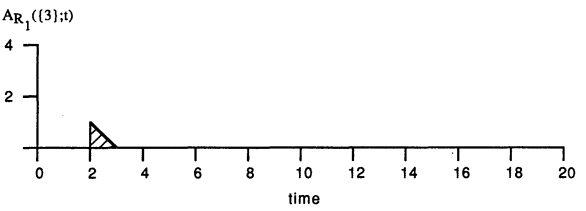

FIGURE 2. Determination of $A^*(\{3\})$.

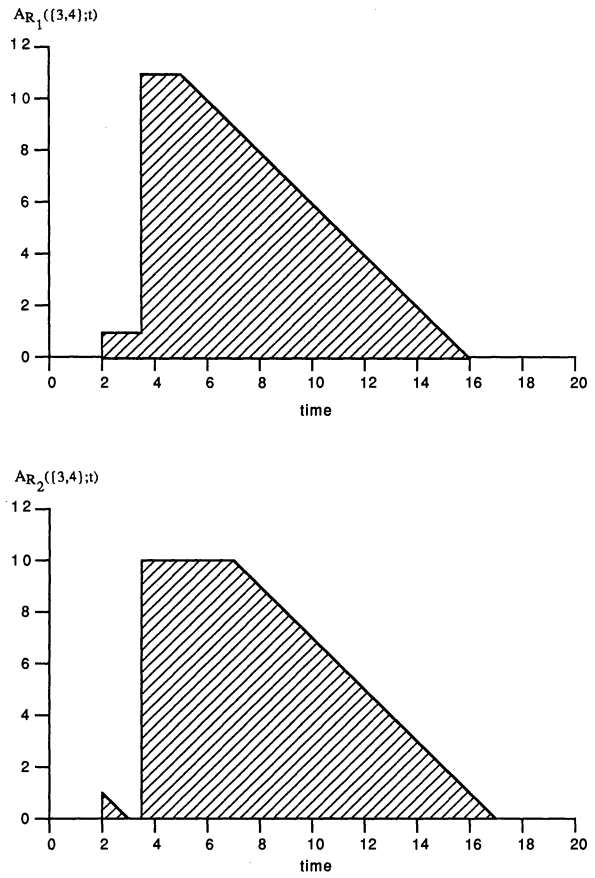

FIGURE 3. Determination of $A^*(\{3, 4\}; t)$.

Any such extreme point has $W_3 = 0$ and $W_4 = 2.8$. However, the only rule under which $W_3 = 0$ is R_1 and under this rule $W_4 = 3$. We conclude that some of the extreme points of W^* are not achievable.

Characterizing the performance space in systems with non-Poisson arrivals remains an open question.[1]

[1] We are greatly indebted to Mark Broadie, William Cook, Lex Schrijver and Tom McCormick for providing us with the basic outline of the proof of Theorem 3. We are equally indebted to Stephen Robinson for referring us to the perturbation results in Fiacco (1976).

References

BAGCHI, U. AND R. SULLIVAN, "Dynamic, Nonpreemptive Priority Queues with General, Linearly Increasing Priority Function," *Oper. Res.*, 33 (1985), 1278–1299.

BAZARAA, M. AND C. SHETTY, *Nonlinear Programming; Theory and Algorithms*, John Wiley, New York, 1979.

BOXMA, O. J., J. W. COHEN AND N. HUFFELS, "Approximations of the Mean Waiting Time in $M/G/s$ Queueing Systems," *Oper. Res.*, 27 (1979), 1115–1127.

EDMONDS, J., "Submodular Functions, Matroids and Certain Polyhedra, in *Combinatorial Structures and Their Applications*, R. Guy et al. (Eds.), Gordon and Breach, New York, 1970, 69–87.

FAYOLLE, G., R. IASNOGORODSKI AND I. MITRANI, "On the Sharing of a Processor Among Many Job Classes," *J. Assoc. Comput. Mach.*, (1980).

FEDERGRUEN, A. AND H. GROENEVELT, "The Greedy Procedure for Resource Allocation Problems: Necessary and Sufficient Conditions for Optimality," *Oper. Res.*, 34 (1986), 909–919.

——— AND ———, "The Impact of the Composition of the Customer Base in General Queueing Models," *J. Appl. Prob.*, (1987), 709–724.

——— AND ———, "Characterization and Control of Achievable Performance in General Queueing System," *Oper. Res.*, (1988).

FIACCO, A., "Sensitivity Analysis for Nonlinear Programming Using Penalty Methods," *Math. Programming*, 10 (1976), 287–311.

FIFE, D., "Scheduling with Random Arrivals and Linear Loss Functions," *Management Sci.*, 11 (1965), 429–437.

GELENBE, E. AND I. MITRANI, *Analysis and Synthesis of Computer Systems*, Academic Press, New York, 1980.

GREEN, L., "A Limit Theorem on Subintervals of Interrenewal Times," *Oper. Res.*, 30 (1982), 210–216.

GROENEVELT, H., "Two Algorithms for Maximizing a Separable Concave Function over a Polymatroid Feasible Region," Graduate School of Management Working Paper, University of Rochester, Rochester, NY, 1985.

———, M. VAN HOORN AND H. TIJMS, "Tables for $M/G/c$ Queueing Systems with Phasetype Service," *European J. Oper. Res.*, 16 (1984), 257–260.

GROSS, D. AND C. HARRIS, *Fundamentals of Queueing Theory*, John Wiley, New York, 1974.

HEYMAN, D. AND M. SOBEL, *Stochastic Models in Operations Research*, Vol. 1, McGraw-Hill, New York, 1982.

——— AND S. STIDHAM, "A Note on the Relation between Customer and Time Averages in Queues," *Oper. Res.*, 28 (1980), 943–944.

HOKSTAD, P., "Approximations for the $M/G/m$ Queue," *Oper. Res.*, 26 (1978), 510–523.

HUDSON, J., *Piecewise Linear Topology*, Benjamin, Inc., New York, 1969.

JACKSON, J., "Some Problems in Queueing with Dynamic Priorities," *Naval Res. Logist. Quart.*, 7 (1960), 235–249.

KLEINROCK, L., "A Delay Dependent Queue Discipline," *Naval Res. Logist. Quart.*, 11 (1964), 329–341.

———, *Queueing Systems*. Vol. 2, John Wiley, New York, 1976.

——— AND R. FINKELSTEIN, "Time Dependent Priority Queues," *Oper. Res.*, 15 (1967), 104–116.

———, R. MUNTZ AND J. HSU, "Tight Bounds on Average Response Time for Processor Sharing Models of Time-Shared Computer Systems," *Inform. Process.*, 71, TA-2 (1971), 50–58.

KRAMPE, H., J. KUBAT AND W. RUNGE, *Bedienungsmodelle*, Oldenburg, München, 1973.

LAWLER, E. AND C. MARTEL, "Computing Maximal Polymatroidal Network Flows," *Math. Oper. Res.*, 7 (1982), 334–346.

MAALØE, E., "Approximation Formula for Estimation of Waiting-time in Multiple-Channel Queueing Systems," *Management Sci.*, 19 (1973), 703–710.

MITRANI, I., "On the Delay Functions Achievable by Non-Preemptive Scheduling Strategies in $M/G/1$ Queues," in *Deterministic and Stochastic Scheduling*, M. Dempster et al. (Eds.), D. Reidel, Dordrecht, Netherlands, 1982.

——— AND J. HINE, "Complete Parameterized Families of Job Scheduling Strategies," *Acta Informat.*, 8 (1977), 61–73.

NETTERMAN, A. AND I. ADIRI, "A Dynamic Priority Queue with General Concave Priority Functions," *Oper. Res.*, 27 (1979), 1088–1100.

NOZAKI, S. A. AND S. M. ROSS, "Approximations in Finite-Capacity Multi-Server Queues with Poisson Arrivals," *J. Appl. Probab.*, 15 (1978), 826–834.

SCHRAGE, L., "An Alternative Proof of a Conservation Law for the Queue $G/G/1$," *Oper. Res.*, 18 (1970), 185–187.

SEELEN, L. AND H. TIJMS, "Approximations to the Waiting Time Percentiles in the $M/G/c$ Queue," *Proceed. 11th Internat. Teletraffic Congress*, Kyoto, Japan, 1985.

———, ——— AND M. VAN HOORN, *Tables for Multi-Server Queues*, North Holland, Amsterdam, 1985.

SMITH, W., "Various Optimizers for Single-Stage Production," *Nav. Res. Logist. Quart.*, 3 (1956), 59–66.

STOYAN, D., "Approximations for $M/G/s$ Queues," *Math. Operations-forsch. Statist.*, 7 (1976), 587–594.

TIJMS, H., M. VAN HOORN AND A. FEDERGRUEN, "Approximations for the Steady-State Probabilities in the $M/G/c$ queue," *Adv. in Appl. Probab.*, 13 (1981), 186–206.

VAN HOORN, M., "Algorithms and Approximations for Queueing Systems," CWI Tract No. 8, CWI, Amsterdam, 1984.

VEINOTT, A., "Least d-Majorized Network Flows with Inventory and Statistical Applications," *Management Sci.*, 17 (1971), 547–567.

WELSH, D., *Matroid Theory*. Academic Press, London, 1976.

WHITT, W., "Existence of Limiting Distributions in the $GI/G/s$ Queue, *Math. Oper. Res.*, 7 (1982), 88–94.

WOLFF, R., "Conditions for Finite Ladder Height and Delay Moments," *Oper. Res.*, 32 (1984), 909–916.

WOOD, D. AND R. SARGENT, "The Synthesis of Multiclass Single Server Queueing Systems, Unpublished manuscript, 1984.