J. Appl. Prob. **19**, 167–182 (1982)
Printed in Israel
0021-9002/82/010167-16 $01.85

SCHEDULING JOBS WITH STOCHASTIC PROCESSING REQUIREMENTS ON PARALLEL MACHINES TO MINIMIZE MAKESPAN OR FLOWTIME

RICHARD R. WEBER,* *University of Cambridge*

Abstract

A number of identical machines operating in parallel are to be used to complete the processing of a collection of jobs so as to minimize either the jobs' makespan or flowtime. The total processing required to complete each job has the same probability distribution, but some jobs may have received differing amounts of processing prior to the start. When the distribution has a monotone hazard rate the expected value of the makespan (flowtime) is minimized by a strategy which always processes those jobs with the least (greatest) hazard rates. When the distribution has a density whose logarithm is concave or convex these strategies minimize the makespan and flowtime in distribution. These results are also true when the processing requirements are distributed as exponential random variables with different parameters.

DYNAMIC PROGRAMMING; FLOWTIME; MAKESPAN; MONOTONE HAZARD RATE; OPTIMAL CONTROL; SIGN-CONSISTENT FREQUENCY; STOCHASTIC SCHEDULING

1. Scheduling to minimize makespan or flowtime

1.1. *Stochastic processing requirements.* A number of identical machines operating in parallel are available for processing a collection of jobs. The total processing required to complete each job has the same probability distribution, but some jobs may have received differing amounts of processing prior to the start. The objective is to complete the jobs so as to minimize either their makespan or flowtime. Let C_i be the time at which job i is completed. The *makespan* is the time of the last job completion, $\max\{C_i\}$, and the *flowtime* is the sum of the job completion times, $\Sigma\, C_i$. Both are random variables which depend on the strategy used to order the processing of the jobs on the machines. Preemptive scheduling is permitted, and thus any job may instantaneously be removed from a machine and another job processed instead. A single machine may process several jobs simultaneously, provided its rates of processing those jobs sum to no more than 1, the maximum rate at which a single machine can work.

Received 10 September 1980.

* Postal address: Queens' College, Cambridge CB3 9ET, U.K.

Although scheduling must proceed without knowing exactly how much more processing each unfinished job requires, it may take account of how much processing each job has already received. When job i has already received an amount of processing x_i the probability that the further processing required to complete it is less than s is $\{F(x_i+s)-F(x_i)\}/\{1-F(x_i)\}$. $F(s)$ is a known distribution function with density $f(s)$. The *hazard rate* of job i is defined as $\rho(x_i)=f(x_i)/\{1-F(x_i)\}$, and thus the probability that δ further processing will be sufficient to complete job i is $\delta\rho(x_i)+o(\delta)$.

1.2. *Results for monotone hazard rates.* Theorem 1 states that when $\rho(s)$ is a monotone function of s (increasing or decreasing) then the LHR and HHR scheduling strategies minimize the expected values of the makespan and flowtime respectively. The LHR strategy begins by assigning machines to the job(s) of *lowest hazard rate*, any remaining machines to the job(s) of second-lowest hazard rate, and continues in this manner until all machines are assigned or all jobs are allocated to machines. If at any stage in this procedure the number of unassigned machines is less than the number of jobs of lowest hazard rate amongst those still unallocated to machines, then LHR shares the effort of those machines equally amongst such jobs if $\rho(s)$ is increasing, and assigns them one by one to the jobs of smallest indices amongst such jobs if $\rho(s)$ is decreasing. In the latter case, the choice of jobs of smallest indices is an arbitrary convention ensuring that LHR is uniquely defined. Strategy HHR is the reverse procedure, which begins by allocating machines to the job(s) of *highest hazard rate.* If the number of unassigned machines is ever less than the number of jobs of highest hazard rate amongst those still unallocated to machines, then HHR shares the effort of those machines equally amongst such jobs if $\rho(s)$ is decreasing, and assigns them one by one to the jobs of smallest indices amongst such jobs if $\rho(s)$ is increasing. The LHR and HHR strategies are optimal even if the number of available machines is not constant, but an arbitrary non-decreasing function of time. It is worth observing that when $\rho(s)$ is decreasing LHR is non-preemptive, as is HHR when $\rho(s)$ is increasing. When $\rho(s)$ is a monotone hazard rate we say it is MHR.

Remark. Although we shall find it convenient to require $\rho(s)$ to be strictly monotone, all our theorems are still true for non-increasing or non-decreasing hazard rates. In such cases LHR becomes the strategy of processing those jobs with the longest expected processing times (LEPT), and HHR becomes the strategy of processing those jobs with the shortest expected processing times (SEPT) (where the processing time is the time needed to finish a job if it is processed continuously by a single machine).

1.3. *Results for sign-consistent densities.* Stronger results can be proved for a second class of distributions. Theorem 2 states that when $\log\{f(s)\}$ is a concave or convex function of s then the LHR and HHR scheduling strategies minimize in distribution the makespan and flowtime respectively. For any γ they minimize the probabilities that the makespan and flowtime are greater than γ. Moreover, LHR minimizes the distribution of the makespan even when the number of available machines is an arbitrary function of time, and when some of the jobs are not available for further processing until random times after the start.

When $f(s)$ has a concave or convex logarithm it is called a sign-consistent density of order two (SC$_2$). The name comes from the fact that the determinant of the 2×2 matrix with elements $f(s_i + t_j)$ has the same sign for all $s_1 < s_2$ and $t_1 < t_2$. It is simple to show that a distribution with an SC$_2$ density also has a MHR hazard rate, and that $\rho(s)$ is increasing or decreasing as $\log\{f(s)\}$ is concave or convex. Karlin (1968) has made a detailed study of sign-consistent densities. He and other authors have described their importance in areas of statistical theory, reliability, game theory and mathematical economics (Pólya densities, the concave case, are especially important). The uniform, exponential, hyperexponential, gamma, and folded-normal distributions all have SC$_2$ densities.

Other processing-time distributions can be represented as the limit of sequences of distributions with SC$_2$ densities, and we can thereby establish the results of this subsection for these distributions as well. Suppose n jobs have processing requirements distributed as exponential random variables with different parameters, say $\lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_n$ (λ_i is the constant hazard rate of job i when processed by a single machine). We consider a distribution having an SC$_2$ density and non-decreasing, continuous hazard rate, such that the hazard rate has n plateaus over which it is successively constant at $\lambda_1, \lambda_2, \cdots$, and λ_n, joined by increasing sections. By imagining that the n jobs have received amounts of processing prior to the start such that their hazard rates at the start are just at the beginnings of the relevant plateaus, and then letting the lengths of the plateaus become very large, we can approximate exponentially distributed processing requirements arbitrarily closely. A deterministic distribution, for which $F(s) = 0$, $(0 \leqq s < a)$ and $F(a) = 1$, can also be approximated arbitrarily closely by a distribution with an SC$_2$ density and increasing hazard rate. This gives the well-known result that SEPT minimizes the flowtime when jobs have differing deterministic processing requirements (see Conway, Maxwell and Miller (1967) and Schrage (1968)).

1.4. *Results for special cases.* The above results generalize previous work on scheduling jobs whose processing requirements are distributed as exponential random variables with different means. Glazebrook (1976), (1979) proved the earliest result, showing that HHR minimizes the expected value of the flowtime. Bruno (1976) proved this for just two machines. Weiss and Pinedo (1979) and Bruno, Downey and Frederickson (1981) proved it for any number of machines.

Bruno and Downey (1977) showed that LHR minimizes the expected value of makespan for two machines, and then with Frederickson extended it to any number. Van der Heyden (1981) also proved this result. Pinedo and Weiss (1979) proved a result for non-exponentially distributed processing requirements, showing that LHR and HHR are expected value optimal when $F(s)$ is a mixture of two exponential distributions (hyperexponential). This distribution has an SC_2 density and a decreasing hazard rate.

2. The proof of LHR and HHR optimality

2.1. *Theorem statements and preliminaries.* In previous articles on parallel machine stochastic scheduling problems we have formulated results in discrete time (Weber (1978), (1980a,b); Weber and Nash (1979)). Here we present them in continuous time, adopting the style of optimal control theory. Nash (1973), (1979), Glazebrook (1976) and Nash and Gittins (1977) have used optimal control formulations in proving results for single-machine stochastic scheduling problems, and special cases of parallel-machine problems in which none of the jobs has received any processing prior to the start and the hazard rate is monotone.

Suppose that at the start, time 0, there are n jobs to be processed. The *state* of the jobs at time t is defined as the vector of the amounts of processing they have so far received and is denoted by $x(t) = (x_1(t), \cdots, x_n(t))$. Writing $x^I(t)$ denotes that the jobs $I = \{i_1 \cdots i_l\}$ have already been completed. An admissible scheduling strategy, is a measurable function $v(x^I, t)$ such that for all $t \geqq 0$,

$$v(x^I, t) \in \Omega^I(t) \triangleq \left\{\omega \in [0,1]^n : \sum_{i=1}^{n} \omega_i \leqq m(t), \text{ and } \omega_i = 0 \text{ if } i \in I\right\}.$$

Between job completions the state is controlled by $\dot{x}^I = v(x^I, t)$. At any instant the processing effort applied to a single job can be no more than 1. If job i receives effort at a rate v_i throughout the interval $[t, t+\delta)$ the probability that it is completed within the interval is $\delta v_i p(x_i^I) + o(\delta)$, and if it is completed the state changes to x^{Ii}. The total effort available at time t is $m(t)$. The function $m(t)$ is continuous on the right and it is restricted to the integers to be consistent with the idea of discrete machines (without this restriction the results are still true but the proofs require more complicated notation).

Remark. It has already been noted that when the hazard rate is increasing LHR may share the effort of a single machine amongst several jobs. For example, if the hazard rate is increasing and three jobs, which have had identical amounts of previous processing, are to be completed on two machines, then LHR processes each job at rate $\frac{2}{3}$ until one job is completed. In practice sharing is approximated by very frequently changing the set of jobs being processed, so that the amounts of processing the three jobs have received remain nearly equal. Similarly, when the hazard rate is decreasing HHR may share machine effort. For this reason the definition of an admissible strategy is framed to permit fractional allocations of effort. The allocation $(\frac{2}{3}, \frac{2}{3}, \frac{2}{3})$ is admissible. LHR and HHR are admissible strategies.

Unless otherwise stated, we assume throughout this section that the number of available machines is non-decreasing in time and that all the jobs are available for processing from the start onwards. With this model defined we state our main results.

Theorem 1. If $\rho(s)$ is MHR then the expected values of the makespan and flowtime are minimized by LHR and HHR respectively.

Theorem 2. If $f(s)$ is SC_2 then for any γ the probabilities that the makespan and flowtime are greater than γ are minimized by LHR and HHR respectively. Moreover, the makespan is minimized in distribution even when $m(t)$ is arbitrary and some of the jobs only become available for further processing at random times after the start.

The proofs of Theorems 1 and 2 will be completed at the end of this section where they will be obtained from Lemma 1 and Theorems 3 and 4. All the theorems are proved by induction on the number of jobs not yet complete. Assuming that Theorems 1 and 2 are true when there are less than n jobs unfinished we will show that they are true when the number of unfinished jobs is n. Rather than give eight separate proofs for each of the possible cases that arise from considering whether the distribution has MHR or SC_2 properties, whether $\rho(s)$ is increasing or decreasing, and whether we are seeking to minimize the makespan or flowtime, we shall as far as possible explain the proofs in a way that will hold for all cases, and comment on variations where necessary. To this end, observe that the probability that a random variable is greater than γ is equal to the expected value of an indicator function which is equal to 1 if the variable is greater than γ and equal to 0 otherwise. Let G be one of the following four functions: the makespan, the flowtime, or one of the two indicator functions for the makespan or flowtime being greater than γ. Henceforth a strategy which is LHR or HHR will be denoted by $u = u(x, t)$, and u should be interpreted as denoting LHR or HHR as we are considering a problem of minimizing makespan or flowtime respectively. We shall find it convenient to let (LHR) denote the assumption that G is a function of makespan and u is LHR, and let (HHR) denote the assumption that G is a function of flowtime and u is HHR.

Let $V^I(x^I, t, c \mid v)$ represent the expected value of G, given that starting in state x at time t we employ a scheduling strategy which is identical to v until the time of the next job completion and identical to u thereafter, where for (HHR) the flowtime is to be computed as c plus the completion times of the remaining jobs (we may think of c as the sum of the times at which jobs in I were completed). Assume that when there are just $n-1$ jobs to finish Theorems 1 and 2 are true and u is optimal. To show u is optimal when there are n jobs to finish, we shall show that $v = u$ minimizes $V(x, t, c \mid v)$ for all (x, t, c). $V(x, t, c \mid u)$ will be abbreviated to $V(x, t, c)$. Writing $\rho_i(x)$ for $\rho(x_i)$, we find that for all (x, t, c),

$$x(s) = x(t) + \int_t^s u(x(z), z)\,dz, \tag{1}$$

and

$$\begin{aligned} &V(x, t, c)Q(x(t)) \\ &\quad = \int_t^\infty \sum_{i=1}^n u_i(x(s), s)\rho_i(x(s))V^i(x^i(s), s, c+s)Q(x(s))\,dx, \end{aligned} \tag{2}$$

where we define

$$Q(x(s)) = \prod_{i=1}^n \{1 - F(x_i(s))\}, \qquad (s \geqq t).$$

Equation (2) comes from observing that $Q(x(s))/Q(x(t))$ is the probability that no job completion has occurred before time s and that, conditional on no prior job completion, $u_i(x(s), s)\rho_i(x(s))ds$ is the probability that job i is completed in the interval $[s, s+ds)$.

2.2. *A sufficient condition for u optimality.* We shall show that u is optimal starting from any state x at any time t by considering the effect of giving a small amount of processing to a single job. Assume that $f(s)$ is twice differentiable. From its definition and the nature of u it should be clear that $V(x, t, c)$ has partial derivatives up to the second order in components of x (we shall justify this further during the proof of Lemma 2(b)). Hence for all (x, t, c) it is possible to define

$$V_i(x, t, c)Q(x) = \rho_i(x)V^i(x, t, c+t)Q(x) + \frac{\partial}{\partial x_i}V(x, t, c)Q(x).$$

Starting from x at time t, $\delta V_i(x, t, c)$ is to first order in δ the amount by which the expected value of G would change from $V(x, t, c)$ if we were to give job i an extra δ of processing just before continuing with u over $[t, \infty)$. We may define V_i^j similarly.

The following lemma states a condition which is sufficient to ensure that u is the optimal strategy. It is a particular case of a result in the Hamilton–Jacobi–Bellman theory of continuous dynamic programming, and is equivalent to the result in discrete time that if a strategy and its value function satisfy a dynamic programming equation then it is optimal (for a general formulation of this result within a description of continuous dynamic programming see Varaiya (1972), p. 192, Theorem 1). The proof will be left to the appendix.

Lemma 1. Suppose that for every (x, t, c) and $\omega \in \Omega(t)$,

$$\sum_{i=1}^{n} u_i(x,t)V_i(x,t,c) \leqq \sum_{i=1}^{n} \omega_i V_i(x,t,c).$$

Then $v = u$ minimizes $V(x, t, c \mid v)$ for every (x, t, c).

The remaining proofs are simplified by making an assumption about the starting state. We say that *the ‡ property holds for a state* if the hazard rates of the uncompleted jobs are monotone in the job indices: such that if (LHR) then $\rho_1 \leqq \cdots \leqq \rho_n$, and if (HHR) then $\rho_1 \geqq \cdots \geqq \rho_n$, amongst uncompleted jobs. Recalling that u prefers jobs of smaller indices when making a choice amongst identical jobs, it is clear that by employing u starting from in a state for which ‡ holds, the hazard rates of the uncompleted jobs remain in the same order, ‡ continues to hold, and $i < j$ implies $u_i \geqq u_j$ (i, j uncompleted) for all subsequent time.

2.3. *Proof of Theorems* 1 *and* 2. Define the difference $D_{ij}(x, t, c) = V_i(x, t, c) - V_j(x, t, c)$. Since indexing is arbitrary it is only necessary to consider starting states for which ‡ holds. It is a consequence of Lemma 1 and the inductive hypothesis that to prove u is optimal it is sufficient to demonstrate that $\omega_i = u(x, t)$ minimizes $\Sigma\, \omega_i V_i(x, t, c)$ for all (x, t, c). This is the same as showing that ‡ and $i < j$ implies $D_{ij} = (V_i - V_j)$ is non-positive. Statements (6) and (15) of Theorems 3 and 4 assert this is the case, so assuming their truth, Theorems 1 and 2 are proved. It only remains to prove Theorems 3 and 4 by an inductive argument.

3. Two theorems for LHR and HHR scheduling

3.1. *The basis of an inductive proof.* For all of this section, u is the only scheduling strategy considered. The following lemma gives an expression for $V_i(x, t, c)$ that is the key to the inductive proofs of Theorems 3 and 4. Let $k(t) = (m(t) + 1)$ and observe that job $k(t)$ is a job which would receive additional processing effort if one of the jobs of smaller index than $k(t)$ were already complete, rather than still uncompleted. When $m(t)$ is n or more then $k(t)$ is undefined, but all the expressions that follow are still correct if we simply delete terms like V_k^i and replace terms like D_{kj}^i by $(0 - V_j^i)$. Throughout what follows, $x(s)$ is given by (1), and if ‡ holds for $x(t)$ it also holds for $x(s)$. Let $\xi(t)$ be an n-component vector in which the first $m(t)$ components are 1 and the remaining components 0.

Lemma 2. For all (x, t, c), ‡

(a)
$$V(x,t,c)Q(x(t)) = \int_t^\infty \sum_{h=1}^n \xi_h(s)\rho_h(x(s))V^h(x^h(s),s,c+s)Q(x(s))ds,$$

(b)
$$V_i(x,t,c)Q(x(t)) = \int_t^\infty \Big[\sum_{h\neq i} \xi_h(s)\rho_h(x(s))V_i^h(x^h(x),s,c+s) + \xi_i(s)\rho_i(x(s))V_k^i(x^i(s),s,c+s) - \rho_i(x)\frac{\partial}{\partial c}\{V^i(x^i(s),s,c+s)\}Q(x(s))\Big]ds.$$

Remark. The derivative of V^i with respect to c is computed on the right. We can show this derivative exists for all (x,t,c) by differentiating through (2) with respect to c and using an inductive argument. In proving Theorems 3 and 4 we shall calculate the right-hand $\partial V/\partial c$ when $n=1$ and show that it exists. The restriction to the right-hand derivative is necessary since when G is an indicator function the left-hand derivative may not exist for some value of c. It will be convenient and somewhat clearer if we display the identities of the lemma without arguments as

(3)
$$VQ = \int_t^\infty \sum_{h=1}^n \xi_h\rho_h V^h Q ds,$$

(4)
$$V_iQ = \int_t^\infty \Big\{\sum_{h\neq i} \xi_h\rho_h V_i^h + \xi_i\rho_i V_k^i - \rho_i\frac{\partial V^i}{\partial c}\Big\} Q ds.$$

Proof of Lemma 2(*a*). Identity (3) is a simple consequence of ‡ and it is obtained by noting that in (2) we can replace every $u_h(x(s),s)$ by 1 provided we then compute the sum over only $1 \leqq h \leqq m = m(s)$. For if $u_h > 0$ and $h > m$, or $u_h < 1$ and $h < m$ (machines are shared), then $x_h = x_m$, and u_h and u_m multiply equal equantities.

We shall give an intuitive interpretation of (b), while leaving the formal proof to the appendix. If we consider the portion of the integral over $[t, t+\delta)$ and the interpretation given to V_i above, the lemma states that to first order in δ the following procedures result in the same expected value of G: (a) giving job i an extra δ of processing just before time t, increasing the value recorded for its completion time by δ if it is completed, and then proceeding with u over $[t,\infty)$, and (b) processing the $m(t)$ jobs of lowest index over $[t, t+\delta)$, giving job i an extra δ of processing once $t+\delta$ is reached, and then continuing with u over $[t+\delta,\infty)$ — unless job i has been completed in $[t, t+\delta)$, in which case the extra δ of processing is given to job k.

Let $H(t)$ denote the set of indices less than or equal to $m(t)$ but excluding i and j. Using the lemma we also obtain for $D_{ij}(x,t,c)Q(x)$,

$$D_{ij}Q = \int_t^\infty \left\{ \sum_{h\in H} \rho_h D_{ij}^h + \xi_i\rho_i \left(D_{kj}^i - \frac{\partial V^i}{\partial c} \right) + \xi_j\rho_j \left(D_{ik}^j + \frac{\partial V^j}{\partial c} \right) \right. \tag{5}$$
$$\left. + (1-\xi_j)\rho_j \frac{\partial V^j}{\partial c} - (1-\xi_i)\rho_i \frac{\partial V^i}{\partial c} \right\} Qds.$$

3.2. *Theorems* 3 *and* 4. We now prove that D_{ij} is non-positive for ‡ and $i<j$. Let $f_i(x)$, $\rho_i(x)$, $f_i'(x)$ and $\rho_i'(x)$ denote $f(x_i)$, $\rho(x_i)$ and their derivatives with respect to x_i.

Theorem 3. Suppose $\rho(s)$ is MHR and G is either the makespan or the flowtime. Then for all (x,t,c), ‡ and $i<j$,

$$D_{ij}(x,t,c) \leqq 0. \tag{6}$$

If (LHR) and $j \leqq m(t)$ then

$$\rho_i'(x)\frac{\partial}{\partial x_i} D_{ij}(x,t,c) \geqq 0. \tag{7}$$

If (HHR) and $j \leqq m(t)$ then

$$D_{ij}(x,t,c)+1 \geqq 0, \tag{8}$$

$$\rho_i'(x)\frac{\partial}{\partial x_i} D_{ij}(x,t,c) \leqq 0, \tag{9}$$

and

$$\rho_i'(x)\{D_{kj}^i(x,t,c) - 1 - D_{ij}(x,t,c)\} \leqq 0. \tag{10}$$

Proof. The proofs of (6)–(10) are by induction on the number of uncompleted jobs. Assuming that they are true when rewritten to apply to starting states with just $n-1$ unfinished jobs, we show they are true as written above for n unfinished jobs. This follows from the identities (11)–(14) that follow below. The identities are all produced by straightforward manipulation of (3)–(5) using the fact that when we wish we can differentiate these with respect to x_i simply by taking the derivative inside the integral. This is justified within the proof of Lemma 2(b). Taking a right-hand derivative of (3) with respect to c we get

$$Q\frac{\partial V}{\partial c} = \int_t^\infty \left\{ \sum_{h\in H} \rho_h \frac{\partial V^h}{\partial c} + \xi_i\rho_i \frac{\partial V^i}{\partial c} + \xi_j\rho_j \frac{\partial V^i}{\partial c} \right\} Qds. \tag{11}$$

We can simplify (5) as follows. Let z be the first time greater than or equal to t for which ξ_j is 1. The following identities can then be produced from (5) and (11) for $D_{ij}(x,t,c)Q(x)$.

$$(12)\qquad \begin{aligned} D_{ij}Q = &\int_t^z \left\{ \sum_{h\in H} \rho_h D^{h}_{ij} + \xi_i\rho_i D^{i}_{kj} - \rho_i \frac{\partial V^i}{\partial c} + \rho_j \frac{\partial V^j}{\partial c} \right\} Qds \\ &+ \int_z^\infty \left\{ \sum_{h\in H} \rho_h D^{h}_{ij} + \rho_i \left(D^{i}_{kj} - \frac{\partial V^i}{\partial c} \right) + \rho_j \left(D^{j}_{ik} + \frac{\partial V^j}{\partial c} \right) \right\} Qds. \end{aligned}$$

For $z = t$, $j \leqq m(t)$,

$$(13)\qquad \begin{aligned} Q\frac{\partial}{\partial x_i}(D_{ij}) = &\int_t^\infty \left\{ \sum_{h\in H} \rho_h \frac{\partial}{\partial x_i}(D^{h}_{ij}) + \rho'_i \left(D^{i}_{kj} - \frac{\partial V^i}{\partial c} - D_{ij} \right) \right. \\ &\left. + \rho_j \frac{\partial}{\partial x_i} \left(D^{j}_{ik} + \frac{\partial V^j}{\partial c} \right) \right\} Qds, \end{aligned}$$

$$(14)\qquad \begin{aligned} \left(D^{i}_{kj} - \frac{\partial V^i}{\partial c} - D_{ij} \right) Q = &\int_t^\infty \left\{ \sum_{h\in H} \rho_h \left(D^{hi}_{kj} - \frac{\partial V^{hi}}{\partial c} - D^{h}_{ij} \right) + \rho_j \left(D^{ji}_{kk'} - D^{j}_{ik} - \frac{\partial V^j}{\partial c} \right) \right. \\ &\left. + \rho_k \left(D^{ik}_{k'j} - \frac{\partial V^{ik}}{\partial c} \right) \right\} Qds. \end{aligned}$$

In (14) we denote $(k(s)+1)$ by k'. Identities (11)–(14) are now sufficient to establish (6)–(10) by induction. Assume $n \geqq 2$ and that the theorem is true for $n-1$. It is trivially true that terms in $\partial/\partial c$ are 0 for (LHR) and 1 for (HHR).

By the inductive hypothesis for (6) we deduce that within the first integral on the right-hand side of (12) D^{h}_{ij} is non-positive, as is D^{i}_{kj} (since $k \leqq j$), and also the term $(\rho_j - \rho_i)$ if (HHR). So the value of this integral is non-positive. The second integral on the right-hand side of (12) is more complicated. The term D^{h}_{ij} is again non-positive. Since we have assumed $m(s)$ is non-decreasing and that all jobs are available from the start the inductive hypotheses for (7)–(10) can be applied within the integral as well as within the integrals on the right-hand sides of (13) and (14). If (LHR) then $\rho_i \leqq \rho_j$, and by the inductive hypotheses for (6)–(7) we can deduce $D^{j}_{ik} \leqq -D^{i}_{kj} \leqq 0$ by letting x_i tend to x_j in D^{j}_{ik} (re-indexing if necessary as the limit is taken). This gives $\rho_i D^{i}_{kj} + \rho_j D^{j}_{ik} \leqq 0$, and hence every term within the integral is non-positive. Similarly, if (HHR) then $\rho_i \geqq \rho_j$, and by the inductive hypotheses for (6) and (9) we deduce $D^{j}_{ik} \leqq -D^{i}_{kj} \leqq 0$. This gives $0 \leqq (D^{j}_{ik}+1) \leqq (-D^{i}_{kj}+1)$ by (8). Hence $\rho_i(D^{i}_{kj}-1) + \rho_j(D^{j}_{ik}+1) \leqq 0$, and every term within the integral is non-positive. The two integrals in (12) have been shown to be non-positive and this completes the inductive step for (6).

If (LHR) (7) follows from (13) using the inductive hypotheses for (6) and (7). If (HHR) (8) follows from adding (11) to (12), (10) follows from (14), and (9) follows from (13) using (10) in the second term of the (13) integral after the inductive step for (10) has already been established.

It only remains to check the theorem when $n = 1$. Suppose only job i is to be completed and $m(s) \geqq 1$ for $s \geqq t$. Then setting $c = 0$ if (LHR) we can calculate

$$V(x, t, c) = c + t + \int_{x_i}^{\infty} \{1 - F(s)\} ds / \{1 - F(x_i)\},$$

and

$$V_i(x, t, c) = -1.$$

Substituting these in (6)–(10) checks that the theorem is true for $n = 1$ (terms in j and k are deleted; for example, D_{ij} becomes V_i).

Theorem 4. Suppose $f(x)$ is SC$_2$ and G is an indicator function for either the makespan or the flowtime being greater than γ. Then for all (x, t, c), ‡ and $i < j$,

$$D_{ij}(x, t, c) \leqq 0. \tag{15}$$

$$\rho_i'(x) \frac{\partial}{\partial x_i} \left\{ \frac{1}{\rho_i(x)} \frac{\partial}{\partial c} V(x, t, c) \right\} \leqq 0. \tag{16}$$

If (LHR) then

$$\rho_i'(x) \frac{\partial}{\partial x_i} \left\{ \frac{1}{\rho_i(x)} D_{ij}(x, t, c) \right\} \geqq 0. \tag{17}$$

If (HHR) and $j \leqq m(t)$ then

$$D_{ij}(x, t, c) + \frac{\partial}{\partial c} V(x, t, c) \geqq 0, \tag{18}$$

and

$$\rho_i'(x) \frac{\partial}{\partial x_i} \left[\frac{1}{\rho_i(x)} \left\{ D_{ij}(x, t, c) + \frac{\partial}{\partial c} V(x, t, c) \right\} \right] \leqq 0. \tag{19}$$

Proof. The proof is similar to the proof of Theorem 3. From (11) and (4) we get the following:

$$\begin{aligned} &\rho_i Q \frac{\partial}{\partial x_i} \left(\frac{1}{\rho_i} \frac{\partial V}{\partial c} \right) \\ &\quad = \int_t^{\infty} \left\{ \sum_{h \in H} \rho_h \frac{\partial}{\partial x_i} \left(\frac{1}{\rho_i} \frac{\partial V^h}{\partial c} \right) + \xi_i \frac{1}{\rho_i} \frac{\partial V}{\partial c} \frac{\partial}{\partial x_i} \left(\frac{f_i'}{f_i} \right) \right\} \rho_i Q ds. \end{aligned} \tag{20}$$

$$\begin{aligned} \rho_i Q \frac{\partial}{\partial x_i} \left(\frac{D_{ij}}{\rho_i} \right) = \int_t^{\infty} &\left\{ \sum_{h \in H} \rho_h \frac{\partial}{\partial x_i} \left(\frac{D_{ij}^h}{\rho_i} \right) + \xi_j \rho_j \frac{\partial}{\partial x_i} \left(\frac{D_{ik}^j}{\rho_i} \right) \right. \\ &\left. + \xi_i \frac{D_{ij}}{\rho_i} \frac{\partial}{\partial x_i} \left(\frac{f_i'}{f_i} \right) + \rho_j \frac{\partial}{\partial x_i} \left(\frac{1}{\rho_i} \frac{\partial V^j}{\partial c} \right) \right\} \rho_i Q ds. \end{aligned} \tag{21}$$

Assume $n \geqq 2$ and that the theorem is true for $n - 1$. If (LHR) we put $\partial / \partial c$ terms equal to 0 in all equations. The inductive step for (16) comes from (20), using the fact that for an SC$_2$ density,

$$(22) \qquad \frac{\partial}{\partial x_i}\left(\frac{f'_i}{f_i}\right) \text{ has the opposite sign to } \rho'_i.$$

For (LHR) the inductive step for (17) follows from (21)–(22) and the inductive hypotheses. The step for (15) follows from (17) by observing that D_{ij} / ρ_i increases as x_i tends to x_j and is 0 when x_i equals x_j. We have not had to consider whether or not j is less than $m(t)$. If some jobs only become available for processing after the start the notation must be changed, but all that will happen is that H may shrink, and ξ_j may decrease at some points in time. But the sign of (21) does not depend on the value of ξ_j. Similarly no difficulty is caused if $m(s)$ decreases.

For (HHR) we must consider the two integrals on the right hand side of (12) separately. The $\partial/\partial c$ terms are clearly non-negative. Using (22) and the inductive hypothesis for (16) in (20) we establish (16). The first integral in (12) is non-positive using the inductive hypotheses for (15) and (16). The second integral is non-positive from the inductive hypotheses for (15) and (19). We establish (18) by adding (11) to (12) and checking that the sign of the integrand is non-negative. Similarly (19) is established by adding (20) to (21).

To check the theorem when $n = 1$, suppose only job i is to be completed and its remaining processing requirement is given by the random variable X. Assume $m(s) \geqq 1$ for $s \geqq t$ (if (LHR) then $m(s)$ may perhaps decrease to 0 at some points; the expressions below require minor adjustments, but these do not affect the conclusions). Setting $c = 0$ if (LHR) let $\alpha = (\gamma - c - t)$. Then $V(x, t, c) = \Pr(X > \alpha)$ and we can calculate

$$V_i(x, t, c) = 0, \qquad \alpha < 0$$

$$\frac{V_i(x, t, c)}{\rho(x_i)} = -\frac{f(x_i + \alpha)}{f(x_i)}, \qquad \alpha \geqq 0,$$

and

$$\frac{\partial}{\partial c}\left\{\frac{V(x, t, c)}{\rho(x_i)}\right\} = \frac{f(x_i + \alpha)}{f(x_i)}, \qquad \alpha \geqq 0 \text{ (HHR)}.$$

Note that the derivative $\partial V/\partial c$ exists on the right for all c, but not on the left for $c = \gamma - t$ ($\alpha = 0$). This is why we stated that derivatives with respect to c (and t) are calculated on the right. It is easy to check that $f(s + \alpha)/f(s)$ is monotone in the opposite direction to $\rho(s)$ for $\alpha > 0$. Thus as ρ_i increases, we find that V_i / ρ_i increases and $(\partial V/\partial c)/\rho_i$ decreases. These facts and the above expressions are used in (15)–(19) to show that the theorem is true for $n = 1$.

4. Concluding remarks

The proofs in this paper are necessarily intricate. They become shorter and easier to follow when specialized to the case of exponentially distributed processing times. The reader is refered to Weber (1982) where this has been done.

The conditions under which we have shown the optimality of LHR and HHR for the makespan and flowtime problems are the most general for which the results can be obtained. It should be clear from the proofs that these strategies are not necessarily optimal for hazard rates that are not monotone, and that the assumptions that $m(t)$ is non-decreasing and that all jobs are available for processing from the start can only be relaxed in the SC_2 (LHR) case (see Weber (1980a) for appropriate counterexamples).

By similar methods it can also be shown that for any γ, LHR minimizes the probability that the time to first idleness is less than γ, where the time to first idleness is defined as the first time that $m(t)$ is greater than the number of jobs still to be completed. This is true for any distribution with a MHR hazard rate, and as before $m(t)$ must be non-decreasing and all jobs must be available for processing at the start unless the density is SC_2. The proof is almost identical to that for LHR minimizing makespan if we simply redefine V as the probability that when using an LHR strategy the time to first idleness is less than γ (Weber and Nash (1979) and Weber (1980a) give further details). The result shows that an LHR strategy maximizes in distribution the length of time for which a machine that needs m components to operate can be kept in operation using a stock of n ($n > m$) components. For $m = 2$ this is the 'lady's nylon stocking problem' for which Cox (1959) hypothesised LHR optimality.

Weiss and Pinedo (1979) consider m non-identical machines which process jobs at different rates, $s_1 \geqq \cdots \geqq s_m$. The jobs have exponentially distributed processing times with parameters $\lambda_1, \cdots, \lambda_n$, and when job i is processed on machine j its instantaneous hazard rate is $s_j\lambda_i$. They show that the expected value of the flowtime is minimized by following a version of HHR which always assigns the job of greatest λ to machine 1, the job of second greatest λ to machine 2, and so on. A version of LHR that assigns the job of lowest λ to machine 1, and so on, minimizes the expected value of the makespan. These results can be generalized to the models of this paper, although the equivalent of Lemma 2 requires a more complicated notation.

5. Appendix

5.1. *Proof of Lemma* 1. Suppose v is an admissible strategy. We get the following series of relationships:

$$0 = \frac{d}{ds}\{V(x,s,c)Q(x)\} + \sum_{i=1}^{n} u_i(x,s)\rho_i(x)V^i(x,s,c+s)Q(x) \tag{23}$$

$$= \frac{\partial}{\partial s}\{V(x,s,c)Q(x)\} + \sum_{i=1}^{n} u_i(x,s)\frac{\partial}{\partial x_i}\{V(x,s,c)Q(x)\}$$

$$+ \sum_{i=1}^{n} u_i(x,s)\rho_i(x)V^i(x,s,c+s)Q(x) \tag{24}$$

$$=\frac{\partial}{\partial s}\{V(x,s,c)Q(x)\}+\sum_{i=1}^{n} u_i(x,s)V_i(x,s,c)Q(x) \tag{25}$$

$$\leqq\frac{\partial}{\partial s}\{V(x,s,c)Q(x)\}+\sum_{i=1}^{n} v_i(x,s)V_i(x,s,c)Q(x) \tag{26}$$

$$=\frac{\partial}{\partial s}\{V(x,s,c)Q(x)\}+\sum_{i=1}^{n} v_i(x,s)\frac{\partial}{\partial x_i}\{V(x,s,c)Q(x)\}$$

$$+\sum_{i=1}^{n} v_i(x,s)\rho_i(x)V^i(x,s,c+s)Q(x)$$

$$=\frac{d}{ds}\{V(x,s,c)Q(x)\}+\sum_{i=1}^{n} v_i(x,s)\rho_i(x)V^i(x,s,c+s)Q(x). \tag{27}$$

Equation (23) is obtained by differentiating (2), and (24) by observing that $d/ds=\partial/\partial s+\Sigma u_i\partial/\partial x_i$. The derivative $\partial V/\partial s$ is taken on the right and can be shown to exist by an inductive proof based on differentiating through (2) and noting that $u(x,t)$ is piecewise constant and continuous on the right. The existence of the derivative $\partial V/\partial x_i$ will be justified in the proof of Lemma 2(b). The definition of V_i gives (25), and (26) follows from the hypothesis of the lemma. Note that the time derivative in (23) is along the trajectory $\dot{x}=u(x,s)$, whereas it is along the trajectory $\dot{x}=v(x,s)$ in (27). Integrating (27) along the trajectory $\dot{x}=v(x,s)$ we find

$$0\leqq -V(x,t,c)Q(x)+\int_t^{\infty}\sum_{i=1}^{n} v_i(x,s)\rho_i(x)V^i(x,s,c+s)Q(x)ds$$

$$=-V(x,t,c\mid u)Q(x)+V(x,t,c\mid v)Q(x).$$

The lemma is therefore proved.

5.2. *Proof of Lemma* 2(*b*). For convenience we omit arguments from the expressions below. By the definition of V_i, ‡ and Lemma 2(a)

$$V_iQ=\rho_iV^iQ+\frac{\partial}{\partial x_i}\int_t^{\infty}\sum_{h=1}^{n}\xi_h\rho_hV^hQds. \tag{28}$$

The integral on the right of (28) is equal to VQ, and for the following reason the derivative of this integral can be found by taking the derivative with respect to x_i inside. Suppose the starting state x is perturbed to y by changing x_i to $x_i+\delta$. Without loss of generality we suppose that amongst indices j such that $x_j=x_i$, i is chosen as the one for which ‡ will hold for y as well as for x. When δ is positive (negative) this can been done by letting i be the smallest (largest) such j. Assume all components of x not equal to x_i are further from x_i then δ. Starting from y, let $y(s)$ be the perturbed state reached at time s by employing u. By considering the nature of the LHR and HHR strategies it is not hard to check that $y(s)=$

$x(s)+\delta(s)$ where $\Sigma\,\delta_j(s)=\delta$ and $\delta_j(s)>0$ only if $|x_i(s)-x_j(s)|<\delta$. Hence the first-order change from VQ when the integral is evaluated along $y(s)$ can be shown to be just δ times the value obtained by differentiating inside the integral, as claimed above. Differentiating the integral once more gives a proof by induction that V is twice differentiable with respect to all components of x, by assuming this for V_h and remembering that $\rho(s)$ is twice differentiable. In particular this shows that V_i is once differentiable, as was assumed in the proofs of Theorems 3 and 4. We carry out the differentiation in (28) to obtain

$$V_iQ=\rho_iV^iQ+\int_t^\infty\left\{\sum_{h\neq i}\xi_h\rho_h\frac{\partial}{\partial x_i}(V^hQ)+\xi_i\frac{\partial}{\partial x_i}(\rho_iV^iQ)\right\}ds. \tag{29}$$

The total time derivative of ρ_iV^iQ satisfies

$$\begin{aligned}&\frac{d}{ds}(\rho_iV^iQ)\\ &=\xi_i\frac{\partial}{\partial x_i}(\rho_iV^iQ)+\sum_{h\neq i}\xi_h\frac{\partial}{\partial x_i}(\rho_iV^iQ)+\frac{\partial}{\partial c}(\rho_iV^iQ)+\frac{\partial}{\partial s}(\rho_iV^iQ).\end{aligned} \tag{30}$$

The V^i terms in (30) are evaluated at $(x, s, c+s)$, but the partial derivative $\partial V^i/\partial s$ is calculated with respect to the second argument only (the derivatives $\partial V^i/\partial c$ and $\partial V^i/\partial s$ are taken on the right). Using (30) to substitute for the last term of (29) and integrating the $d(\rho_iV^iQ)/ds$ term we get

$$\begin{aligned}V_iQ=\int_t^\infty&\left[\sum_{h\neq i}\xi_h\left\{\rho_h\frac{\partial}{\partial x_i}(V^hQ)-\rho_i\frac{\partial}{\partial x_h}(V^iQ)\right\}\right.\\ &\left.-\rho_i\frac{\partial}{\partial c}(V^iQ)-\rho_i\frac{\partial}{\partial s}(V^iQ)\right]ds.\end{aligned} \tag{31}$$

Adding and subtracting $\rho_i\rho_hV^{ih}Q$ within the summation of (31) makes this sum become $\Sigma\,\xi_h\{\rho_hV_i^h-\rho_iV_h^i\}Q$. Let u^i denote u when job i has been completed. Let $\xi^i(s)$ be the n vector in which the first $m(s)$ components amongst the set not including the ith component are 1, and all the other components 0. For the same reasons as (25) in Lemma 1, and Lemma 2(a), u^i and ξ^i must satisfy the following equation:

$$0=\frac{\partial}{\partial t}(V^i)+\sum_{h\neq i}u_h^iV_h^i=\frac{\partial}{\partial t}(V^i)+\sum_{h\neq i}\xi_h^iV_h^i. \tag{32}$$

Multiplying (32) by ρ_iQ and substituting for the final term of (31) we obtain

$$V_iQ=\int_t^\infty\left[\sum_{h\neq i}\{\xi_h\rho_hV_i^h+(\xi_h^i-\xi_h)\rho_iV_h^i\}-\rho_i\frac{\partial V^i}{\partial c}\right]Qds. \tag{33}$$

Observe that $(\xi_h^i-\xi_h)$ is 1 for $i\leqq m$ and $h=k$, and 0 otherwise. Thus the middle term of (33) becomes just $\xi_i\rho_iV_k^i$ and this completes the proof of the lemma by giving identity (4).

References

BRUNO, J. (1976) Sequencing tasks with exponential service times on parallel machines. Technical report, Department of Computer Science, Pennsylvania State University.

BRUNO, J. AND DOWNEY, P. (1977) Sequencing tasks with exponential service times on parallel machines. Technical report, Department of Computer Sciences, University of California at Santa Barbara.

BRUNO, J., DOWNEY, P. AND FREDERICKSON, G. N. (1981) Sequencing tasks with exponential service times to minimize the expected flowtime or makespan. *J. Assoc. Comput. Mach.* **28**, 100–113.

CONWAY, R. W., MAXWELL, W. L. AND MILLER, L. W. (1967) *The Theory of Scheduling.* Addison-Wesley, Reading, Ma.

COX, D. R. (1959) A renewal problem with bulk ordering of components. *J.R. Statist. Soc.* B **21**, 180–189.

GLAZEBROOK, K. D. (1976) Stochastic Scheduling. Ph.D. Thesis, University of Cambridge.

GLAZEBROOK, K. D. (1979) Scheduling tasks with exponential service times on parallel processors. *J. Appl. Prob.* **16**, 685–689.

KARLIN, S. (1968) *Total Positivity*, Vol. I. Stanford University Press, Stanford.

NASH, P. (1973) Optimal Allocation of Resources to Research Projects. Ph.D. Thesis, University of Cambridge.

NASH, P. (1979) Controlled jump process models for stochastic scheduling problems. *Internat. J. Control* **30**, 1011–1026.

NASH, P. AND GITTINS, J. C. (1977) A Hamiltonian approach to optimal stochastic resource allocation. *Adv. Appl. Prob.* **9**, 55–68.

PINEDO, M. AND WEISS, G. (1979) Scheduling stochastic tasks on two parallel processors. *Naval Res. Logist. Quart.* **27**, 528–536.

SCHRAGE, L. E. (1968) A proof of the shortest remaining process time discipline. *Operat. Res.* **16**, 687–689.

VAN DER HEYDEN, J. (1981) Scheduling jobs with exponential processing and arrival times on identical processors so as to minimize expected makespan. *Math. Operat. Res.* **6**, 305–312.

VARAIYA, P. P. (1972) *Notes on Optimization.* Van Nostrand Reinhold, New York.

WEBER, R. R. (1978) On the optimal assignment of customers to parallel servers. *J. Appl. Prob.* **15**, 406–413.

WEBER, R. R. AND NASH, P. (1979) An optimal strategy in multi-server stochastic scheduling. *J. R. Statist. Soc.* B **40**, 323–328.

WEBER, R. R. (1980a) Optimal Organization of Multi-server Systems. Ph.D. Thesis, University of Cambridge.

WEBER, R. R. (1980b) On the marginal benefit of adding servers to *G/GI/m* queues. *Management Sci.* **26**, 946–951.

WEBER, R. R. (1982) Scheduling stochastic jobs on parallel machines to minimize makespan or flowtime. Proceedings of the ORSA-TIMS Special Interest Meeting: Applied Probability — Computer Science, the Interface. To appear.

WEISS, G. AND PINEDO, M. (1979) Scheduling tasks with exponential service times on non-identical processors to minimize various cost functions. *J. Appl. Prob.* **17**, 187–202.