A STOCHASTIC APPROACH TO HIERARCHICAL PLANNING AND SCHEDULING

M.A.H. Dempster
Dalhousie University, Halifax, Nova Scotia, Canada
and
Balliol College, Oxford, England

RR-84-6
March 1984

Reprinted from *Deterministic and Stochastic Scheduling*, NATO Advanced Study Institute Proceedings Series, M.A.H. Dempster *et al.* (Eds) (1982)

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
Laxenburg, Austria

Reprinted with permission from *Deterministic and Stochastic Scheduling*, NATO Advanced Study Institute Proceedings Series, pp. 271–296.

Printed by Novographic, Vienna, Austria

PREFACE

This paper reports on a project involving an international group of researchers. It continues recent trends in IIASA research which concern studies of hierarchical systems and optimization of stochastic systems and is a sequel to two earlier papers RR-84-4 and RR-84-5.

The purpose of this sequence of reports is to demonstrate that the efficiency of hierarchical planning systems can be analyzed in a more rigorous fashion than has been customary so far. These systems are, after all, nothing more than appropriate heuristics to solve multistage stochastic programs. Given the obvious intractability of such problems, precise statements about the performance of approximation algorithms that mirror the top-down sequential nature of actual hierarchical decision making (i.e., based on averaging and aggregation until more refined data become available) are of immediate interest to researchers and practitioners.

In this paper the author relates the earlier research to the latest results in parallel machine stochastic scheduling and stochastic programming and treats in detail some two-level machine shop design/scheduling problems and a three-level distribution planning/vehicle routing problem which is currently an object of study by the group.

All the members of this group are active in the development of computer software for planning and operations management in various environments, so that in a very real sense this paper describes theoretical research stemming from practice.

M.A.H. DEMPSTER

A STOCHASTIC APPROACH TO HIERARCHICAL PLANNING AND SCHEDULING

M.A.H. Dempster

Dalhousie University, Halifax
Balliol College, Oxford

This paper surveys recent results for stochastic discrete programming models of hierarchical planning problems. Practical problems of this nature typically involve a sequence of decisions over time at an increasing level of detail and with increasingly accurate information. These may be modelled by multistage stochastic programmes whose lower levels (later stages) are stochastic versions of familiar NP-hard deterministic combinatorial optimization problems and hence require the use of approximations and heuristics for near-optimal solution. After a brief survey of distributional assumptions on processing times under which SEPT and LEPT policies remain optimal for m-machine scheduling problems, results are presented for various 2-level scheduling problems in which the first stage concerns the *acquisition* (or assignment) of machines. For example, heuristics which are asymptotically optimal in expectation as the number of jobs in the system increases are analyzed for problems whose second stages are either *identical* or *uniform* m-machine scheduling problems. A 3-level location, distribution and routing model in the plane is also discussed.

1. INTRODUCTION

Practical hierarchical planning problems typically involve a sequence of decisions over time at an increasing level of detail and with increasingly accurate information. For example, a 3-level hierarchy of planning decisions in terms of increasingly finer time units is often utilized for manufacturing operations (see Figure 1). The first level concerns medium term planning, which works with projected quarterly or monthly averages and is primarily concerned with the acquisition of certain resources. The next level treats

M. A. H. Dempster et al. (eds.), Deterministic and Stochastic Scheduling, 271–296.

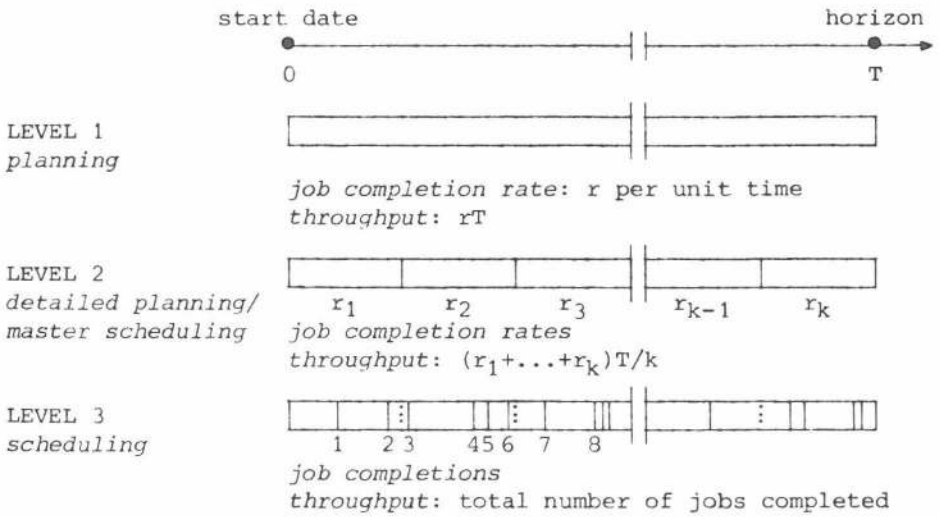

Figure 1. 3-Level scheme for hierarchical production planning/ scheduling.

weekly production scheduling, while the third level is concerned with the real-time sequencing of jobs through various machine centres on the shop floor. The first two levels can currently be handled adequately by respectively deterministic linear programming and combinatorial permutation procedures, but the third realistically invclves a network of stochastic m-machine scheduling problems whose natural setting is in continuous time.

More generally, many hierarchical planning problems can be modelled by multistage stochastic programmes whose later stages (lower levels) are stochastic versions of familiar NP-hard deterministic combinatorial optimization problems. Hence they usually require the use of approximations and heuristics for near-optimal solution. For these systems, in which at the higher levels - as in their practical counterparts - details are suppressed and instead replaced by *approximate* aggregates, one would hope to demonstrate that the instances of the data for which these higher level assumptions are *severely* violated occur with increasingly negligible probability as the number of tasks in the system becomes large. Asymptotic probabilistic analysis of heuristics has therefore an important role to play in the analysis of hierarchical stochastic programming models (*cf.* [Dempster *et al.* 1981A]).

Recently, computer-based planning systems have become popular for practical multilevel decision problems in a variety of applications including manufacturing production planning and scheduling, trade training school planning and scheduling, distribution planning and vehicle scheduling, manpower planning, crew routing and scheduling and computer utilization and scheduling (see, for

example, [Dempster *et al.* 1981A; Dempster & Whittington 1976; Dirickx & Jennergren 1979; Giannessi & Nicoletti 1979; Kao & Queyranne 1981; Kleinrock 1976]). In principle, the performance of such systems can be evaluated relative to optimality for the appropriate multistage stochastic programming model.

This paper primarily reports on a programme of research conducted jointly with M.L. Fisher, B.J. Lageweg, J.K. Lenstra, A.H.G. Rinnooy Kan and L. Stougie.

Section 2 contains a brief survey of distributional assumptions on processing times under which *shortest expected processing time* (SEPT) and *longest expected processing time* (LEPT) policies remain optimal for m-machine scheduling problems with appropriate expected value criteria. Section 3 sets out various 2-level scheduling problems as 2-stage stochastic programmes with recourse in which the first stage concerns the *acquisition* (or assignment) of machines and the second stage is an m-machine scheduling problem. The difficulty of exact solution of such problems is also discussed in §3 and representative results are quoted. In §4 results are presented which analyze heuristics for the 2-level scheduling problems in §3. For example, heuristics which are almost surely asymptotically optimal as the number of jobs in the system increases are analyzed for problems whose second stages are either *identical* or *uniform* m-machine scheduling problems. A 3-level location, distribution and routing model in the plane is discussed in §5 and some conclusions drawn in §6. Open problems and directions for further research are indicated throughout the paper.

2. RECENT RESULTS IN PARALLEL MACHINE STOCHASTIC SCHEDULING

This section surveys recent results for the following basic m-machine scheduling problem.

Problem 2.1. Schedule n *jobs* $j \in J := \{1,\ldots,n\}$ with independent random *processing requirements* $\tilde{p}_j$ on m *uniform machines* $i \in M := \{1,\ldots,m\}$ with *speeds* s_i, subject to the usual constraints that at any moment at most one job can be processed by any machine and at most one machine can process any job.

We may think of the processing requirements as defined relative to standard time units so that if job j is assigned for processing solely to machine i it will be completed in the random (clock) *time* $\tilde{p}_j/s_i$. The *machine set* M will be assumed to be ordered in decreasing order of speed so that $s_1 \geq s_2 \geq \ldots \geq s_m \geq 0$. When $s_i \equiv 1$, we speak of m *identical* machines.

In order to complete the definition of Problem 2.1 various alternative assumptions may be made. These concern the nature of the *time* set, the possibility of *preemption* of running jobs and the possibility and nature of *release dates* or *arrivals* of some of the jobs *after* time zero.

The time set for the problem may be either discrete ($\mathbb{N}$) or continuous ($\mathbb{R}_+$). Scheduling problems are most naturally set in *continuous time* (CT); usually a *discrete time* (DT) setting is generated by a discrete time step, for the purposes of approximation or simplification. It is usually not an entirely trivial or even straightforward matter to extend discrete time results to their continuous time analogues.

Although finer classifications are possible (see, *e.g.*, [Pinedo & Schrage 1982]) we shall be interested simply in whether or not any job currently being processed may, at any point, be interrupted and set aside for later processing or immediate assignment to a different machine. In the affirmative situation we say *preemption* is allowed, and otherwise we say the problem allows *no preemption*.

In case all n jobs are available for processing at time zero we say that *no arrivals* are allowed. If some, say ℓ ($0 < \ell \leq n$), of the n jobs are not available for processing at time zero and are released at subsequent random times $\underset{\sim}{r}_j$ according to some (possibly labelled) stochastic point process independent of the processing requirements we speak of a problem with *arrivals*. In this case we condition the problem on the occurrence of exactly ℓ events of the arrival process.

We shall be interested in constructing schedules which are "optimal" in terms of two schedule measures. Let $\underset{\sim}{C}_j$ denote the (random) *completion time* of job $j \in J$ under a given scheduling policy. Then the schedule *makespan* $\underset{\sim}{C}_{max}$ is the earliest time at which all jobs are completed, defined by

$$\underset{\sim}{C}_{max} := \max_{j \in J} \{\underset{\sim}{C}_j\},$$

while the schedule *flowtime* $\Sigma\underset{\sim}{C}_j$ is the sum of the job completion times given by

$$\Sigma\underset{\sim}{C}_j := \Sigma_{j=1}^{n} \underset{\sim}{C}_j.$$

In the deterministic case, minimization of makespan optimizes the completion time of the *last* job, while minimization of flowtime is equivalent to optimizing the completion time of the *average* job. In the stochastic case, a natural schedule minimization criterion is in terms of the *expected value* of makespan or flowtime. For these two measures we shall also be interested in minimality in *distribution*, *i.e.* the probability of achieving a given makespan or flowtime level γ is, uniformly in γ, at least as great for the schedule resulting from the optimal policy π^o as for any other π. For example, for all $\gamma \in \mathbb{R}_+$,

$$P\{\underset{\sim}{C}_{max}^{\pi^o} \leq \gamma\} \geq P\{\underset{\sim}{C}_{max}^{\pi} \leq \gamma\}, \text{ i.e.}$$

$$\bar{F}_{\underset{\sim}{C}_{max}^{\pi^o}}(\gamma) \leq \bar{F}_{\underset{\sim}{C}_{max}^{\pi}}(\gamma), \tag{2.1}$$

where $\bar{F}_{\underset{\sim}{C}_{max}} := 1 - F_{\underset{\sim}{C}_{max}}$ denotes the *survivor function* of $\underset{\sim}{C}_{max}$ under

the appropriate policy. If $\underset{\sim}{x}$ and $\underset{\sim}{y}$ are two random variables, then $\underset{\sim}{x}$ is *stochastically dominated* by $\underset{\sim}{y}$, written $\underset{\sim}{x} <_D \underset{\sim}{y}$, if, and only if, $\bar{F}_{\underset{\sim}{x}}(t) \leq \bar{F}_{\underset{\sim}{y}}(t)$ for all t. It is easy to see that $\underset{\sim}{x} <_D \underset{\sim}{y}$ implies $E\underset{\sim}{x} \leq E\underset{\sim}{y}$ (when the expectations exist), but not conversely.

Choosing one of the possibilities discussed above regarding the time set, preemption and job arrivals - and specifying an optimality criterion - generates from the basic Problem 2.1 a stochastic scheduling problem. Since we shall allow scheduling decisions at t = 0 and at the epochs of subsequent job arrivals and completions, the resulting stochastic scheduling problems can be formulated as *semi-Markov* decision problems over an infinite horizon (see, *e.g.*, [Ross 1970, Ch.4]). In this section we are interested in conditions on the processing requirement distributions under which optimal scheduling policies for these problems can be specified in a simple form which utilizes dynamic *priority indices* (*cf.* Gittins' "dynamic allocation" indices [Gittins 1979]). At any moment these policies assign to each unfinished job a number - its priority index - and at decision epochs unfinished jobs are assigned to (speed ordered) free machines in monotonic order of their current indices. (When preemptions are allowed all m machines are considered free at job completion epochs.) Policies of similar form have recently been found applicable to a large class of related *Markov* decision problems in discrete time including 1-machine scheduling, search problems and multiarmed bandit and superbandit processes (see [Presman & Sonin 1979; Gittins 1979; Whittle 1980; Nash 1980]).

More formally, let U_t ($\subset J$) denote the set of *unfinished jobs* at time $t \geq 0$. Then a *priority policy* π defines for each decision epoch t a permutation of the elements of U_t,

$$\pi_t : U_t \xrightarrow{1-1} U_t,\ j \longrightarrow \pi_t(j),$$

and assigns jobs to free machines in (speed) order according to their permutation (priority) order. Let $p_j(t)$ be the amount of processing already received at time t by job $j \in U_t$ with processing requirement $\underset{\sim}{p}_j$ and denote by $\mu_j(t) := E\{\underset{\sim}{p}_j | \underset{\sim}{p}_j > p_j(t)\}$ the expectation of the processing requirement remaining at time t for job j. The *longest expected processing time* (LEPT) policy is the priority policy which at a decision epoch t reorders jobs in *decreasing* order of $\mu_j(t)$. The *shortest expected processing time* (SEPT) policy is the priority policy which at a decision epoch t reorders jobs in *increasing* order of $\mu_j(t)$.

Table 1 sets out currently known results concerning the optimality of LEPT and SEPT policies for the variants of Problem 2.1 resulting from the possible alternatives cited above. Since, as previously mentioned, makespan criteria concern minimizing the completion time of the *last* job, it is intuitively obvious that potentially long jobs should be processed first and hence LEPT is a candidate for optimizing makespan. Similarly, since flowtime criteria require the minimization of the completion time of the *average* job, potentially short jobs should be processed first and

LEPT/Makespan $C_{\max}$		SEPT/Flowtime ΣC_j	
Deterministic			
$P\|pmtn\|C_{\max}$ (LPT)	(see text)	$P\|\|\Sigma C_j$ (SPT)	[Conway *et al.* 1967, Ch.4.4]
Exponential			
$P\|\|E C_{\max}$ $Q\|pmtn\|E C_{\max}$	[Weiss & Pinedo 1980]	$P\|\|E\Sigma C_j$ $Q\|pmtn\|E\Sigma C_j$	[Weiss & Pinedo 1980]
$P\|\|\bar{F}_{C_{\max}}$	?	$P\|\|\bar{F}_{\Sigma C_j}$	?
$P\|pmtn,r_j\|\bar{F}_{C_{\max}}$	?	$P\|pmtn,r_j\|\bar{F}_{\Sigma C_j}$	?
Geometric			
$P\|\|E C_{\max}$	?	$P\|\|E\Sigma C_j$	[Gittins 1981]
$Q\|pmtn\|E C_{\max}$	?	$Q\|pmtn\|E\Sigma C_j$	?
$P\|\|\bar{F}_{C_{\max}}$	?	$P\|\|\bar{F}_{\Sigma C_j}$	?
$P\|pmtn,r_j\|\bar{F}_{C_{\max}}$	?	$P\|pmtn,r_j\|\bar{F}_{\Sigma C_j}$	?
Log Convex Similar			
$P\|pmtn\|\bar{F}_{C_{\max}}$ $P\|pmtn,r_j\|\bar{F}_{C_{\max}}$	DT [Weber 1979] CT [Weber 1981]	$P\|\|\bar{F}_{\Sigma C_j}$	DT [Weber 1979] CT [Weber 1981]
ICR *Similar*			
$P\|pmtn\|E C_{\max}$	DT [Weber 1979] CT [Weber 1981]	$P\|\|E\Sigma C_j$	DT [Weber 1979] CT [Weber 1981]
Log Concave Similar			
$P\|\|\bar{F}_{C_{\max}}$ $P\|pmtn,r_j\|\bar{F}_{C_{\max}}$	DT [Weber 1979] CT [Weber 1981]	$P\|pmtn\|\bar{F}_{\Sigma C_j}$	DT [Weber 1979] CT [Weber 1981]
DCR *Similar*			
$P\|\|E C_{\max}$	DT [Weber 1979] CT [Weber 1981]	$P\|pmtn\|E\Sigma C_j$	DT [Weber 1979] CT [Weber 1981]

Table 1. Summary of independent processing requirement distributions under which priority policies optimize makespan and flowtime criteria for multimachine scheduling.

SEPT is the obvious candidate for an optimal policy regarding flowtime.

Problems in Table 1 are specified by a natural modification for stochastic problems of the 3-field problem classification $\alpha|\beta|\gamma$ currently in use for deterministic scheduling problems. The fields α, β and γ refer respectively to the *machine environment*, *job characteristics* and *optimality criterion*. (The reader is referred to [Lawler *et al.* 1982] for more details.) As mentioned above, we are interested here only in identical (P) and uniform (Q) parallel machine environments. The job characteristics of interest are whether preemption is permitted (*pmtn*) or not (blank field) and whether random *release dates* corresponding to an arrival process are specified for all jobs ($\tilde{r}_j$) or all jobs are available at t = 0 (blank field). Results for two stochastic optimality criteria are reported for both makespan and flowtime, *viz.* minimization in expectation (*e.g.* $E\Sigma\tilde{C}_j$) and in distribution, *i.e.* with respect to the partial ordering of *stochastic dominance* (*e.g.* $\bar{F}_{\tilde{C}_{max}}$) (*cf.* (2.1)). Since the families of processing requirement distributions for which results have been obtained have fairly complex specifications for which no acronyms - or even terminology - have been generally agreed, distributional assumptions have not been incorporated in the symbolic problem classifications (for the opposite approach see [Pinedo & Schrage 1982]). Table 1 reports only the best results obtained to date; no attempt has been made to supply complete references on a problem (but in this regard see [Weber 1981; Weiss 1982]). In order to appreciate the information contained in Table 1, some remarks are in order.

First notice that for the deterministic problem $P||\Sigma C_j$ the priority policy *shortest processing time first* (SPT) is actually a *list scheduling* policy - jobs may be placed in order (of increasing processing requirement) at t = 0 and assigned to machines as they become free in this order without subsequent permutation - and hence it may be implemented in O(n log n) running time. Since remaining processing requirements decrease linearly with processing, the SPT order of unfinished jobs never changes and hence reordering of U_t and preemption are never required to optimize flowtime. On the other hand, the *largest processing time first* (LPT) order of unfinished jobs will of course change with processing, and hence the LPT *list scheduling* policy is easily seen to be suboptimal for the *NP-hard* [Karp 1972] nonpreemptive problem $P||C_{max}$.

The preemptive problem $P|pmtn|C_{max}$ is usually solved by McNaughton's *wrap-around rule* (see, *e.g.*, [Baker 1974, Ch.5.2.1]) which yields the optimal value

$$C^{o}_{max} = \max\{P_n/m, p_{max}\}, \tag{2.2}$$

where

$$P_n := \Sigma_{j=1}^{n} p_j, \quad p_{max} := \max_{j\in J}\{p_j\}, \tag{2.3}$$

in O(n) time. This algorithm gives only one of many optimal schedules and, although it creates at most m-1 preemptions, makes no attempt to minimize this number. The problem of minimizing the number of preemptions is in fact NP-hard. Alternatively, an optimal schedule can be obtained by a simple preemptive LPT priority policy, which is based on *processor sharing*. The algorithm may be described as follows. Arrange the jobs such that $p_1 \geq \ldots \geq p_n$. At time zero start processing jobs $1,\ldots,m'$, where $m' = \max\{j|p_j = p_m\}$; if $m' > m$, a number of jobs with processing requirement p_m must equally share a (smaller) number of machines. The next decision epoch occurs when the remaining processing requirement of another job becomes equal to that of a job with initial processing requirement p_m. Then repeat, with remaining rather than original processing requirements. Apply McNaughton's rule in each of the intervals generated to resolve processor sharing. All this requires $O(n^2)$ time.

In the case of deterministic *uniform* machine problems, $Q||\Sigma C_j$ and $Q|pmtn|\Sigma C_j$ can be solved in polynomial time by appropriately modified SPT policies, and $Q|pmtn|C_{max}$ is still solvable in O(n) time given an LPT ordering of the jobs (see [Lawler *et al.* 1982]).

By virtue of the *memoryless* property of the exponential distribution, for exponentially distributed processing requirements the expectation of the remaining processing requirement is always equal to the expectation of the original requirement. Hence preemption may be expected to be irrelevant and, in the parallel machine case, jobs may be initially monotonically ordered in terms of expected processing requirement and both LEPT and SEPT implemented as list scheduling (*i.e.* nondynamic priority) policies in O(n log n) time. This has been established in [Weiss & Pinedo 1980]. They have also given the only treatment to date of (optimal) stochastic scheduling for uniform machine models. For exponential processing requirements, preemption will only be necessary in optimal LEPT and SEPT priority policies for these models to move running jobs to faster machines. Preemption and priority policies are required for *all* problems involving random release dates, since preemption and job reordering may be needed at job arrival epochs. Results involving optimality in distribution for problems with job arrival processes and exponential processing requirements are currently open, as (with the sole exception of the treatment of $P||E\Sigma\tilde{C}_j$ in [Gittins 1981]) are discrete time - *i.e. geometric* processing requirement distribution - analogues of the Weiss-Pinedo results. Nevertheless, sufficient is known about various stochastic scheduling problems with exponential processing requirements to begin the analysis of their computational complexity, see [Pinedo 1982] where several apparently anomalous results are presented.

Some definitions are in order to continue the discussion of Table 1. The *completion rate* $h_{\tilde{p}}$ of a job with processing requirement $\tilde{p}$, density $f_{\tilde{p}}$ and survivor function $\bar{F}_{\tilde{p}}$ is given by

$$h_{\underset{\sim}{p}}(t) := \begin{cases} f_{\underset{\sim}{p}}(t)/\bar{F}_{\underset{\sim}{p}}(t) & \text{if } \underset{\sim}{p} \text{ is absolutely continuous,} \\ f_{\underset{\sim}{p}}(t+1)/\bar{F}_{\underset{\sim}{p}}(t) & \text{if } \underset{\sim}{p} \text{ is discrete.} \end{cases} \tag{2.4}$$

The distribution of $\underset{\sim}{p}$ is *increasing* (*decreasing*) *completion rate* (ICR, respectively DCR) if, and only if, $h_{\underset{\sim}{p}}$ is a nondecreasing (nonincreasing) function on $\mathbb{R}_+$. If $\underset{\sim}{p}$ is absolutely continuous, its distribution is *log(arithmically) convex* (*concave*) if, and only if, log $f_{\underset{\sim}{p}}$ is convex (concave). Alternatively, the distribution of $\underset{\sim}{p}$ is said to be *increasing* (*decreasing*) *likelihood ratio* (ILR, respectively DLR). If $\underset{\sim}{p}$ is discrete, its distribution is ILR (DLR) if, and only if, the function given by

$$h_{\underset{\sim}{p}}(t+1)[1-h_{\underset{\sim}{p}}(t)]/h_{\underset{\sim}{p}}(t) \tag{2.5}$$

is nonincreasing (nondecreasing). This allows a definition of log convexity (concavity) for discrete $\underset{\sim}{p}$. Since their completion rates are constant and the logarithm, respectively (2.5), of its density is linear, both the exponential and geometric distributions are simultaneously ICR, DCR, log convex and log concave. The uniform, hyperexponential, gamma, beta, Gaussian and folded-normal distributions all have either log convex or log concave densities.

Log convex and ICR processing requirement distributions correspond to practical situations (such as are found, for example, in manufacturing) in which processing tends to accelerate job completion. Log concave and DCR distributions, on the other hand, correspond to situations in which *work hardening* of jobs occurs and processing tends to delay job completion (as, for example, with some types of faulty software running on a computer system). It may be shown (*cf.* [Weiss 1982]) that a log convex (concave) processing requirement distribution is necessarily ICR (DCR), but not conversely. (For more details on these concepts see [Barlow & Proschan 1975; Karlin 1968].)

Counterexamples to the optimality of LEPT and SEPT priority policies for multimachine problems with arbitrary processing requirement distributions are easily constructed, see, *e.g.*, [Sevcik 1974; Weber 1979; Weiss 1982]. What is needed to obtain the optimality in *expectation* of these policies is that at any moment current processing requirements can be compared in terms of the stochastic ordering $<_D$ introduced above and hence in terms of expectations (which generate a corresponding order). To obtain optimality in *distribution* (and entertain the possibility of job arrivals) current processing requirements must be comparable in terms of the stronger *likelihood ratio ordering* $<_{LR}$. (If $\underset{\sim}{x}$ and $\underset{\sim}{y}$ are two random variables with densities $f_{\underset{\sim}{x}}$ and $f_{\underset{\sim}{y}}$ respectively, then $\underset{\sim}{x}$ is *likelihood ratio dominated* by $\underset{\sim}{y}$, written $\underset{\sim}{x} <_{LR} \underset{\sim}{y}$, if, and only if, $f_{\underset{\sim}{y}}/f_{\underset{\sim}{x}}$ is a nondecreasing function.) This order will again correspond to expectation (and stochastic) order (both of which it implies). It follows that given the expectation functions of remaining processing requirements $O(n^2 \log n)$ running time is

needed to implement LEPT and SEPT as preemptive priority policies.

For a processing requirement $\tilde{p}$, the processing requirement $\tilde{p}(s)$ remaining after s units of processing has distribution function $[F_{\tilde{p}}(\cdot+s)-F_{\tilde{p}}(s)]/\bar{F}_{\tilde{p}}(s)$ and completion rate $h_{\tilde{p}}(\cdot+s)$. The remaining processing requirements for ICR and DCR processing requirement distributions are always comparable in stochastic order, *i.e.* for all s,t either $\tilde{p}(s) <_D \tilde{p}(t)$ or $\tilde{p}(s) >_D \tilde{p}(t)$ (or both). A similar statement can be made with regard to the likelihood ratio order $<_{LR}$ for the remaining processing requirements generated by log convex (ILR) and log concave (DLR) distributions.

The processing requirements of a set J of jobs are *similar* if, and only if, they are given by an independent collection $\tilde{p}(s_j)$, $j \in J$, of the remaining processing requirements generated by a processing requirement $\tilde{p}$. Thus similar jobs have identical processing requirements, but may have received differing amounts of processing prior to the problem.

Weiss [Weiss 1982] has shown that when the completion rates $h_{\tilde{p}j}$, $j \in J$, are continuous, it is necessary for current processing requirement comparability as discussed above to have either similar processing requirements, or processing requirement distributions whose completion ratio $h_{\tilde{p}}$ may be ordered in the sense of uniform pointwise order. Since these conditions are easily shown to be sufficient for current processing requirement comparability, in order to obtain a best possible result (subsuming all previous ones and settling affirmatively the open problems in Table 1) a direct proof is needed which is based only on current requirement comparability in the appropriate sense and which is equally applicable *mutatis mutandis* to both continuous and discrete time. The most promising approach is through the Bellman-Hamilton-Jacobi sufficiency condition for optimal stochastic control problems along the lines of the arguments from [Weber 1981] for similar processing requirement problems in continuous time. (In fact, Weber defines current priority orderings in terms of completion rates rather than expectations, but under the assumptions necessary for current processing requirement comparability, as we have seen, the two orderings are identical.)

Notice that with similar log convex and ICR (log concave and DCR) processing requirement distributions and makespan (flowtime) criteria, preemption is necessary for LEPT (SEPT) priority policies to be optimal since - analogous to the situation for the deterministic $P|pmtn|C_{max}$ problem - the remaining processing requirements of running jobs tend to diminish (increase) and at decision epochs LEPT (SEPT) reordering of the set of unfinished jobs may be required. For such problems processor sharing - as discussed above for the deterministic problem $P|pmtn|C_{max}$ - is introduced in [Weber 1981] for remaining processing requirements equal in priority. However, processor sharing may be resolved here - as in the preemptive LPT priority algorithm for the deterministic problem - as a consequence of the fact that unfinished job reordering is only necessary at permitted decision (job arrival and completion) epochs.

Observe also that only the LEPT priority policy remains optimal for problems with random release dates. Intuitively this is so because available jobs should be processed in LEPT order to minimize makespan, regardless of job arrival events in the future, whereas SEPT order may need to be violated to minimize flowtime in order to take advantage of future job arrival events, *cf.* [Weber 1979].

Finally, it should be mentioned that in [Weber 1979] a series of counterexamples is given to show that the discrete time results for similar processing requirement distributions in Table 1 are best possible.

3. STOCHASTIC PROGRAMMING MODELS OF 2-LEVEL SCHEDULING

In this section we shall consider some alternative (multistage) dynamic stochastic programming models of 2-level planning and scheduling in continuous time. In these models, the set of machines to be acquired or assigned must be decided at the first level (stage) *before* any processing begins at $t = 0$. This decision must be made so as to minimize the sum of machine costs and the expected criterion value of an appropriate variant of the stochastic multimachine scheduling problem (Problem 2.1 treated in §2) which forms the second level (second and subsequent stages) of the problem. We consider second stage (dynamic stochastic) scheduling problems which are variants of $Q|pmtn|E\underset{\sim}{C}_{max}$ and $Q|pmtn|E\Sigma\underset{\sim}{C}_j$. Even under distributional assumptions on processing requirements which guarantee the optimality of LEPT or SEPT preemptive priority policies for these problems, we shall see that a closed form expression of second stage cost - which is required to calculate the optimal first stage decision utilizing the usual dynamic programming method of backwards recursion - is not readily available.

Let $\mathcal{M}$ denote a set of m uniform machines with ordered speeds s_i (as in §2) and costs $c_i \geq 0$, and suppose that the n jobs of the set J of jobs to be processed have independent processing requirements $\underset{\sim}{p}_j$ with means $E\underset{\sim}{p}_j$. Denote by $\underset{\sim}{p} := (\underset{\sim}{p}_1,\ldots,\underset{\sim}{p}_n)$ the *vector* of nonnegative processing requirements and by $c(M) := \Sigma_{i\in M}\, c_i$ the *cost* of employing the $|M|$ machines in $M \subset \mathcal{M}$. At the second level we shall permit (as before) preemptive scheduling policies π with decision epochs at $t = 0$ and subsequent job completions. All jobs are assumed available at $t = 0$.

Consider the following planning decision problems. Choose a subset $M \subset \mathcal{M}$ of the available machines to be applied so as to:

$$(P)\quad \min_{M\subset\mathcal{M}}\{c(M) + \inf_{\pi} E\{C_{max}(M,\pi,\underset{\sim}{p})\}\};$$

$$(P')\quad \min_{M\subset\mathcal{M}}\{c(M) + \inf_{\pi} E\{\Sigma_{j=1}^{n}\, C_j(M,\pi,\underset{\sim}{p})\}\};$$

$$(P'')\quad \min_{M\subset\mathcal{M}}\{c(M) + \inf_{\pi} E\{(C_{max}(M,\pi,\underset{\sim}{p})-T)_+\}\}.$$

Here $C_{max}(M,\pi,p)$ and $\Sigma_{j=1}^{n} C_j(M,\pi,p)$ denote respectively the makespan and flowtime of the jobs in J with processing requirements p performed on the machines in M under scheduling policy π. Without loss of generality, total machine allocation cost c(M) may be assumed to be expressed in terms of schedule delay costs in time units. In stochastic programming terminology (see, *e.g.*, [Dempster 1980]), these problems are multistage *recourse* problems and the second terms in their objective functions are the total expected costs of *recourse* to the first stage decision M through the scheduling policy π. The policy π is a *complete* recourse decision - *i.e.* its choice imposes only considerations of cost on the choice of M - and hence ordinary dynamic programming methods are applicable to the problems at hand. For the problem (P') the *(total) recourse cost* is linear, while for (P) and (P") it is piecewise linear, in job completion times. These costs are clearly monotonically decreasing in $|M| = k$ for (speed ordered) machine sets of the form $\{1,\ldots,k\}$, but depend in a complicated nonlinear manner on the scheduling policy π. The more realistic recourse cost of problem (P") is a piecewise linear function of makespan which represents *overtime cost* incurred when the schedule makespan exceeds the scheduling horizon T (such as occurs, for example, when weekend working is necessary to finish work planned for a given week).

The results from [Weiss & Pinedo 1980] may be interpreted to show that when processing times are independently exponentially distributed the optimal recourse decisions for (P) and (P') are (preemptive) list scheduling policies.

THEOREM 3.1. [Weiss & Pinedo 1980] *Let* $\underset{\sim}{p}_j \sim \lambda_j e^{-\lambda_j t}$, $j \in J$, *and assume* $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. *Then the optimal recourse policies are:*

π^o: $j \to j$ (LEPT) *for* (P);

π^o: $j \to n-j+1$ (SEPT) *for* (P').

Further, assume without loss of generality that $m' := |M| \geq n$ *(for otherwise machines of cost and speed* 0 *can be added to* M*). Then the optimal expected recourse cost* $E\underset{\sim}{C}^o_{max} = G_{\pi^o}(J,s)$ *for* (P) *can be computed from the backwards recursion over unfinished job sets* U *(with initial value* $G_{\pi^o}(\emptyset,s) := 0$*) given by*

$$G_{\pi^o}(U,s) = \frac{1_U + \Sigma_{j\in U} \lambda_j s_j G_{\pi^o}(U/\{j\},s)}{\Sigma_{j\in U} \lambda_j s_j} \tag{3.1}$$

where $s := (s_1,\ldots,s_{m'})$ *is the vector of machine speeds and* 1_U *denotes the (binary) indicator function of* U. *Similarly, for* (P'), $E\Sigma\underset{\sim}{C}^o_j = G_{\pi^o}(J,s)$, *with*

$$G_{\pi^o}(U,s) = \frac{|U| + \Sigma_{j\in U} \lambda_j s_j G_{\pi^o}(U/\{j\},s)}{\Sigma_{j\in U} \lambda_j s_j}. \quad \square \tag{3.2}$$

Thus even with independent exponential processing requirements, although the optimal recourse policies for (P) and (P') are known *explicitly*, the optimal total expected recourse costs must be determined *computationally*.

Algorithmic determination of the optimal expected recourse cost of (P) or (P') by means of (3.1) or (3.2) for a fixed machine set M is of course of exponential (time and space) complexity. Moreover, even if these expected recourse costs were known for each M, an argument similar to that from [Dempster *et al.* 1981B, Lemma 4], shows that the problem (P) of minimizing total expected costs over all $M \subset \bar{M}$ is NP-hard. The situation is even worse for the more realistic 2-level planning and scheduling problem (P") in which all jobs have a common due date T and the problem is to minimize expected tardiness of the last job. Hence we turn in the next section to consideration of approximate solutions for these problems through the use of heuristics.

4. PROBABILISTIC ANALYSIS OF HEURISTICS FOR 2-LEVEL SCHEDULING

First notice, for example, that since makespan is nonnegative, problem (P) is *value equivalent* to

$$\min_{M \subset \bar{M}} \{c(M) + E(\min_{\pi} \{C_{max}(M, \pi, \underset{\sim}{p})\})\}. \tag{4.1}$$

That is, given M, we may find the expected recourse cost corresponding to an optimal stochastic scheduling policy $\pi^o(M)$ by finding an optimal *deterministic* scheduling policy $\pi^o(M,p)$ for each realization $\underset{\sim}{p} = p$ of the processing requirement data (actually for a set of realizations of the data which occurs with probability one). However, since we are considering stochastic scheduling policies which allow preemptions only at $t = 0$ and job completions, it follows that we must solve an instance of $Q|pmtn|C_{max}$ in which preemptions are limited to moving running jobs to faster machines for each realization $\underset{\sim}{p} = p$ of the data. But as is well known [Karp 1972] even $P||C_{max}$ is NP-hard for $m \geq 2$.

Therefore let us first consider a version of (4.1) involving *identical* machines with identical assignment cost $c_i \equiv c > 0$, *viz.*

$$\begin{aligned} E\underset{\sim}{z}^o(m^o) &:= \min_{m \in \mathbb{N}} \{cm + E(\min_{\pi} \{C_{max}(m, \pi, \underset{\sim}{p})\})\} \\ &=: \min_{m \in \mathbb{N}} \{cm + EC^o_{max}(m, \underset{\sim}{p})\} \end{aligned} \tag{4.2}$$

where $EC^o_{max}(m,\underset{\sim}{p})$ denotes the expectation of the minimum makespan for $P||C_{max}$ (a random variable) for a random data instance $\underset{\sim}{p}$. We have thus reduced our multistage problem (P) to the value equivalent 2-stage recourse problem (4.2) in which at the first level the number m of identical machines to be assigned must be chosen, *before* the processing requirement data $\underset{\sim}{p} = p$ is realized at $t = 0$, after which, at the second level, an instance of $P||C_{max}$ must be

solved. At the second stage we are in the realm of *probabilistic analysis* of algorithms - or, in the terminology of stochastic programming, the *distribution problem* - for $P||C_{max}$.

Similar to the situation in §3, however, manifold difficulties are attendant on solving (4.2) for the optimal m by backwards recursion - *i.e.* by solving first the NP-hard lower level combinatorial optimization problem for each m. Moreover, this is not the natural order of decisions in practice. Therefore, consider applying an idea fundamental to planning at the higher levels of a hierarchy: namely, suppression of detailed lower level structure and its replacement by aggregates. Replace the optimal recourse cost $C^{O}_{max}(m,p)$ by its obvious lower bound $P/m := \Sigma_{j=1}^{n} p_j/m$, *cf.* (2.3), and - in order to determine the *approximately* optimal higher level decision - solve the easy *lower bounding* problem to (4.2),

$$E\tilde{z}^{LB}(m^{LB}) := \min_{m \in \mathbb{N}} \{cm + E\tilde{P}/m\}. \tag{4.3}$$

The solution m^{LB} of (4.3) minimizes $cm+E\tilde{P}/m$ subject to $m \in \mathbb{N} \cap \{\lfloor\sqrt{E\tilde{P}/c}\rfloor, \lceil\sqrt{E\tilde{P}/c}\rceil\}$.

In practice, the scheduling decisions corresponding to a foreman's dispatching task are also not explicitly optimally planned, but rather are handled *ad hoc* in real time through the use of heuristics. Consider the solution of the (deterministic) second stage problem of (4.2) using the *list scheduling* heuristic - *i.e.* place the n jobs in an *arbitrary* order and at each step assign the next job on the list to the earliest available machine.

Let $z^{LS}(m) := cm+C^{LS}_{max}(m,p)$, where for given m and p $C^{LS}_{max}(m,p)$ denotes the earliest time by which jobs are completed under this heuristic. The *2-stage heuristic* procedure defined for problem (4.2) produces a total expected cost of

$$E\tilde{z}^{LS}(m^{LB}) := cm^{LB} + EC^{LS}_{max}(m^{LB},\tilde{p}). \tag{4.4}$$

Notice that in the more realistic dynamic stochastic situation of problem (P) in which the realization $\tilde{p}_j = p_j$ becomes known only upon the completion of job $j \in J$, list scheduling may be implemented in an *on line* manner. At t = 0 a job is assigned to each of the m^{LB} machines in list (*e.g.* LEPT) order, as soon as a job is completed on a machine the next job on the list is assigned to that machine, and so on, until $C^{LS}_{max}(m^{LB},\tilde{p})$ is realized. Hence our 2-stage heuristic procedure is also applicable to the version of the original multistage recourse problem (P) value equivalent to (4.2), *viz.*

$$\min_{m \in \mathbb{N}} \{cm + \inf_{\pi} E\{C_{max}(m,\pi,\tilde{p})\}\} = E\tilde{z}^{O}(m^{O}). \tag{4.5}$$

We are thus in a position to study the stochastic performance of our 2-stage stochastic programming heuristic for this problem through the more familiar problem (4.2). (No notational distinction will be made in the sequel between the common optimal value of the two problems.)

Let us first briefly review known results on the *performance ratio* of list scheduling relative to the minimum makespan $C^{o}_{max}(m,p)$ for the deterministic problem $P||C_{max}$. An easy demonstration yields, for given m and p,

$$P/m \leq C^{o}_{max}(m,p) \leq C^{LS}_{max}(m,p) \leq P/m + p_{max}. \tag{4.6}$$

It follows that for the performance ratio we have

$$1 \leq C^{LS}_{max}(m,p)/C^{o}_{max}(m,p) \leq 1 + \frac{p_{max}}{P/m}. \tag{4.7}$$

Graham [Graham 1966, 1969] has obtained the *data independent* worst case bound

$$C^{LS}_{max}(m,p)/C^{o}_{max}(m,p) \leq 2 - \frac{1}{m} \tag{4.8}$$

and has shown that for list scheduling in LPT order this bound can be considerably improved to

$$C^{LPT}_{max}(m,p)/C^{o}_{max}(m,p) \leq \frac{4}{3} - \frac{1}{3m}. \tag{4.9}$$

The bound (4.8) is tight for maximum processing requirement ratios $p_{max}/p_{min} \geq 4$ and in [Achugbue & Chin 1981] tight bounds are given for lower values of this ratio. Observe that (4.9) is of little use in analyzing list scheduling heuristics for the multistage recourse problem (4.5) since LPT order requires full knowledge of the data realization $\underset{\sim}{p} = p$ and cannot be implemented in an on line manner when the data is realized sequentially. Moreover, we shall see below (Proposition 4.7) that data independent bounds such as (4.8) or (4.9) do not produce asymptotically tight bounds for the 2-stage heuristic as the number of jobs in the system becomes large.

Although our proper concern is the distribution of the *relative performance ratio* $\underset{\sim}{z}^{LS}(m^{LB})/\underset{\sim}{z}^{o}(m^{o})$ of the 2-stage heuristic for a random data instance $\underset{\sim}{p}$, an easy consequence of (4.6) and the definitions yields a bound on the ratio of the expected value of this heuristic relative to the expected optimum [Dempster *et al.* 1981B].

THEOREM 4.1. $1 \leq E\underset{\sim}{z}^{LS}(m^{LB})/E\underset{\sim}{z}^{o}(m^{o}) \leq 1 + E\underset{\sim}{p}_{max}/(2\sqrt{cE\underset{\sim}{P}})$. □

In order to study the asymptotic performance of the 2-stage heuristic for (4.5) as the number of jobs (and machines) in the system becomes large, the following assumption on the processing requirement data was made in [Dempster *et al.* 1981B].

Assumption A. The processing requirements $\underset{\sim}{p}_j$, $j \in J$, are independent identically distributed random variables with two moments finite: $\mu := E\underset{\sim}{p}$ and $\sigma^2 := V\underset{\sim}{p}$.

Thus, from the modelling point of view, Assumption A only allows

consideration of random variations in the processing requirements of identical jobs. Under Assumption A, asymptotic extreme value theory may be invoked to conclude that

$$E\underset{\sim}{p}_{max}/\sqrt{n} \longrightarrow 0, \tag{4.10}$$

and

$$\underset{\sim}{p}_{max}/\sqrt{n} \xrightarrow{a.s.} 0, \tag{4.11}$$

read $\underset{\sim}{p}_{max}/\sqrt{n}$ tends *almost surely* to 0 as the number n of jobs in the system tends to infinity (*i.e.* $P\{\lim_{n\to\infty} \underset{\sim}{p}_{max}/\sqrt{n} = 0\} = 1$). Analogous results to those obtained in [Dempster *et al.* 1981B] under Assumption A follow from the observation that (4.11) and (4.10) continue to hold (by a slight extension of the arguments given in [Dempster *et al.* 1981B, Appendix]) under the following weaker assumption.

Assumption A'. The processing requirements $\underset{\sim}{p}_j$, $j \in J$, are independent random variables with two moments finite: $\mu_j := E\underset{\sim}{p}_j$ and $\sigma_j^2 := V\underset{\sim}{p}_j$; $\mu := \lim_{n\to\infty} \sum_{j=1}^{n} \mu_j/n$ and $\sigma^2 := \lim_{n\to\infty} \sum_{j=1}^{n} \sigma_j^2/n$ are finite.

For fixed n, Assumption A' permits more realistic processing requirement distributions, for example, as considered in §2. The asymptotic requirements on processing requirement means and variances ensure that no large (*i.e.* infinite) set of processing requirements dominate. Put another way, Assumption A' ensures that the contributions of individual jobs to long run processing requirement statistics are negligible - exactly the preconditions for aggregation in higher level planning.

An immediate consequence of Theorem 4.1, (4.10) and the observation that under our assumption

$$E\underset{\sim}{P} \to n\mu \quad \text{as } n \to \infty, \tag{4.12}$$

is the asymptotic optimality in expectation of the 2-stage heuristic.

THEOREM 4.2. *Under Assumption* A', $\lim_{n\to\infty} E\underset{\sim}{z}^{LS}(m^{LB})/E\underset{\sim}{z}^{O}(m^{O}) = 1$. □

To obtain the analogue for the performance ratio, two lemmas will be needed. Minor modifications of the arguments in [Dempster *et al.* 1981B] to accommodate the extra passage to the limit entailed in (4.12) yield the required common asymptotic characterizations of heuristic and optimal first stage decisions and expected recourse costs.

LEMMA 4.3. *Under Assumption* A',

$$\lim_{n\to\infty} m^{LB}/\sqrt{n\mu/c} = \lim_{n\to\infty} m^{O}/\sqrt{n\mu/c} = 1. \ \square$$

LEMMA 4.4. *Under Assumption* A',

$$C^{LS}(m^{LB},\underset{\sim}{p})/(n\mu/m^{LB}) \xrightarrow{a.s.} 1,\ C^{o}_{max}(m^{o},\underset{\sim}{p})/(n\mu/m^{o}) \xrightarrow{a.s.} 1.$$

Proof. Use Kolmogorov's strong law of large numbers for *nonidentical* independent random variables (see, *e.g.* [Tucker 1967, Theorem 1, p.124]) to conclude that $(\underset{\sim}{p}-n\mu)/n \xrightarrow{a.s.} 0$ in the argument of [Dempster *et al.* 1981B, pp.8,10]. □

Combining the lemmas yields asymptotic optimality of the 2-stage heuristic (in terms of the performance ratio) with probability one, *i.e.* for almost every instance of the requirements data $\underset{\sim}{p}$.

THEOREM 4.5. *Under Assumption* A', $\underset{\sim}{z}^{LS}(m^{LB})/\underset{\sim}{z}^{o}(m^{o}) \xrightarrow{a.s.} 1$. □

Theorem 4.5 constitutes a justification for hierarchical planning procedures - as represented by our 2-stage heuristic - in the 2-level planning and scheduling situation modelled by (4.5). Indeed, it may be interpreted loosely to state that, for large parallel machine shops and many jobs with arbitrary processing requirements, aggregation and approximation at the higher level and *ad hoc* heuristics at the lower level are approximately optimal.

We note also that for the multistage problem (4.5) with a second stage stochastic scheduling problem involving *identical* machines this result comes about through a common asymptote for the optimal and heuristic value.

COROLLARY 4.6. *Under Assumption* A',

$$\underset{\sim}{z}^{LS}(m^{LB})/(2\sqrt{cn\mu}) \xrightarrow{a.s.} 1,\ \underset{\sim}{z}^{o}(m^{o})/(2\sqrt{cn\mu}) \xrightarrow{a.s.} 1.$$ □

Before considering problem (P) and its value equivalent (4.1) which involve *uniform* machine scheduling problems at the second level, it is worth noting that the above results are delicate, in that an attempt to use Graham's *data independent* bound (4.8) results in an asymptotic expectation ratio bound greater than one.

PROPOSITION 4.7. *Using* (4.8) *rather than* (4.6) *to bound* $C^{LS}_{max}(m^{LB},p)$ *yields only* $1 \le E\underset{\sim}{z}^{LS}(m^{LB})/E\underset{\sim}{z}^{o}(m^{o}) \le 3/2$ *as* $n \to \infty$.

Proof. A simple argument using (4.8) and (4.6) shows that

$$1 \le E\underset{\sim}{z}^{LS}(m^{LB})/E\underset{\sim}{z}^{o}(m^{o}) \le 1 + \left(1-\frac{1}{m^{LB}}\right)\frac{\sqrt{E\underset{\sim}{p}/c}}{2m^{LB}} + \left(2-\frac{1}{m^{LB}}\right)\frac{E\underset{\sim}{p}_{max}}{2\sqrt{cE\underset{\sim}{p}}}.$$

Lemma 4.3, (4.10) and (4.12) imply that the right hand side of the second inequality $\to 3/2$ as $n \to \infty$. □

Analysis of 2-stage heuristics for the 2-level uniform machine problem (P) is more difficult than the analysis of the identical machine special case presented above. The first problem lies with

the obvious extension of the lower bounding problem (4.3) to the general case, *viz.*

$$E\tilde{w}^{LB}(M^{LB}) := \min_{M \subset \bar{M}} \{c(M) + E\tilde{p}/s(M)\}, \tag{4.13}$$

where $s(M) := \Sigma_{i \in M} s_i$. By a reduction from the NP-complete *partition* problem it is shown in [Dempster *et al.* 1981B] that the lower bounding problem (4.13) is NP-hard! To ensure polynomial time determination of an appropriate first level decision for (P) we shall therefore employ a heuristic for the approximate solution of (4.13). To this end, reorder $\bar{M}$ in increasing order of $q_i := c_i/s_i$ and define $C_i := \Sigma_{h=1}^{i} c_h$, $S_i := \Sigma_{h=1}^{i} s_h$ and $W_i := C_i + E\tilde{p}/S_i$. The *greedy heuristic* chooses the machine set $M^G := \{1,\ldots,g\} \subset \bar{M}$ so that g is the largest index such that $W_{g-1} > W_g$.

Making the obvious definition of extreme machine costs, speeds and q ratios, it may be shown [Dempster *et al.* 1981B] that the greedy decision M^G satisfies

$$W_g = \min_{i \in \bar{M}} \{W_i\}, \tag{4.14}$$

$$E\tilde{w}^{LB}(M^G) \leq E\tilde{w}^{LB}(M^{LB}) + c_{max}, \tag{4.15}$$

and that in the present case the analogue of Theorem 4.1 for the expectation ratio follows from (4.15) by a simple argument (*cf.* (4.6)).

THEOREM 4.8. $1 \leq E\tilde{w}^{LS}(M^G)/E\tilde{w}^{O}(M^O) \leq 1 + (c_{max} + E\tilde{p}_{max}/s_{min})/(2\sqrt{q_{min} E\tilde{p}})$. □

In order to obtain the analogues of Theorems 4.2 and 4.5 in the uniform machine case, some reasonable assumptions are needed about the growth of the available machine set $\bar{M}$ as the number of jobs in the system tends to infinity.

Assumption B. The bounds $c_{min} \leq c_i \leq c_{max}$ and $s_{min} \leq s_i \leq s_{max}$ ($i \in \bar{M}$) are fixed constants. Moreover, there exist constants $D, D' > 0$ and $\varepsilon' \geq \varepsilon > 0$ such that $Dn^{\frac{1}{2}+\varepsilon} \leq |\bar{M}| \leq D'n^{\frac{1}{2}+\varepsilon'}$.

Assumption B allows an efficient implementation of the greedy heuristic (in $O(n \log n)$ time) in that the number of *available* machines remains polynomially bounded in n. We shall see that it ensures that the number of *selected* machines grows as $\sqrt{n}$, as in the identical machine case.

Theorem 4.8, (4.10) and (4.12) yield immediately the asymptotic optimality in expectation of the 2-stage heuristic for (P) defined as the first stage greedy heuristic followed by arbitrary (on line) list scheduling for the uniform machine set M^G chosen. (Recall that in the present context list scheduling is implemented so that at job completion epochs preemption may be applied to move running jobs to faster machines in list order, as described in §2).

THEOREM 4.9. *Under Assumptions* A' *and* B, $\lim_{n\to\infty} E\underset{\sim}{w}^{LS}(M^G)/E\underset{\sim}{w}^{O}(M^O) = 1$. □

Let $g(n) = \Theta(f(n))$ denote the existence of constants $C,C' > 0$ such that $Cf(n) \leq |g(n)| \leq C'f(n)$ for all n sufficiently large. Then an easy extension of the argument given in [Dempster *et al.* 1981B] (to account for the extra passage to the limit necessitated by Assumption A') yields the analogue of Lemma 4.3 for uniform machines.

LEMMA 4.10. *Under Assumptions* A' *and* B,

$$s(M^{LB}) = \Theta(\sqrt{n}),\ s(M^G) = \Theta(\sqrt{n}),\ s(M^O) = \Theta(\sqrt{n}).\ \square$$

The extension of Theorem 4.5 to the performance ratio of the 2-stage heuristic for the uniform machine case under Assumption A is due to [Stougie 1981]. He has given a direct proof which is (trivially) extended below to accommodate Assumption A'.

THEOREM 4.11. *Under Assumptions* A' *and* B, $\underset{\sim}{w}^{LS}(M^G)/\underset{\sim}{w}^{O}(M^O) \xrightarrow{a.s.} 1$.

Proof. For every realization $\underset{\sim}{p} = p$ of the data it may be shown (*cf.* [Dempster *et al.* 1981B, pp.15-16]) that

$$w^{LS}(M^G) \leq w^{O}(M^O) + |P - \Sigma_{i=1}^{n}\mu_i|\left(\frac{1}{s(M^G)} + \frac{1}{s(M^O)}\right) + c_{max} + \frac{p_{max}}{s_{min}} \quad (4.16)$$

and

$$w^{O}(M^O) \geq q_{min}s(M^O) + P/s(M^O). \quad (4.17)$$

The existence of an $\alpha > 0$ such that $w^O(M^O) \geq \alpha\sqrt{n}$, for sufficiently large n, follows from (4.17) and Lemma 4.10. Hence, combining (4.16) and (4.17) and observing that $w^O(M^O) \leq w^{LS}(M^G)$ for every realization $\underset{\sim}{p} = p$, we have that

$$1 \leq \underset{\sim}{w}^{LS}(M^G)/\underset{\sim}{w}^{O}(M^O) \leq$$
$$1 + \left|\frac{\underset{\sim}{P} - \Sigma_{i=1}^{n}\mu_i}{\alpha\sqrt{n}}\right|\left(\frac{1}{s(M^G)} + \frac{1}{s(M^O)}\right) + \frac{c_{max}}{\alpha\sqrt{n}} + \frac{\underset{\sim}{p}_{max}}{s_{min}\alpha\sqrt{n}} \quad (4.18)$$

surely. But Assumptions A' and B imply that $c_{max}/(\alpha\sqrt{n}) \to 0$ and $\underset{\sim}{p}_{max}/(s_{min}\alpha\sqrt{n}) \xrightarrow{a.s.} 0$. Moreover, Lemma 4.10 implies that there are constants α' and α'' for which, for sufficiently large n, $1/s(M^G) \leq \alpha'/\sqrt{n}$ and $1/s(M^O) \leq \alpha''/\sqrt{n}$. It follows that the second term of the right hand side of (4.18) tends almost surely to 0 with $|\underset{\sim}{P}/n - \Sigma_{i=1}^{n}\mu_i/n|$. Applying Kolmogorov's strong law of large numbers (*op. cit.*) to this expression yields the desired result. □

It is perhaps worth observing that Assumption A' could be weakened to be necessary and sufficient through the use of canonically truncated processing requirements and Kolmogorov's three series theorem (see, *e.g.*, [Tucker 1967, Theorem 4, p.113]). From

a modelling point of view however little would be gained but unnecessary mathematical complexity.

Also note that no analogue of Corollary 4.6 appears to be possible in the uniform machine case. The most precise statement about the asymptotic form of the optimal and heuristic values we can give is the existence of constants $C,C' > 0$ such that, for sufficiently large n,

$$C\sqrt{n} \leq \underset{\sim}{w}^{O}(M^{O}) \leq \underset{\sim}{w}^{LS}(M^{G}) \leq C'\sqrt{n} \tag{4.19}$$

almost surely.

Consider next the problem (P') of §3 involving a flowtime recourse cost. This problem is difficult for reasons which illuminate the intricacies of hierarchical problems. As noticed in §2, the second stage problem $Q|pmtn|\Sigma C_j$ for the value equivalent problem to (P') is easy! In the identical machine case it can be solved (nonpreemptively) in $O(n \log n)$ running time by SPT, while in the uniform machine case it can be solved in $O(n \log n + mn)$ time by an extension of SPT due to Gonzalez, which only preempts running jobs to move them to faster machines (see [Lawler *et al.* 1982]). An explicit expression for the optimal flowtime of the nonpreemptive problem $Q||\Sigma C_j$ is given [Conway *et al.* 1967, p.97] by

$$\Sigma_{j=1}^{n} C_j^{O}(M,p) = \Sigma_{i\in M} \Sigma_{k=1}^{n_i} (n_i-k+1)p_{i[k]}/s_i, \tag{4.20}$$

where n_i is the number of jobs assigned to machine $i \in M$ by the optimal schedule and $p_i[1],\ldots,p_i[n_i]$ are their processing requirements in SPT order. This expression could be used to form an upper bound on the optimal flowtime of the preemptive problem $Q|pmtn|C_{max}$. Although the asymptotic expectation of (4.20) could in principle be evaluated (under Assumption A) using the theory of order statistics, it appears to be of little use in developing a lower bounding problem for (P'). More generally (and unlike the situation for the NP-hard problem $Q||C_{max}$) bounds for the criterion value produced by suboptimal schedules for the easy problem $Q|pmtn|\Sigma C_j$ useful in the analysis of heuristics for its stochastic counterpart $Q|pmtn|E\Sigma\underset{\sim}{C}_j$ - which forms the second stage of (P') - are not readily apparent.

The more realistic problem (P") of §3 involving a schedule tardiness recourse cost is also difficult. If, in the identical machine case, one uses the obvious lower bounding problem

$$\min_{m\in\mathbb{N}} \{cm + E(\max\{\underset{\sim}{P}/m-T,0\})\} \tag{4.21}$$

to determine the first stage heuristic decision m^{LB}, then it can only be determined as a nearest integer to the solution of the integral equation

$$m^2 - \int_{mT}^{\infty} P dF_{\underset{\sim}{P}}(P)/c = 0. \tag{4.22}$$

(If $T = 0$, (4.22) yields m^{LB} as a nearest integer to $\sqrt{E\tilde{P}/c}$ as before.) Under the realistic assumption that $T = \theta(\sqrt{n})$ - which models the idea of many small jobs whose processing requirements are small relative to the schedule horizon T - asymptotic analysis of the performance ratio of the above first stage heuristic followed by list scheduling as the number of jobs in the system tends to infinity appears complicated. This is in no small measure because - unlike the above higher level heuristics based on expected values and aggregation - distributional information on the processing requirements must be taken into account at the higher level due to the nonnegativity restriction on the lower bound for the second stage cost. Unfortunately, a complete analysis of (P") is a prerequisite to the analysis of a realistic *multiperiod* planning and scheduling model in which work is allowed either to overflow from one period to the next or to be finished in overtime at a higher recourse cost.

In the presence of random nonnegative job release dates $\tilde{r}_j$ in the 2-level models (P) and (P'), list scheduling will no longer suffice and priority policies become necessary (as noted in §2). Unfortunately, the list scheduling bounds (4.6) and its uniform machine extension fundamental to our asymptotic analysis then cease to hold and a more careful analysis is required.

5. A 3-LEVEL DISTRIBUTION PLANNING MODEL

To illustrate the complexities involved, this section briefly sets out a realistic 3-level hierarchical spatial planning and scheduling model for which suitable heuristics are currently under investigation. The problem concerns the location of distribution facilities and delivery vehicles in a region in order to ultimately route the vehicles at the facilities through the customers in the region in a cost effective way. As in the 2-level stochastic machine scheduling problems treated in this paper, the random data is realized successively at each level *after* decisions are taken at the previous level.

More precisely, suppose given a random natural number $\tilde{n}$ of *customers* and a random finite sequence $\tilde{x}(\tilde{n})$ of the Cartesian coordinates $(\tilde{x}_{11},\tilde{x}_{12}),(\tilde{x}_{21},\tilde{x}_{22}),\ldots,(\tilde{x}_{\tilde{n}1},\tilde{x}_{\tilde{n}2})$ of their *locations* in a planar (simply connected) region $\tilde{\Omega}$ of area A. Assume that $\tilde{n} = n$ will be realized before $\tilde{x}(n)$ is known and consider the following 3-level hierarchical distribution planning and vehicle routing problem. At level:

1. Choose the *number* k and *locations* $y(k) := ((y_{11},y_{12}), (y_{21},y_{22}),\ldots,(y_{k1},y_{k2}))$ of identical distribution *facilities* to be placed in the region Ω before $\tilde{n}$ is realized.
2. Observe $\tilde{n} = n$. At each facility i, choose the *territory* $\Omega_i \subset \Omega$ ($\Omega_i \cap \Omega_j = \emptyset$, $i \neq j$, $\cup_{i=1}^{k} \Omega_i = \Omega$) to be served and the *number* ℓ_i of identical *vehicles* of unlimited capacity to service customers in the territory before $\tilde{x}(n)$ is realized.

3. Observe $\underset{\sim}{x}(n) = x(n)$. At each facility, *allocate* realized *customers* to vehicles and *route vehicles* so as to minimize the length of the longest vehicle tour through the allocated customer locations in Ω_i.

If C denotes the cost of a distribution facility and c denotes the cost of a vehicle (in transportation cost units), this problem may be given the following 3-stage complete recourse stochastic programming formulation:

$$\inf_{k,y(k)}\{Ck + \Sigma_{i=1}^{k} E(\inf_{\ell_i,\Omega_i}\{ck_i + E(V^{o}(\ell_i,\Omega_i;\underset{\sim}{n},\underset{\sim}{x})\,|\,\underset{\sim}{n})\})\} \quad (5.1)$$

where $V^{o}(\ell_i,\Omega_i;n,x)$ denotes the minimal longest vehicle tour length (in terms of Euclidean distance) for the ℓ_i vehicles servicing territory Ω_i, $i = 1,\ldots,k$, when the n customers in Ω have locations x.

As in the machine scheduling models, it is prudent first to attempt to analyze very simple special cases of (5.1). Even these raise some intriguing and nontrivial questions. For example, suppose Ω is the unit disk $\{(x_1,x_2) \in \mathbb{R}^2 : x_1^2+x_2^2 \leq 1\}$, $\underset{\sim}{n}$ is geometric on $[N,\infty)$ for some large N, *i.e.* $f_{\underset{\sim}{n}}(n) = p(1-p)^{i-N+1}$, $i = N,N+1,N+2,\ldots$ $(0 < p < 1)$, and $\underset{\sim}{x}(n)$ is spatial Poisson on Ω, *i.e.* the n customer locations are distributed uniformly at random in Ω. It is an obvious advantage in analysis to have all second level problems identical. But is it even approximately optimal at the first level to choose and partition Ω into pie shaped territories Ω_i of equal area with the i-th facility located at, say, the centroid of Ω_i, $i = 1,\ldots,k$? More generally, what is the effect of ignoring the partition constraints $\Omega_i \cap \Omega_j = \emptyset$, $i \neq j$, $i,j = 1,\ldots,k$ (while maintaining $\cup_{i=1} \Omega_i = \Omega$) - *cf.* U.S. national oil distribution in 1975 and 1979 - on the optimal choice of territories? Answers to these questions of course depend on the nature of the metric imposed on the higher level problems by the minimal longest vehicle tour cost measure, and results in random graph theory (see, *e.g.*, [Erdös & Spencer 1973]) can be expected to be helpful.

A single second level problem has been analyzed for a fixed circular Ω_i of area π in [Marchetti Spaccamela *et al.* 1982] building on earlier work reported in [Beardwood *et al.* 1959; Karp 1977; Steele 1980]. They observe that the length of the longest of the optimal tours of the vehicles through the customers assigned to them exceeds $1/\ell_i$ times the length of an optimal travelling salesman tour. Using a theorem from [Steele 1980], which gives an almost sure asymptote for this tour involving a constant β, they define the lower bounding problem

$$\min_{\ell_i \in \mathbb{N}}\{c\ell_i + \beta\sqrt{n\pi}/\ell_i\}$$

to yield a second stage heuristic decision $\ell^{LB} = O(n^{\frac{1}{4}})$ for sufficiently large n. Their third level multivehicle routing heuristic is based on a modification appropriate to a circular region of Karp's [Karp 1977] "divide and conquer" polynomial time approximation algorithm for the NP-hard Euclidean travelling salesman

problem posed in a rectangle. (Such approximation algorithms - unfortunately sometimes termed *probabilistic* in the literature - have the property of arbitrary ε-optimality for sufficiently large finite n with a probability which has a precisely known lower bound tending to 1 with n tending to ∞ and hence are almost surely asymptotically optimal.) Marchetti Spaccamela *et al.* demonstrate that the expectation ratio of this 2-stage heuristic relative to the optimal value approaches 1 and that the heuristic is optimal in performance ratio almost surely for random data instances as the number n of customers in the system tends to infinity. They also obtain similar results for the case of random $\underset{\sim}{n}$ and the realistic third level *repetitive* vehicle routing situation in which customers in given locations require a (Bernoulli) random delivery with probability p.

The first level problem defined by (5.1) is essentially a *non*-Euclidean *planar* k-*median problem*. Thus there is some hope in extending the analysis of the *Euclidean* k-median problem given in [Fisher & Hochbaum 1980] to the metric defined on the n customers by the sum of the second and third stage costs of (5.1). These authors give an asymptotic probabilistic analysis of a polynomial time approximation algorithm for the NP-hard [Papadimitriou 1980] Euclidean problem in a planar region of area A (including the almost sure asymptote $n\sqrt{A/k}$ for the sum of the minimal Euclidean distances to each point from the k centres) whose extension would provide a suitable first stage heuristic for (5.1). We are currently working in this direction.

Although it retains the essence of practical hierarchical planning in the distribution field, the simplified model set out here could be usefully extended in many directions to improve its realism. It is however already sufficiently difficult and, for example, addition of vehicle capacity constraints (as in deterministic models) would complicate matters even more.

6. CONCLUSIONS

Open problems and directions for further research have been indicated throughout this paper. Rather than collect them here, some remarks on the nature of stochastic models for hierarchical planning and scheduling decisions seem more appropriate.

First, it is worth observing that many of the parallel machine scheduling problems of §2 provide instances of NP-hard deterministic problems for which simple suboptimal heuristics (*e.g.* LEPT) become optimal when the problem data is (more realistically) taken to be suitably random. The implication - a central thesis of this paper - is that in a practical situation suboptimality of relatively simple heuristics can be the erroneous conclusion of the wrong model, which has been taken to be deterministic for analytic convenience rather than stochastic for realism.

More generally, multistage recourse stochastic programming

models appear to provide a realistic representation of hierarchical planning and scheduling decision problems in several fields of application. Heuristics for such problems are necessitated by their analytic and computational complexity and the sequential availability of data and can be made to mirror the top down sequential nature of actual hierarchical decision making based on averaging and aggregation until more refined data becomes available. Analyses which demonstrate the asymptotic optimality of these heuristics with the growth of random instances of the problem data tend to reinforce the long held views of practical persons faced with difficult decisions - in sufficiently complex environments *suitable* rules of thumb can be highly efficient.

Finally, the project described in this paper can be seen as part of a current general trend in mathematical sciences. Driven by the exigencies of numerical computation, approximation methods are moving from applications to functions, equations and other relatively simple deterministic structures to the approximation of more and more complex stochastic problems.

ACKNOWLEDGEMENTS

It is a pleasure to acknowledge the support of IIASA, where much of the preparation of this paper was completed. The wider work of my colleagues (named in §1) and myself on this topic has been partially supported by IIASA, by NSF Grant ENG-7826500 to the University of Pennsylvania, by NATO Special Research Grant 9.2.02 (SRG.7) and by NATO Research Grant 1575. I am indebted to my colleagues for stimulating discussions and for providing me with manuscripts of work in progress which permitted this attempt at a comprehensive survey of our collective effort. In particular, I would like to thank Jan Karel Lenstra and Leen Stougie for their helpful comments on the manuscript.

Many thanks are also due to Marjolein Roquas for very rapidly producing the typescript to her usual exacting standards.

REFERENCES

J.O. ACHUGBUE, F.Y. CHIN (1981) Bounds on schedules for independent tasks with similar execution times. *J. Assoc. Comput. Mach.* 28,81-99.

K.R. BAKER (1974) *Introduction to Sequencing and Scheduling*, Wiley, New York.

R.E. BARLOW, F. PROSCHAN (1975) *Statistical Theory of Reliability and Life Testing: Probability Models*, Holt, Rinehart and Winston, New York.

J. BEARDWOOD, H.J. HALTON, J.M. HAMMERSLEY (1959) The shortest path through many points. *Proc. Cambridge Phil. Soc.* 55, 299-327.

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

M.A.H. DEMPSTER (ed.) (1980) *Stochastic Programming*, Academic, London.

M.A.H. DEMPSTER, M.L. FISHER, L. JANSEN, B.J. LAGEWEG, J.K. LENSTRA, A.H.G. RINNOOY KAN (1981A) Analytical evaluation of hierarchical planning systems. *Oper. Res.* 29,707-716.

M.A.H. DEMPSTER, M.L. FISHER, L. JANSEN, B.J. LAGEWEG, J.K. LENSTRA, A.H.G. RINNOOY KAN (1981B) Analysis of heuristics for stochastic programming: results for hierarchical scheduling problems. Report BW 142, Mathematisch Centrum, Amsterdam.

M.A.H. DEMPSTER, C.H. WHITTINGTON (1976) *Computer Scheduling of REME Training*, FS 3/01 Final Report I & II, Council for Educ. Tech., London.

Y.M.I. DIRICKX, L.P. JENNERGREN (1979) *Systems Analysis by Multilevel Methods: With Applications to Economics and Management*, International Series on Applied Systems Analysis 6, Wiley, Chichester.

P. ERDÖS, J. SPENCER (1973) *Probabilistic Methods in Combinatorics*, Academic, New York.

M.L. FISHER, D.S. HOCHBAUM (1980) Probabilistic analysis of the planar K-median problem. *Math. Oper. Res.* 5,27-34.

F. GIANNESSI, B. NICOLETTI (1979) The crew scheduling problem: a travelling salesman approach. In: N. CHRISTOFIDES *et al.* (eds.) (1979) *Combinatorial Optimization*, Wiley, Chichester, 389-408.

J.C. GITTINS (1979) Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. Ser. B* 41,148-177.

J.C. GITTINS (1981) Multiserver scheduling of jobs with increasing completion rates. *J. Appl. Probab.* 18,321-324.

R.L. GRAHAM (1966) Bounds for certain multiprocessing anomalies. *Bell System Tech. J.* 45,1563-1581.

R.L. GRAHAM (1969) Bounds on multiprocessing timing anomalies. *SIAM J. Appl. Math.* 17,263-269.

E.P.C. KAO, M. QUEYRANNE (1981) Aggregation in a two-stage stochastic program for manpower planning in the service sector. Research Report, Department of Quantitative Management Science, University of Houston.

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

R.M. KARP (1972) Reducibility among combinatorial problems. In: R.E. MILLER, J.W. THATCHER (eds.) (1972) *Complexity of Computer Computations*, Plenum, New York, 85-103.

R.M. KARP (1977) Probabilistic analysis of partitioning algorithms for the traveling-salesman problem in the plane. *Math. Oper. Res.* 2,209-224.

L. KLEINROCK (1976) *Queueing Systems, Vol. II: Computer Applications*, Wiley, New York.

E.L. LAWLER, J.K. LENSTRA, A.H.G. RINNOOY KAN (1982) Recent developments in deterministic sequencing and scheduling: a survey. This volume.

A. MARCHETTI SPACCAMELA, A.H.G. RINNOOY KAN, L. STOUGIE (1982) Hierarchical vehicle routing. To appear.

P. NASH (1980) A generalized bandit problem. *J. Roy. Statist. Soc. Ser. B* 42,165-169.

C.H. PAPADIMITRIOU (1980) Worst-case analysis of a geometric location problem. Technical Report, Laboratory for Computing Science, Massachusetts Institute of Technology, Cambridge, MA.

M.L. PINEDO (1982) On the computational complexity of stochastic scheduling problems. This volume, 355.

M.L. PINEDO, L. SCHRAGE (1982) Stochastic shop scheduling: a survey. This volume, 181.

E.L. PRESMAN, I.M. SONIN (1979) On the asymptotic value function of the many-armed bandit problem. In: V.I. ARKIN, H.YA. PETRAKOV (eds.) (1979) *Theoretical Probabilistic Methods for Problems of Economic Process Control*, Central Economic Mathematical Institute, USSR Academy of Science, Moscow. (In Russian.)

S.M. ROSS (1970) *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco.

K.C. SEVCIK (1974) Scheduling for minimum total loss using service time distributions. *J. Assoc. Comput. Mach.* 21,66-75.

J.M. STEELE (1980) Subadditive Euclidean functionals and non-linear growth in geometric probability. Research Report, Department of Statistics, Stanford University.

L. STOUGIE (1981) Private communication.

H.G. TUCKER (1967) *A Graduate Course in Probability*, Academic, New York.

R.R. WEBER (1979) Optimal organization of multi-server systems. Ph.D. Thesis, University of Cambridge.

R.R. WEBER (1981) Scheduling jobs with stochastic processing requirements on parallel machines to minimize makespan or flowtime. *J. Appl. Probab.*, to appear.

G. WEISS (1982) Multiserver stochastic scheduling. This volume.

G. WEISS, M.L. PINEDO (1980) Scheduling tasks with exponential service times on non-identical processors to minimize various cost functions. *J. Appl. Probab.* 17,187-202.

P. WHITTLE (1980) Multi-armed bandits and the Gittins index. *J. Roy. Statist. Soc. Ser. B* 42,143-149.