# Recent Asymptotic Results in the Probabilistic Analysis of Schedule Makespans

*E. G. Coffman, Jr.* and *Ward Whitt*
AT&T Bell Laboratories
Murray Hill, New Jersey 07974-0636

## ABSTRACT

Makespan scheduling problems are in the mainstream of operations research, industrial engineering, and computer science. A basic multiprocessor version requires that $n$ tasks be scheduled on $m$ identical processors so as to minimize the makespan, i.e., the latest task finishing time. In the standard probability model considered here, the task durations are i.i.d. random variables with a distribution $F$, and the objective is to estimate the distribution of the makespan as a function of $m$, $n$, and $F$. This paper surveys probabilistic results for the multiprocessor scheduling problem and an important variant known as the permutation flow-shop problem. Several of the results are new; the others have appeared in the last few years.

Because of the difficulty of exact analysis, the results take the form of limits as $n \to \infty$ or as both $m \to \infty$ and $n \to \infty$ with $m < n$. Some highlights of the survey are: a new asymptotic analysis of the on-line greedy scheduling policy, the resolution of a longstanding open problem in the analysis of off-line policies, new applications of central limit theorems to makespan scheduling, and limit theorems giving the asymptotic behavior under the greedy and optimal policies for the flow-shop problem. Open problems and modeling issues are also discussed.

October 12, 1993

# Recent Asymptotic Results in the Probabilistic Analysis of Schedule Makespans

*E. G. Coffman, Jr.* and *Ward Whitt*
AT&T Bell Laboratories
Murray Hill, New Jersey 07974-0636

## 1. Introduction

An integer $m \geq 2$ together with positive task running times $T_1, \ldots, T_n$ defines an instance of the *multiprocessor scheduling problem*: Schedule $T_1, \ldots, T_n$ on $m$ identical processors $P_1, \ldots, P_m$ so as to minimize the latest task finishing time or *makespan*; i.e., partition the set $\{T_1, \ldots, T_n\}$ into subsets $P_1, \ldots, P_m$ so as to minimize the maximum subset sum

$$L_{m,n} \equiv \max_{1 \leq j \leq m} \sum_{\{i: T_i \in P_j\}} T_i \ .$$

To avoid trivialities, we assume that $n \geq m$ unless stated otherwise. The problem finds application in operations research as a model of scheduling parallel machines in industrial job shops. It has also had a prominent role in computer science, where the term multiprocessor originates. Along with a number of other fundamental NP-complete problems, it has served as a theoretical testbed for the development of new ideas in the design and analysis of algorithms (see Garey and Johnson (1979)).

Because of the problem's complexity, several heuristic policies have been studied. Our interest here is in simple, but effective techniques (as illustrated in the next paragraph) rather than elaborate heuristic search techniques. The combinatorial worst-case analysis of such policies dates back over 25 years (see Graham (1966), and for a general treatment, Blazewicz, et al. (1993)). More recently, the competitive analysis of on-line algorithms has been applied to the problem (see Phillips and Westbrook (1993) for recent results and references to others). In this setting, the problem has also been called *load balancing*, a term that suggests broader applications. For example, in computer storage allocation, it may be necessary to distribute $n$ files among $m$ identical storage units so as to minimize the maximum of the total file sizes.

For the purposes of defining multiprocessor scheduling policies, it is convenient to assume that the tasks are presented in the form of a list $(T_1, \ldots, T_n)$. The on-line *greedy* policy is arguably the simplest (and fastest) heuristic for finding approximate solutions to the multiprocessor scheduling problem. This policy uses no advance information on the number or

durations of tasks. The policy begins by assigning the first $m$ tasks $T_1, \ldots, T_m$ to the $m$ processors $P_1, \ldots, P_m$; the processors start running these tasks at time 0, while the remaining tasks wait. Thereafter, whenever a processor finishes its current task, the next waiting task, if any, is assigned to the idle processor. In queueing terminology the system operates as an $m$-server queue with a first-come first-served service discipline; $n$ customers arrive to an empty system at time 0, and the latest of their departure times is the makespan. The rule for resolving ties among processors is immaterial, so we leave it unspecified. The off-line greedy policy operates just as the on-line version, except that the list $(T_1, \ldots, T_n)$ is first sorted into decreasing order. The off-line version is also called the *largest-processing time* (LPT) policy, a term we use hereafter; the term greedy by itself refers to the on-line policy.

Understandably, the greedy and LPT policies were among the first policies studied when the probabilistic analysis of scheduling algorithms began some 15 years ago. In the standard probability model considered here, the task durations $T_i$ are independent and identically distributed (i.i.d.) with distribution $F(t) = P(T_i \leq t)$. The problem is to find the distribution of the makespan $L_{m,n}$ as a function of the number $m$ of processors, the number $n$ of tasks and the distribution $F$. The general aim is to bring out typical behavior rather than the worst-case behavior, which can be highly unlikely. With explicit formulas in mind, probabilistic analysis is usually quite difficult, so research has often turned to large-$n$ asymptotics.

This paper surveys new probabilistic results, concentrating on those of the past few years; earlier research is covered in Coffman and Lueker (1991). We do not claim that our survey is exhaustive; rather, our goal is to illustrate current directions, mathematical approaches, and open problems in a field that is quite active.

Section 2 covers the greedy policy, presenting new results of the authors in collaboration with L. Flatto, A. Weiss, and P. E. Wright. The analysis here is self-contained, but Coffman et al. (1993) study theoretical questions in more depth. Section 3 discusses off-line policies, concentrating on the differencing methods of Karmarkar and Karp (1982). Yakir (1993) recently solved an intriguing open problem set by Karmarkar and Karp's original analysis. The principal new insight in Yakir's approach is described.

Central limit theorems are natural tools for asymptotic makespan analysis. Section 4 applies these tools in a policy-free set-up, i.e., limit theorems are proved which hold simultaneously for all scheduling policies. These results are new.

Research on the permutation flow-shop problem, a fundamental variant of makespan schedul-

ing, is surveyed in Section 5. In this problem, each task consists of $m$ *operations*, one for each processor, i.e., $T_i \equiv (T_{i1}, \ldots, T_{im})$, $1 \leq i \leq n$. A permutation of the task indices $(1, 2, \ldots, n)$ defines a schedule because the operations of each task must be performed on processors in the sequence $P_1, \ldots, P_m$, and because the $n$ operations must be performed in the same task order on every processor. Now the $nm$ operation times $T_{ij}$ are regarded as i.i.d. with distribution $F$. Thus, under the greedy policy the queueing system with analogous dynamics is a network with $m$ single-server queues in tandem, the first-come first-served service discipline and $n$ customers initially at the first queue ready to begin service. In this case all service times are i.i.d. Asymptotic behavior is described for the greedy policy when either or both of $m$ and $n$ are large, and for an optimal policy when $m = 2$ and $n$ is large. The results for the greedy policy are contained in Glynn and Whitt (1991), Greenberg, Schlunk and Whitt (1993) and Srinivasan (1993); the results for the optimal policy when $m = 2$ are due to Ramudhin, Bartholdi, Calvin, Vande Vate and Weiss (1993). Section 6 caps off the paper with a discussion of open problems and modeling issues.

This section concludes with matters of convention. Probabilistic results have varied widely in the classes of distributions $F$ allowed. *A convenient, common subset consists of those distributions supported on a finite interval, with a positive continuous density $f$. In what follows, $F$ has these properties, unless stated otherwise.* The mean and variance of $F$ are denoted by $\tau$ and $\sigma^2$. Our uniform treatment simplifies the presentation of the basic ideas; but while such distributions are adequate as models of most practical situations, many of the results hold for broader classes of distributions. Details on these technical matters can be found in the references.

Because of the form of the results in Sections 2–4, policies will also be assessed in terms of the *error* $\alpha_{m,n} = L_{m,n} - n\tau/m$, where $n\tau/m$ is an obvious lower bound on $E[L_{m,n}]$. An analogous normalization is introduced in Section 5 for the flow-shop problem. The notation $L_{m,n}$, $\alpha_{m,n}$ will be used generically; in any given instance, the problem and policy being considered will be clear in context.

## 2. The On-Line Greedy Policy

In the past decade, several papers have been devoted to an asymptotic analysis of the greedy policy for general $m$; see Boxma (1985), Bruno and Downey (1986), Coffman and Gilbert (1985), Han, Hong, and Leung (1992) and Loulou (1984). None of this work has led to

limiting behavior as precise as that found for $m = 2$ in an early result of Feller (1971), p. 208; for $m \geq 3$, the analysis has resorted to various bounding techniques. However, Feller's result can in fact be generalized, as shown below. (See Coffman et al. (1993) for extensions.)

Fix $m \geq 2$, and for convenience, extend the greedy process to the infinite time horizon; i.e., construct a greedy schedule from the infinite sequence $T_1, T_2, \ldots$. Let $C_n$ denote the $n^{\text{th}}$ completion time; let $R_i^n$, $1 \leq i \leq m-1$, denote the residual times of those tasks still running at time $C_n$, ordered by increasing processor index; and let $R_{(1)}^n \leq \cdots \leq R_{(m-1)}^n$ denote the order statistics of the $R_i^n$. Without loss of generality, assume that $C_n \neq C_{n+1}$ for all $n \geq 1$. At time $C_{n-m+1}$, tasks $T_1, \ldots, T_n$ have all started, $n-m+1$ of them have finished, and $m-1$ are still running with residual times $R_i^{n-m+1}$, $1 \leq i \leq m-1$. Then the tasks $T_1, \ldots, T_n$ have a latest finishing time

$$L_{m,n} = C_{n-m+1} + R_{(m-1)}^{n-m+1} \ , \tag{2.1}$$

and a sum of running times that can be expressed as

$$\sum_{i=1}^{n} T_i = mC_{n-m+1} + \sum_{i=1}^{m-1} R_{(i)}^{n-m+1} = mC_{n-m+1} + \sum_{i=1}^{m-1} R_i^{n-m+1} \ . \tag{2.2}$$

Combining (2.1) and (2.2) gives

$$L_{m,n} = \frac{1}{m}\left[\sum_{i=1}^{n} T_i + mR_{(m-1)}^{n-m+1} - \sum_{i=1}^{m-1} R_i^{n-m+1}\right] \ . \tag{2.3}$$

To proceed, we need information on the random variables $R_{(i)}^n$. A direct approach analyzes the Markov chain $\{(R_{(1)}^n, \ldots, R_{(m-1)}^n),\ n \geq 1\}$. Let $p_n$ denote the density at epoch $C_n$ of the Markov chain, and define

$$\begin{aligned} \mathbf{t}_0(y) &= (t_1 + y, \ldots, t_{m-1} + y) \ , \\ \mathbf{t}_i(y) &= (y, t_1 + y, \ldots, t_{i-1} + y, t_{i+1} + y, \ldots, t_{m-1} + y), \quad 1 \leq i \leq m-1 \ , \end{aligned}$$

with $\mathbf{t} \equiv \mathbf{t}_0(0)$. Let $t_0 = 0$. A straightforward analysis then shows that, with $p_0$ given,

$$p_{n+1}(\mathbf{t}) = \sum_{i=0}^{m-1} \int_0^\infty p_n(\mathbf{t}_i(y)) f(t_i + y) dy, \quad n \geq 0 \ . \tag{2.4}$$

An application of the theory of Harris-recurrent Markov chains (see Asmussen (1987), pp. 150–158) proves convergence of the distributions to a proper limit independent of the starting state. In terms of random variables, we can write

$$(R_{(1)}^n, \ldots, R_{(m-1)}^n) \Rightarrow (R_{(1)}^*, \ldots, R_{(m-1)}^*) \quad \text{as} \quad n \to \infty \ , \tag{2.5}$$

where $\Rightarrow$ denotes convergence in distribution. Indeed, there is convergence of the probability measures in total variation as $n \to \infty$.

To obtain an explicit formula for the limiting distribution, i.e., the distribution of $(R^*_{(1)}, \ldots, R^*_{(m-1)})$ in (2.5), the Markov chain analysis now requires that we solve the stationary version of the rather awkward recurrence in (2.4). A key observation allows us to side-step this difficulty by applying the theory of stationary point processes, in an argument that makes no direct use of properties already established by the Markov chain approach. The observation is that the sequence $\{C_n\}$ generated by the greedy rule is equal stochastically to the superposition of $m$ i.i.d. ordinary renewal processes; i.e., each is defined independently by $F$ and each starts with a point at 0. The time-stationary version of each renewal process is the familiar equilibrium renewal process, in which the distance to the first point has the equilibrium residual-life distribution $G$ with density $g(t) = [1 - F(t)]/\tau$. Each original renewal process is the Palm (or synchronous) version of its time-stationary version.

Now consider the superposition of $m$ i.i.d. copies of this time-stationary renewal process. This is a time-stationary point process with the distance to the first point from each component stream having distribution $G$. Since we want to look at the superposition process at completion times, we are interested in the Palm (synchronous) version of this stationary point process. Section 5.1 of Baccelli and Brémaud (1987) characterizes this Palm version in terms of the Palm and stationary versions of the component processes. However, from this superposition process alone we cannot extract the stationary distribution of $(R^n_{(1)}, \ldots, R^n_{(m-1)})$ directly. To do this, we mark the points of each component stream with the index of the processor on which it occurs, and then apply the corresponding superposition result for stationary marked point processes in Section 1.3.5 of Franken et al (1982). This result shows that each of the $m-1$ streams is equally likely to produce the current point (i.e., the $m$ possible marks of the current point are equally likely), and that the residual time to the next point in each of the remaining $m-1$ streams has the distribution $G$. It follows that the stationary version of $(R^n_{(1)}, \ldots, R^n_{(m-1)})$ at completion times coincides with the order statistics of $m-1$ i.i.d. random variables with distribution $G$. This implies that a stationary solution to (2.4) is

$$p(\mathbf{t}) = (m-1)! \prod_{i=1}^{m-1} g(t_i) \ , \tag{2.6}$$

which a substitution into (2.4) will verify.

As a consequence, we have the limit

$$L_{m,n} - \frac{1}{m}\sum_{i=1}^{n} T_i \Rightarrow \alpha_m \equiv R^*_{(m-1)} - \frac{1}{m}\sum_{i=1}^{m-1} R^*_i \quad \text{as} \quad n \to \infty \ , \tag{2.7}$$

where $R^*_1, \ldots, R^*_{m-1}$ are i.i.d. random variables with distribution $G$.

We now consider expected values. Note that the residual-life distribution $G$ has mean $(\sigma^2 + \tau^2)/2\tau = \tau(\nu^2 + 1)/2$, where $\nu \equiv \sigma/\tau$ is the coefficient of variation of $F$ (recall that $\tau$ and $\sigma$ are the mean and standard deviation of $F$). Uniform integrability follows from our assumptions on $F$, so from (2.7) we obtain

$$E[\alpha_{m,n}] \to E[\alpha_m] = E[R^*_{(m-1)}] - \frac{1}{m}\sum_{i=1}^{m-1} E[R^*_i] \quad \text{as} \quad n \to \infty \tag{2.8}$$

or, equivalently,

$$E[\alpha_{m,n}] = \int_0^\infty [1 - G^{m-1}(x)]dx - \left(\frac{m-1}{m}\right)\tau\left(\frac{\nu^2+1}{2}\right) + o(1) \quad \text{as} \quad n \to \infty \ . \tag{2.9}$$

An important special case is the uniform distribution, $F(t) = t$, with $g(t) = 2(1-t)$, $0 \le t \le 1$. From (2.9), we obtain

$$E[\alpha_m] = \int_0^1 [1 - (2t - t^2)^{m-1}]dt - \frac{(m-1)}{3m} \ . \tag{2.10}$$

A direct calculation gives

$$E[\alpha_{m,n}] = \frac{2m+1}{3m} - \frac{\Gamma\left(\frac{3}{2}\right)\Gamma(m)}{\Gamma\left(m+\frac{1}{2}\right)} + o(1) \quad \text{as} \quad n \to \infty \ . \tag{2.11}$$

Coffman et al. (1993) also consider rates of convergence. For our distributions $F$, it is easily verified that, from any point of the (compact) state space to any other such point, the $r$-step transition density is strictly positive for at least one $r \le 2(m-1)$. Then Doeblin's condition holds and convergence to the stationary distribution is geometrically fast (see, for example, Meyn and Tweedie (1993), Section 16.2). Similarly, it can be shown that the $o(1)$ term in (2.9) can be sharpened to $O(\rho^n)$ for some $\rho, 0 < \rho < 1$.

A simplified analysis applies to the exponential distribution $F(t) = 1 - e^{-t/\tau}$, $t \ge 0$, which falls outside our standard class of distributions. In this case the $C_n$, $n \ge m$, are the epochs of a Poisson process at rate $m/\tau$, so that for all $i$ and $n \ge m$, the $R^n_i$ and thus $R^*_i$ are $m-1$ i.i.d. random variables with the distribution $G = F$. Then (2.3) gives

$$E[\alpha_{m,n}] = \tau[H_m - 1] \ \text{ for all } m \text{ and } n \ge m \ , \tag{2.12}$$

where $H_m = \sum_{j=1}^{m} 1/j$ (see also Coffman and Gilbert (1985)).

The exponential case with $m = 2$ was examined by Coffman and Wright (1992) under more general assumptions, namely, an initial delay $x$ (release time) on one of the processors and a random number $N$ of tasks having either a geometric or a Poisson distribution with mean $n$. Explicit, though complicated, expressions for the moments $E[L_{2,n}^k(x)]$ were studied by computing a variety of asymptotics as $n \to \infty$ and $x \to \infty$ at different rates.

Because of the increased use of massively parallel computers, it is natural to consider asymptotics as $m \to \infty$. From expressions like (2.11) and (2.12), large-$m$ asymptotics for the mean of the time-stationary random variable $\alpha_m$ can be obtained directly. For example, (2.11) and asymptotics for the gamma function give

$$E[\alpha_m] = \frac{2}{3} - \frac{\sqrt{\pi}}{2\sqrt{m}} + O\left(\frac{1}{m}\right) \quad \text{as} \quad m \to \infty \tag{2.13}$$

when $F$ is the uniform distribution on $[0, 1]$; similarly, when $F$ is the exponential distribution, (2.12) and asymptotics for $H_m$ give

$$E[\alpha_m] = \tau(\ln m - 1 - \gamma) + o(1) \quad \text{as} \quad m \to \infty \ , \tag{2.14}$$

where $\gamma$ is Euler's constant $(0.5772\ldots)$.

More generally, we can obtain asymptotic properties of $E[\alpha_m]$ from (2.9). When $F$ has support $[0, b]$, (2.7) and the strong law of large numbers implies that

$$\alpha_m \to \alpha \equiv b - \tau\frac{(\nu^2 + 1)}{2} \quad \text{w.p.1} \quad \text{as} \quad m \to \infty \ . \tag{2.15}$$

From (2.15) we can see how $F$ influences the asymptotic error $\alpha$. For a given bound $b$, $\alpha$ *decreases* in $\tau$ and $\nu^2$. For given $b$ and $\tau$, the lowest value of $\alpha$ is $b/2$, which is approached by the two-point distribution with mass $\tau/b$ on $b$ and mass $(b - \tau)/b$ on 0; e.g., see p. 120 of Whitt (1984).

It is interesting that for this extremal two-point distribution the greedy policy is optimal for all $m$ and $n$; i.e., *there is a distribution with finite positive variance for which greedy gives the minimum expected error.* The optimality of the greedy policy in this case is easy to see because the makespan is the same as for a random number of tasks, each with a constant running time $b$. In this case, all work-conserving policies (in which no processor is idle when there is a task that has not started) are optimal. This two-point distribution is not in our class of distributions, but it is approached by such distributions.

The term $R^*_{(m-1)}$ in $\alpha_m$ obviously becomes even more important if $F$ does *not* have finite support. The asymptotic behavior of $R^*_{m-1}$ is described by the classical extreme-value theory; see Leadbetter, Lindgren and Rootzén (1983) and Reiss (1989). This extreme value theory applies to the iterated limit as first $n \to \infty$ and then $m \to \infty$ provided (2.9) is still valid.

Since the superposition of $m$ i.i.d. renewal processes, appropriately scaled, converges to a Poisson process as $m \to \infty$, e.g., see Çinlar (1972), one might expect that the general formula for $E[\alpha_m]$ in (2.8) and (2.9) would in some sense approach the formulas for the exponential distribution in (2.12) and (2.14), but this is *not* the case. For the question here, the superposition limit theorem does not apply. The superposition limit theorem implies that the distribution of $R^*_{(1)} \equiv R^{*m}_{(1)}$ is asymptotically exponential as $m$ gets large, but in (2.8) we focus on $R^*_{(m-1)}$ and $\sum_{i=1}^{m-1} R^*_{(i)}$.

An interesting open problem is the joint limiting behavior as $m \to \infty$ and $n \to \infty$. Above, we considered only the iterated limit in which first $n \to \infty$ and then $m \to \infty$. If $m = n$, then the extreme-value theory for i.i.d. random variables with distribution $F$ describes the makespan. It would be interesting to develop different asymptotics in intermediate cases.

## 3. Off-Line Policies

The off-line component of the LPT policy introduced in Section 1 is simply an initial sorting of the list $(T_1, \ldots, T_n)$ into decreasing order. Results for LPT as precise as those for the greedy policy in Section 2 have not been obtained. On the other hand, asymptotic behavior for fixed $m$ is rather well understood. For comparison with Section 2, we illustrate the main results by the following, taken from Rinnooy Kan and Frenk (1986) and Frenk and Rinnooy Kan (1987). For any fixed $m$, $\alpha_{m,n} \Rightarrow 0$ a.s., $n \to \infty$, so long as $F$ is strictly positive in a neighborhood of the origin. Moreover, if $F(t) = t^a$, $0 \le t \le 1$, for some $0 < a < \infty$, then the convergence rate is $O(\log\log n/n)^{1/a}$, and the moments are bounded by $E[\alpha^k_{m,n}] = O(n^{-k/a})$. Results in a similar vein were presented by Boxma (1985), Coffman, Frederickson, and Lueker (1984), and Coffman, Flatto, and Lueker (1984).

Consider next the superior differencing methods of Karmarkar and Karp (1982), with the restriction to $m = 2$, for simplicity. We describe the largest-first differencing method (LDM); it is a particularly simple differencing method, and no other such method is known to have a better asymptotic performance. Other methods that have been successfully analyzed are either much more elaborate (Karmarkar and Karp (1982)) or have worse performance (Lueker

(1987)); see also Coffman and Lueker (1991) for a discussion of these methods.

LDM starts by computing the absolute difference $d$ between the largest two tasks in the current list; it then replaces these two tasks by a single task of duration $d$, leaving a list with one fewer task. LDM iterates this procedure $n-2$ more times until a single task remains; the duration of this task is $2\alpha_{2,n} = 2L_{2,n} - S_n$. Noting that the largest two tasks being differenced at each step are to be put on different processors, it is a simple exercise to work backward through the differencing sequence to determine a partition of $\{T_1, \ldots, T_n\}$ that gives the final difference $2\alpha_{2,n}$. An intriguing problem set by Karmarkar and Karp over 10 years ago was a proof that, under LDM with $F$ the uniform distribution on $[0,1]$,

$$E[\alpha_{2,n}] = O(n^{-c\log n}) \tag{3.1}$$

for some constant $c > 0$. The recent work of Yakir (1993) provides an elegant solution to this problem. We give below Yakir's important new insight into the structure of LDM.

The new insight is based on Lueker's (1987) initial transformation of the problem. This transformation uses the well-known fact that, if $S_j = \sum_{i=1}^{j} X_i$, $1 \le j \le n+1$, are the partial sums of $n+1$ i.i.d. exponentials $X_i$ with parameter 1, then the ratios $\frac{S_1}{S_{n+1}}, \ldots, \frac{S_n}{S_{n+1}}$ are independent of $S_{n+1}$ and equal in distribution to $T_{(1)}, \ldots, T_{(n)}$, i.e., to the order statistics of $n$ independent samples from the uniform distribution on $[0,1]$. Let $\alpha_{2,n}, \widehat{\alpha}_{2,n}$ denote the errors produced by LDM from the respective lists $(T_1, \ldots, T_n)$, $(S_1, \ldots, S_n)$, and let $\stackrel{\mathrm{d}}{=}$ denote equality in distribution. Then one obtains $\widehat{\alpha}_{2,n} \stackrel{\mathrm{d}}{=} S_{n+1}\alpha_{2,n}$, with $S_{n+1}$ and $\alpha_{2,n}$ independent, so $E[S_{n+1}] = n+1$ gives $E[\alpha_{2,n}] = \frac{1}{n+1}E[\widehat{\alpha}_{2,n}]$. Thus, to prove (3.1) it is sufficient to prove that $E[\widehat{\alpha}_{2,n}] = O(n^{-c\log n})$ for some $c > 0$.

Let $X_1^r, \ldots, X_r^r$ denote the spacings between the tasks $S_j^r = \sum_{i=1}^{j} X_i^r$, $1 \le j \le r$, just after the $(n-r)^{\text{th}}$ iteration of LDM, $1 \le r \le n-1$; the initial spacings $X_i^n = X_i$ are i.i.d. parameter-1 exponentials. The $(n-r)^{\text{th}}$ iteration of LDM inserts the difference $X_r^r = S_r^r - S_{r-1}^r$ into the sequence $S_1^r, \ldots, S_{r-2}^r$ to form the sequence $S_1^{r-1}, \ldots, S_{r-1}^{r-1}$. The key result, easily proved by induction on $r$, is that, if the $X_i^r$, $1 \le i \le r$, are independent exponentials with parameters $\lambda_i^r$, then the following holds:

(i) Given the event $\{S_{i-1}^r \le X_r^r < S_i^r\}$, $i = 1, \ldots, r-2$, the spacings in the list $(S_1^{r-1}, \ldots, S_{r-1}^{r-1})$, i.e.,

$$\{X_k^r\}_{1\le k\le i-1},\ X_r^r - S_{i-1}^r,\ S_i^r - X_r^r,\ \{X_k^r\}_{i+1\le k\le r-2}\ ,$$

are (conditionally) independent and exponential with parameters $\{\lambda_k^r\}_{1\le k\le i}$, $\{\lambda_k^r+\lambda_r^r\}_{i+1\le k\le r-2}$. Given the remaining possibility $\{X_r^r \ge S_{r-2}^r\}$, the $X_i^r$, $1 \le i \le r-2$, and $X_r^r - S_{r-2}^r$ are independent exponentials with parameters $\{\lambda_k^r + \lambda_r^r\}_{1\le k\le r-2}$ and $\lambda_r^r$.

(ii) The event probabilities are

$$Pr\{S_{i-1} \le X_r^r < S_i\} = \frac{\lambda_r^r}{\lambda_i^r + \lambda_r^r} \prod_{j=1}^{i-1} \frac{\lambda_i^r}{\lambda_j^r + \lambda_r^r}\ , \quad 1 \le i \le r-2\ ,$$

(3.2)

$$Pr\{X_r^r \ge S_{r-2}\} = \prod_{j=1}^{r-2} \frac{\lambda_j^r}{\lambda_j^r + \lambda_r^r}\ .$$

It follows immediately that the sequence $(\lambda_1^n, \ldots, \lambda_n^n), (\lambda_1^{n-1}, \ldots, \lambda_{n-1}^{n-1}), \ldots, (\lambda_1^1)$ has the Markov property with transition probabilities given by (3.2). By (i) and the initial state $\lambda_1^n = \cdots = \lambda_n^n = 1$, we note that $\lambda_1^n \le \lambda_1^{n-1} \le \cdots \le \lambda_1^1$, and for each $r = 1, \ldots, n$, we have $\lambda_1^r \ge \lambda_2^r \ge \cdots \ge \lambda_r^r$. The proof of (3.1) now reduces to an estimate of the growth rate of $\lambda_1^r$. Yakir's (1993) bounds establish that, to within a constant factor, $\lambda_1^r$ grows as $n^{c\log(n-r)}$ for some $c > 0$, as desired.

Recent simulations of Bentley (1993) suggest that the value of $c$ in (3.1) is close to $1/2$, with 2 being the base of the logarithm. Johnson et al. (1991) ran experiments comparing LDM with simulated-annealing approaches to the makespan minimization problem. The clear superiority of LDM in this comparison contributed to their conclusion that the local optimization framework of simulated annealing was not well suited to makespan minimization problems.

With $m = 2$ and $F$ the uniform distribution on $[0, 1]$, a rough summary of the expected-error results is that, for the greedy, LPT, LDM, and an optimal policy, we have respectively,

$$E[\alpha_{2,n}] = O(1), O(n^{-1}), O(n^{-c\log n}), O(\rho^n)$$

for some $c > 0$, $0 < \rho < 1$. The $O(\rho^n)$ result for optimal scheduling is not discussed here (see Coffman and Lueker (1991), Section 4.3), and in fact remains a conjecture. The strongest results of this type are those of Karmarkar et al. (1986) who showed that the median of the final difference, $2\alpha_{2,n}$, is bounded by a constant times $n/2^n$.

## 4. Policy-Free Error Asymptotics

From (2.7) it is clear that the relative size of the error $\alpha_{m,n}$ compared to the makespan $L_{m,n}$ itself is asymptotically negligible as $n \to \infty$ for the greedy policy. For large $n$, obviously

the dominant part of $L_{n,m}$ for any policy is the normalized sum of all the processing times. In this section we establish a stronger result. We show that *for any policy the limiting behavior of the error $\alpha_{m,n}$ as $n \to \infty$ is independent of the policy.* In particular, the expected error $E[\alpha_{m,n}]$ for a given policy is asymptotically negligible compared to the standard deviation of the makespan (which is the same as the standard deviation of the error). In a probabilistic setting, what we can gain from a good policy is asymptotically negligible as $n \to \infty$ compared to our degree of uncertainty about the makespan.

Central limit theorems (CLTs) and functional central limit theorems (FCLTs) provide asymptotics that exhibit this property for general distributions and a policy-free set-up, i.e., a model yielding results simultaneously valid for all policies. The policies to be considered in the illustrations below are those in the class of *list scheduling* (LS) policies. Such a policy begins by computing a permutation $\pi_n = (\pi(1), \ldots, \pi(n))$ of the integers $1, \ldots, n$, and then schedules the ordered list $(T_{\pi(1)}, \ldots, T_{\pi(n)})$ by the greedy rule. For any given sequence $T_1, T_2, \ldots$, an LS policy defines a sequence of permutations $\{\pi_n, n \geq 1\}$. Note that the policies of earlier sections are all LS policies.

Let $S_n$ and $M_n$ be the sum and maximum of $T_1, \ldots, T_n$, and note that both quantities are invariant under permutations of $T_1, \ldots, T_n$. Let the number $m_n$ of processors be a nondecreasing function of $n$, and denote the makespan and error under permutation $\pi_n$ by $L_n^{\pi_n}$ and $\alpha_{m,n}^{\pi_n}$. From (2.3) we obtain the basic inequality

$$|S_n - m_n L_n^{\pi_n}| \leq m_n M_n \quad \text{for all} \quad \pi_n \ , \tag{4.1}$$

from which we see that the limiting behavior of $L_n^{\pi_n}$, $\alpha_{m,n}^{\pi_n}$ is determined by the asymptotics of $(S_n, M_n)$. Typically, when a CLT holds for $S_n$, $M_n$ is asymptotically negligible compared to $S_n$. (See §4.5 of Resnick (1986) for further discussion of the asymptotic behavior of $(S_n, M_n)$.) In our case, we have the CLT

$$n^{-1/2}(S_n - n\tau) \Rightarrow N(0, \sigma^2) \quad \text{as} \quad n \to \infty \ , \tag{4.2}$$

where $\Rightarrow$ denotes convergence in distribution and $N(a, b)$ denotes a normally distributed random variable with mean $a$ and variance $b$. It then follows from (4.1) and Theorem 4.1 of Billingsley (1968) that, if

$$m_n n^{-1/2} M_n \Rightarrow 0 \quad \text{as} \quad n \to \infty \ ,$$

then for any sequence of permutations $\{\pi_n, n \geq 1\}$,

$$n^{-1/2} m_n \alpha_n^{\pi_n} \Rightarrow N(0, \sigma^2) \quad \text{as} \quad n \to \infty \ . \tag{4.3}$$

For example, suppose $T$ is exponentially distributed with mean $\tau$. Since $M_n / \ln n \Rightarrow \tau$, $n \to \infty$ (e.g. see Leadbetter, Lindgren, and Rootzén (1983)), then (4.3) holds if $m_n = o(n^{1/2}/\ln n)$, $n \to \infty$.

For a fixed number $m_n = m$, $n \geq 1$, of processors, no explicit assumption about $M_n$ needs to be made. This can be seen in the general setting of the following FCLT for $S_n$. In terms of the usual diffusion-limit scalings, define the normalized processes

$$\mathbf{S}_n \equiv \mathbf{S}_n(t) = \frac{S_{\lfloor nt \rfloor} - \tau nt}{n^{1/2}} \ , \quad t \geq 0$$

$$\boldsymbol{\alpha}_n^{\pi_n} \equiv \boldsymbol{\alpha}_n^{\pi_n}(t) = \frac{L_{\lfloor nt \rfloor}^{\pi_n} - (\tau/m) nt}{n^{1/2}} \ , \quad t \geq 0 \ .$$

If $\mathbf{B}$ denotes standard (zero drift, unit diffusion) Brownian motion, then we have

$$\mathbf{S}_n \Rightarrow \sigma \mathbf{B} \quad \text{as} \quad n \to \infty \ , \tag{4.4}$$

where $\Rightarrow$ denotes weak convergence in the Skorohod space $D \equiv D([0,1], \mathbf{R})$ (see Ethier and Kurtz (1986) for technical details). By the continuous mapping theorem with the maximum jump functional, we deduce from (4.4) that $n^{-1/2} M_n \Rightarrow 0$ as $n \to \infty$. Hence, for any sequence of permutations $\{\pi_n, n \geq 1\}$,

$$\boldsymbol{\alpha}_{m,n}^{\pi_n} \Rightarrow \frac{\sigma}{m} \mathbf{B} \ \text{ in } \ D \quad \text{as} \quad n \to \infty \ . \tag{4.5}$$

This gives the approximation

$$L_{m,n}^{\pi_n} \approx \frac{n\tau}{m} + n^{1/2} N(0, \sigma^2/m^2) \ , \tag{4.6}$$

in which $\pi_n$ does not appear, since the effect of the permutation $\pi_n$ is of order $M_n$, which is asymptotically negligible compared to $n^{1/2}$.

We remark that the setting for the above limit laws can be broadened considerably, covering interesting cases where the independence assumption or the identical-distribution assumption does not hold.

## 5. Flow Shops

In a variant of multiprocessor scheduling, called *permutation flow-shop scheduling*, the processors are connected in tandem and tasks consist of ordered sets of *operations* $(T_{i1}, \ldots, T_{im})$, $1 \le i \le n$, to be done in sequence on $P_1, \ldots, P_m$. A schedule is determined by a permutation $\pi_n = (\pi(1), \ldots, \pi(n))$; on processor $j$ for each $j$ the operation of $T_{\pi(i),j}$ must precede the operation of $T_{\pi(i+1),j}$, $1 \le i \le n-1$. Thus, we have a tandem queueing system with $m$ single-server queues and $n$ arrivals, all available at time 0. The makespan $L_{m,n}^{\pi_n}$ is the finishing time of $T_{\pi(n),m}$. In the problem considered here all $mn$ operations are i.i.d. random variables; for simplicity the distribution $F$ and its properties will be carried over to the flow shop problem, but it will refer to operation times rather than entire task times.

The combinatorial problem of selecting an optimal permutation $\pi_n$ is NP-complete for $m \ge 3$, as shown by Garey, Johnson, and Sethi (1976), but Johnson (1954) proved that the following simple rule is optimal for $m = 2$: if $\min(T_{i1}, T_{j2}) < \min(T_{i2}, T_{j1})$ then schedule $T_i$ before $T_j$. We return to an analysis of Johnson's rule after discussing the greedy rule for general $m$.

As before, the greedy rule sequences tasks in the order given, i.e., with $\pi_n = (1, \ldots, n)$. The makespan is again denoted by $L_{m,n}$; the error is now defined to be $\alpha_{m,n} \equiv L_{m,n} - (n+m-1)\tau$, where $(n+m-1)\tau$ is a trivial lower bound to the expected makespan. Glynn and Whitt (1991), motivated by earlier work of Srinivasan (1993), studied in depth the asymptotic behavior of $L_{m,n}$ as $m$, $n$, or both tend to infinity. An immediate dual is obtained for each limit by the easily proved symmetry,

$$\{L_{i,j} : 1 \le i \le m, 1 \le j \le n\} \stackrel{\mathrm{d}}{=} \{L_{j,i} : 1 \le j \le n, 1 \le i \le m\} \; . \tag{5.1}$$

Theorem 2 of Iglehart and Whitt (1970) gives an FCLT for each $m$,

$$n^{-1/2} \boldsymbol{\alpha}_{m,n} \Rightarrow \sigma \widehat{\boldsymbol{\alpha}}_m \quad \text{as} \quad n \to \infty \; , \tag{5.2}$$

where $\widehat{\boldsymbol{\alpha}}_m$ is a functional of $m$-dimensional Brownian motion and convergence is in the appropriate Skorohod space. By applying the subadditive ergodic theorem (Liggett (1985), p. 277), Glynn and Whitt established that

$$m^{-1/2} \widehat{\boldsymbol{\alpha}}_m \Rightarrow \gamma \quad \text{as} \quad m \to \infty \; , \tag{5.3}$$

where $\gamma$ is a positive constant. Analysis has so far yielded little information about $\widehat{\boldsymbol{\alpha}}_m$ or $\gamma$.

This prompted a simulation study by Greenberg, Schlunk, and Whitt (1993), which provided further insights, e.g., the simulations suggest that $\gamma = 2$.

By means of a strong approximation theorem, Glynn and Whitt proved a result more general than the above iterated limit: If for any $\epsilon$, $0 < \epsilon < 1$, $m_n = n^{1-\epsilon}$, then

$$(nm_n)^{-1/2}\alpha_{m_n,n} \Rightarrow \gamma \quad \text{as} \quad n \to \infty \ , \tag{5.4}$$

with $\gamma$ as in (5.3).

If $m, n \to \infty$ at comparable rates, then hydrodynamic limits for $L_{m,n}$ emerge. By applying the results in Section 4.2 of Srinivasan (1993), Glynn and Whitt showed that, if the $T_i$ are exponentially distributed with mean 1, then

$$n^{-1}L_{\lfloor xn \rfloor,n} \to (1+\sqrt{x})^2 \ \text{ w.p.1 as } \ n \to \infty \ . \tag{5.5}$$

for any $x > 0$. Glynn and Whitt also extend this result to any distribution with an exponential tail; they get a deterministic function $\gamma(x)$ as the limit which, according to simulations, depends on the distribution. They verify that $\gamma(x)$ is strictly increasing and concave and provide upper and lower bounds.

We return now to an analysis of optimal scheduling, and discuss results recently given by Ramudhin, et al. (1993) for the case $m = 2$ with $F$ uniform on a finite interval. Ramudhin, et al. begin by introducing the following more easily analyzed, stochastically symmetric version of the optimal policy. Partition $\{T_1, \ldots, T_n\}$ into the sets $\mathcal{T}_1, \mathcal{T}_2$ of tasks with shorter operations on $P_1$, $P_2$, respectively, i.e., $\mathcal{T}_1 = \{T_i : T_{i1} \le T_{i2}\}$, $\mathcal{T}_2 = \{T_i : T_{i2} < T_{i1}\}$. An optimal policy first schedules the tasks in $\mathcal{T}_1$ in increasing order of the $T_{i1}$ and then schedules the tasks of $\mathcal{T}_2$ in decreasing order of the $T_{i2}$. An asymptotic analysis of this schedule shows that $L_{2,n} \stackrel{\mathrm{d}}{=} T_{11} + \sum_{i=1}^{n} T_{i2} + I$, if $\sum_{i=1}^{n} T_{i1} < \sum_{i=1}^{n} T_{i2}$, and $L_{2,n} \stackrel{\mathrm{d}}{=} \sum_{i=1}^{n} T_{i1} + T_{n2} + I$, otherwise, where $I \Rightarrow 0$ a.s. as $n \to \infty$. The asymptotic makespan then becomes $\max\{\sum_{i=1}^{n} T_{i1} + T_{n2}, T_{11} + \sum_{i=1}^{n} T_{i2}\}$, in the sense of the limit law

$$\alpha_{2,n} \Rightarrow \sigma \ \max\{N_1, N_2\} \quad \text{a.s.}, \quad n \to \infty \ , \tag{5.6}$$

where $N_1$, $N_2$ are i.i.d. standard normal random variables. This yields the estimate

$$E[\alpha_{2,n}] = \sigma\sqrt{n/\pi} + o(\sqrt{n}) \quad \text{as} \quad n \to \infty \ , \tag{5.7}$$

which differs only in the multiplicative constant from that obtainable from (5.2).

Ramudhin et al. present several other results on queue lengths, workloads, and waiting times under the optimal policy and a simpler, near optimal policy. A nontrivial lower bound on expected optimal makespans for $m \geq 3$ remains an open problem. In particular, it would be interesting to see whether $E[\alpha_{m,n}]$ grows at least as fast as $\sqrt{mn}$ under an optimal policy. The upper bound provided by the greedy policy would then show that this growth rate is exact within a constant factor.

## 6. Final Remarks

When viewed against the much broader and more varied background of combinatorial makespan scheduling problems, probabilistic analysis appears to be in its infancy. A few of the many variants, scarcely touched at present, are task precedence constraints, dedicated processors, set-up times, interprocessor transfer times, variable profiles, and preemptive sequencing policies (see Lawler, et al. (1992)). Modeling issues are often an initial hurdle in problems with additional structure. To obtain a tractable model, the uniform or exponential is often the distribution of choice. A case in point is the recent work of Dell'Olmo, Speranza, and Tuza (1993) on dedicated three-processor systems. In this variant of multiprocessor scheduling, each task specifies a nonempty subset of the processors that it requires throughout its running time. In a uniform model of the processors required by tasks, Dell'Olmo, Speranza, and Tuza show that, for $m = 3$ and for all $n$ sufficiently large, optimal schedules can be computed in linear time for over 95% of the instances; this property of an instance is checkable in advance.

However, uniform assumptions do not always lead to interesting structures, as the following classical model with precedence constraints illustrates. Add to the problem instance of Sections 2–4 a random irreflexive partial order $\prec$ on $\{T_1, \ldots, T_n\}$ representing precedence relations ($T_i \prec T_j$ means that $T_j$ can not begin until $T_i$ is finished). To define the term "random," a natural first assumption would be that $\prec$ is chosen uniformly at random among all partial orders on $\{T_1, \ldots, T_n\}$. However, a typical such partial order has height 3 (see Kleitman and Rothschild (1975)), and hence yields a simplistic model for many applications of makespan scheduling. A more promising model is that studied by Winkler (1985) in which random orders are constructed from the intersection of $k \geq 2$ random linear orders, or equivalently, the random orders induced by the ordinary product order on $n$ points chosen independently and uniformly at random from the unit hypercube $[0,1]^k$. In the induced order, $(x_1, \ldots, x_k) \prec (y_1, \ldots, y_k)$ if and only if $x_i \leq y_i$ for all $i = 1, \ldots, k$, with strict inequality holding for at least one $i$. For large

$n$, typical such random orders have the properties: (i) there are no isolated tasks, (ii) there are $(\ln n)^{k-1}/(k-1)!$ minimal and (by symmetry) maximal tasks, (iii) the height is $c_k n^{1/k}$ for some $c_k$, $0 < c_k < e$, and (iv) the width is approximately $c'_k n^{(k-1)/k}$ for some constant $c'_k$; in the balanced case, with $k = 2$, $c_k = c'_k = 2$. An interesting sample problem might be to determine asymptotic expected makespans under a greedy policy with random partial orders as above and all $T_i = 1$. In this case, a greedy policy could be highest-level-first (the next task to be scheduled is one that dominates a longest chain), with some rule for resolving ties amongst highest-level tasks.

Many other open problems exist in the same settings as discussed here, but with different performance metrics, such as the sum, possibly weighted, of task finishing times, and tardiness measures; again see Lawler et al. (1992). One fundamental variant that has received a great deal of attention is the following dual of the multiprocessor scheduling problem: for a fixed makespan (deadline) that exceeds the longest task running time, determine the least $m$ such that $\{T_1, \ldots, T_n\}$ can be scheduled on $P_1, \ldots, P_m$ with all tasks finishing by the deadline. This is the one-dimensional bin packing problem; the monograph by Coffman and Lueker (1991) covers much of the probabilistic analysis of this problem.

## References


Asmussen, S. (1987), *Applied Probability and Queues*, Wiley, New York.

Baccelli, F. and Brémaud, P. (1987), *Palm Probabilities and Stationary Queueing Systems*, Springer, New York.

Bentley, J. L. (1993), private communication.

Billingsley, P. (1968), *Convergence of Probability Measures*, Wiley, New York.

Blazewicz, J., Ecker, K., Schmidt, G., and Weglarz, J. (1993), *Scheduling in Computer and Manufacturing Systems*, Springer-Verlag, Berlin.

Boxma, O. J. (1984), "A Probabilistic Analysis of the LPT Scheduling Rule," in E. Gelenbe, editor, *Performance '84: Proceedings of the Tenth International Symposium on Models of Computer System Performance*, North-Holland.

Boxma, O. J. (1985), "A Probabilistic Analysis of Multiprocessor List Scheduling: the Erlang Case," *Stochastic Models*, 1:209–220.

Bruno, J. L. and Downey, P. J. (1986), "Probabilistic Bounds on the Performance of List Scheduling," *SIAM J. Comput.*, 15:409–417.

Çinlar, E. (1972), "Superposition of Point Processes," in *Stochastic Point Processes: Statistical Analysis, Theory and Applications*, P. A. W. Lewis (ed.), Wiley, New York, 549–606.

Coffman, E. G., Jr., Flatto, L., Weiss, A., Whitt, W., and Wright, P. E. (1993), "Limit Laws for Multiprocessor Scheduling," in preparation.

Coffman, E. G., Jr., Flatto, L., and Lueker, G. S. (1984), "Expected Makespans for Largest-First Multiprocessor Scheduling," *Tenth International Symposium on Models of Computer System Performance*, North-Holland.

Coffman, E. G., Jr., Frederickson, G. N., and Lueker, G. S. (1984), "A Note on Expected Makespan for Largest-First Sequences of Independent Tasks on Two Processors," *Math. Oper. Res.*, 9(2):260–266.

Coffman, E. G., Jr. and Gilbert, E. N. (1985), "On the Expected Relative Performance of List Scheduling," *Oper. Res.*, 33:548–561.

Coffman, E. G., Jr. and Lueker, G. S. (1991), *Probabilistic Analysis of Packing and Related Partitioning Problems*, Wiley, New York.

Coffman, E. G., Jr. and Wright, P. E. (1992), "Load Balancing on Two Identical Facilities," Tech. Memo., AT&T Bell Laboratories, Murray Hill, NJ 07974.

Dell'Olmo, P., Speranza, M. G., and Tuza, Zs. (1993), "Polynomial Instances and Approximation Results for the Scheduling Problem on Three Dedicated Processors," Tech. Rep. I.A.S.I.-C.N.R., Viale Manzorci, 30, I-00185 Rome.

Ethier, S. N. and Kurtz, T. G. (1986), *Markov Processes*: *Characterization and Convergence*, Wiley, New York.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, Volume II, second edition, John Wiley & Sons, New York.

Franken, P., König, D., Arndt, U., and Schmidt, V. (1982), *Queues and Point Processes*, Wiley, New York.

Frenk, J. B. G. and Rinnooy Kan, A. H. G. (1987), "The Asymptotic Optimality of the LPT Rule," *Math. Oper. Res.*, 12(2):241–254.

Garey, M. R. and Johnson, D. S. (1979), *Computers and Intractability*: *A Guide to the Theory of NP-Completeness*, W. H. Freeman, New York.

Garey, M. R., Johnson, D. S., and Sethi, R. (1976), "The Complexity of Flow Shop and Job Shop Scheduling," *Math. Oper. Res.*, 1:117–129.

Glynn, P. W. and Whitt, W. (1991), "Departures from Many Queues in Series," *Ann. Appl. Prob.*, 1:546-572.

Graham, R. L. (1966), "Bounds for Certain Multiprocessing Anomalies," *Bell. Sys. Tech. J.*, 45:1563–1581.

Greenberg, A. G., Schlunk, O. and Whitt, W. (1993), "Using Distributed-Event Parallel Simulation to Study Departures from Many Queues in Series," *Prob. Eng. Inf. Sci.*, 7:159–186.

Han, S., Hong, D., and Leung, J. Y.-T. (1992), "On the Asymptotic Optimality of Multiprocessor Scheduling Heuristic Algorithms," Tech. Rep., Computer Science Dept., University of Nebraska, Lincoln, NE 68588-0115.

Iglehart, D. L. and Whitt, W. (1970), "Multiple Channel Queues in Heavy Traffic. II, Sequences, Networks, and Batches," *Adv. Appl. Prob.*, 2:355–369.

Johnson, S. M. (1954), "Optimal Two and Three Stage Production Schedules with Set-Up Times Included," *Nav. Res. Log. Quart.*, 1:61–68.

Johnson, D. S., Aragon, C. R., McGeoch, L. A., and Schevon, C. (1991), "Optimization by Simulated Annealing: An Experimental Evaluation; Part II, Graph Coloring and Number Partitioning," *Oper. Res.*, 39:378–406.

Karmarkar, N. and Karp, R. M. (1982), "The Differencing Method of Set Partitioning," Technical Report UCB/CSD 82/113, Computer Science Division (EECS), University of California, Berkeley.

Karmarkar, N., Karp, R. M., Lueker, G. S., and Odlyzko, A. M (1986), "Probabilistic Analysis of Optimum Partitioning," *J. Appl. Prob.*, 23(3):626–645.

Kleitman, D. J. and Rothschild, B. L. (1975), "Asymptotic Enumeration of Partial Orders on a Finite Set," *Trans. Am. Math. Soc.*, 205:205–210.

Lawler, E. L. Lenstra, J. K., Rinnooy Kan, A. H. G., and Shmoys, D. B. (1992), "Sequencing and Scheduling: Algorithms and Complexity," Rep. BS-R8909, Centre for mathematics and Computer Science, 1009AB Amsterdam. (To appear in *Handbook of Operations Research and Management Science*, North-Holland, Amsterdam.)

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983), *Extremes and Related Properties of Random Sequences and Processes*, Springer, New York.

Liggett, T. M. (1985), *Interacting Particle Systems*, Springer, New York.

Loulou, R. (1984), "Tight Bounds and Probabilistic Analysis Two Heuristics for Parallel Processor Scheduling," *Math. Oper. Res.*, 9:142–150.

Lueker, G. S. (1987), "A Note on the Average-Case Behavior of a Simple Differencing Method for Partitioning," *Oper. Res. Lett.*, 6:285-287.

Meyn, S. P. and Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability*, Springer-Verlag, London.

Phillips, S. and Westbrook, J. (1993), "Online Load Balancing and Network Flow," *Proc. 25th Ann. ACM Symp. Th. Comput*, ACM Press, New York, 402–411.

Ramudhin, A., Bartholdi, J. J., III, Calvin, J. M., Vande Vate, J. H., and Weiss, G. (1993), "A Probablistic Analysis of 2-Machine Flowshops," Tech. Rep., School of Industrial and System Engineering, Georgia Institute of Technology, Atlanta, GA 30332.

Reiss, R.-D. (1989), *Approximate Distributions of Order Statistics*, Springer-Verlag, New York.

Resnick, S. I. (1986), "Point Processes, Regular Variation and Weak Convergence," *Adv. Appl. Prob.*, 18:66–138.

Rinnooy Kan, A. H. G. and Frenk, J. B. G. (1986), "On the Rate of Convergence to Optimality of the LPT Rule," *Discrete Applied Mathematics*, 14:187–198.

Srinivasan, R. (1993), "Queues in Series Via Interacting Particle Systems," *Math. Oper. Res.*, 18:39–50.

Whitt, W. (1984), "On Approximations for Queues, I: Extremal Distributions," *AT&T Bell. Lab. Tech. J.*, 63:115–138.

Winkler, P. (1985), "Random Orders," *Order*, 4:317–331.

Yakir, B. (1993), "The Differencing Algorithm LDM for Partitioning: A Proof of Karp's Conjecture," Tech. Rep. 93/03, Dept. of Biostatistics, University of Rochester, Rochester, NY.