ON THE ASYMPTOTIC OPTIMALITY OF THE SPT RULE FOR THE FLOW SHOP AVERAGE COMPLETION TIME PROBLEM

CATHY H. XIA

IBM Thomas J. Watson Research Center, PO Box 704, Yorktown Heights, New York 10598
hhcx@watson.ibm.com

J. GEORGE SHANTHIKUMAR

Department of IEOR and the Walter A. Haas School of Business, University of California at Berkeley, Berkeley, California 94720
jgshant@ieor.berkeley.edu

PETER W. GLYNN

Department of EES & OR, Stanford University, Stanford, California 94305
glynn@leland.stanford.edu

(Received September 1997; revisions received October 1998, December 1998; accepted December 1998)

Consider a flow shop with M machines in series, through which a set of jobs are to be processed. All jobs have the same routing, and they have to be processed in the same order on each of the machines. The objective is to determine such an order of the jobs, often referred to as a permutation schedule, so as to minimize the total completion time of all jobs on the final machine. We show that when the processing times are statistically exchangeable across machines and independent across jobs, the Shortest Processing Time first (SPT) scheduling rule, based on the total service requirement of each job on all M machines, is asymptotically optimal as the total number of jobs goes to infinity. This extends a recent result of Kaminsky and Simchi-Levi (1996), in which a crucial assumption is that the processing times on all M machines for all jobs must be i.i.d.. Our work provides an alternative proof using martingales, which can also be carried out directly to show the asymptotic optimality of the weighted SPT rule for the Flow Shop Weighted Completion Time Problem.

1. INTRODUCTION

Flow shop scheduling, as modeled by a tandem queue, is a rich classical topic motivated mainly by practical needs in the manufacturing context. With recent developments in communication networks, the role of tandem queue scheduling becomes increasingly significant. Research has concentrated on different issues such as buffering, service effort allocation, and completion time minimization. Results such as reversibility, duality, and the bowl-shaped phenomenon are all well known in the literature, as established by Muth (1979), Pinedo (1995a), and Hillier and Boling (1979). A more detailed literature review of these issues can be found in Buzacott and Shanthikumar (1993), Pinedo (1995b), and Weber (1992).

In this paper, we restrict our attention to the issue of completion time. Consider a flow shop with M machines in series, through which a set of jobs are to be processed. All jobs have the same routing, and they have to be processed in the same sequence on each of the machines. The objective is to determine a sequence of jobs, often referred to as a permutation schedule, that minimizes the average—or equivalently, the total—of the completion times of all jobs on the final machine. It is assumed that the buffer space between machines is unlimited and that the jobs are processed individually without preemption. This problem is well known to be NP-hard even in the two-machine case, as pointed out by Garey et al. (1976).

Research on completion time can be divided into two categories. One focuses on *makespan*, the time needed to complete processing all jobs, which takes a social point of view from the system side. The other concentrates on the *average completion time* of all jobs on the final machine, which focuses more on the individual jobs. Although the latter captures important real-life managerial scheduling concerns that are not reflected in makespan related objectives, the analysis is often very complicated and difficult. So far, most results in this area have focused on minimizing the makespan. For a complete survey of makespan related work see, for example, Pinedo (1995b).

For average completion time related scheduling, most of the previous research has studied deterministic problems using branch-and-bound strategies and is often limited to a small number of jobs and machines. Examples include Krone and Steiglitz (1974) and Van de Velde (1990), along with many others. In this paper, we look at the stochastic case instead and search for an asymptotically optimal schedule. We adopt the notion of asymptotically optimal scheduling that has been recently considered by Shanthikumar and Xu (1997), and Xia (1999), in queueing control problems. The limiting performance of tandem queues

Subject classifications: Asymptotically optimal scheduling: flow shop average completion time problem. Flow shop scheduling: average completion time, asymptotic optimality. Tandem queueing: asymptotic optimality of SPT.
Area of review: STOCHASTIC MODELS.

has been studied extensively by, for example, Glynn and Whitt (1991). In this study, we combine ideas from both and apply them to the *Flow Shop Average Completion Time Problem*. We show that as the number of jobs to be scheduled becomes larger and larger, which is often demanded in real-life applications, the Shortest total Processing Time (SPT) first rule is asymptotically optimal, provided that the processing times are statistically exchangeable across machines and independent across jobs.

It should be noted that this research was motivated by a recent result of Kaminsky and Simchi-Levi (1996b), in which it was first pointed out that the SPT rule was asymptotically optimal when the processing times are i.i.d. across machines and jobs. We provide a simplified proof using martingales and extend the result to a more general realm, where the jobs need not be identical—an assumption crucial to their paper. We show that for the SPT rule to be asymptotically optimal, the jobs can have different processing time distributions, as long as they are independent across jobs and the processing times of each job on all machines are statistically exchangeable (not necessarily independent). In addition, our argument can be easily extended to show the asymptotic optimality of SWPT—Shortest Weighted total Process Time first rule—for the *Flow Shop Weighted Completion Time Problem*. This extends another result of Kaminsky and Simchi-Levi (1996a) and simplifies their argument.

The remainder of the paper is organized as follows. Section 2 introduces the basic model. Section 3 reviews the well-known critical path approach using an activity network in determining the final completion time for each job. The total completion time for all jobs is then evaluated in §§4 and 5, where a simple lower bound and an upper bound are presented, respectively. In §6 we show, using martingales, that when scheduling is based on the total service requirement on all machines for each job, the difference between the upper bound and the lower bound converges to zero under an *exchangeability* hypothesis. The main result is then presented in §7.

2. THE MODEL

We consider the traditional Flow Shop Average Completion Time Problem, which is well known in the field of scheduling. In this problem, a flow shop consisting of M machines, each with unlimited buffer space, must sequentially process n jobs. Each machine can handle at most one job at a time, and each job can be processed on only one machine at a time without preemption. Job j, $j = 1,2,\ldots,n$ has a processing time $t_{j,m}$ on machine m, $m = 1,2,\ldots,M$, and the processing times must satisfy the following assumptions.

ASSUMPTION A1. *Processing times of different jobs are independent, i.e.,* $((t_{j,1},\ldots,t_{j,M}): j \geq 1)$ *is a sequence of independent random vectors.*

ASSUMPTION A2. *For each job, the processing times on the* M *machines are exchangeable random variables, i.e., for any given* j (≥ 1), $t_{j,1},\ldots,t_{j,M}$ *are exchangeable.*[1]

We restrict our attention to the set of *permutation* scheduling policies, where the order in which the jobs are processed on the first machine is maintained throughout the system; that is, once the order of processing the n jobs on the first machine is scheduled, all other machines process the jobs in a *first come first serve* (FCFS) manner. It should also be noticed that permutation schedules are not necessarily optimal for the general flow shop scheduling problem where reordering is allowed at each machine. For simplicity, we consider only the permutation schedules.

The objective is to determine π—a *schedule*, or sequence of the jobs, such that $Z_M^{\pi}(n)$, the total completion times of all jobs on the last machine M is minimized. Let $Z_M^*(n)$ denote the optimal objective function value.

This paper provides a proof of the asymptotic optimality of the well-known rule in scheduling theory, known as the *Shortest Processing Time* (SPT) first policy. According to this policy, jobs are sequenced in increasing order of their *total service requirements* on all M machines.

Note that there may exist different SPT schedules associated with the same total processing time sequence (when there is a tie in the total processing time). For the general flow shop problem (without Assumptions A1 and A2), these SPT schedules are not necessarily optimal. See Kaminsky and Simchi-Levi (1996b) for such a counterexample. Nevertheless, given Assumptions A1, A2, and some finite moment conditions on the processing times, we show that SPT is asymptotically optimal in the sense that

$$\frac{Z_M^{\mathrm{SPT}}(n)}{Z_M^*(n)} \to 1 \quad \text{a.s. as } n \to \infty,$$

where $Z_M^{\mathrm{SPT}}(n)$ denotes the corresponding total completion time for processing the n jobs under the SPT policy.

3. CRITICAL PATH

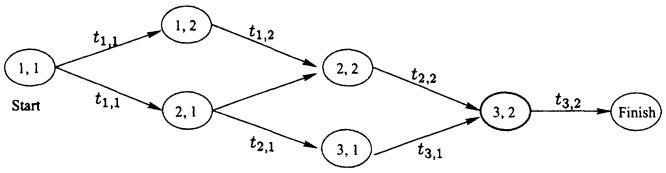
It is well known that an *Activity Network*, composed of nodes and directed arcs, is helpful in calculating completion times in a flow shop. We consider the model in which jobs are processed according to their nominal order $1,2,\ldots,n$. Let node (j, m) be associated with the activity of processing job j on machine m. Activity node (j, m) is connected with its immediate consequent activities, i.e., node $(j+1, m)$ and node $(j, m+1)$ with a directed arc of length $t_{j,m}$. This represents the fact that only when job j is completed on machine m, which should take time $t_{j,m}$, can the service of job $j+1$ on machine m and the service of job j on machine $m+1$ be initiated. Figure 1, for example, illustrates the activity network graph for processing jobs 1, 2, and 3 sequentially on machines 1 and 2.

Let $F_{j,m}$ be the completion time of job j on machine m. A basic recursion for the completion times is

$$F_{k+1,m} = \max\{F_{k,m}, F_{k+1,m-1}\} + t_{k+1,m}. \tag{1}$$

From (1), it is easy to establish the following lemma by induction.

Figure 1. Activity network graph.



LEMMA 1. *The completion time of job k on the last machine M is given by*:

$$F_{k,M} = \max_{1 \leq l_1 \leq l_2 \leq \cdots \leq l_{M-1} \leq k} \cdot \left\{ \sum_{j=1}^{l_1} t_{j,1} + \sum_{j=l_1}^{l_2} t_{j,2} + \cdots + \sum_{j=l_{M-1}}^{k} t_{j,M} \right\} \tag{2}$$

for $k = 1, \ldots, n$.

Lemma 1 is a well-known result; see, for example, Conway et al. (1967), Pinedo (1995b), and Glynn and Whitt (1991). Graphically, based on the activity network, this is equivalent to saying that the completion time of the last activity must be the directed path with maximum length from the starting node to the finishing node, which is often identified as the *Critical Path.*

4. A SIMPLE LOWER BOUND FOR THE TOTAL COMPLETION TIME

Associated with the M-machine Flow Shop Average Completion Time problem introduced in §2 is the following *single machine* Average Completion Time problem: n jobs are to be processed on a single machine where job j has processing time $T_j := \Sigma_{m=1}^{M} t_{j,m}$, $j = 1, \ldots, n$. The objective is to determine a sequence of the jobs so as to minimize the average completion time of all jobs. Let $\bar{Z}_1^*(n)$ be the corresponding optimal total completion time.

It turns out that the optimal total completion time of a single machine scheduling problem naturally gives a lower bound for the total completion time of the M-machine scheduling problem. This was first observed by Kaminsky and Simchi-Levi (1996b), and we state their result in the following lemma.

LEMMA 2. *Consider the M-machine Flow Shop Average Completion Time Problem and the Associated Single Machine Scheduling Problem. Suppose that the corresponding optimal total completion times are* $Z_M^*(n)$ *and* $\bar{Z}_1^*(n)$, *respectively. We then have*

$$\frac{1}{M} \bar{Z}_1^*(n) \leq Z_M^*(n). \tag{3}$$

It is well known that the *Shortest Processing Time* first rule is the optimal solution to the single machine scheduling problem; see for example Conway et al. (1967) and Pinedo (1995b). Therefore, if we order jobs $1, 2, \ldots, n$ in the increasing order of T_js, such that $T_1 \leq T_2 \leq \cdots \leq T_n$, then the optimal total completion time is given by $\bar{Z}_1^*(n) = \Sigma_{k=1}^{n} \Sigma_{j=1}^{k} T_j$.

In later sections we will see that the optimality of the SPT rule for single machine scheduling is directly related to the asymptotically optimal scheduling of M-machine scheduling. To be more specific, we will show that the lower bound given by (3) is asymptotically tight, which then yields the asymptotic optimality of the SPT rule based on the T_js for the original problem.

5. UPPER BOUND FOR THE TOTAL COMPLETION TIME

From (2) we obtain

$$\begin{aligned}
&F_{k,M} - \sum_{j=1}^{k} t_{j,1} \\
&\leq \max_{1 \leq l_1 \leq l_2 \leq \cdots \leq l_{M-1} \leq k} \left\{ \left[\sum_{j=l_1}^{l_2} (t_{j,2} - t_{j,1}) + \cdots + \sum_{j=l_{M-1}}^{k} (t_{j,M} - t_{j,1}) \right] + \sum_{m=1}^{M-1} t_{l_m,1} \right\} \\
&\leq \sum_{m=2}^{M} \max_{1 \leq l_a \leq l_b \leq k} \left| \sum_{j=l_a}^{l_b} (t_{j,m} - t_{j,1}) \right| \\
&\quad + \max_{1 \leq l_1 \leq l_2 \leq \cdots \leq l_{M-1} \leq k} \sum_{m=1}^{M-1} t_{l_m,1} \\
&\leq \sum_{m=2}^{M} 2 \max_{1 \leq l \leq k} \left| \sum_{j=1}^{l} (t_{j,m} - t_{j,1}) \right| + M \max_{1 \leq l \leq k} t_{l,1}.
\end{aligned}$$

Given that the n jobs are processed in an order defined by a permutation σ_n, then the total completion time $Z_M^{\sigma_n}(n) = \Sigma_{k=1}^{n} F_{\sigma_n(k),M}$ will satisfy

$$\begin{aligned}
&\frac{Z_M^{\sigma_n}(n) - \sum_{k=1}^{n} \sum_{j=1}^{k} t_{\sigma_n(j),1}}{n^2} \\
&\leq \frac{1}{n^2} \sum_{k=1}^{n} \sum_{m=2}^{M} 2 \max_{1 \leq l \leq k} \left| \sum_{j=1}^{l} (t_{\sigma_n(j),m} - t_{\sigma_n(j),1}) \right| \\
&\quad + M \frac{1}{n^2} \sum_{k=1}^{n} \max_{1 \leq l \leq k} t_{\sigma_n(l),1} \\
&\leq 2 \sum_{m=2}^{M} \frac{1}{n} \max_{1 \leq l \leq n} \left| \sum_{j=1}^{l} (t_{\sigma_n(j),m} - t_{\sigma_n(j),1}) \right| \\
&\quad + M \frac{1}{n} \max_{1 \leq l \leq n} t_{\sigma_n(l),1}.
\end{aligned} \tag{4}$$

For a sequence of jobs $1, 2, \ldots$, let $\mathcal{G} := \sigma(T_1, T_2, \ldots)$, i.e., $\mathcal{G}$ is the σ-algebra consisting of all the information associated with the total service requirement on all M machines

for each job. Note that $\mathcal{G}$ includes all the information that is needed for the initial scheduling using SPT. We next show that if σ_n is $\mathcal{G}$-measurable, meaning we schedule the jobs only based on the T_js, then the right-hand side of (4) converges to zero as $n \to \infty$ under some mild conditions.

6. CONVERGENCE THEOREM

LEMMA 3. *Given a probability space* $(\Omega, \mathcal{F}, P)$, *assume that* $((t_{j,m} : 1 \leqslant m \leqslant M) : j \geqslant 1)$ *satisfies Assumptions A1 and A2. Then*

(*i*) $E[t_{\sigma_n(j),m} - t_{\sigma_n(j),1} | \mathcal{G}] = 0$.

(*ii*) *Conditional on* $\mathcal{G}$, $t_{\sigma_n(j),m} - t_{\sigma_n(j),1}$ $(1 \leqslant j \leqslant n)$ *are independent r.v.s.*

PROOF. We fix $M = 2$ in this proof. The same argument works for the general case. Note that

$$P(t_{j,m} \in dx_{jm}, 1 \leqslant j \leqslant n,\ m = 1,2 | \mathcal{G}) = \prod_{j=1}^{n} P(t_{j,m} \in dx_{jm},\ m = 1,2 | \mathcal{G}) = \prod_{j=1}^{n} P(t_{j,m} \in dx_{jm},\ m = 1,2 | T_j).$$

Because the processing times are exchangeable, i.e., $(t_j^1, t_j^2) \stackrel{d}{=} (t_j^2, t_j^1)$, we have

$$P(t_{j,1} \in dx_{j1},\ t_{j,2} \in dx_{j2} | T_j) = P(t_{j,2} \in dx_{j1},\ t_{j,1} \in dx_{j2} | T_j).$$

The results then follow immediately. □

THEOREM 1. *Assume the conditions of Lemma* 3. *If* σ_n *is* $\mathcal{G}$*-measurable, and*

$$\sup_{j \geqslant 1} \operatorname{var}(t_{j,1}) < \infty, \tag{5}$$

then, for $\delta > 0$,

$$\frac{\max_{1 \leqslant l \leqslant n} |\sum_{j=1}^{l} (t_{\sigma_n(j),m} - t_{\sigma_n(j),1})|}{n^{1/2+\delta}} \Rightarrow 0,$$

as $n \to \infty$, $\forall m = 2, \ldots, M$, *where* $\Rightarrow$ *means convergence in distribution.*

PROOF. We fix $m = 2$ in this proof. The same argument works for the general case.

Based on Lemma 3, $t_{\sigma_n(j),2} - t_{\sigma_n(j),1}$ $(1 \leqslant j \leqslant n)$ are conditionally independent random variables with zero mean. Hence, by *Kolmogorov's inequality*, (refer to, e.g., Williams, 1991) we have:

$$P\left(\max_{1 \leqslant l \leqslant n} \left| \sum_{j=1}^{l} (t_{\sigma_n(j),2} - t_{\sigma_n(j),1}) \right| > x | \mathcal{G} \right) \leqslant \frac{\sum_{j=1}^{n} E[(t_{\sigma_n(j),2} - t_{\sigma_n(j),1})^2 | \mathcal{G}]}{x^2} = \frac{E[\sum_{j=1}^{n} (t_{\sigma_n(j),2} - t_{\sigma_n(j),1})^2 | \mathcal{G}]}{x^2} = \frac{E[\sum_{j=1}^{n} (t_{j,2} - t_{j,1})^2 | \mathcal{G}]}{x^2} \leqslant \frac{4 \sum_{j=1}^{n} \operatorname{var}(t_{j,1} | \mathcal{G})}{x^2}.$$

Hence it follows that

$$P\left(\max_{1 \leqslant l \leqslant n} \left| \sum_{j=1}^{l} (t_{\sigma_n(j),2} - t_{\sigma_n(j),1}) \right| > x \right) \leqslant \frac{2}{x^2} \sum_{j=1}^{n} E[\operatorname{var}(t_{j,1} | \mathcal{G})] \leqslant \frac{2}{x^2} \sum_{j=1}^{n} \operatorname{var}(t_{j,1}).$$

The result follows immediately from this inequality. □

REMARK. In general, condition (5) is valid when the processing times have uniformly finite expected residual lives, i.e.,

$$R := \sup_j \sup_{t \geqslant 0} E[t_{j,m} - t | t_{j,m} > t] < \infty. \tag{6}$$

To see this, we note that

$$\begin{aligned} E[t_{j,m}^2] &= 2E \int_0^{t_{j,m}} (t_{j,m} - u)\, du \\ &= 2E \int_0^{\infty} (t_{j,m} - u) I(t_{j,m} > u)\, du \\ &= 2 \int_0^{\infty} E[t_{j,m} - u | t_{j,m} > u] P(t_{j,m} > u)\, du \\ &\leqslant 2R \int_0^{\infty} P(t_{j,m} > u)\, du \\ &= 2RE[t_{j,m}] \leqslant 2R^2, \end{aligned}$$

where the exchange of the expectation and the integral in third equality is justified by Fubini's theorem.

Examples satisfying condition (6) include cases where:

1. The $t_{j,m}$s are uniformly bounded.
2. $t_{j,m} \sim \exp(\mu_j)$ with $\inf_j \mu_j > 0$.
3. The $t_{j,m}$s are new better than used in expectation (NBUE) with $\sup_j E[t_{j,m}] < \infty$.

Our next theorem deals with the *almost sure* convergence to zero of the right-hand side of (4) as n goes to infinity, which requires the following stronger condition.

THEOREM 2. *Assume the condition of Lemma* 3. *If*

$$\sup_{j \geqslant 1} E[|t_{j,1}|^p] < \infty, \tag{7}$$

for some $p > 2$, *and* $(\sigma_n,\ n \geqslant 1)$ *is a sequence of* $\mathcal{G}$*-measurable permutations, then*

$$\frac{\max_{1 \leqslant l \leqslant n} |\sum_{j=1}^{l} (t_{\sigma_n(j),m} - t_{\sigma_n(j),1})|}{n} \to 0 \quad a.s.\ as\ n \to \infty.$$

PROOF. Again, we look only at $m=2$ and extend the same argument to general m.

For $1 \leqslant l \leqslant n$, let $\mathscr{H}_l := \mathscr{G} \vee \sigma(t_{\sigma_n(j),2} - t_{\sigma_n(j),1} : 1 \leqslant j \leqslant l)$. Then $(\Sigma_{j=1}^{l}(t_{\sigma_n(j),2} - t_{\sigma_n(j),1}) : 0 \leqslant l \leqslant n)$ is a martingale with respect to $(\mathscr{H}_l : 1 \leqslant l \leqslant n)$. This is obvious because, given $\mathscr{G}$, the increments are conditionally independent with zero conditional expectation, as established in Lemma 3. Hence for $p \geqslant 1$,

$$\left(\left| \sum_{j=1}^{l} (t_{\sigma_n(j),2} - t_{\sigma_n(j),1}) \right|^p : 0 \leqslant l \leqslant n \right)$$

is a submartingale, provided that

$$E\left(\left| \sum_{j=1}^{l} (t_{\sigma_n(j),2} - t_{\sigma_n(j),1}) \right|^p \right) < \infty, \tag{8}$$

for $1 \leqslant j \leqslant n$.

In order to verify (8), note that under condition (7) we have

$$\infty > E\left(\sum_{j=1}^{n} |t_{j,2} - t_{j,1}|^p \right) = E\left(\sum_{j=1}^{n} |t_{\sigma_n(j),2} - t_{\sigma_n(j),1}|^p \right) = \sum_{j=1}^{n} E(|t_{\sigma_n(j),2} - t_{\sigma_n(j),1}|^p).$$

So, it follows that $E(|t_{\sigma_n(j),2} - t_{\sigma_n(j),1}|^p) < \infty$ for $1 \leqslant j \leqslant n$. Note that $|x|^p$ is a convex function in x; therefore, by Jensen's inequality,

$$E\left(\left| \sum_{j=1}^{l} (t_{\sigma_n(j),2} - t_{\sigma_n(j),1}) \right|^p \right) \leqslant \sum_{j=1}^{l} E(|t_{\sigma_n(j),2} - t_{\sigma_n(j),1}|^p) < \infty.$$

We now apply the *maximal inequality for submartingales* (see Chung 1974, p. 303):

$$P\left(\max_{1 \leqslant l \leqslant n} \left| \sum_{j=1}^{l} (t_{\sigma_n(j),2} - t_{\sigma_n(j),1}) \right| > x \right) = P\left(\max_{1 \leqslant l \leqslant n} \left| \sum_{j=1}^{l} (t_{\sigma_n(j),2} - t_{\sigma_n(j),1}) \right|^p > x^p \right) \leqslant x^{-p} E\left(\left| \sum_{j=1}^{n} (t_{\sigma_n(j),2} - t_{\sigma_n(j),1}) \right|^p \right) = x^{-p} E\left(\left| \sum_{j=1}^{n} (t_{j,2} - t_{j,1}) \right|^p \right).$$

Hence,

$$P\left(\max_{1 \leqslant l \leqslant n} \left| \sum_{j=1}^{l} (t_{\sigma_n(j),2} - t_{\sigma_n(j),1}) \right| > x \right) \leqslant x^{-p} E\left(\left| \sum_{j=1}^{n} (t_{j,2} - t_{j,1}) \right|^p \right).$$

Under condition (7), we know that for $p > 2$,

$$\sup_{j \geqslant 1} E[|t_{j,2} - t_{j,1}|^p] < \infty.$$

This implies, using the *Burkholder-Gundy square inequality* for martingales (see Burkholder 1972), that

$$E\left[\left| \sum_{j=1}^{n} (t_{j,2} - t_{j,1}) \right|^p \right] = O(n^{p/2}). \tag{9}$$

Therefore, $\forall \varepsilon > 0$,

$$\sum_{n=1}^{\infty} P\left(\max_{1 \leqslant l \leqslant n} \left| \sum_{j=1}^{l} (t_{\sigma_n(j),2} - t_{\sigma_n(j),1}) \right| > \varepsilon n \right) \leqslant \varepsilon^{-p} \sum_{n=1}^{\infty} n^{-p} E\left(\left| \sum_{j=1}^{n} (t_{j,2} - t_{j,1}) \right|^p \right) = \varepsilon^{-p} \sum_{n=1}^{\infty} n^{-p} O(n^{p/2}) < \infty.$$

Consequently, the *Borel-Cantelli lemma* establishes that if σ_n is $\mathscr{G}$-measurable, then

$$\frac{\max_{1 \leqslant l \leqslant n} |\sum_{j=1}^{l} (t_{\sigma_n(j),2} - t_{\sigma_n(j),1})|}{n} \to 0 \quad \text{a.s. as } n \to \infty.$$

□

REMARK. In general, condition (7) is valid given that the processing times have uniformly finite $(p-1)$th moment for the residual lives, i.e.,

$$\sup_j \sup_{t \geqslant 0} E[(t_{j,m} - t)^{p-1} | t_{j,m} > t] < \infty.$$

Examples include:

1. The $t_{j,m}$s are uniformly bounded.
2. $t_{j,m} \sim \exp(\mu_j)$ with $\inf_j \mu_j > 0$.

7. ASYMPTOTIC OPTIMALITY

Based on (4), Theorems 1 and 2 then establish, respectively, the weak convergence and almost sure convergence of

$$\frac{Z_M^{\sigma_n}(n) - \sum_{k=1}^{n} \sum_{j=1}^{k} t_{\sigma_n(j),1}}{n^2} \quad \text{to zero as } n \to \infty.$$

Returning to §5, if instead we reformulate (2) as

$$F_{k,M} \leqslant \sum_{j=1}^{k} t_{j,m} + \max_{1 \leqslant l_1 \leqslant l_2 \leqslant \cdots \leqslant l_{M-1} \leqslant k} \cdot \left\{ \left[\sum_{s=1,\ldots,M;\, s \neq m} \sum_{j=l_{s-1}}^{l_s} (t_{j,s} - t_{j,m}) \right] + \sum_{s=1}^{M-1} t_{l_s,m} \right\},$$

where $l_0 = 1$ and $l_M = k$, similar arguments can then be carried out to obtain the convergence of

$$\frac{Z_M^{\sigma_n}(n) - \sum_{k=1}^{n} \sum_{j=1}^{k} t_{\sigma_n(j),m}}{n^2} \quad \text{to zero as } n \to \infty,$$

for $m = 1, 2, \ldots, M$.

The next result then follows immediately.

THEOREM 3. *In an M-machine flow shop, assume the conditions of Lemma 3. If jobs $1, 2, \ldots, n$ are processed according to a $\mathcal{G}$-measurable permutation σ_n, and $Z_M^{\sigma_n}(n)$ is the corresponding total completion time, then*

(i) under condition (5),

$$\frac{Z_M^{\sigma_n}(n) - \frac{1}{M} \sum_{k=1}^{n} \sum_{j=1}^{k} T_{\sigma_n(j)}}{n^2} \Rightarrow 0, \tag{10}$$

as $n \to \infty$,

(ii) under condition (7),

$$\frac{Z_M^{\sigma_n}(n) - \frac{1}{M} \sum_{k=1}^{n} \sum_{j=1}^{k} T_{\sigma_n(j)}}{n^2} \to 0 \quad \text{a.s. as } n \to \infty. \tag{11}$$

If we process the n jobs based on the SPT rule, i.e., the permutation σ_n is defined such that

$$T_{\sigma_n(1)} \leqslant T_{\sigma_n(2)} \leqslant \cdots \leqslant T_{\sigma_n(n)}, \tag{12}$$

and call the corresponding performance $Z_M^{\mathrm{SPT}}(n)$, then

$$\frac{1}{M} \bar{Z}_1^*(n) \leqslant Z_M^*(n) \leqslant Z_M^{\mathrm{SPT}}(n),$$

where $\bar{Z}_1^*(n) = \bar{Z}_1^{\mathrm{SPT}}(n) = \sum_{k=1}^{n} \sum_{j=1}^{k} T_{\sigma_n(j)}$. Therefore,

$$\left| \frac{Z_M^{\mathrm{SPT}}(n) - Z_M^*(n)}{n^2} \right| \leqslant \left| \frac{Z_M^{\mathrm{SPT}}(n) - \frac{1}{M} \bar{Z}_1^*(n)}{n^2} \right| = \left| \frac{Z_M^{\mathrm{SPT}}(n) - \frac{1}{M} \sum_{k=1}^{n} \sum_{j=1}^{k} T_{\sigma_n(j)}}{n^2} \right|.$$

From Theorem 3 it then follows immediately that, as $n \to \infty$,

$$\left| \frac{Z_M^{\mathrm{SPT}}(n) - Z_M^*(n)}{n^2} \right| \Rightarrow 0 \quad \text{under condition (5),}$$

and

$$\left| \frac{Z_M^{\mathrm{SPT}}(n) - Z_M^*(n)}{n^2} \right| \to 0 \quad \text{a.s. under condition (7).}$$

LEMMA 4. *If there exists a $\delta > 0$ such that*

$$\inf_{j \geqslant 1} P\{t_{j,m} \geqslant \delta\} > 0, \tag{13}$$

then

$$\liminf_{n \to \infty} \frac{\bar{Z}_1^*(n)}{n^2} > 0 \quad \text{a.s.}$$

PROOF. For notation simplicity, let $\alpha_j := P\{T_j \geqslant \delta\}$ and $\gamma := \inf_j \alpha_j$. Clearly (13) implies $0 < \gamma \leqslant 1$.

For a given increasing sequence $a_1 \leqslant a_2 \leqslant \cdots \leqslant a_n$, suppose a_l is the first a_i larger than δ, then $n + 1 - l$ is the total number of a_is larger than δ, and we have

$$\sum_{k=1}^{n} \sum_{j=1}^{k} a_j = \sum_{j=1}^{n} (n + 1 - j) a_j \geqslant \delta \sum_{j=l}^{n} (n + 1 - j) = \frac{\delta}{2} (n + 1 - l)(n + 2 - l).$$

Applying this to the increasing sequence in (12) then gives

$$\bar{Z}_1^*(n) = \sum_{k=1}^{n} \sum_{j=1}^{k} T_{\sigma_n(j)} \geqslant \frac{\delta}{2} N_n (N_n + 1),$$

where $N_n := \sum_{j=1}^{n} I_{\{T_j \geqslant \delta\}}$.

Let $Y_j = I_{\{T_j \geqslant \delta\}} I_{\{U_j \leqslant \gamma/\alpha_j\}}$, where $\{U_j : j \geqslant 1\}$ is a sequence of independent uniform random variables, independent of the processing time processes. Note that $P(Y_j = 1) = \gamma$ for all js, therefore the Y_js are i.i.d. r.v.s. Hence, based on the SLLN,

$$\frac{N_n}{n} \geqslant \frac{\sum_{j=1}^{n} Y_j}{n} \to E[Y_j] = \gamma > 0.$$

It then follows that

$$\frac{\bar{Z}_1^*(n)}{n^2} \geqslant \frac{N_n (N_n + 1)}{2n^2} \delta \to \frac{1}{2} \gamma^2 \delta > 0,$$

as $n \to \infty$. Thus the claim in Lemma 4 follows. □

REMARK. If $\inf_{j \geqslant 1} E[t_{j,m}] > 0$ and condition (5) is assumed, condition (13) holds.

To see this, let c be the upper bound on the variances implied by (5), and let b be the assumed lower bound on the means of the $t_{j,m}$s. Then, for $0 < \varepsilon < 1$, Chebyshev's inequality implies that

$$b \leqslant E[t_{j,m}] = \int_0^{\infty} P(t_{j,m} > u)\, du = \int_0^{\varepsilon} P(t_{j,m} > u)\, du + \int_{\varepsilon}^{1/\varepsilon} P(t_{j,m} > u)\, du + \int_{1/\varepsilon}^{\infty} P(t_{j,m} > u)\, du$$

$$\leq \varepsilon + \left(\frac{1}{\varepsilon} - \varepsilon\right) P(t_{j,m} > \varepsilon) + \int_{1/\varepsilon}^{\infty} \frac{\text{var}(t_{j,m})}{u^2}\, du$$

$$\leq \left(\frac{1}{\varepsilon} - \varepsilon\right) P(t_{j,m} > \varepsilon) + (1 + c)\varepsilon.$$

Hence,

$$P(t_{j,m} > \varepsilon) \geq \frac{b - (1 + c)\varepsilon}{1/\varepsilon - \varepsilon}.$$

By choosing ε sufficiently small, we find the required positive lower bound on $P(t_{j,m} > \varepsilon)$. Examples of distributions satisfying the lower bound property are:

1. the $t_{j,m}$s are uniformly bounded away from zero, and
2. $t_{j,m} \sim \exp(\mu_j)$ with $\sup_j \mu_j < \infty$.

We then conclude with the following main result.

THEOREM 4. *For the M-machine flow shop total completion time scheduling problem, the SPT rule is asymptotically optimal in the sense that*

(i) under the conditions of Lemma 3 *and* (5),

$$\frac{Z_M^{\text{SPT}}(n) - Z_M^*(n)}{n^2} \Rightarrow 0 \quad \text{as } n \to \infty,$$

(ii) under the conditions of Lemma 3 *and* (7),

$$\frac{Z_M^{\text{SPT}}(n) - Z_M^*(n)}{n^2} \to 0 \quad \text{a.s. as } n \to \infty.$$

If, in addition, all the processing times $t_{j,m}$s are uniformly bounded away from zero, then

$$\frac{Z_M^{\text{SPT}}(n)}{Z_M^*(n)} \to 1 \quad \text{a.s. as } n \to \infty.$$

REMARK. The same argument can be applied to the *Flow Shop Weighted Completion Time Problem. In this problem*, the processing times of job j, $j = 1, \ldots, n$, are given an associated weight ω_j, with $0 < \omega_j < 1$, and the objective is to determine a permutation schedule that minimizes the total weighted completion time of all jobs on the final machine. For example, if the n jobs are processed in the nominal order $1, \ldots, n$, then the total weighted completion time is $Z = \Sigma_{j=1}^n \omega_j F_{j,M}$. Under the assumption that the processing times are i.i.d., Kaminsky and Simchi-Levi (1996a) have established the asymptotic optimality of the *Shortest Weighted Completion Time* (SWPT) first rule, which sequences the jobs in the increasing order of the weighted total processing times $\omega_j T_j$, $j = 1, \ldots, n$.

By applying the convergence theorems in §6 to the weighted sum and with Assumptions A1 and A2, we may prove theorems analogous to Theorems 3 and 4, showing that the SWPT rule is asymptotically optimal. This extends Kaminsky and Simchi-Levi (1996a) to a more general setting than i.i.d. processing times.

8. CONCLUSION

In this paper we have provided a martingale approach that proves the asymptotic optimality of the SPT rule for the *Flow Shop Average Completion Time Problem* when the processing times on the machines are i.i.d., or more generally, statistically exchangeable across machines and independent across jobs. The same argument can also be carried out to show the asymptotic optimality of the SWPT rule for the *Flow Shop Weighted Completion Time Problem.* This extends the recent results of Kaminsky and Simchi-Levi (1996a, 1996b).

END NOTES

Random variables $X_1, \ldots, X_k$ are said to be *exchangeable* if $(X_1, \ldots, X_k) \stackrel{d}{=} (X_{\sigma_k(1)}, \ldots, X_{\sigma_k(k)})$ for any deterministic permutation σ_k of $\{1, \ldots, k\}$.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Army Research Office under contract/grant DAAG 55-97-1-0377.

REFERENCES

Burkholder, D., D. Davis, R. Gundy. 1972. Integral inequalities for convex functions of operators on martingales. *Proc. Sixth Berkeley Sympos. Math. Statist. Probab., Vol. 2.* Univ. of California Press, Berkeley, CA, 223–240.

Buzacott, J., J.G. Shanthikumar. 1993. *Stochastic Models of Manufacturing Systems.* Prentice Hall, Englewood Cliffs, NJ.

Chung, K. L. 1974. *A Course in Probability Theory.* Academic Press, New York.

Conway, R. W., W. L. Maxwell, L. W. Miller. 1967. *Theory of Scheduling.* Addison-Wesley, Reading, MA.

Garey, M. R., D. S. Johnson, R. Sethi. 1976. The complexity of flowshop and jobshop scheduling. *Math. Oper. Res.* **1** 117–129.

Glynn, P. W., W. Whitt. 1991. Departures from many queues in series. *Ann. Appl. Probab.* **4** 546–572.

Hillier, F. S., R. M. Boling. 1979. On the optimal allocation of work in symmetric balanced production line systems with variable operation times. *Management Sci.* **25** 721–728.

Kaminsky, P., D. Simchi-Levi. 1996a. Probabilistic analysis and practical algorithms for the flow shop weighted completion time problem. *Oper. Res.* **46** 872–882.

——,——. 1996b. The asymptotic optimality of the SPT rule for the flow shop mean completion time problem. To appear in *Oper. Res.*

Krone, M. J., K. Steiglitz. 1974. Heuristic programming solutions of a flowshop scheduling problem. *Oper. Res.* **22** 629–638.

Muth, E. J. 1979. The reversibility property of production lines. *Management Sci.* **25** 152–158.

Pinedo, M. 1995a. Minimizing the expected makespan in stochastic flow shops. *Oper. Res.* **30** 148–162.

——. 1995b. *Scheduling: Theory, Algorithms and Systems.* Prentice Hall, Englewood Cliffs, NJ.

Shanthikumar, J. G., S. Xu. 1997. Asymptotically optimal routing and service rate allocation in a multi-server queueing system. *Oper. Res.* **45** 464–469.

Van de Velde, S. L. 1990. Minimizing the sum of the job completion times in the two-machine flow shop by Lagrangean relaxation. *Ann. Oper. Res.* **26** 257–268.

Weber, R. R. 1992. The interchangeability of tandem queues with heterogeneous customers and dependent service times. *Adv. Appl. Prob.* **24** 727–737.

Williams, D. 1991. *Probability with Martingales*. University Press, Cambridge, UK.

Xia, C. H. 1999. Dynamic scheduling of queuing systems with applications to computer networks and flexible manufacturing. Ph.D. Thesis. Stanford University.

Xia, C. H., J. G. Shanthikumar, 1998. Optimal and asymptotically optimal scheduling—an extension of the $c\mu$ rule, Working Paper, Stanford University.