Online Scheduling to Minimize Average Stretch

S. Muthukrishnan* Rajmohan Rajaraman† Anthony Shaheen‡ Johannes E. Gehrke§

Abstract

We consider the classical problem of online job scheduling on uniprocessor and multiprocessor machines. For a given job, we measure the quality of service provided by an algorithm by the **stretch** *of the job, which is defined as the ratio of the amount of time that the job spends in the system to the processing time of the job. For a given sequence of jobs, we measure the performance of an algorithm by the* **average stretch** *achieved by the algorithm over all the jobs in the sequence. The average stretch metric has been used to evaluate the performance of scheduling algorithms in many applications arising in databases, networks and systems; however, no formal analysis of scheduling algorithms is known for the average stretch metric.*

The main contribution of this paper is to show that the **shortest remaining processing time** *algorithm (*SRPT*) is* $O(1)$*-competitive with respect to average stretch for both uniprocessors as well as multiprocessors. For uniprocessors, we prove that* SRPT *is* 2*-competitive; we also establish an essentially matching lower bound on the competitive ratio of* SRPT. *For multiprocessors, we show that the competitive ratio of* SRPT *is at most* 14. *Furthermore, we establish constant-factor lower bounds on the competitive ratio of any on-line algorithm for both uniprocessors and multiprocessors.*

1 Introduction

Many servers, such as web and database servers, receive a continuous stream of requests with processing times that vary over several orders of magnitude. The main challenge in such a scenario is to schedule the stream of requests so as to provide various service guarantees such as response time, throughput, and fairness. In the combinatorial scheduling literature, this problem is typically abstracted as one of optimizing a suitably defined performance measure. Two classical performance measures are the *makespan*, which is the maximum completion time of any job, and the *total completion time*, which is the sum of the completion times of all the jobs; these are suitable for applications where the jobs are executed in batches. For jobs arriving continuously, a relevant parameter is the time that a job spends in the system, namely, the *response time* of the job, which can be defined as the difference between the completion time and release time of the job. For over two decades, the *average response time* has been a popular performance measure in online scheduling research.

Recently, alternative performance measures have been considered. The premise underlying these studies is that an important parameter of performance for a job is the *factor by which it is slowed down relative to the time it takes on a unloaded system*. Formally, the *stretch* (also known as the *slowdown*) of a job is the ratio of the response time of the job to the processing time of the job. Intuitively, the stretch measure relates the users' waiting times to their demands since it measures the quality of service provided to a job in proportion to its demand. It may also reflect users' psychological expectation that in a system with highly variable job sizes, users are willing to wait longer for larger requests. Thus it is a reasonably "fair" measure of the service that an individual job gets in the system.

The stretch measure, more precisely, the *average stretch* (or equivalently the *total stretch*) has been used to empirically study the performance of several applications, e.g., in databases [13], operating systems [9], and parallel systems [7]. Average stretch is used for measuring the performance of scheduling algorithms in the context of web servers and database servers or clusters thereof [8, 17]; The average stretch is also a good indicator of the total load of the system; a system with large average stretch is necessarily overloaded while one with a large average response time is not necessarily so (the large response time may be because of the large processing time of a job).

To summarize, average stretch is a natural measure that fits many applications, and is a better indicator of sys-

*AT&T Shannon Laboratories, Florham Park, NJ 07932. Email: `muthu@research.att.com`.

†College of Computer Science, Northeastern University, Boston MA 02115. Email: `rraj@ccs.neu.edu`. Part of this work was done when the author was at DIMACS.

‡Loyola Marymount University, Los Angeles, CA 90045. Part of this work was done when the author was at DIMACS as part of the Research Experiences for Undergraduates program.

§Department of Computer Science, University of Wisconsin, Madison, WI 53706. Email: `johannes@cs.wisc.edu`.

tem performance than the traditional average response time measure. Surprisingly, very little is known about scheduling algorithms that optimize (minimize) average stretch; in fact, the rather basic problem of minimizing average stretch on a single processor has not been studied (in contrast, optimal online algorithms are known for minimizing the average response time [2]).

In this paper, we study the problem of online scheduling to minimize average stretch on both single and multiple processor systems. Our main contribution is to show that the *shortest remaining processing time* (SRPT) heuristic is $O(1)$ competitive with respect to average stretch for both uniprocessors as well as multiprocessors. Our results show that the problem of optimizing the average stretch is fundamentally different from that of optimizing the average response time in uniprocessor as well as multiple processor case, although, interestingly, the difference is contrasting. (See Section 1.2.) Our results, taken together with previous work on analyzing average response time [2, 12], imply that SRPT simultaneously optimizes the average response time and average stretch measures, upto constant factors, for both uniprocessors and multiprocessors.

1.1 Problem and definitions

We consider the following online scheduling scenario. A sequence of jobs arrives online and the processing time of each job is known at the time of its arrival. Our goal is to produce a schedule to process the continuously arriving stream of such jobs. We focus on the *preemptive* version of scheduling; that is, jobs may be stopped before their completion and resumed later after processing other jobs in the interim. The quality of service we consider is the *stretch* of a job, which is defined as the ratio of the amount of time that the job spends in the system to the processing time of the job. Thus, if a job j with processing time $p(j)$ arrives at time $r(j)$ and is completed at time $C(j)$, then the stretch of the job is given by $s(j) = (C(j) - r(j))/p(j)$. We seek to minimize the *average stretch*, or equivalently, the total stretch, of all the jobs in a given instance.

We study the problem of minimizing average stretch for both uniprocessor as well as multiprocessor scheduling. In the case of multiprocessor scheduling, we assume that all of the processors are identical. Thus, in the standard 3-field scheduling notation [11], the problems that we study may be listed as $1 \,|\, pmtn, r(j) \,|\, \sum s(j)$ and $Pm \,|\, pmtn, r(j) \,|\, \sum s(j)$, respectively.

Since our focus is on online algorithms, we perform a competitive analysis [16] of the algorithms under consideration. An online algorithm $\mathcal{A}$ is *r-competitive* if there exists a constant c such that for every instance $\mathcal{J}$, $S(\mathcal{J})$ is at most $rS^*(\mathcal{J}) + c$, where $S(\mathcal{J})$ and $S^*(\mathcal{J})$ represent the total stretch achieved by $\mathcal{A}$ and an optimal offline algorithm, respectively, on instance $\mathcal{J}$. The *competitive ratio* of an algorithm $\mathcal{A}$ is the infimum over all r such that $\mathcal{A}$ is r-competitive.

Much of our work concerns the analysis of the well-known SRPT algorithm. For uniprocessors, SRPT is simply defined as follows: In each time step, process a job with the shortest remaining processing time among all the unfinished jobs. For an m-processor machine each time step consists of the following operation: If there are at least m unfinished jobs, a set of m jobs that have the m shortest remaining processing times are processed, where we break ties arbitrarily; otherwise, every unfinished job is processed. (Since the processors are assumed to be identical, the particular assignment of an unfinished job to a processor at any step is irrelevant [1].)

1.2 Our results

Our main results are threefold:

- [UNIPROCESSOR SRPT] We show that for any instance $\mathcal{J}$ on a uniprocessor machine, the total stretch $S(\mathcal{J})$ achieved by SRPT is at most $S^*(\mathcal{J}) + |\mathcal{J}|$; since $S^*(\mathcal{J}) \geq |\mathcal{J}|$, this implies that SRPT is 2-competitive. We also establish an essentially matching lower bound on SRPT for uniprocessors. These results appear in Section 2.

  Our upper bound for SRPT makes it an attractive algorithm for uniprocessor scheduling since it is nearly optimal with respect to average stretch while it is known to be truly optimal with respect to average response time. Indeed, recent research on scheduling algorithms for web servers [8] has shown that SRPT outperforms several other scheduling policies currently used in web servers.

- [MULTIPROCESSOR SRPT] For multiprocessors, we show that for any instance $\mathcal{J}$, the total stretch $S(\mathcal{J})$ achieved by SRPT is at most $3S^*(\mathcal{J}) + 11|\mathcal{J}|$, which implies an upper bound of 14 on the competitive ratio, independent of the number of processors $m \geq 2$. These results appear in Section 3.

  Our constant-factor competitiveness result for multiprocessors provides a surprising contrast to the recent lower bound established on the competitive ratio of any online algorithm with respect to average response time. It is shown in [12] that the best competitive ratio achievable for average response time on an m-processor machine is $\Omega(\max\{\log P, \log(n/m)\})$, where P is the ratio of the maximum to the minimum processing time of any job and n is the number of jobs. In [12], it is also shown that the competitive ratio of SRPT with respect to average response time on multiprocessors is optimal to within a

[1] A job may be preempted at one processor and later resumed at a different processor.

constant factor. This indicates that SRPT is a compelling algorithm for multiprocessor scheduling as well since it matches the best competitive ratio up to constant factors for both average stretch as well as average response time.

- [LOWER BOUNDS] We show that there exists no on-line algorithm that minimizes average stretch for every instance; specifically, we show that the competitive ratio of every online algorithm is at least 1.036. We also show a lower bound of at least 1.186 (resp., 1.093) on the competitive ratio for minimizing average stretch on any 2-processor machine (resp., $2m$-processor machine, for any positive integer m). These results appear in Section 4.

The average response time and average stretch metrics have the following interesting contrasting characteristics. For uniprocessors, while average response time can be optimized by an on-line algorithm, the best competitive ratio achievable for average stretch is strictly greater than 1. On the other hand, for multiprocessors, while a constant-factor competitive online algorithm exists for average stretch, the best competitive ratio for average response time grows logarithmically with the number of jobs.

1.3 Overview of our analyses

Our analysis of SRPT relies on a careful comparison of the queue of unfinished jobs in SRPT with the corresponding queue associated with any other scheduling algorithm at every time step. For the case of uniprocessors, we derive a tight bound on the total stretch of SRPT by placing a separate bound, for *every* unfinished job at *every* time step, of the contribution of the job to the total stretch of SRPT. A key component behind this approach is to appropriately map the jobs in SRPT's queue to the jobs in the queue associated with an arbitrary scheduling algorithm at every time step. The particular mapping enables us to analyze the total stretch of SRPT time step by time step. The relevant property of this mapping is illustrated in Figure 1 of Section 2 and formally established in Lemma 2.4.

Our analysis for multiprocessors is more involved. The difficulty in the analysis arises due to the following observation made in [12]: there exist multiprocessor scheduling instances that may force SRPT to have many more unfinished jobs at certain time steps than an alternative schedule has at that time. As a result, the contribution to the total stretch corresponding to such time steps may be disproportionately larger for SRPT than for the other schedule. We overcome this hurdle by showing the existence of an appropriate partial mapping from the jobs in SRPT's queue to the jobs in the queue associated with the other schedule. In addition to placing an upper bound on the contribution of the mapped jobs in each step, the particular mapping that enables us to place an upper bound on the exact or amortized contribution of the unmapped jobs in SRPT's queue at each time step. The mapping is illustrated in Figure 1 of Section 3 and formally established in Lemma 3.3.

1.4 Related work

While average response time has been studied extensively in the literature [2, 10, 15], the analytical study of stretch as a performance measure was initiated only recently [4]. All of the results presented in [4], however, concern the metric of *maximum stretch* (that is, $\max_i S_i$) and do not yield any useful bounds for average stretch. Recently, max-stretch was also studied in the context of performing file transfers over a network with given bandwidth constraints on the underlying links [5].

A measure closely related to average stretch is *average weighted completion time*, which equals $\sum_i w_i C(i)$, where w_i is the prespecified weight of job i. Any schedule that optimizes weighted completion time with $w_i = 1/p_i$ also optimizes total stretch. From the point of view of approximation, however, the two metrics are different. A randomized online algorithm that achieves a competitive ratio of $4/3$ with respect to average completion time is given in [14] (also see [6] for more results in this area); this result, however, does not directly yield any useful upper bound on the competitive ratio for average stretch. In fact, it is easy to construct schedules that have a constant factor approximation ratio with respect to weighted completion time, with $w_i = 1/p_i$, and yet have a linear approximation ratio with respect to average stretch.

2 Analysis of SRPT for uniprocessor scheduling

In this section, we analyze the competitiveness of SRPT on uniprocessors with respect to total stretch. An instance $\mathcal{J}$ of our scheduling problem consists of a set of jobs with arbitrary release times and processing times. For job $j \in \mathcal{J}$, let $r(j)$ and $p(j)$ denote the release time and processing time, respectively, of j. The performance of a schedule is measured by the total stretch achieved. Recall that the total stretch of a schedule for a given instance is the sum, taken over all jobs $j \in \mathcal{J}$, of the stretch of j under the schedule, where the stretch of a job is the ratio of the response time of the job to the processing time of the job. Let $S(\mathcal{J})$ and $S^*(\mathcal{J})$ denote the total stretch of SRPT and an optimal schedule, respectively, for instance $\mathcal{J}$. The main result of this section is the following theorem that places an upper bound on the total stretch achieved by SRPT.

Theorem 1 *For any instance $\mathcal{J}$, $S(\mathcal{J})$ is at most $S^*(\mathcal{J}) + |\mathcal{J}|$.*

The proof of Theorem 1 is given in Section 2.1. In terms of competitive ratio, the following corollary is immediate since the stretch of each job is at least 1.

Corollary 1.1 SRPT *is* 2*-competitive with respect to average stretch.* ■

We also show an essentially matching lower bound in Theorem 2, which is proved in Section 2.2.

Theorem 2 *For any real* $\varepsilon > 0$ *and any positive integer* $n \geq 3$*, there exists an instance* $\mathcal{J}$ *with* n *jobs such that* $S(\mathcal{J})$ *is at least* $S^*(\mathcal{J}) + n - 1 - \varepsilon$.

2.1 Upper bound

In this section, we prove Theorem 1. We begin by introducing some notation to characterize the execution of SRPT on a given instance $\mathcal{J}$. For any job j in $\mathcal{J}$ and time $t \geq r(j)$, let $\rho_t(j)$ denote the remaining processing time of job j at time t under SRPT. For any time t, let Q_t denote the set of all jobs in $\mathcal{J}$ that have been released at some time less than or equal to t and have not been completed by SRPT; that is, Q_t is the set of all jobs j for which $r(j) \leq t$ and $\rho_t(j) > 0$. We rank all of the jobs in Q_t in nondecreasing order of their remaining processing times, breaking ties according to an arbitrary ordering of the job IDs whenever necessary. For $i \geq 1$, let $q_{t,i}$ denote the job of rank i in Q_t; if the job of rank i is not defined, we set $q_{t,i}$ to be a job with processing time ∞. We refer to a job as *active* (resp., *waiting*) at time t if the rank of the job is equal to 1 (resp., greater than 1). Note that in time step t, SRPT processes one unit of the job that is active at that time. Let U_t denote the set of waiting jobs at time t.

We now express the total stretch $S(\mathcal{J})$ in SRPT as a sum, taken over all time steps t and over all jobs in Q_t, of the reciprocal of the processing time of the job.

$$S(\mathcal{J}) = \sum_{j \in \mathcal{J}} \sum_{t : j \in Q_t} \frac{1}{p(j)} = \sum_{t} \sum_{j \in Q_t} \frac{1}{p(j)}.$$

We now introduce two new variables for notational convenience. Let α_t denote the sum of the reciprocals of the processing times of the jobs in Q_t. Thus, the total stretch $S(\mathcal{J})$ is simply $\sum_t \alpha_t$. Finally, let β_t denote the sum of the reciprocals of the remaining processing times of the jobs in Q_t. Since the remaining processing time of a job at any time is at most its processing time, we obtain that $\alpha_t \leq \beta_t$ for all t.

We split the total stretch $S(\mathcal{J})$ into two parts $S_1(\mathcal{J})$ and $S_2(\mathcal{J})$, which represent the total contribution of the active jobs and the waiting jobs, respectively, in SRPT.

$$S_1(\mathcal{J}) = \sum_t \frac{1}{p(q_{t,1})} \quad \text{and} \quad S_2(\mathcal{J}) = \sum_t \sum_{j \in U_t} \frac{1}{p(j)}.$$

The total contribution of the active jobs can be calculated easily. Whenever a job j is active, it contributes $1/p(j)$ to the term $S_1(\mathcal{J})$. Since an active job is always processed by SRPT, there are exactly $p(j)$ time steps when j is active. Thus, j contributes $p(j)/p(j) = 1$ to $S_1(\mathcal{J})$. We thus obtain the following lemma.

Lemma 2.1 *For any instance* $\mathcal{J}$, $S_1(\mathcal{J}) = |\mathcal{J}|$. ■

We place an upper bound on the contribution of the waiting jobs by a careful comparison of SRPT's queue Q_t with the corresponding queue of an arbitrary scheduling algorithm $\mathcal{A}$. Just as we have characterized the execution of SRPT for a given instance $\mathcal{J}$ using the variables ρ_t, Q_t, and q_t, we can characterize the execution of $\mathcal{A}$ for $\mathcal{J}$. Let $\widetilde{Q}_t$, $\widetilde{\rho}_t$, $\widetilde{\alpha}_t$, and $\widetilde{\beta}_t$ denote the equivalent of Q_t, ρ_t, α_t, and β_t, respectively, for the execution of $\mathcal{A}$. Furthermore, whenever we henceforth introduce a new notation for a variable associated with SRPT, it is implicit that the "tilde" version of the notation denotes the corresponding variable for $\mathcal{A}$.

We obtain the desired upper bound on $S_2(\mathcal{J})$ by showing the following key lemma that relates β_t and $\widetilde{\alpha}_t$.

Lemma 2.2 *For any time t, we have* $\beta_t \leq \widetilde{\alpha}_t + \frac{1}{\rho_t(q_{t,1})}$.

Corollary 2.2.1 *For any instance* $\mathcal{J}$, $S_2(\mathcal{J}) \leq S^*(\mathcal{J})$.

Proof: Since $\alpha_t \leq \beta_t$, we obtain the following inequality by invoking Lemma 2.2.

$$\begin{aligned} S_2(\mathcal{J}) &= \sum_t \left(\alpha_t - \frac{1}{\rho_t(q_{t,1})} \right) \leq \sum_t \left(\beta_t - \frac{1}{\rho_t(q_{t,1})} \right) \\ &\leq \sum_t \widetilde{\alpha}_t \leq \widetilde{S}(\mathcal{J}). \end{aligned}$$

Since the above inequality holds for every scheduling algorithm $\mathcal{A}$, the desired claim follows. ■

Theorem 1 follows directly from Lemma 2.1 and Corollary 2.2.1. So all that remains to prove is Lemma 2.2. A crucial component of our proof of Lemma 2.2 is to establish the following relationship between the queues of SRPT and $\mathcal{A}$: for any $i > 1$, the remaining processing time of the job of rank i is *at least* the remaining processing time of the job of rank $i - 1$. The preceding claim is formally stated in Lemma 2.4 and illustrated in Figure 1.

We establish Lemma 2.4 by relating appropriately defined "prefixes" of the two queues Q_t and $\widetilde{Q}_t$. For any time t and positive integer i, let $P_t(i)$ denote the set of jobs in Q_t with remaining processing time at most i. For any time t and positive integer i, let $V_t(i)$ denote the sum of the remaining processing times of all jobs in $P_t(i)$. We call $V_t(i)$ the *volume* of jobs in Q_t with remaining processing time at most i.

Lemma 2.3 *For any time t and any positive integer i,* $V_t(i)$ *is at most* $\widetilde{V}_t(i) + i$.

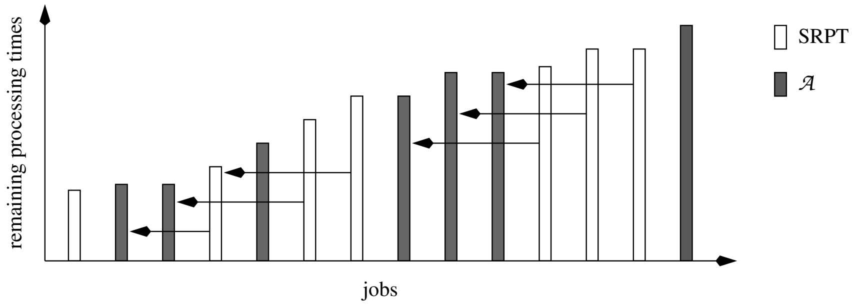

Figure 1. The remaining processing times of the jobs in the queues of SRPT and $\mathcal{A}$ at time t. The shaded jobs belong to $\mathcal{A}$'s queue $\widetilde{Q}_t$ while the unshaded jobs belong to SRPT's queue Q_t. The arrows indicate that for any $i > 1$, the remaining processing time of the job of rank i in Q_t is at least that of the job of rank $i-1$ in $\widetilde{Q}_t$.

Proof: The proof is by induction on t. The induction basis trivially holds since for all i, $V_0(i) = \widetilde{V}_0(i)$. For the induction hypothesis, we assume that the claim is true at the end of step $t-1$. We now establish the induction step for time step t. Let i be any positive integer. We first note that the release of new jobs at the start of step t does not change $V_t(i) - \widetilde{V}_t(i)$ for any i. Now let X denote the union of the set $P_{t-1}(i)$ and the set of jobs with processing time at most i that are released at time t.

We consider two cases depending on whether X is empty. In the case where X is nonempty, SRPT executes one job from X. Since $\mathcal{A}$ performs at most one unit of work, we can invoke the induction hypothesis and obtain that $V_t(i) \leq \widetilde{V}_t(i) + i$. We now consider the case where X is empty. In this case, at most one job in Q_t can have remaining processing time at most i at the end of step t. Therefore, we have $V_t(i) \leq i \leq \widetilde{V}_t(i) + i$. This completes the proof of the desired claim. ■

Lemma 2.4 *For any time t and any integer $k > 1$, we have $\rho_t(q_{t,k}) \geq \widetilde{\rho}_t(\widetilde{q}_{t,k-1})$.*

Proof: We first prove that for all $k > 1$, $\rho_t(q_{t,k}) \geq \widetilde{\rho}_t(\widetilde{q}_{t,k-1})$. The proof is by contradiction. Let, if possible, $k > 1$ be the smallest integer such that $\rho_t(q_{t,k}) < \widetilde{\rho}_t(\widetilde{q}_{t,k-1})$. Let b denote $\rho_t(q_{t,k})$. It follows that $|P_t(b)| \geq k$, while $|P_t(b)| = k-2$. We thus have the following inequality.

$$\begin{aligned} V_t(b) &= \sum_{1 \leq i \leq k} \rho_t(q_{t,i}) \\ &= \rho_t(q_{t,1}) + \rho_t(q_{t,k}) + \sum_{2 \leq i < k} \rho_t(q_{t,i}) \\ &\geq \rho_t(q_{t,1}) + b + \sum_{1 \leq i < k-1} \widetilde{\rho}_t(\widetilde{q}_{t,i}) \\ &= \rho_t(q_{t,1}) + b + \widetilde{V}_t(b) \\ &> b + \widetilde{V}_t(b), \end{aligned}$$

thus contradicting Lemma 2.3. (In the third step, we use the assumption that for $1 < i < k$, $\rho_t(q_{t,i}) \leq \widetilde{\rho}_t(\widetilde{q}_{t,i-1})$. In the final step, we use the fact that $\rho_t(q_{t,1})$ is positive.) ■

Corollary 2.4.1 *For any time t, we have $\beta_t \leq \widetilde{\beta}_t + \frac{1}{\rho_t(q_{t,1})}$.*

Proof: The desired claim follows from Lemma 2.4.

$$\begin{aligned} \beta_t - \widetilde{\beta}_t &= \sum_{k \geq 1} \left(\frac{1}{\rho_t(q_{t,k})} - \frac{1}{\widetilde{\rho}_t(\widetilde{q}_{t,k})} \right) \\ &\leq \frac{1}{\rho_t(q_{t,1})} + \sum_{k \geq 2} \left(\frac{1}{\rho_t(q_{t,k})} - \frac{1}{\widetilde{\rho}_t(\widetilde{q}_{t,k-1})} \right) \\ &\leq \frac{1}{\rho_t(q_{t,1})}, \end{aligned}$$

since $\rho_t(q_{t,k}) \geq \widetilde{\rho}_t(\widetilde{q}_{t,k-1})$ for all $k \geq 2$. ■

We are now ready to prove Lemma 2.2.

Proof of Lemma 2.2: Consider an arbitrary time step t. Let $\mathcal{B}$ be a scheduling algorithm that minimizes the sum of the reciprocals of the processing times of the jobs in the queue at time t. Let α'_t denote the sum of the reciprocals of the processing times of the jobs in $\mathcal{B}$'s queue at time t.

We now argue that $\mathcal{B}$ can be chosen such that for every job j in $\mathcal{B}$'s queue, the remaining processing time at time t is the same as the processing time $p(j)$. If not, we can define a schedule $\mathcal{C}$ that mimics $\mathcal{B}$ except that $\mathcal{C}$ never processes any of the jobs that remain in $\mathcal{B}$'s queue at time t. Clearly, $\mathcal{C}$'s queue at time t contains exactly the same jobs as $\mathcal{B}$. Moreover, none of the jobs in $\mathcal{C}$'s queue at time t have been processed prior to time t. Since the remaining processing time of every job in $\mathcal{C}$'s queue is the same as the actual processing time, we obtain the desired inequality from Lemma 2.4.1.

$$\beta_t \leq \alpha'_t + \frac{1}{\rho_t(q_{t,1})} \leq \widetilde{\alpha}_t + \frac{1}{\rho_t(q_{t,1})}.$$

■

2.2 Lower bound

In this section, we prove Theorem 2. Consider an instance in which a job of size $\ell > 1$ arrives in time step 0 and one job of size 1 arrives in each of the time steps $\ell - 1$ through $\ell + n - 3$. If SRPT does not discriminate between two jobs when they have the same remaining processing time, then we can see that an SRPT schedule may finish the job of size ℓ first and then finish all of the jobs of size 1. The total stretch of this schedule is $2(n-1)+1 = 2n-1$. An alternative schedule is to complete each of the size-1 jobs immediately upon arrival, and complete the size-ℓ job at time $n+\ell$. The total stretch of this alternative schedule is $n+\ell/n$. By setting $\ell = n/\varepsilon$, we get the desired lower bound. In the full paper, we establish the same lower bound under a more general scenario, where SRPT may discriminate between jobs with identical remaining processing times.

3 Analysis of SRPT for multiprocessor scheduling

This section analyzes the competitiveness of SRPT on an m-processor machine, where m is any integer greater than 1. For any instance $\mathcal{J}$, let $S(\mathcal{J})$ and $S^*(\mathcal{J})$ denote the total stretch of SRPT and an optimal schedule, respectively. The main result of this section is the following upper bound on the total stretch achieved by SRPT.

Theorem 3 *For any instance $\mathcal{J}$ on multiprocessors, $S(\mathcal{J})$ is at most $3S^*(\mathcal{J}) + 11|\mathcal{J}|$. Thus,* SRPT *is 14-competitive for multiprocessors.*

We can also show a constant-factor lower bound on the competitive ratio of SRPT for multiprocessors. We defer the proof of the following theorem to the full version of the paper.

Theorem 4 *For any even integer $m > 1$ and any positive real ε, there exists a scheduling instance $\mathcal{J}$ for an m-processor machine such that $S(\mathcal{J})$ is at least $(2.5 - \varepsilon)S^*(\mathcal{J})$.* ■

The remainder of this section is devoted to the proof of Theorem 3. A crucial component of our upper bound proof for uniprocessors in Section 2 is the property of SRPT's queue established in Lemma 2.4. Lemma 2.4 implies that given any scheduling algorithm $\mathcal{A}$ and any time t, we can map each job j in SRPT's queue except the one with the smallest remaining processing time to a unique job in $\mathcal{A}$'s queue at time t that has remaining processing time at most that of j. For multiprocessors, however, a similar claim does not hold. In fact, even for a two-processor system, one can construct an instance for every positive integer k such that the following claim holds: there exists a time step t when the number of jobs in SRPT's queue at time t is k while the queue in the optimal schedule at time t is empty. A nice example of such an instance is presented in [12], where it is used to derive a lower bound on the competitive ratio of SRPT with respect to total flow time.

While establishing an upper bound on the total flow of SRPT, [12] provides a characterization of the queue of SRPT that implies an upper bound on the *number of jobs* in the queue of SRPT at each step. We note that the total flow time of a schedule is the sum, taken over every time step t, of the number of jobs in the queue at time t. The analysis of [12] does not suffice for our purposes, however, because the total stretch takes into account the sum of the *reciprocals of the processing times* of the jobs in the queue at any time step. In our analysis of SRPT that follows, we perform a more careful comparison between SRPT and an optimal schedule in terms of the relative distribution of the remaining processing times of the jobs in the two queues at each time step. We are able to establish a *partial* mapping from the jobs in SRPT's queue to the jobs in another schedule such that following property of the mapping holds: the remaining processing time of a mapped job in SRPT's queue is at most that of the job to which it is mapped. This mapping, together with an appropriate characterization of the unmapped jobs in SRPT's queue, then enables us to place an upper bound on the exact or "amortized" contribution of the jobs in the queue at each time step.

As in the case of uniprocessor scheduling, we rank all of the jobs in Q_t in nondecreasing order of their remaining processing times, breaking ties according to an arbitrary ordering of the job IDs whenever necessary. Let F_t denote the set of jobs in Q_t ranked at most m. Let U_t denote the set $Q_t \setminus F_t$.

Analogous to the uniprocessor case, we split the sum $S(\mathcal{J})$ into two parts $S_1(\mathcal{J})$ and $S_2(\mathcal{J})$.

$$S_1(\mathcal{J}) = \sum_t \sum_{j \in F_t} \frac{1}{p(j)} \quad \text{and} \quad S_2(\mathcal{J}) = \sum_t \sum_{j \in U_t} \frac{1}{p(j)}.$$

By invoking the same argument that is used in proving Lemma 2.1, we show that $S_1(\mathcal{J})$ equals $|\mathcal{J}|$.

Lemma 3.1 *For any instance $\mathcal{J}$, $S_1(\mathcal{J}) = |\mathcal{J}|$.* ■

We now consider the jobs in U_t. For any time t and positive integer i, let $P_t(i)$ denote the set of jobs in Q_t with remaining processing time at most i. We define the *volume* of any subset X of Q_t as the sum of the remaining processing time of the jobs in X at time t. For any time t and positive integer i, let $V_t(i)$ denote the volume of $P_t(i)$.

The following lemma is an easy generalization of Lemma 2.3. (As in Section 2 we follow the notational convention that the "tilde" version of a variable defined for SRPT denotes the corresponding variable for $\mathcal{A}$.)

Lemma 3.2 *For any time t and any positive integer i, $V_t(i)$ is at most $\widetilde{V}_t(i) + mi$.*

Proof: The proof is by induction on t. The induction basis trivially holds since for all i, $V_0(i) = \widetilde{V}_0(i)$. For the induction hypothesis, we assume that claim is true at the end of step $t-1$. We now establish the induction step for time step t. Let i be any positive integer. We first note that the release of new jobs at the start of step does not change $V_t(i) - \widetilde{V}_t(i)$ for any i. Let X denote the union of the set $P_{t-1}(i)$ and the set of jobs with processing time at most i that are released at time t.

We consider two cases depending on whether $|X|$ is at least m. In the case where $|X|$ is at least m, all of the jobs processed by SRPT are in X. Since $\mathcal{A}$ can perform at most m units of work, we invoke the induction hypothesis and obtain that $V_t(i) \le \widetilde{V}_t(i) - mi$. We now consider the case where $|X|$ is less than m. In this case, the number of jobs not in X that are processed in step t is at most $m - |X|$. Therefore, at most $m - |X|$ jobs outside of X may have remaining processing time at most i after time step t. We thus obtain that $V_t(i) \le mi \le \widetilde{V}_t(i) + mi$. ■

Given any schedule $\mathcal{A}$ and any time t, we define a partial mapping f_t from Q_t to $\widetilde{Q}_t$ that satisfies properties P1 and P2 defined below. Let MPD_t (resp., UMPD_t) denote the set of jobs in Q_t that are mapped (resp., unmapped) by f_t.

- **(P1)** For each job j in MPD_t, $\widetilde{\rho}_t(f_t(j)) \le \rho_t(j)$.
- **(P2)** For any $i > 0$, the total volume of the jobs in $\text{UMPD}_t \cap P_t(i)$ is at most mi.

Lemma 3.3 (Partial mapping) *For any schedule $\mathcal{A}$ and any time t, there exists an injective mapping f_t from Q_t to $\widetilde{Q}_t$ that satisfies properties P1 and P2.*

Proof: We describe a procedure for defining the mapping f_t. This procedure runs in phases; in phase i, $i > 0$, we map a subset of the jobs in Q_t with remaining processing time i to jobs in $\widetilde{P}_t(i)$. While defining the mapping, we ensure that for every i, the following two properties hold at the start of phase i: (i) for each job j in Q_t that has been mapped, $\widetilde{\rho}_t(f_t(j)) \le \rho_t(j)$; (ii) for $0 < k < i$, the total volume of the unmapped jobs in $P_t(k)$ is at most mk. Clearly, if the two properties hold at the end of the procedure when all of the jobs in Q_t have been considered, then the desired claim follows.

The proof is by induction on i. For the induction basis, $i = 1$ and the properties (i) and (ii) vacuously hold. We now assume that properties (i) and (ii) hold at the start of phase i and consider the effect of phase i. Let X denote the set of jobs in $P_t(i)$ with remaining processing time equal to i. Let Y denote the set of jobs in $\widetilde{P}_t(i)$ to which no job in Q_t has been mapped at the start of phase i. In phase i, we map as many jobs in X to jobs in Y as possible. Clearly, property (i) holds at the start of phase $i+1$.

We now property (ii) for the induction step. By property (ii) of the induction hypothesis, it follows that for $0 < k < i$, the total volume of the unmapped jobs in $P_t(k)$ is at most mk. Thus, it suffices to prove that at the start of phase $i+1$, the total volume of the unmapped jobs in $P_t(i)$ is at most mi. We consider two cases. The first case is when every job in $\widetilde{P}_t(i)$ is the image of some job after phase i. By property (ii), each job in MPD_t is mapped to a job in $\widetilde{Q}_t$ with no larger remaining processing time. Therefore, it follows from Lemma 3.2 that the volume of unmapped jobs in $P_t(i)$ is at most mi. The second case is when there exists a job in $\widetilde{P}_t(i)$ to which no job is mapped at the end of phase i. In this case, we note that every job in Q_t with remaining processing time equal to i is mapped in phase i. Therefore, we invoke property (i) of the induction hypothesis to obtain that the total volume of the jobs in $P_t(i)$ that are unmapped at the end of phase i is at most $m(i-1) \le mi$. This completes the induction step and the proof of the lemma. ■

Properties P1 and P2 help in placing upper bounds on the contribution of the mapped jobs and unmapped jobs, respectively, to $S_2(\mathcal{J})$. While Lemma 3.3 holds for any schedule $\mathcal{A}$, the particular mapping f_t that we employ for the remainder of this proof is with respect a schedule $\mathcal{B}$ that minimizes the sum of the reciprocals of the processing times of the jobs in the queue at time t. Let α'_t denote the sum of the reciprocals of the processing times of the jobs in $\mathcal{B}$'s queue at time t. As in Lemma 2.2 we can assume without loss of generality that none of the jobs that belong to $\mathcal{B}$'s queue at time t have been processed until time t. Therefore, for each job j in $\mathcal{B}$'s queue at time t, the remaining processing time equals $p(j)$. By property P1 of Lemma 3.3, we know that $\rho_t(j) \ge p(f_t(j))$. Since f is injective, we obtain the following lemma about the mapped jobs in Q_t.

Lemma 3.4 (Mapped jobs) *For any time t, we have:* $\sum_{j \in \text{MPD}_t} \frac{1}{\rho_t(j)} \le \alpha'_t$. ■

We now consider the unmapped jobs. In the following lemma, we invoke an inequality established in Appendix A.

Lemma 3.5 *For any positive integer i and any step t, we have:*

$$\sum_{j \in \text{UMPD}_t, \rho_t(j) \ge i} \frac{1}{\rho_t(j)} \le \frac{2m}{i}.$$

Proof: Let X denote the set $\{j \in \text{UMPD}_t : \rho_t(j) \ge i\}$. We invoke Lemma A.1 to prove the desired claim. We note that the condition of Lemma A.1 holds due to property P2 of Lemma 3.2. The claim of Lemma A.1 is the desired inequality. ■

In order to account for the contribution of the unmapped jobs, we further classify unmapped jobs. A job j is *marked* at time t if $j \in Q_t$ and the number of mapped jobs in

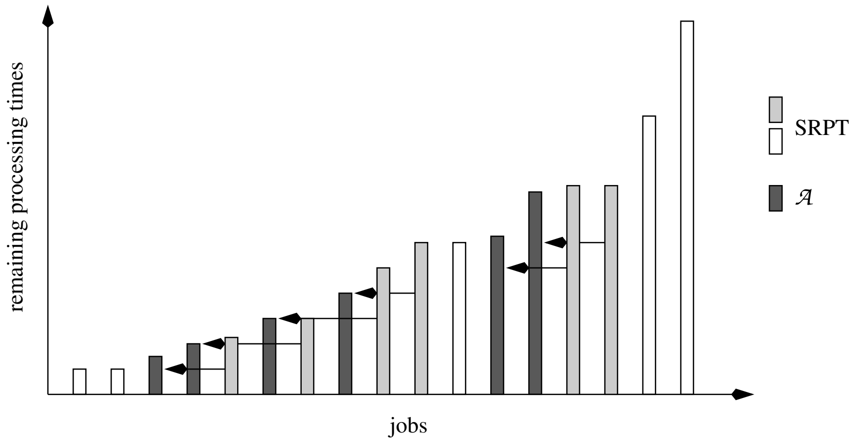

Figure 2. The remaining processing times of the jobs in the queues of SRPT and $\mathcal{A}$ at time t. The arrows depict the mapping f_t. The darkly shaded jobs belong to $\mathcal{A}$'s queue, while the lightly shaded and unshaded jobs belong to SRPT's queue. The lightly shaded jobs are the mapped jobs in SRPT's queue.

$P_t(\rho_t(j))$ is at most m. Note that whether a job is marked depends on the particular time t; thus, for example, a job may be marked at one time step and may be unmarked at a subsequent step. We call the first time when a job gets marked as the *marking time* of the job; the marking time of a job that is never marked may be set to ∞. Let MKD_t and UMKD_t denote the set of marked and unmarked jobs, respectively, at time t. We first account for the contribution of unmarked jobs by relating it to the contribution of mapped jobs of lesser rank.

Lemma 3.6 (Unmarked jobs) *For any time t, we have:*

$$\sum_{j \in \text{UMKD}_t} \frac{1}{p(j)} \leq \sum_{j \in \text{MPD}_t} \frac{2}{\rho_t(j)}.$$

Proof: For any mapped job j in Q_t, let $X(j)$ denote the set of unmarked jobs in Q_t that have rank greater than that of j. Since each job in $X(j)$ is unmapped, the following inequality follows from Lemma 3.5.

$$\sum_{j_1 \in X(j)} \frac{1}{\rho_t(j_1)} \leq \frac{2m}{\rho_t(j)} \tag{1}$$

By definition, for each unmarked job j_1, there exist at least m jobs j such that j_1 is in $X(j)$. Therefore, by adding Equation 1 over all the mapped jobs, we obtain the following inequality.

$$\sum_{j_1 \in \text{UMKD}_t} \frac{1}{p(j_1)} \leq \sum_{j_1 \in \text{UMKD}_t} \frac{1}{\rho_t(j_1)} \leq \sum_{j \in \text{MPD}_t} \frac{2}{\rho_t(j)}.$$

∎

We now consider the marked jobs. We call a marked job j *dependent* at time t, if the number of jobs in F_t with release time at least the marking time of j is greater than $m/2$. A marked job that is not dependent at time t is said to be *independent*. Let DEP_t and IND_t denote the set of dependent and independent jobs, respectively, at time t. At a high level, we account for the the contributions of the jobs in IND_t and DEP_t in the following manner. In Lemma 3.7, we show that any job j is independent during $O(p(j))$ steps only. This implies that the total contribution of the independent jobs during the execution of SRPT is $O(|\mathcal{J}|)$. Next, in Lemma 3.8, we show that the total contribution of the marked dependent jobs is within a constant factor of the contribution of the jobs in F_t, which we have already shown in Lemma 3.1 to be equal to $|\mathcal{J}|$.

Lemma 3.7 (Independent jobs) *Any job j is independent during at most $4p(j)$ time steps.*

Proof: Since any job j is independent only it is marked, we only need to consider time steps beginning the marking time of j. Let t denote the marking time of j. Let X denote the set of jobs in Q_t that have rank less than that of j. Since j is marked at time t, it follows that the number of mapped jobs in X is at most m. Moreover, by property P2 of Lemma 3.3, the volume of all of the unmapped jobs in X at time t is at most $m\rho_t(j) \leq mp(j)$. Therefore, the volume of of all the jobs in X is at most $2mp(j)$.

Consider any time step $t' \geq t$ when j is independent. It follows from the definition of independence that there are at most $m/2$ jobs in $F_{t'}$ with release time at least t. Since j is not in $F_{t'}$, it follows that there are at least $m/2$ jobs in X that are in $F_{t'}$. Thus, the volume of the jobs in X decreases by at least $m/2$ in time step t_1. Once all of the jobs in X are completed, j can be never be independent. Since the total volume of X at time t is at most $2mp(j)$, the number of time steps when j is independent is at most $4p(j)$. ■

We next consider the dependent jobs.

Lemma 3.8 (Dependent jobs) *For any time t, we have:*

$$\sum_{j \in \mathrm{DEP}_t} \frac{1}{p(j)} \leq \sum_{j \in F_t} \frac{6}{p(j)}.$$

Proof: Consider a job j in F_t. Let $X(j)$ be the set of jobs in DEP_t that were marked before the release time of j. Since every job in $X(j)$ has rank greater than m and hence greater than that of j, it follows that at the time of the release of j, the remaining processing time of every job in $X(j)$ is at least $p(j)$; that is, for any j_1 in $X(j)$, we have $\rho_{r(j)}(j_1) \geq p(j)$. Moreover, if $Y(j)$ denotes the set of jobs in $X(j)$ that have been processed at any time during the time interval $[r(j), t]$, then $|Y(j)|$ is at most $m-1$.

Consider the jobs in $X(j) \setminus Y(j)$. For every job j_1 in $X(j) \setminus Y(j)$, we have $\rho_t(j_1) = \rho_{r(j)}(j_1) \geq p(j)$. Moreover, since every job in $X(j)$ is unmapped, we obtain the following inequality from Lemma 3.5:

$$\sum_{j_1 \in X(j) \setminus Y(j)} \frac{1}{p(j_1)} \leq \frac{2m}{p(j)} \tag{2}$$

We now consider the jobs in $Y(j)$. Since there are at most $m-1$ jobs in $Y(j)$, each with processing time at least $p(j)$, we have the following inequality.

$$\sum_{j_1 \in Y(j)} \frac{1}{p(j_1)} \leq \frac{m-1}{p(j)} \tag{3}$$

Equations 2 and 3 together yield the following inequality for each job j in F_t.

$$\sum_{j_1 \in X(j)} \frac{1}{p(j_1)} \leq \frac{3m}{p(j)} \tag{4}$$

Now consider a job j_1 in DEP_t. Since j_1 is dependent, there are at least $m/2$ jobs j in F_t such that j_1 is in $X(j)$. Thus, if we add up both the left-hand and right-hand sides of Equation 4 over all jobs in F_t, j_1 appears at least $m/2$ times on the left-hand side. Therefore, we obtain the following inequality, that establishes the desired claim.

$$\sum_{j_1 \in \mathrm{DEP}_t} \frac{1}{p(j_1)} \leq \sum_{j \in F_t} \frac{6}{p(j)}$$

■

Putting Lemmas 3.4, 3.7, 3.8, and 3.6 together, we obtain the following main lemma.

Lemma 3.9 *For any instance $\mathcal{J}$, we have:*

$$\sum_t \sum_{j \in U_t} \frac{1}{p(j)} \leq 3S^*(\mathcal{J}) + 10|\mathcal{J}|.$$

Proof: We establish the desired inequality as follows.

$$\begin{aligned}
\sum_t \sum_{j \in U_t} \frac{1}{p(j)} &\leq \sum_t \Bigg(\sum_{j \in \mathrm{MPD}_t} \frac{1}{p(j)} + \sum_{j \in \mathrm{IND}_t} \frac{1}{p(j)} \\
&\quad + \sum_{j \in \mathrm{DEP}_t} \frac{1}{p(j)} + \sum_{j \in \mathrm{UMKD}_t} \frac{1}{p(j)} \Bigg) \\
&\leq \sum_t \sum_{j \in \mathrm{MPD}_t} \frac{3}{\rho_t(j)} + \sum_j \sum_{t : j \in \mathrm{IND}_t} \frac{1}{p(j)} \\
&\quad + \sum_t \sum_{j \in F_t} \frac{6}{p(j)} \\
&\leq \sum_t 3\alpha'_t + \sum_t \sum_{j \in F_t} \frac{10}{p(j)} \\
&\leq 3S^*(\mathcal{J}) + 10|\mathcal{J}|
\end{aligned}$$

(In the second step, we invoke Lemmas 3.8 and 3.6. In the third step, we invoke Lemmas 3.4 and 3.7. The final step follows from Lemma 3.1.) ■

The main theorem now directly follows from Lemmas 3.1 and 3.9.

4 Lower bounds

We first establish a lower bound on the competitive ratio of any online uniprocessor scheduling algorithm with respect to average stretch.

Theorem 5 *If there are only two distinct job sizes, there exists an optimal online algorithm for minimizing average stretch. If there are three or more distinct job sizes, the competitive ratio of any online algorithm is at least* 1.036.

Proof: Consider the case when there are only two distinct job sizes. We first observe that there is an optimal schedule in which all large (respectively, small) jobs are processed in First-in-First-out (FIFO) order. Thus at any time t, there is at most one partially processed job of either category, and that is the job at the head of the FIFO order in that category at that time; let ℓ and s denote the job at the head of the FIFO order in the large and small categories, respectively. We need to decide which one of these two jobs must be processed. If the large job is completed first, followed by the small job, then the total stretch contributed by the two jobs is $(t+\rho_t(\ell)-r(\ell))/p(\ell)+(t+\rho_t(\ell)+\rho_t(s)-r(s))/p(s)$, while if the small job is completed first, followed by the the large job, then the stretch contributed by the two jobs is $(t+\rho_t(s)-r(s))/p(s)+(t+\rho_t(\ell)+\rho_t(s)-r(\ell))/p(\ell)$. The former is desirable when $\rho_t(\ell)/p(s) \leq \rho_t(s)/p(\ell)$, that is, when $\rho_t(\ell)p(\ell) \leq \rho_t(s)p(s)$. Furthermore, arrivals of new large or small jobs in the interim does not affect the stretch of jobs ℓ and s. Therefore, the optimal algorithm is one that in each time step t processes a job i with the smallest $\rho_t(i)p(i)$.

We next consider the lower bound when there are three or more distinct job sizes. We first give the proof of a weaker claim first. Consider three jobs j_1, j_2, and j_3, with release times 0, 4, and 5, respectively, and processing times 5, 2, and 1, respectively. Let instance $\mathcal{J}_1$ be the set $\{j_1, j_2\}$ and instance $\mathcal{J}_2$ be the set $\{j_1, j_2, j_3\}$. Given $\mathcal{J}_1$, the optimal solution is to process j_1 for 4 steps, then process j_2 for 2 steps and then complete j_1. The optimal total stretch is $1 + 7/5 = 2.4$. Given $\mathcal{J}_2$, the optimal solution is to process j_1 for 5 steps and thus complete it, then process j_3 for 1 step, and finally, process j_2. The optimal stretch value for this instance is 4. Given instance $\mathcal{J}_1$, if the given on-line algorithm completes j_1 before j_2, then its stretch is at least 2.5, implying a competitive ratio of at least 1.04. If, on the other hand, the algorithm preempts j_1 at time 4 (i.e., at the start of the fifth step), then the two best scenarios for the algorithm with regard to instance $\mathcal{J}_2$ is to finish the jobs in one of the following orders: (i) j_2, j_3, j_1 or (ii) j_3, j_2, j_1. The stretch for case (i) is 4.6 and for case (ii) is 4.1. Therefore, the competitive ratio is at least 1.025.

We can generalize the above example as follows. Let ℓ, m, and s be the jobs with processing times $p(\ell)$, $p(m)$, and $p(s)$, where $p(\ell) > p(m) > p(s)$. Consider the instance in which ℓ arrives at time 0 and m arrives at the end of time step $p(\ell) - k$, where $k \leq p(m)$. If the algorithm does not preempt the first job when the second job arrives, the competitive ratio is at least $(2 + (k/p(m)))/(2 + (p(m)/p(\ell)))$. Otherwise, job s is presented at the end of time step $p(\ell)$. Then we can argue as above that any such algorithm has average stretch at least $\min\{3 + (p(s)/p(m)) + (p(m) + p(s))/p(\ell), 3 + (p(m) + p(s))/p(\ell) + (p(m) - k)/p(s)\}$ while the optimum average stretch is at most $3 + (k + p(s))/p_m$. By choosing the parameters $p(\ell)$, $p(m)$, and $p(s)$ appropriately, we can now conclude that every algorithm has a competitive ratio of at least 1.036. ∎

For multiprocessors, we can establish a slightly stronger lower bound on the competitive ratio of any online algorithm. The following theorem is proved in the full paper.

Theorem 6 *The competitive ratio of any online algorithm for a 2-processor machine is at least 1.186. For any $m > 2$, the competitive ratio of any online algorithm for a $2m$-processor machine is at least 1.093.* ∎

5 Discussion

We have studied the online complexity of scheduling to minimize average stretch. Our results show that SRPT is attractive for both uniprocessor and multiprocessor cases since it simultaneously gives the best known performance for average response time as well as average stretch. Any algorithm that optimizes the average metric (response time or stretch), and SRPT in particular, may starve jobs on worst case inputs. However, for certain applications such as serving web traffic, it appears that the adverse effect of starvations in SRPT is limited [8]. In multiprocessors, a potential drawback of SRPT is the overhead it incurs due to migrations. Recall that SRPT may preempt a job at one processor and later resume the job at a different processor. A natural variant of SRPT that does not perform any migrations has been proposed in [1]. In recent work [3], it has been shown that the "migrationless" scheduling algorithm of [1] also achieves a constant-factor competitive ratio with respect to average stretch.

Many interesting problems remain open. In particular, the lower bounds may quite possibly be strengthened. Also, it will be of interest to analyze the algorithm that schedules the job i with the smallest $p(i)\rho_t(i)$ at time t, for multiple job sizes; for two job sizes, this is optimal as we showed (see Section 4). The problem of scheduling to minimize average stretch may also be studied in other models, such as when preemption is not allowed.

We have also considered the offline complexity of our problem. In the full paper, we establish the NP-hardness of the problem of minimizing average stretch on multiprocessors. On the other hand, our analysis of SRPT immediately gives an $O(1)$ approximation. It would be interesting to determine whether there exists a PTAS for this problem. Also of interest is the offline complexity for the uniprocessor case which we have left open.

References

[1] B. Awerbuch, Y. Azar, S. Leonardi, and O. Regev. Minimizing flow time without migration. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, pages 198–205, May 1999.

[2] K. R. Baker. *Introduction to Sequencing and Scheduling*. Wiley, 1974.

[3] L. Becchetti, S. Leonardi, and S. Muthukrishnan. Scheduling to minimize average stretch without migration. Manuscript, July 1999.

[4] M. Bender, S. Chakrabarti, and S. Muthukrishnan. Flow and stretch metrics for scheduling continuous job streams. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 270–279, January 1998.

[5] A. Goel, M. Henzinger, S. Plotkin, and E. Tardos. Scheduling data transfers in a network and the set scheduling problem. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, Atlanta, Georgia, May 1999. To appear.

[6] L. Hall, A. Schulz, D. Shmoys, and J. Wein. Scheduling to minimize average completion time: Off-line and on-line algorithms. *Mathematics of Operations Research*, 22:513–544, 1997.

[7] M. Harchol-Balter, M. Crovella, and C. Murta. Task assignment in a distributed server. In *Proceedings of the 10th Inter-*

national Conference on Modeling Techniques and Tools for Computer Performance Evaluation, September 1998.

[8] M. Harchol-Balter, M. Crovella, and S. Park. The case for SRPT scheduling in web servers. Technical Report MIT–LCS–TR–767, Laboratory for Computer Science, Massachussets Institute of Technology, October 1998.

[9] R. Jain. *The Art of Computer Systems Performance Analysis*. John Wiley, New York, 1991.

[10] D. Karger, C. Stein, and J. Wein. Scheduling algorithms. In M. Atallah, editor, *Handbook of Algorithms and Theory of Computation*. CRC Press. To appear.

[11] E. Lawler, J. Lenstra, A. Rinnooy Kan, and D. Shmoys. Sequencing and scheduling: Algorithms and complexity. In S. Graves, A. Rinnooy Kan, and P. Zipkin, editors, *Handbooks in Operations Research and Management Science, Vol. 4: Logistics of Production and Inventory*, pages 445–522. North-Holland, 1993.

[12] S. Leonardi and D. Raz. Approximating total flow time on parallel machines. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, pages 110–119, May 1997.

[13] M. Mehta and D. J. DeWitt. Dynamic memory allocation for multiple-query workloads. In *Proceedings of the 19th International Conference on Very Large Data Bases*, pages 354–367, August 1993.

[14] A. Schulz and M. Skutella. Scheduling-LPs bear probabilities: Randomized approximations for min-sum criteria. Technical Report No. 533/1996, Technische Universität Berlin, November 1996. Revised March 1997.

[15] J. Sgall. On-line scheduling – a survey. In A. Fiat and G. Woeginger, editors, *On-Line Algorithms*. Springer-Verlag, Berlin, 1997.

[16] D. D. Sleator and R. E. Tarjan. Amortized efficiency of list update and paging rules. *Communications of the ACM*, 28(2):202–208, 1985.

[17] H. Zhu, B. Smith, and T. Yang. A scheduling framework for web server clusters with intensive dynamic content processing. Technical Report TRCS98-29, Department of Computer Science, University of California, Santa Barbara, November 1998.

A A technical inequality

Lemma A.1 *Let m be a positive integer. Let X be a multiset of positive reals that satisfies the following condition: For any positive real x the sum of the elements in X with value at most x is at most mx. Then the following inequality holds true for any positive real z.*

$$\sum_{x \in X, x \geq z} \frac{1}{x} \leq \frac{2m}{z}.$$

Proof: Let $v(X)$ denote the term $\sum_{x \in X, x \geq z} \frac{1}{x}$. Consider the following greedy algorithm for constructing X with the goal of maximizing $v(X)$. Initially set X to $\emptyset$. Iteratively, add to X the smallest element greater than or equal to z that can be added without violating the condition on X as stated in the lemma. Let x_i be the ith element added to X. We claim that that x_i may be defined as follows:

$$x_i = \begin{cases} z & \text{if } 1 \leq i \leq m, \\ \frac{\sum_{1 \leq j < i} x_j}{m-1} & \text{otherwise.} \end{cases}$$

We first prove the following two claims by induction on i: (i) $x_i \geq x_j$ for $1 \leq j < i$, and (ii) the multiset $\{x_j : 1 \leq j \leq i\}$ is a valid solution for X. For the base case, we consider $1 \leq i \leq m$. Since $x_i = z$, the desired claim holds trivially for the base case. We now assume that the desired claim holds for $i-1$. Consider iteration i. The element x_i equals $(\sum_{1 \leq j < i} x_j)/(m-1)$, which is at least $(\sum_{1 \leq j < i-1} x_j)/(m-1)$ and hence at least x_{i-1}, thus establishing part (i) of the induction step. For part (ii) of the induction step, we note that part (i) implies that we only need to check whether $\sum_{1 \leq j \leq i} x_j \leq m x_i$. By the definition of x_i it follows that $\sum_{1 \leq j \leq i} x_j = m x_i$, thus ensuring that the multiset $\{x_j : 1 \leq j \leq i\}$ is a valid solution for X. This completes the proofs of claims (i) and (ii).

We next show that $X^* = \{x_j : j \geq 1\}$ is an optimal choice for X. Let, if possible, Y be a different optimal solution. Let y_i denote the ith smallest element of Y, where we break ties arbitrarily. Let i be the smallest integer such that $y_i \neq x_i$. We now derive a contradiction in each of the following two cases: $y_i > x_i$ and $y_i < x_i$. If $y_i > x_i$, we can replace the element y_i in Y by the element x_i and obtain a valid solution yielding a value strictly larger than $v(Y)$, thus contradicting the assumption that $v(Y)$ is optimal. If $y_i < x_i$, then since Y is a valid solution, so is the multiset $\{y_j : 1 \leq j \leq i\}$. This implies the following inequality.

$$\begin{aligned} \sum_{1 \leq j \leq i} y_j &= y_i + \sum_{1 \leq j < i} x_j \\ &= y_i + (m-1)x_i \\ &> y_i + (m-1)y_i \\ &= m y_i, \end{aligned}$$

thus violating the condition stated in the lemma.

To complete the proof of the lemma, we now compute $v(X^*)$. We first note that for $i > m$, x_i is equal to $z(m/m-1)^{i-m}$. Therefore, we can calculate $v(X^*)$ as follows.

$$\begin{aligned} v(X^*) &= \frac{m}{z} + \sum_{i>m} \frac{1}{z} \left(\frac{m-1}{m} \right)^{i-m} \\ &= \frac{m}{z} + \sum_{j>0} \frac{1}{z} \left(\frac{m-1}{m} \right)^{j} \\ &= \frac{2m}{z}. \end{aligned}$$

■