Selective preemption strategies for parallel job scheduling

Rajkumar Kettimuthu*
Argonne National Laboratory, Argonne, IL 60439, USA
The University of Chicago, Chicago, IL 60615, USA
E-mail: kettimut@mcs.anl.gov
*Corresponding author

Vijay Subramani and Srividya Srinivasan
Microsoft Corporation, Redmond, WA 98052, USA
E-mail: vijays@microsoft.com E-mail: srivis@microsoft.com

Thiagaraja Gopalsamy
Altera Corporation, San Jose, CA 95134, USA
E-mail: tgopalsa@altera.com

D.K. Panda and P. Sadayappan
The Ohio State University, Columbus, OH 43210, USA
E-mail: panda@cse.ohio-state.edu E-mail: saday@cse.ohio-state.edu

Abstract: Although theoretical results have been established regarding the utility of preemptive scheduling in reducing average job turnaround time, job suspension/restart is not much used in practice at supercomputer centres for parallel job scheduling. A number of questions remain unanswered regarding the practical utility of preemptive scheduling. We explore this issue through a simulation based study, using real job logs from supercomputer centres. We develop a tuneable selective suspension strategy and demonstrate its effectiveness. We also present new insights into the effect of preemptive scheduling on different job classes and deal with the impact of suspensions on worst case response time. Further, we analyse the performance of the proposed schemes under different load conditions.

Keywords: preemptive scheduling; parallel job scheduling; backfilling.

Reference to this paper should be made as follows: Kettimuthu, R., Subramani, V., Srinivasan, S., Gopalsamy, T., Panda, D.K. and Sadayappan, P. (2005) 'Selective preemption strategies for parallel job scheduling', *Int. J. High Performance Computing and Networking*, Vol. 3, Nos. 2/3, pp.122–152.

Biographical notes: Rajkumar Kettimuthu is a Researcher at The University of Chicago's Computation Institute and a Resident Associate at Argonne National Laboratory's Mathematics and Computer Science Division. His research interests include Data Transport in High-bandwidth and High-delay Networks and Scheduling and Resource Management for Cluster Computing and the Grid. He has a Bachelor of Engineering degree in Computer Science and Engineering from Anna University, Madras, India, and a Master of Science in Computer and Information Science from the Ohio State University.

Vijay Subramani received his BE in Computer Science and Engineering from Anna University, India, in 2000 and his MS in Computer and Information Science from the Ohio State University in 2002. His research at Ohio State included Scheduling and Resource Management for Parallel and Distributed Systems. He currently works at Microsoft Corporation in Redmond, WA. His past work experience includes an internship at Los Alamos National Laboratory, where he worked on Buffered Coscheduling.

Srividya Srinivasan currently works as a Software Engineer at Microsoft Corporation in Redmond, WA. She worked as a Software Developer at Bloomberg LP in New York earlier. She received a BE degree in Computer Science and Engineering from Anna University, Chennai, India, in 2000 and her MS in Computer and Information Science from the Ohio State University in 2002. Her research at Ohio State focused on Parallel and Distributed Systems with emphasis on Parallel Job Scheduling.

Thiagaraja Gopalsamy is a Senior Software Engineer with Altera Corporation, San Jose. He received his Bachelor's degree in Computer Science and Engineering in 1999 from Anna University, India and his Master's degree in Computer and Information Science in 2001 from the Ohio State University. His past research interests include Mobile Ad hoc Networks and Parallel Computing. He is currently working on Field Programmable Gate Arrays and Reconfigurable Computing.

D.K. Panda is a Professor of Computer Science at the Ohio State University. His research interests include Parallel Computer Architecture, High Performance Networking, and Network Based Computing. He has published over 150 papers in these areas. His research group is currently collaborating with national laboratories and leading companies on designing various communication and I/O subsystems of next generation HPC systems and datacentres with modern interconnects. The MVAPICH (MPI over VAPI for InfiniBand) package developed by his research group (http://nowlab.cis.ohio-state.edu/projects/mpi-iba/) is being used by more than 160 organisations worldwide to extract the potential of InfiniBand-based clusters for HPC applications. Dr. Panda is a recipient of the NSF CAREER Award, OSU Lumley Research Award (1997 and 2001), and an Ameritech Faculty Fellow Award. He is a senior member of IEEE Computer Society and a member of ACM.

P. Sadayappan received his BTech degree from the Indian Institute of Technology, Madras, India, and an MS and PhD from the State University of New York at Stony Brook, all in Electrical Engineering. He is currently a Professor in the Department of Computer Science and Engineering at the Ohio State University. His research interests include Scheduling and Resource Management for Parallel/Distributed Systems and Compiler/Runtime Support for High Performance Computing.

1 INTRODUCTION

Although theoretical results have been established regarding the effectiveness of preemptive scheduling strategies in reducing average job turnaround time (DasGupta and Palis, 2000; Deng and Dymond, 1996; Deng et al., 1996; Epstein, 2001; Schwiegelshohn and Yahyapour, 2000), preemptive scheduling is not currently used for scheduling parallel jobs at supercomputer centres. Compared to the large number of studies that have investigated nonpreemptive scheduling of parallel jobs (Anastasiadis and Sevcik, 1997; Cirne and Berman, 2000; Feitelson, 2002; Jones and Nitzberg, 1999; Ward, et al., 2002; Lawson and Smirni, 2002; Lawson et al., 2002; Lifka, 1995; Mu'alem and Feitelson, 2001; Sabin et al., 2003; Srinivasan et al., 2002a, 2002c; Subramani et al., 2002a, 2002b; Talby and Feitelson, 1999; Zotkin and Keleher, 1999), little research has been reported on evaluation of preemptive scheduling strategies using real job logs (Chiang et al., 1994; Chiang and Vernon, 2001; Leutenneger and Vernon, 1990; Parsons and Sevcik, 1997). The basic idea behind preemptive scheduling is simple: If a long running job is temporarily suspended and a waiting short job is allowed to run to completion first, the wait time of the short job is significantly decreased, without much fractional increase in the turnaround time of the long job. Consider a long job with run time T_l. After time t, let a short job arrive with run time T_s. If the short job were to run after completion of the long job, the average job turnaround time would be $((T_l + (T_l + T_s - t))/2)$, or $T_l + ((T_s - t)/2)$. Instead, if the long job were suspended when the short job arrived, the turnaround times of the short and long jobs would be T_s and $(T_s + T_l)$, respectively, giving an average of $T_s + (T_l/2)$. The average turnaround time with suspension is less if $T_s < T_l - t$, that is, the remaining run time of the running job is greater than the run time of the waiting job.

The suspension criterion has to be chosen carefully to ensure freedom from starvation. Also, the suspension scheme should bring down the average turnaround times without increasing the worst case turnaround times. Even though theoretical results (DasGupta and Palis, 2000; Deng and Dymond, 1996; Deng et al., 1996; Epstein, 2001; Schwiegelshohn and Yahyapour, 2000) have established that preemption improves the average turnaround time, it is important to perform evaluations of preemptive scheduling schemes using realistic job mixes derived from actual job logs from supercomputer centres, to understand the effect of suspension on various categories of jobs.

The primary contributions of this work are as follows:

- development of a 'selective suspension' strategy for preemptive scheduling of parallel jobs
- characterisation of the significant variability in the average job turnaround time for different job categories
- demonstration of the impact of suspension on the worst case turnaround times of various categories, and development of a tuneable scheme to improve worst case turnaround times.

This paper is organised as follows. Section 2 provides background on parallel job scheduling and discusses prior work on preemptive job scheduling. Section 3 characterises the workload used for the simulations. Section 4 presents the proposed selective preemption strategies and evaluates their performance under the assumption of accurate estimation of job run times. Section 5 studies the impact of inaccuracies in user estimates of run time on the selective preemption strategies. It also models the overhead for job suspension and restart and evaluates the proposed schemes in the presence of overhead. Section 6 describes the performance of the selective preemption strategies under different load conditions. Section 7 summarises the results of this work.

2 BACKGROUND AND RELATED WORK

Scheduling of parallel jobs is usually viewed in terms of a 2D chart with time along one axis and the number of processors along the other axis. Each job can be thought of as a rectangle whose width is the user estimated run time and height is the number of processors requested. Parallel job scheduling strategies have been widely studied in the past (Aida, 2000; Arndt et al., 2000; Cirne, 2003; Perkovic and Keleher, 2000; Keleher et al., 2000; Krallmann et al., 1999; Srinivasan et al., 2002b; Streit, 2001). The simplest way to schedule jobs is to use the first come, first served, (FCFS) policy. This approach suffers from low system utilisation, however, because of fragmentation of the available processors. Consider the scenario where a few jobs are running in the system and many processors are idle, but the next queued job requires all the processors in the system. An FCFS scheduler would leave the free processors idle even if there were waiting queued jobs requiring only a few processors. Some solutions to this problem are to use dynamic partitioning (McCann et al., 1993) or gang scheduling (Feitelson and Jette, 1997). An alternative approach to improve the system utilisation is backfilling.

2.1 Backfilling

Backfilling was developed for the IBM SP1 parallel supercomputer as part of the Extensible Argonne Scheduling System (EASY) (Lifka, 1995) and has been implemented in several production schedulers (Jackson et al., 2001; Skovira et al., 1996). Backfilling works by identifying ‘holes’ in the 2D schedule and moving forward, smaller jobs that fit those holes. With backfilling, users are required to provide an estimate of the length of the jobs submitted for execution. This information is used by the scheduler to predict when the next queued job will be able to run. Thus, a scheduler can determine whether a job is sufficiently small to run without delaying any previously reserved jobs.

It is desirable that a scheduler with backfilling support two conflicting goals. On the one hand, it is important to move forward as many short jobs as possible in order to improve utilisation and responsiveness. On the other hand, it is also important to avoid starvation of large jobs and, in particular, to be able to predict when each job will run. There are two common variants to backfilling – conservative and aggressive (EASY) – that attempt to balance these goals in different ways.

2.1.1 Conservative backfilling

With conservative backfilling, every job is given a reservation (start time guarantee) when it enters the system. A smaller job is allowed to backfill only if it does not delay any previously queued job. Thus, when a new job arrives, the following allocation procedure is executed by a conservative backfilling scheduler. Based on the current knowledge of the system state, the scheduler finds the earliest time at which a sufficient number of processors are available to run the job for a duration equal to the user estimated run time. This is called the ‘anchor point’. The scheduler then updates the system state to reflect the allocation of processors to this job starting from its anchor point. If the job’s anchor point is the current time, the job is started immediately.

An example is given in Figure 1. The first job in the queue does not have enough processors to run. Hence, a reservation is made for it at the anticipated termination time of the longer running job. Similarly, the second queued job is given a reservation at the anticipated termination time of the first queued job. Although enough processors are available for the third queued job to start immediately, it would delay the second job; therefore, the third job is given a reservation after the second queued job’s anticipated termination time.

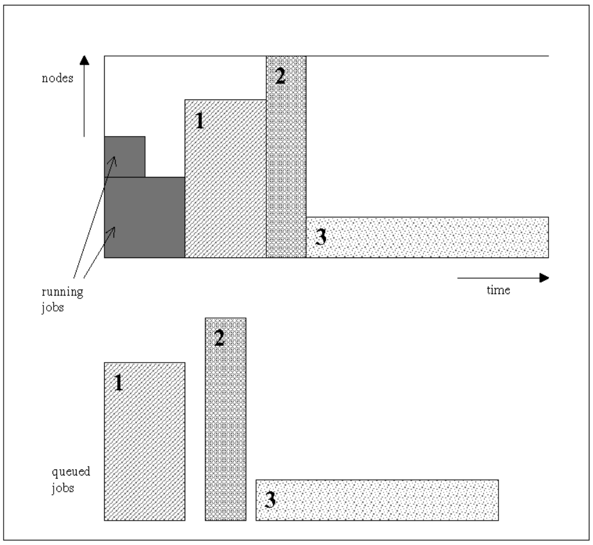

Figure 1 *Conservative backfilling*

Thus, in conservative backfilling, jobs are assigned a start time when they are submitted, based on the current usage profile. But they may actually be able to run sooner if previous jobs terminate earlier than expected. In this scenario, the original schedule is compressed by releasing the existing reservations one by one, when a running job terminates, in the order of increasing reservation start time guarantees and attempting backfill for the released job. If as a result of early termination of some job, 'holes' of the right size are created for a job, then it gets an earlier reservation. In the worst case, each released job is reinserted in the same position it held previously. With this scheme, there is no danger of starvation, since a reservation is made for each job when it is submitted.

2.1.2 Aggressive backfilling

Conservative backfilling moves jobs forward only if they do not delay any previously queued job. Aggressive backfilling takes a more aggressive approach and allows jobs to skip ahead, provided they do not delay the job at the head of the queue. The objective is to improve the current utilisation as much as possible, subject to some consideration for the queue order. The price is that execution guarantees cannot be made, because it is impossible to predict how much each job will be delayed in the queue.

An aggressive backfilling scheduler scans the queue of waiting jobs and allocates processors as requested. The scheduler gives a reservation guarantee to the first job in the queue that does not have enough processors to start. This reservation is given at the earliest time at which the required processors are expected to become free, based on the current system state. The scheduler then attempts to backfill the other queued jobs. To be eligible for backfilling, a job must require no more than the currently available processors and must satisfy either of two conditions that guarantee it will not delay the first job in the queue:

- it must terminate by the time the first queued job is scheduled to commence
- it must use no more nodes than those that are free at the time the first queued job is scheduled to start.

Figure 2 shows an example.

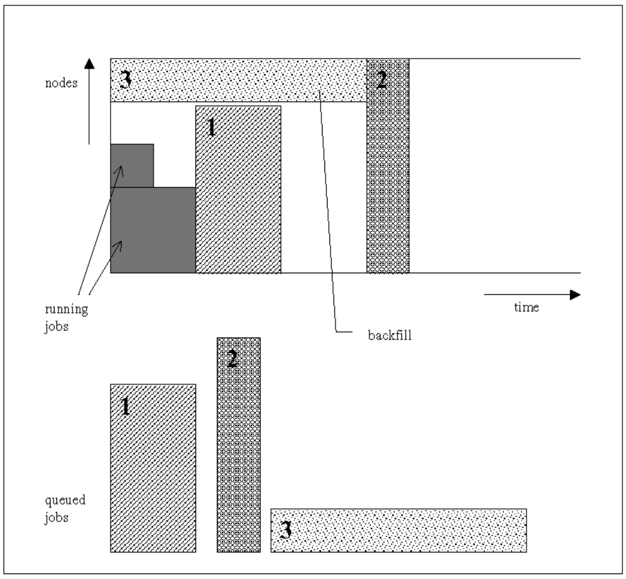

Figure 2 *Aggressive backfilling*

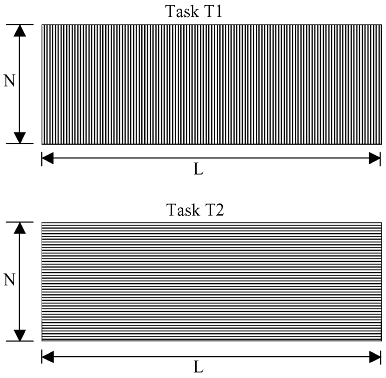

Figure 3 *Two simultaneously submitted tasks T_1 and T_2, each requiring 'N' processors for 'L' seconds*

2.2 Metrics

Two common metrics used to evaluate the performance of scheduling schemes are the average turnaround time and the average bounded slowdown. We use these metrics for our studies. The bounded slowdown (Feitelson et al., 1997) of a job is defined as follows:

$$\text{Bounded slowdown} = (\text{Wait time} + \text{Max}(\text{Run time},10)) / \text{Max}(\text{Run time},10). \quad (1)$$

The threshold of 10 seconds is used to limit the influence of very short jobs on the metric.

Preemptive scheduling aims at providing lower delay to short jobs relative to long jobs. Since long jobs have greater tolerance to delays as compared to short jobs, our suspension criterion is based on the expansion factor (xfactor), which increases rapidly for short jobs and gradually for long jobs.

$$\text{xfactor} = (\text{Wait time} + \text{Estimated run time})/\text{Estimated run time}. \quad (2)$$

2.3 Related work

Although preemptive scheduling is universally used at the operating system level to multiplex processes on single processor systems and shared memory multiprocessors, it is rarely used in parallel job scheduling. A large number of studies have addressed the problem of parallel job scheduling (see Feitelson et al. (1997) for a survey of work on this topic), but most of them address nonpreemptive scheduling strategies. Further, most of the work on preemptive scheduling of parallel jobs considers the jobs to be malleable (Deng et al., 1996; Parsons and Sevcik, 1997; Sevcik, 1994; Zahorjan and McCann, 1990); in other words, the number of processors used to execute the job is permitted to vary dynamically over time.

In practice, parallel jobs submitted to supercomputer centres are generally rigid; that is, the number of processors used to execute a job is fixed. Under this scenario, the various schemes proposed for a malleable job model are inapplicable. Few studies have addressed preemptive scheduling under a model of rigid jobs, where the preemption is 'local', that is, the suspended job must be restarted on exactly the same set of processors on which they were suspended.

Chiang and Vernon (2001) evaluate a preemptive scheduling strategy called 'immediate service (IS)' for shared memory systems. With this strategy, each arriving job is given an immediate timeslice of 10 minutes, by suspending one or more running jobs if needed. The selection of jobs for suspension is based on their instantaneous-xfactor, defined as 'wait time + total accumulated run time'/ 'total accumulated run time'. Jobs with the lowest instantaneous-xfactor are suspended. The IS strategy significantly decreases the average job slowdown for the traces simulated. A potential shortcoming of the IS strategy, however, is that its preemption decisions do not reflect the expected run time of a job. The IS strategy can be expected to significantly improve the slowdown of aborted jobs in the trace. Hence, it is unclear how much, if any, of the improvement in slowdown is experienced by the jobs that completed normally. However, no information is provided on how different job categories are affected.

Chiang et al. (1994) examine the 'run to completion' policy with a suspension policy that allows a job to be suspended, at most, once. Both this approach and the IS strategy, limit the number of suspensions; whereas we use a 'suspension factor' to control the rate of suspensions, without limiting the number of times a job can be suspended.

Parsons and Sevcik (1997) discuss the design and implementation of a number of multiprocessor preemptive scheduling disciplines. They study the effect of preemption under the models of rigid, migratable and malleable jobs. They conclude that their proposed preemption scheme may increase the response time for the model of rigid jobs.

So far, few simulation based studies have been done on preemption strategies for clusters. With no process migration, the distributed memory systems impose an additional constraint that a suspended job should get the same set of processors when it restarts. In this paper, we propose tuneable suspension strategies for parallel job scheduling in environments where process migration is not feasible.

3 WORKLOAD CHARACTERISATION

We perform simulation studies using a locally developed simulator with workload logs from different supercomputer centres. Most supercomputer centres keep a trace file as a record of the scheduling events that occur in the system. This file contains information about each job submitted and its actual execution. Typically the following data are recorded for each job:

- name of job, user name, and so forth
- job submission time
- job resources requested, such as memory and processors
- user estimated run time
- time when job started execution
- time when job finished execution.

From the collection of workload logs available from Feitelson's (2001) archive, subsets of the CTC workload trace, the SDSC workload trace and the KTH workload trace were used to evaluate the various schemes. The CTC trace was logged from a 430 node IBM SP2 system at the Cornell Theory Center, the SDSC trace from a 128 node IBM SP2 system at the San Diego Supercomputer Center, and the KTH trace from a 100 node IBM SP2 system at the Swedish Royal Institute of Technology. The other traces did not contain user estimates of run time. We observed similar performance trends with all the three traces. In order to minimise the number of graphs, we report the performance results for CTC and SDSC traces alone. This selection is purely arbitrary.

Although user estimates are known to be quite inaccurate in practice as explained above, we first studied the effect of preemptive scheduling under the idealised assumption of accurate estimation before studying the effect of inaccuracies in user estimates of job run time. Also, we first studied the impact of preemption under the assumption that the overhead for job suspension and restart were negligible and then studied the influence of the overhead.

Any analysis that is based only on the average slowdown or turnaround time of all jobs in the system cannot provide insights into the variability within different job categories. Therefore, in our discussion, we classify the jobs into various categories based on the run time and the number of processors requested, and we analyse the slowdown and turnaround time for each category.

To analyse the performance of jobs of different sizes and lengths, we classified jobs into 16 categories: considering four partitions for run time – Very Short (VS), Short (S), Long (L) and Very Long (VL) – and four partitions for the number of processors requested – Sequential (Seq), Narrow (N), Wide (W) and Very Wide (VW). The criteria used for job classification are shown in Table 1. The distribution of jobs in the trace, corresponding to the sixteen categories, is given in Tables 2 and 3.

Table 1 *Job categorisation criteria*

	1 Proc	*2–8 Procs*	*9–32 Procs*	*>32 Procs*
0–10 min	VS Seq	VS N	VS W	VS VW
10 min–1 hr	S Seq	S N	S W	S VW
1 hr–8 hr	L Seq	L N	L W	L VW
>8 hr	VL Seq	VL N	VL W	VL VW

Table 2 *Job distribution by category – CTC trace*

	1 Proc (%)	*2–8 Procs* (%)	*9–32 Procs* (%)	*>32 Procs* (%)
0–10 min	14	8	13	9
10 min–1 hr	18	4	6	2
1 hr–8 hr	6	3	9	2
>8 hr	2	2	1	1

Table 3 *Job distribution by category – SDSC trace*

	1 Proc (%)	*2–8 Procs* (%)	*9–32 Procs* (%)	*>32 Procs* (%)
0–10 min	8	29	9	4
10 min–1 hr	2	8	5	3
1 hr–8 hr	8	5	6	1
>8 hr	3	5	3	1

Tables 4 and 5 show the average slowdowns for the different job categories under a nonpreemptive, aggressive backfilling strategy. The overall slowdown for the CTC trace was 3.58, and for the SDSC trace was 14.13. Even though the overall slowdowns are low, from the tables one can observe that some of the Very Short categories have slowdowns as high as 34 (CTC trace) and 113 (SDSC trace). Preemptive strategies aim at reducing the high average slowdowns for the short categories without significant degradation to long jobs.

Table 4 *Average slowdown for various categories with nonpreemptive scheduling – CTC trace*

	1 Proc	*2–8 Procs*	*9–32 Procs*	*>32 Procs*
0–10 min	2.6	4.76	13.01	34.07
10 min–1 hr	1.26	1.76	3.04	7.14
1 hr–8 hr	1.13	1.43	1.88	1.63
>8 hr	1.03	1.05	1.09	1.15

Table 5 *Average slowdown for various categories with nonpreemptive scheduling – SDSC trace*

	1 Proc	*2–8 Procs*	*9–32 Procs*	*>32 Procs*
0–10 min	2.53	14.41	37.78	113.31
10 min–1 hr	1.15	2.43	4.83	15.56
1 hr–8 hr	1.19	1.24	1.96	2.79
>8 hr	1.03	1.09	1.18	1.43

4 SELECTIVE SUSPENSION

We first propose a preemptive scheduling scheme called Selective Suspension (SS), where an idle job may preempt a running job if its 'suspension priority' is sufficiently higher than the running job. An idle job attempts to suspend a collection of running jobs so as to obtain enough free processors. In order to control the rate of suspensions, a suspension factor (SF) is used. This specifies the minimum ratio of the suspension priority of a candidate idle job to the suspension priority of a running job for preemption to occur. The suspension priority used is the xfactor of the job.

4.1 Theoretical analysis

Let T_1 and T_2 (Figure 3) be two tasks submitted to the scheduler at the same time. Let both tasks be of the same length and require the entire system for execution, with the system being free when the two tasks are submitted. Let 's' be the suspension factor. Before starting, both tasks have a suspension priority of 1. The suspension priority of a task remains constant when the task executes, and increases when the task waits. One of the two tasks, say T_1, will start instantly. The other task, say, T_2 will wait until its suspension priority τ_2 becomes s times the priority of T_1 before it can preempt T_1. Now T_1 will have to wait until its suspension priority τ_1 becomes s times τ_2 before it can preempt T_2. Thus, execution of the two tasks will alternate, controlled by the suspension factor. Figures (4–6) show the execution pattern of the tasks T_1 and T_2 for various values of SF. The optimal value for SF, to restrict the number of repeated suspensions by two similar tasks arriving at the same time, can be obtained as follows:

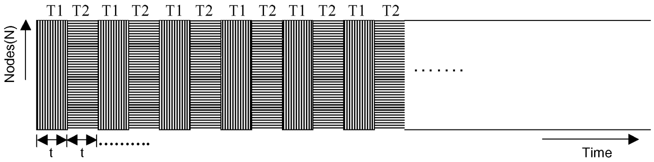

Figure 4 *Execution pattern of the tasks T_1 and T_2 when SF = 1. Here, t represents the minimum time interval between two suspensions*

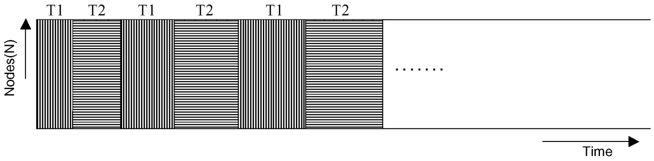

Figure 5 *Execution pattern of the tasks T_1 and T_2 when $1 < SF < \sqrt{2}$*

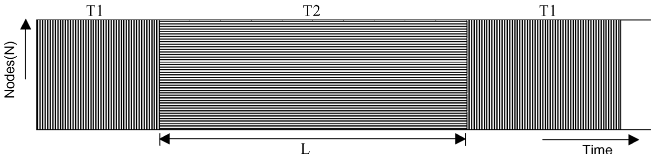

Figure 6 *Execution pattern of the tasks T_1 and T_2 when $SF = \sqrt{2}$*

Let τ_w represent the suspension priority of the waiting job and τ_r represent the suspension priority of the running job.

The condition for the first suspension is

$$\tau_w = s.$$

The preemption swaps the running job and the waiting job. Thus, after the preemption, $\tau_w = 1$ and $\tau_r = s$.

The condition for the second suspension is

$$\tau_w = s\tau_r$$

$$\tau_w = s^2.$$

Similarly, the condition for the nth suspension is $\tau_w = s^n$. The lowest value of s for which at most n suspensions occur is given by $\tau_w = s^{n+1}$, when the running job completes.

When the running job completes,

$$\tau_w = \frac{\text{wait time} + \text{run time}}{\text{run time}};$$

that is, $\tau_w = 2$ since the wait time of the waiting job = the run time of the running job

$$s^{n+1} = 2 \text{ and } s = 2^{(1/(n+1))}.$$

Thus, if the number of suspensions is to be 0, then $s = 2$. For at most one suspension, $s = \sqrt{2}$. With $s = 1$, the number of suspensions is very large, bounded only by the granularity of the preemption routine.

With all jobs having equal length, any suspension factor >2 will not result in suspension and will be the same as a suspension factor of 2. However, with jobs of varying length, the number of suspensions reduces with higher suspension factors. Thus, to avoid thrashing and to reduce the number of suspensions, we use different suspension factors between 1.5 and 5 in evaluating our schemes.

4.2 Preventing starvation without reservation guarantees

With priority based suspension, an idle job can preempt a running job only if its priority is at least SF times greater than the priority of the running job. All the idle jobs that are able to find the required number of processors by suspending lower priority running jobs are selected for execution by preempting the corresponding jobs. All backfilling scheduling schemes use job reservations for one or more jobs at the head of the idle queue as a means of guaranteeing finite progress and thereby avoiding starvation. But start time guarantees do not have much significance in a preemptive context. Even if we give start time guarantees for the jobs in the idle queue, they are not guaranteed to run to completion. Since the SS strategy uses the expected slowdown (xfactor) as the suspension priority, there is an automatic guarantee of freedom from starvation: ultimately any job's xfactor will get large enough so that it will be able to preempt some running job(s) and begin execution. Thus, one can use backfilling without the usual reservation guarantees. We therefore remove guarantees for all our preemption schemes.

Jobs in some categories inherently have a higher probability of waiting longer in the queue than do jobs with comparable xfactors from other job categories. For example, consider a VW job needing 300 processors, and a Sequential job in the queue at the same time. If both jobs have the same xfactor, the probability that the Sequential job finds a running job to suspend is higher than the probability that the VW job finds enough lower priority running jobs to suspend. Therefore, the average slowdown of the VW category will tend to be higher than the Sequential category. To redress this inequity, we impose a restriction that the number of processors requested by a suspending job should be at least half of the number of processors requested by the job that it suspends, thereby preventing the wide jobs from being suspended by the narrow jobs. The scheduler periodically (after every minute) invokes the preemption routine.

4.3 Algorithm

Let τ_i, be the suspension priority for a task t_i which requests n_i processors. Let N_i represent the set of processors allocated to t_i. Let F_t represent the set of free processors and f_t represent the number of free processors at time t when the preemption is attempted.

The set of tasks that can be preempted by task t_i is given by

$$C_i = \left\{ t_j \mid \tau_i \geq (SF)t_j \text{ and } \frac{n_j}{n_i} \leq 2 \right\}.$$

Task t_i can be scheduled by preempting one or more tasks in C_i if and only if

$$n_i \leq \left(f_t + \sum_{j:t_j \in C_i} n_j \right).$$

Let $(t_1,t_2,t_3,\ldots,t_x)$ be the elements of C_i. Let ϕ be a permutation of $(1,2,3,\ldots,x)$ such that $n_{\phi(1)} \geq n_{\phi(2)} \geq n_{\phi(3)}, \ldots, \geq n_{\phi(x-1)} \geq n_{\phi(x)}$. (If $n_{\phi(l)} = n_{\phi(l+1)}$, then $\tau_l \leq \tau_{l+1}$. If $\tau_l = \tau_{l+1}$, then the start time of t_l is less than or equal to the start time of t_{l+1}. If the start time of t_l is equal to the start time of t_{l+1}, then the queue time of t_l less than the queue time of t_{l+1}). So,

$$\exists k_i \mid 1 \leq k_i \leq x \text{ and } \left(\sum_{j=1}^{k_i - 1} n_{\phi(j)} \right) < (n_i - f_t)$$
$$\text{and } \left(\sum_{j=1}^{k_i} n_{\phi(j)} \right) \geq (n_i - f_t).$$

The set of tasks preempted by task t_i, is given by

$$P_i = \{ t_{\phi(r)} \mid 1 \leq r \leq k_i \}.$$

If t_i is a previously suspended task attempting reentry, then it has to get the same set of processors that it was using before it was suspended. Here, we remove the restriction that the number of processors requested by a suspending job should be at least half of the number of nodes requested by the job that it suspends. Otherwise if a VW job happens to suspend a narrow job, then in the worst case, the narrow job has to wait till the VW job completes to get rescheduled. So the set of tasks that can be preempted by t_i in this case is given by

$$C_i = \{ t_j \mid \tau_i \geq (SF)\tau_j \text{ and } N_i \cap Nj \neq \phi \}.$$

Task t_i can be scheduled by preempting one or more tasks in C_i if and only if

$$N_i \subseteq \left(F_t \bigcup_{j:t_j \in C_j} N_j \right).$$

Pseudocode for the selective suspension scheme

```
Sort the list of running jobs in ascending order of suspension priority
Sort the list of idle jobs in descending order of suspension priority
for each idle job
do
    set the candidate_job_set to be the null set
    if  (idle job is a suspended job)
    then
        goto already_suspended
    else
        available_processors = number of free processors
        for each running job
        do
            if  (number of processors requested by the idle job > available_processors)
            then
                if  ((suspension priority of the idle job >= SF * suspension priority of the running job)  &&
                   (number of processors used by the running job <= 2 * number of processors requested by the idle job))
                then
                    available_processors = available_processors + number of processors used by the running job
                    candidate_job_set = {candidate_job_set} u {running job}
                else
                    goto next_idle_job
                end if
            else
                goto suspend_jobs_1
            end if
        done
        end for
    end if
    goto next_idle_job
    already_suspended:
        set available_processor_set to the set of free processors
        for each running job
        do
            if  (set of processors requested by idle job is not a subset of available_processor_set)
            then
                if  (suspension priority of the idle job >= SF * suspension priority of the running job)
                then
                    if  ({set of processors used by the running job}  n {set of processors requested by the idle job} is not empty)
                    then
                        available_processor_set = {available_processor_set} u {set of processors used by running job}
                        candidate_job_set = {candidate_job_set} u {running job}
                    end  if
                else
                    goto next_idle_ job
                end if
            else
                goto suspend_job_2
            end if
        done
        end for
    goto next_idle_job
    suspend_jobs_1:
        sort job(s) in candidate_job_set in descending order of number of processors used
        available_processors = number of free processors
        for each job in candidate_job_set
        do
            if  (number of processors requested by the idle job > available_processors)
            then
                suspend the job
                available_processors = available_processors + number of processors used by the suspended job
            else
                schedule the idle job
                goto next_idle_job
            end if
        done
        end for
    goto next_idle_job
    suspend_jobs_2:
        suspend all jobs in the candidate_job_set
        schedule the idle job
     next_idle_job:
        do nothing
done
end for
```

4.4 Results

We compare the SS scheme run under various suspension factors with the No-Suspension (NS) scheme with aggressive backfilling and the IS scheme. From Figures 7–10, we can see that the SS scheme provides significant improvement for the Very Short (VS) and Short (S) length categories and Wide (W) and Very Wide (VW) width categories. For example, for the VS-VW category, slowdown is reduced from 113 for the NS scheme to 7 for SS with SF = 2 for the SDSC trace (reduced from 34 for the NS scheme to under 3 for SS with SF = 2 for the CTC trace).

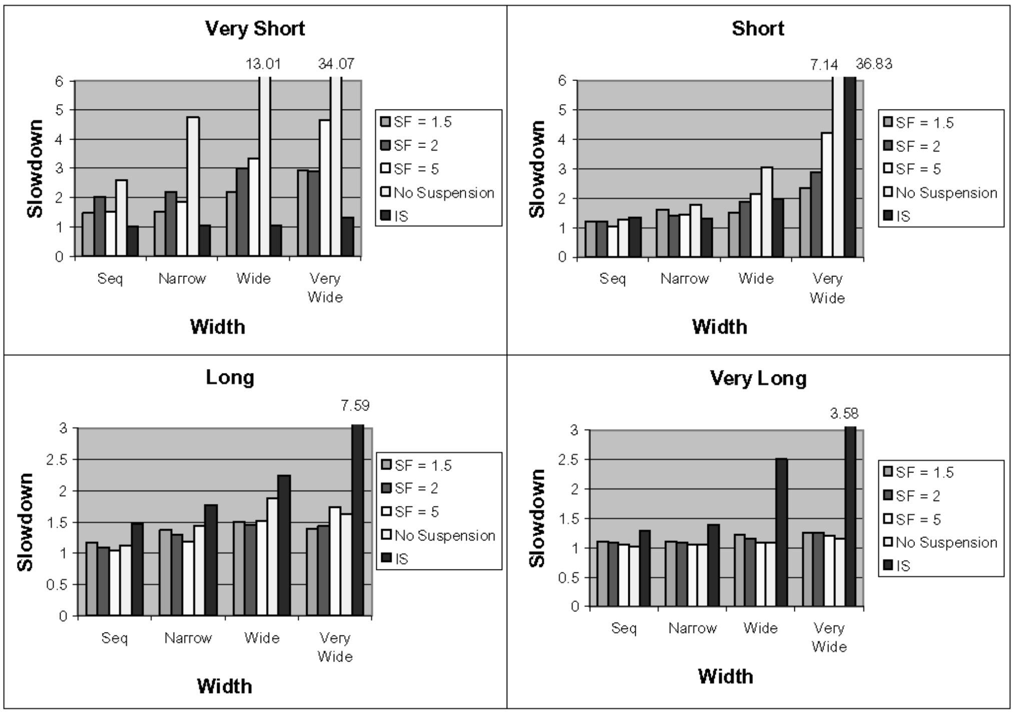

Figure 7 *Average slowdown: SS scheme, CTC trace. Compared to NS, SS provides significant benefit for the VS, S, W, and VW categories; slight improvement for most of L categories; but a slight deterioration for the VL categories. Compared to IS, SS performs better for all the categories except for the VS categories*

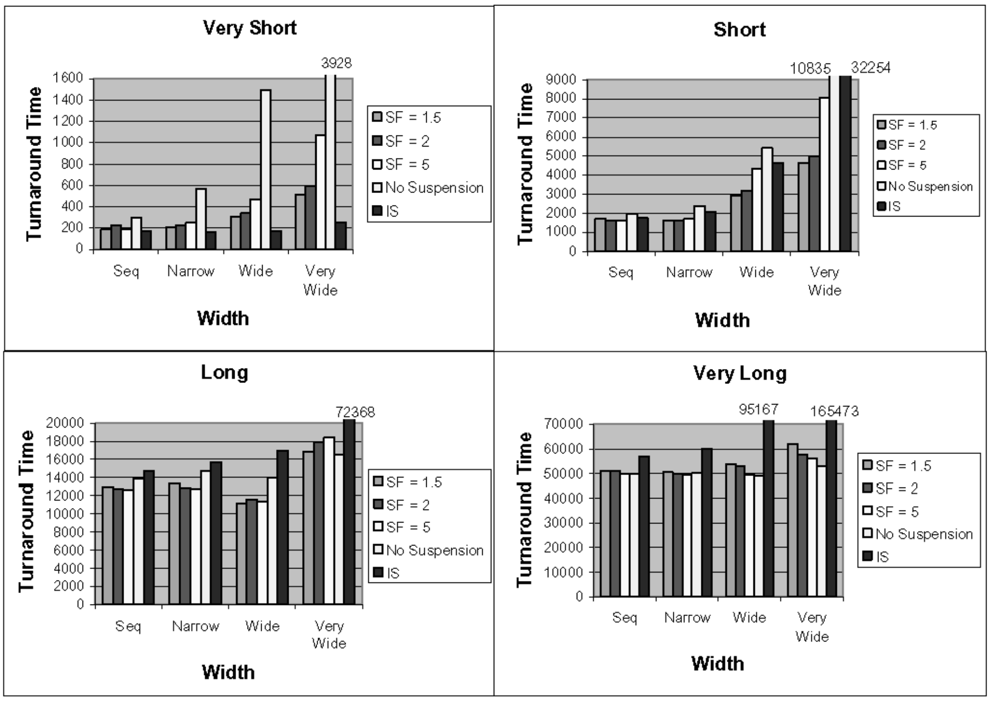

Figure 8 *Average turnaround time: SS scheme, CTC trace. The trends are similar to those with the average slowdown metric (Figure 7)*

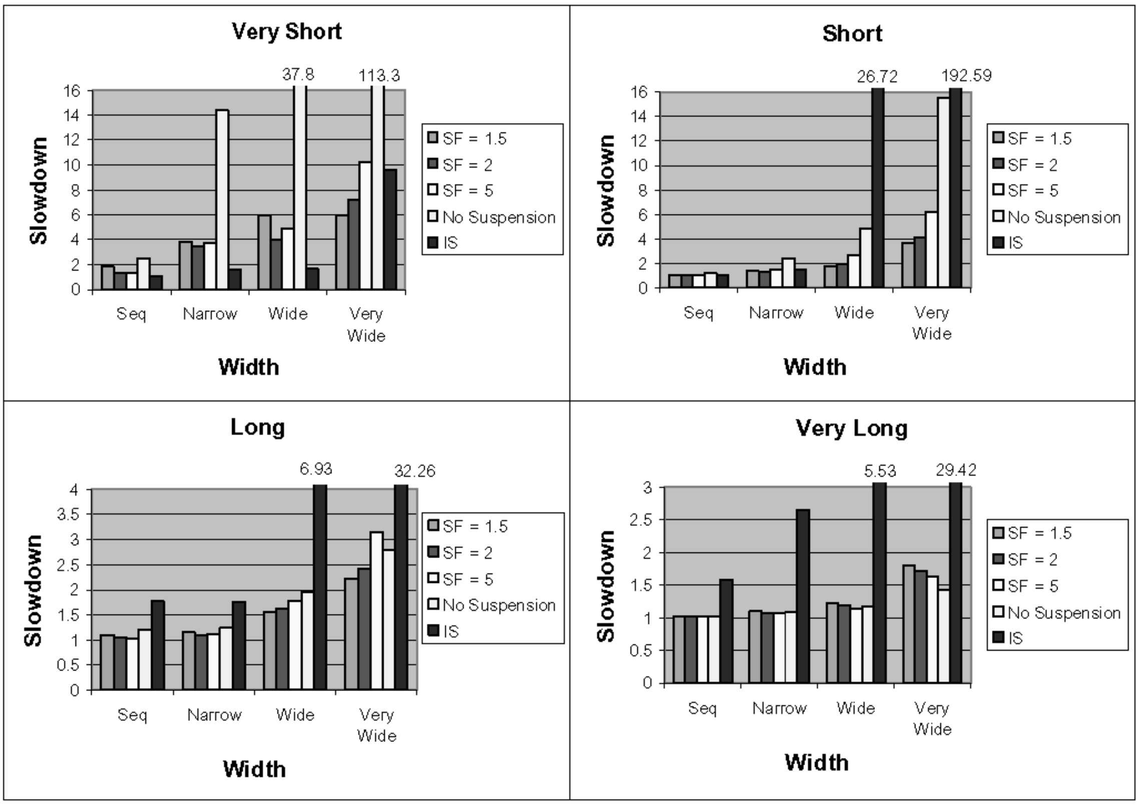

Figure 9 *Average slowdown: SS scheme, SDSC trace. Compared to NS, SS provides significant benefit for the VS, S, W, and VW categories; slight improvement for most of L categories; but a slight deterioration for the VL categories. Compared to IS, SS performs better for all the categories except for the VS categories*

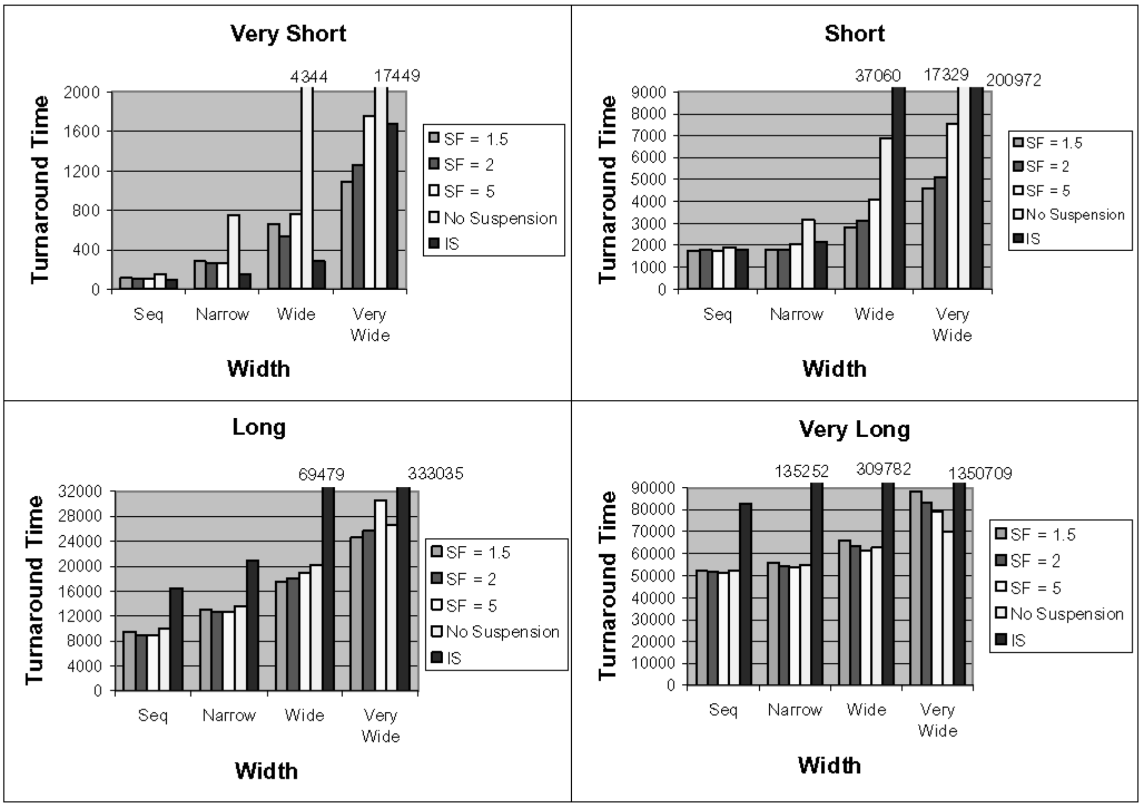

Figure 10 *Average turnaround time: SS scheme, SDSC trace. The trends are similar to those with the average slowdown metric (Figure 9)*

For the VS and S length categories, a lower SF results in lowered slowdown and turnaround time. This is because a lower SF increases the probability that a job in these categories will suspend a job in the Long (L) or Very-Long (VL) category. The same is also true for the L length category, but the effect of change in SF is less pronounced. For the VL length category, there is an opposite trend with decreasing SF: the slowdown and turnaround times worsen. This is due to the increasing probability that a Long job will be suspended by a job in a shorter category as SF decreases. In comparison to the base No Suspension (NS) scheme, the SS scheme provides significant benefits for the VS and S categories and a slight improvement for most of the Long categories but is slightly worse for the VL categories. The performance of the IS scheme is very good for the VS categories. It is better than the SS scheme for the VS categories and worse for the other categories. Although the overall slowdown for IS is considerably less than the No Suspension scheme, it is not better than SS. Moreover, with IS, the VW and VL categories get significantly worse.

4.5 Tuneable selective suspension (TSS)

From the graphs of the previous section, one can observe that the SS scheme significantly improves the average slowdown and turnaround time of various job categories. But from a practical point of view, the worst case slowdowns and turnaround times are very important. A scheme that improves the average slowdowns and turnaround times for most of the categories but makes the worst case slowdown and turnaround time for the long categories worse is not an acceptable scheme. For example, a delay of 1 hr for a 10 minutes job (slowdown = 7) is tolerable, whereas a slowdown of 7 for a 24 hr job is unacceptable. Figure 11 compares the worst case slowdowns for SF = 2 with the worst case slowdowns of the NS scheme and the IS scheme for the CTC trace. One can observe that the worst case slowdowns with the SS scheme are much better than with the NS scheme for most of the cases. But the worst case slowdowns for some of the long categories are worse than for the NS scheme.

Although the worst case slowdown with SS is generally less than that with NS, the absolute worst case slowdowns are much higher than the average slowdowns for some of the short categories. For the IS scheme, the worst case slowdowns for the very short categories are lower, but they are very high for the long jobs, an unacceptable situation. Figure 12 compares the worst case turnaround times for the SS scheme, with worst case turnaround times for the NS scheme and the IS scheme, for the CTC trace. Even though the trends observed here are similar to those with the worst case slowdowns, the categories where SS is the best with respect to worst case turnaround time are not same as the categories for which SS is the best with respect to worst case slowdowns. This is because a job with the worst case turnaround time need not be the one with worst case slowdown. Similar trends can be observed for the SDSC trace from Figures 13 and 14.

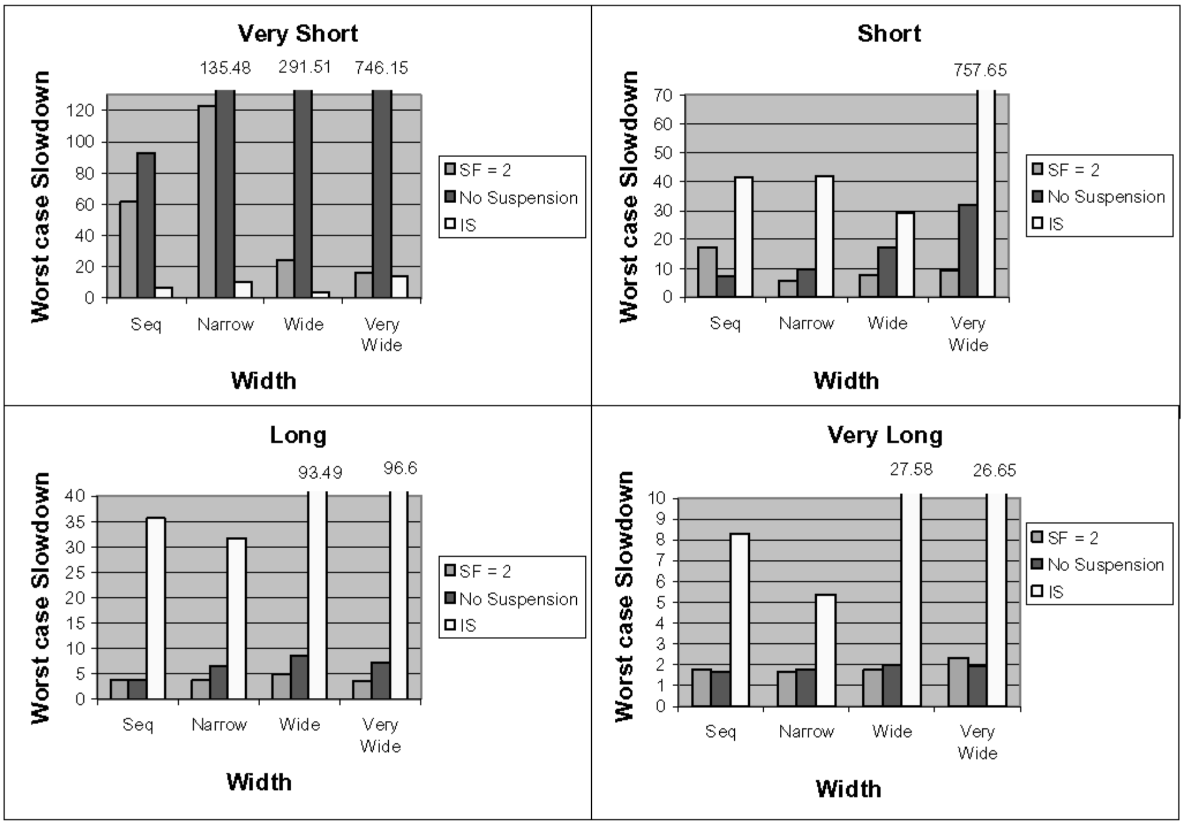

Figure 11 *Worst case slowdown: SS scheme, CTC trace. SS is much better than NS for most of the categories and is slightly worse for some of the VL categories. Compared to IS, SS is much better for all the categories except for the VS categories*

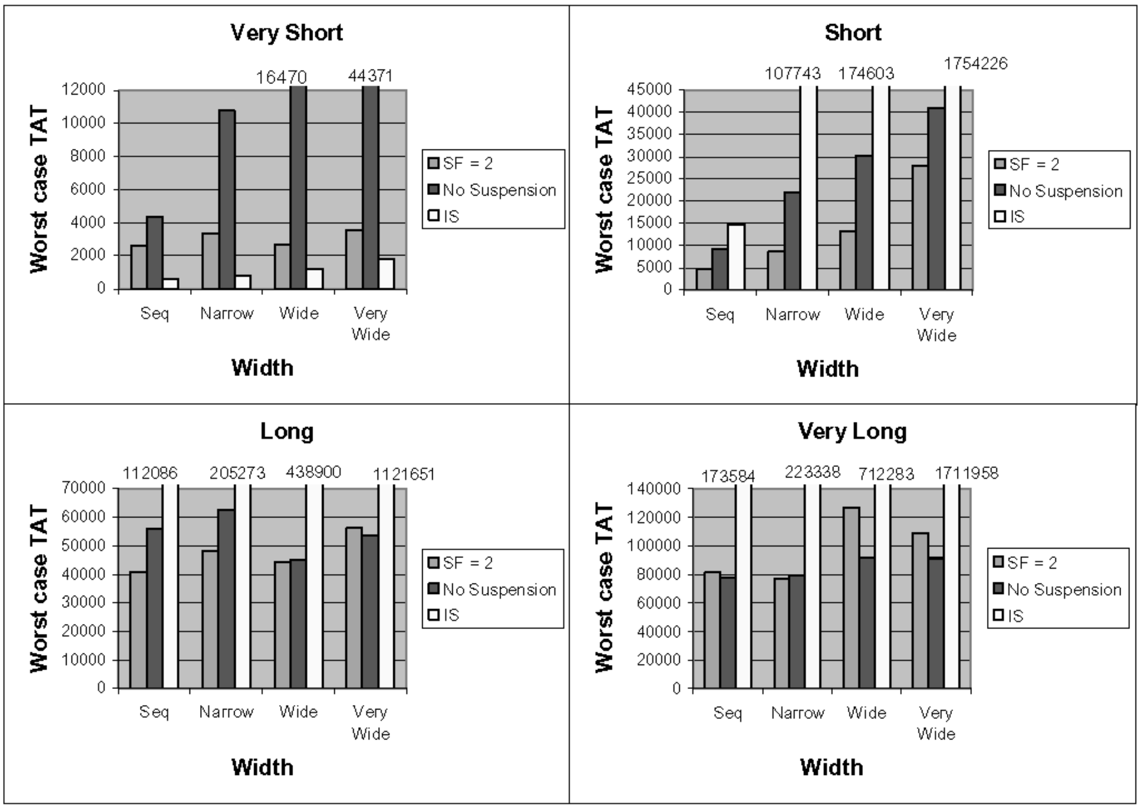

Figure 12 *Worst case turnaround time: SS scheme, CTC trace. The trends are similar to those with the worst case slowdown metric (Figure 11)*

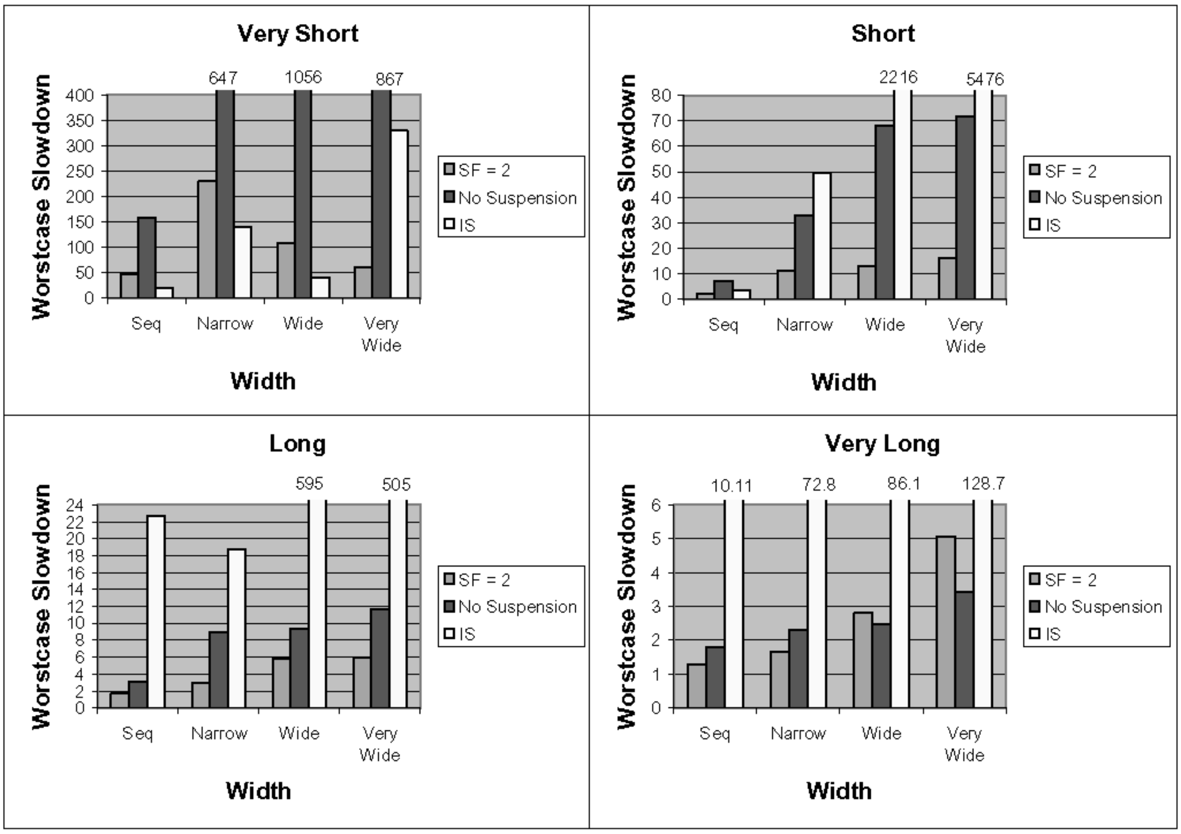

Figure 13 *Worst case slowdown: SS scheme, SDSC trace. SS is much better than NS for most of the categories and is slightly worse for some of the VL categories. Compared to IS, SS is much better for all the categories except for the VS categories*

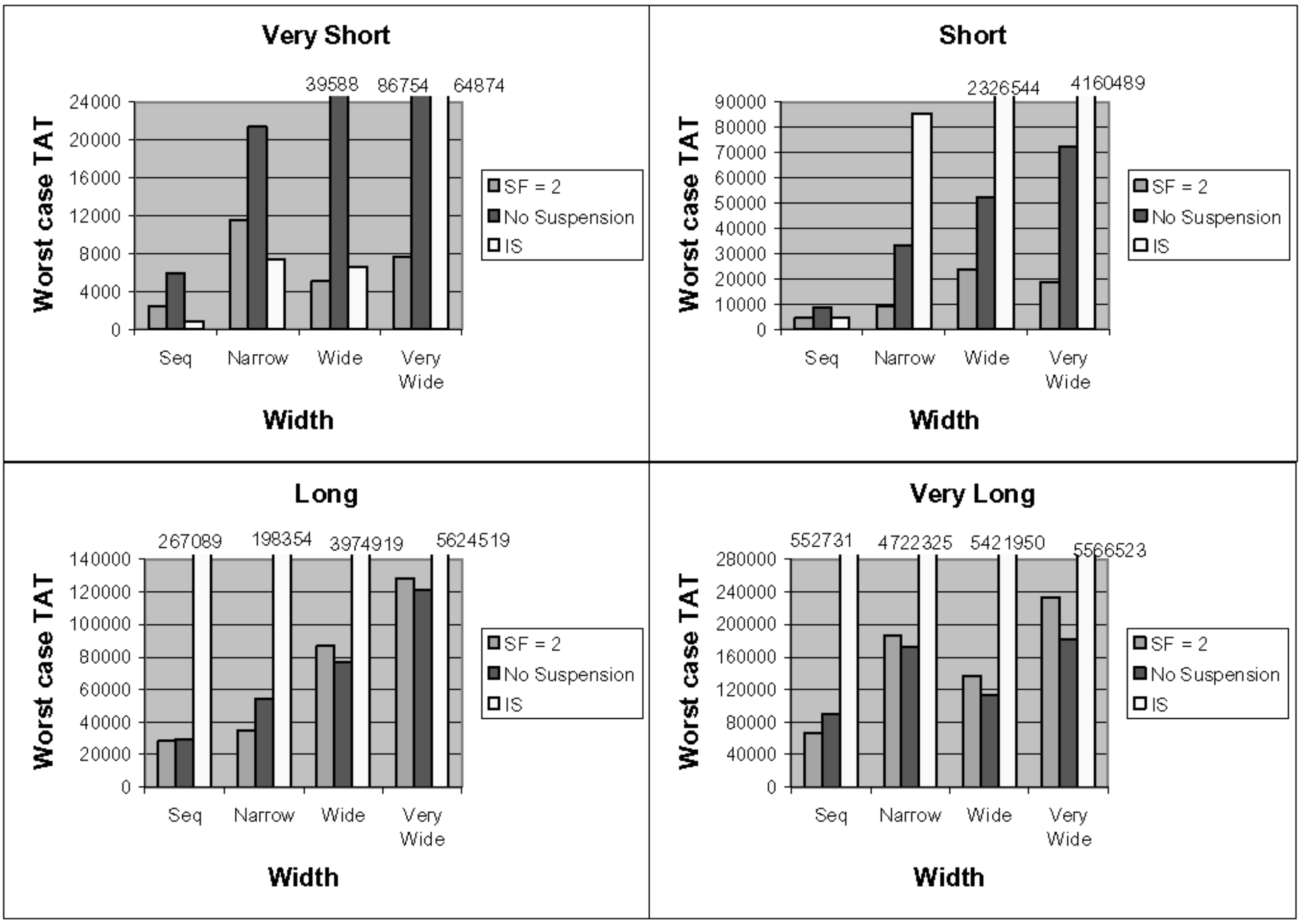

Figure 14 *Worst case turnaround time: SS scheme, SDSC trace. The trends are similar to those with the worst case slowdown metric (Figure 13)*

We next propose a tuneable scheme to improve the worst case slowdown and turnaround time without significant deterioration of the average slowdown and turnaround time. This scheme involves controlling the variance in the slowdowns and turnaround times by associating a limit with each job. Preemption of a job is disabled when its priority exceeds this limit. This limit is set to 1.5 times the average slowdown of the category that the job belongs to.

The candidate set of tasks that can be preempted by a task t_i is given by

$$C_i = \{t_j \mid \tau_i \geq (SF)\tau_i \text{ and } \frac{n_j}{n_i} \leq 2$$
$$\text{and } \tau_i \leq 1.5 \times SD_{\text{avg}}(\text{category}(t_j))\},$$

where SD_{avg} (category(t_j)) represents the average slowdown for the job category to which t_j belongs.

If t_i is a previously suspended task attempting reentry, then

$$C_i = \{t_j \mid \tau_i \geq (SF)\tau_i \text{ and } N_i \cap N_j \neq \phi$$
$$\text{and } \tau_i \leq 1.5 \times SD_{\text{avg}}(\text{category}(t_j))\}.$$

All the other conditions remain the same as mentioned in Section 4.3.

Figures 15 and 16 show the result of this tuneable scheme for the CTC trace. It improves the worst case slowdowns for some long categories (VL W, VL VW, L N) and some short categories (VS Seq, VS N, S Seq) without affecting the worst case slowdowns of the other categories. It improves the worst case turnaround times for most of the categories without affecting the worst case turnaround times of the other categories. Figures 17 and 18 show similar trends for the SDSC trace. This scheme can also be applied to selectively tune the slowdowns or turnaround times for particular categories. The TSS scheme is used for all the subsequent experiments, and the term ‘Selective Suspension’ or ‘SS’ in the following sections refers to ‘Tuneable Selective Suspension’.

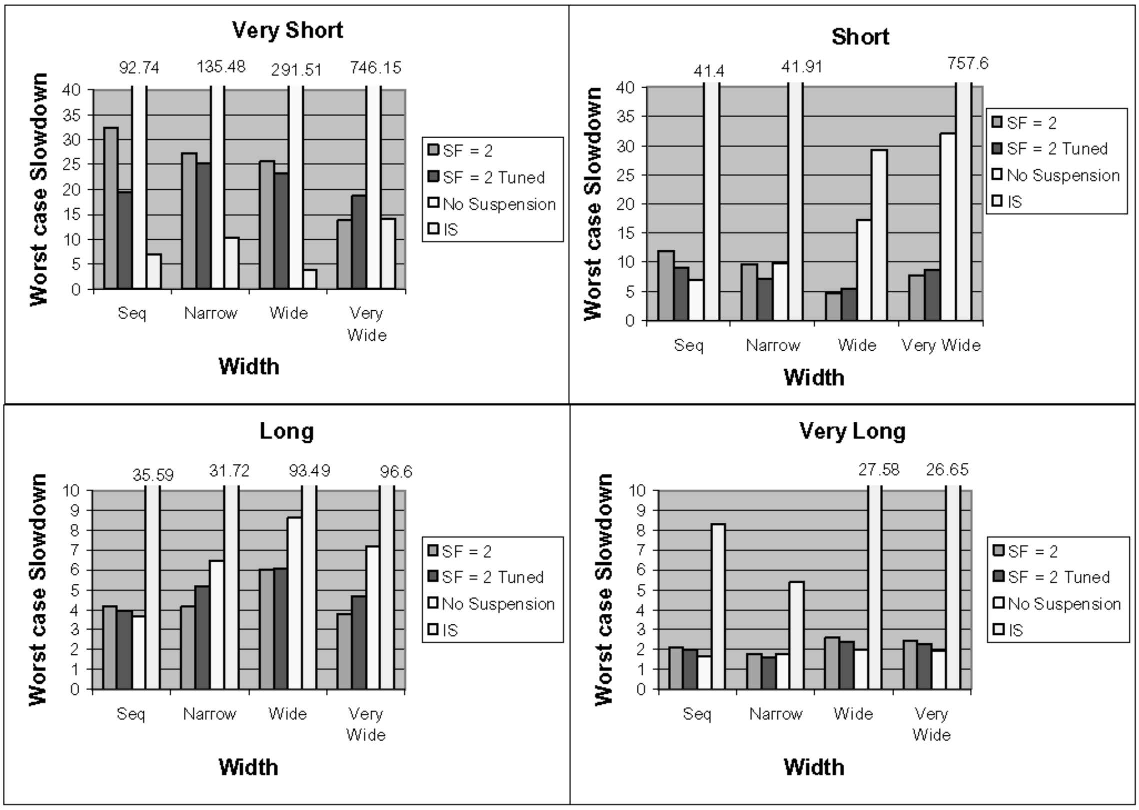

Figure 15 *Worst case slowdown for the TSS scheme: CTC trace. TSS improves the worst case slowdowns for many categories without affecting the worst case slowdowns for other categories*

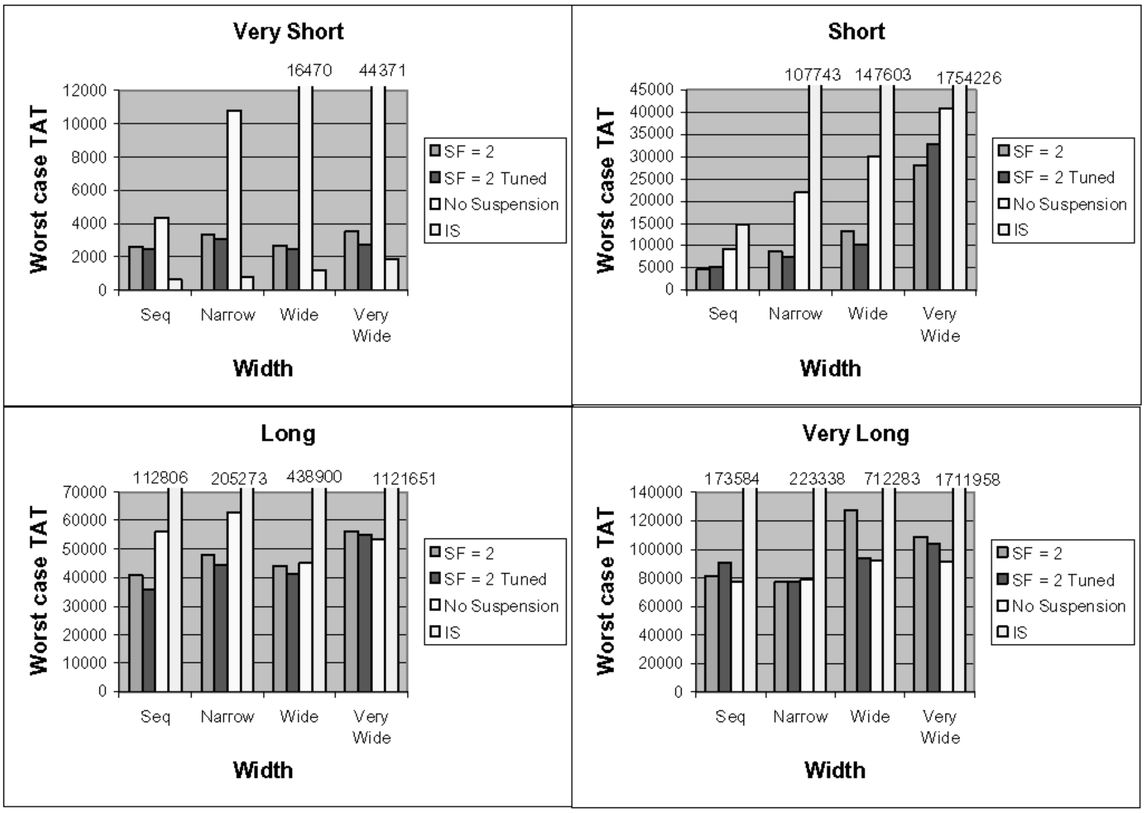

Figure 16 *Worst case turnaround times for the TSS scheme: CTC trace. TSS improves the worst case turnaround times for many categories without affecting the worst case tunraround times for other categories*

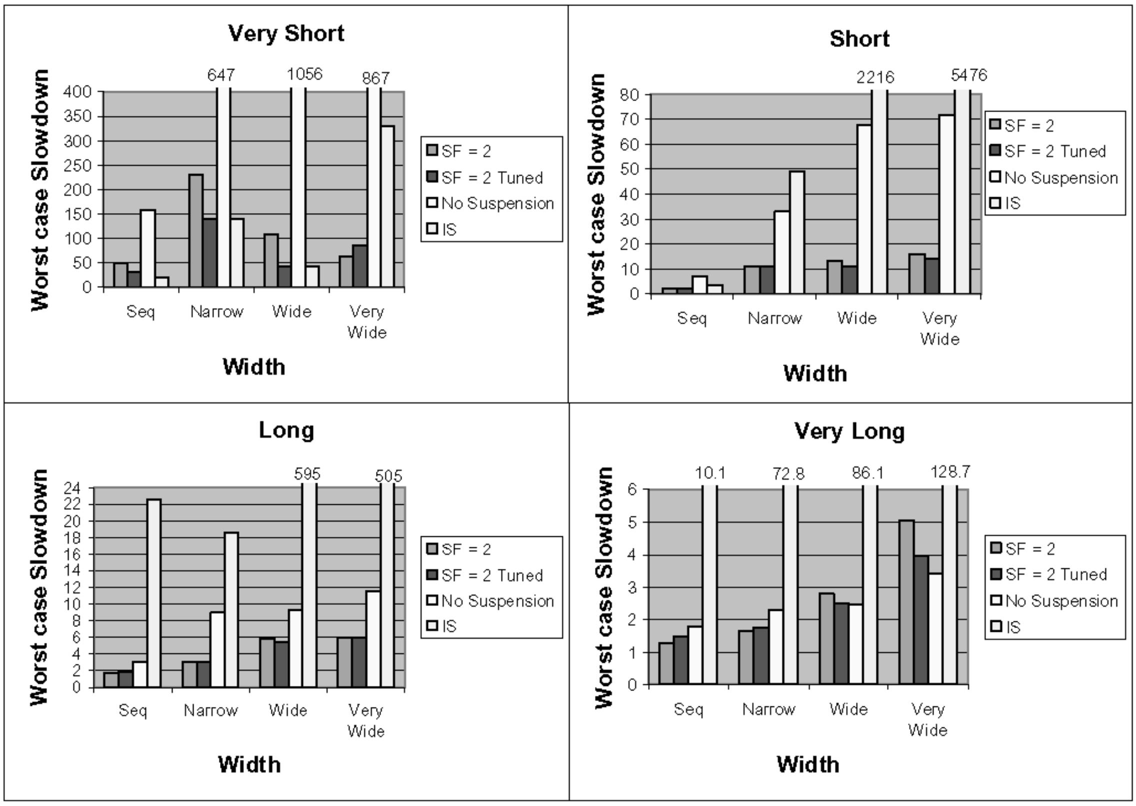

Figure 17 *Worst case slowdown for the TSS scheme: SDSC trace. TSS improves the worst case slowdowns for many categories without affecting the worst case slowdowns for other categories*

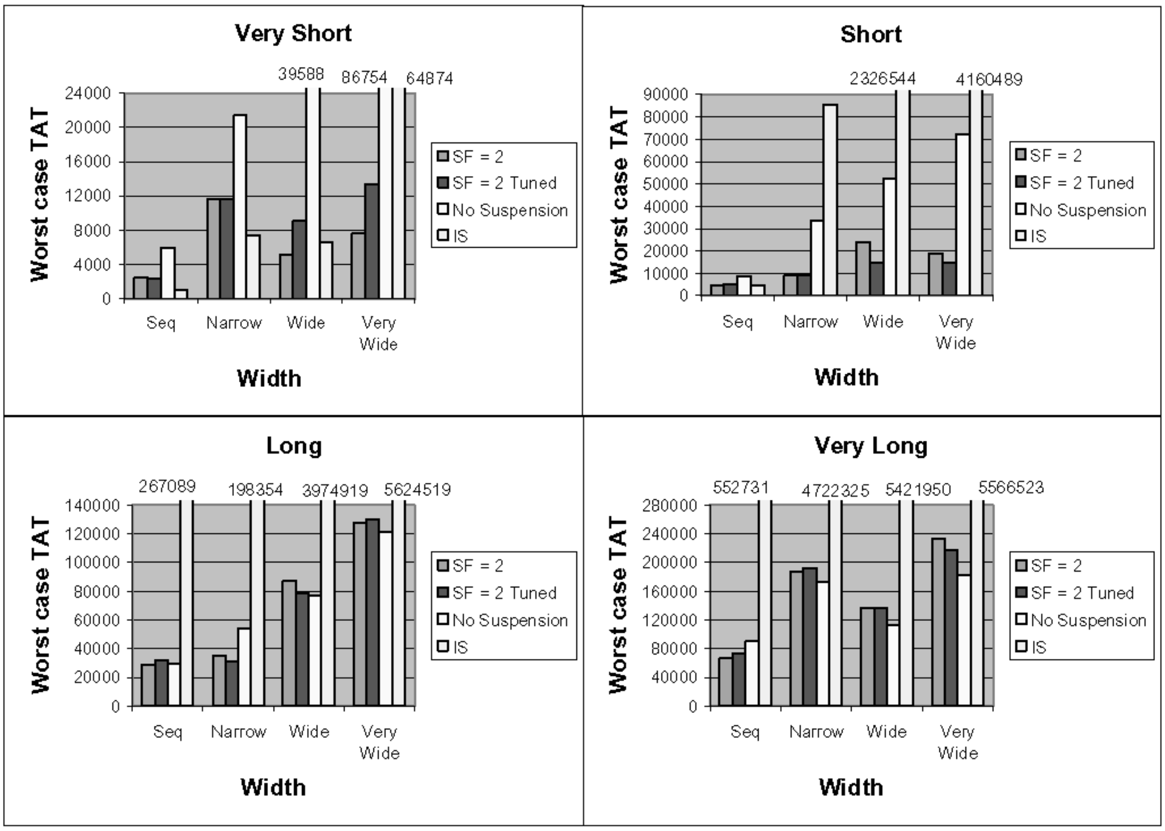

Figure 18 *Worst case turnaround times for the TSS scheme: SDSC trace. TSS improves the worst case turnaround times for many categories without affecting the worst case tunraround times for other categories*

5 IMPACT OF USER ESTIMATE INACCURACIES

We have so far assumed that the user estimates of job run time are perfect. Now, we consider the effect of user estimate inaccuracies on the proposed schemes. This analysis is needed for modelling an actual system workload. In this context, we believe that a problem has been ignored by previous studies when analysing the effect of over estimation on scheduling strategies. Abnormally aborted jobs tend to excessively skew the average slowdown of jobs in a workload. Consider a job requesting a wall clock limit of 24 hrs, which is queued for 1 hr and then aborts within one minute due to some fatal exception. The slowdown of this job would be computed to be 60, whereas the average slowdown of normally completing long jobs is typically under two. If even 5% of the jobs have a high slowdown of 60, while 95% of the normally completing jobs have a slowdown of two, the average slowdown over all jobs would be around five. Now consider a scheme such as the speculative backfilling strategy evaluated in Perkovic and Keleher (2000). With this scheme, a job is given a free timeslot to execute in, even if that slot is considerably smaller than the requested wall clock limit. Aborting jobs will quickly terminate, and since they did not have to be queued till an adequately long window was available, their slowdown would decrease dramatically with the speculative backfilling scheme. As a result, the average slowdown of the entire trace would now be close to two, assuming that the slowdown of the normally completing jobs does not change significantly. A comparison of the average slowdowns would seem to indicate that the speculative backfilling scheme results in a significant improvement in job slowdown from 5 to 2. However, under the above scenario, the change is due only to the small fraction of aborted jobs, and not due to any benefits to the normal jobs. In order to avoid this problem, we group the jobs into two different estimation categories:

- jobs that are well estimated (the estimated time is not more than twice the actual run time of that job)
- jobs that are poorly estimated (the estimated run time is more than twice the actual run time).

Within each group, the jobs are further classified into 16 categories based on their actual run time and the number of processors requested.

One can observe from Figures 19 and 20 that the Selective Suspension scheme improves the slowdowns for most of the categories without adversely affecting the other categories. The slowdowns for the short and wide categories are quite high compared to the other categories, mainly because of the overestimation. Since the suspension priority used by the SS scheme is xfactor, it favours the short jobs. But if a short job was badly estimated, it would be treated as a long job and its priority would increase only gradually. So, it will not be able to suspend running jobs easily and will end up with a high slowdown. This situation does not happen with IS because of the 10 minutes time quantum for each arriving job, irrespective of the estimated run time, and therefore the slowdowns for the very short category (whose length is less than or equal to 10 minutes) are better in IS than other schemes. For the other categories, however, SS performs much better than IS.

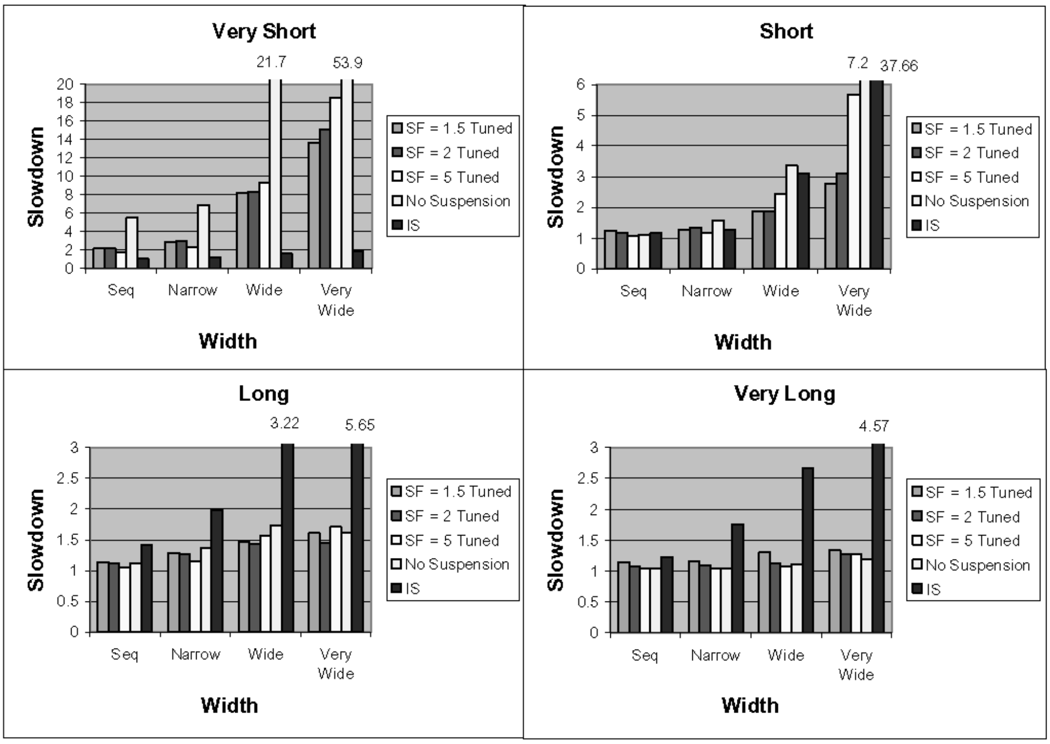

Figure 19 *Average slowdown: Inaccurate estimates of run time; CTC trace. Compared to NS, SS improves the slowdowns for most of the categories with little deterioration to other categories. The performance of IS is bad for the long jobs*

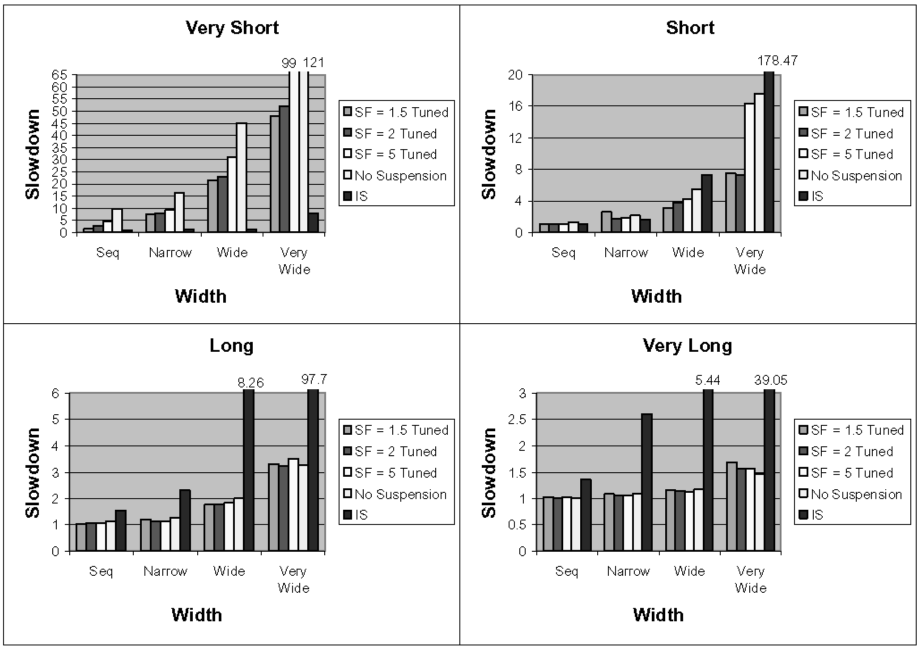

Figure 20 *Average slowdown: Inaccurate estimates of run time; SDSC trace. Compared to NS, SS improves the slowdowns for most of the categories with little deterioration to other categories. The performance of IS is bad for the long jobs*

Figures 21 and 22 compare the average turnaround times for the SS scheme with that of the NS and IS schemes for the CTC and SDSC traces, respectively. The improvement in performance for the short and wide categories is much less when compared to the improvement achieved with the accurate user estimate case. The reasoning provided above for the increase in slowdowns for the short and wide categories holds for this case also. The seemingly long jobs (badly estimated short jobs) are unable to suspend running jobs easily and have to wait in the queue for a longer time, thus ending up with a high turnaround time. From Figures 23–26, the higher slowdowns for the VS categories with SS clearly are due to the badly estimated jobs. Figures 27–30 show that the reduction in the percentage improvement of the average turnaround times for the short and wide categories in SS, is due to the badly estimated jobs. One can also observe that, for the well estimated jobs, SS is better than or comparable to IS for the VS categories and SS outperforms IS in all other categories.

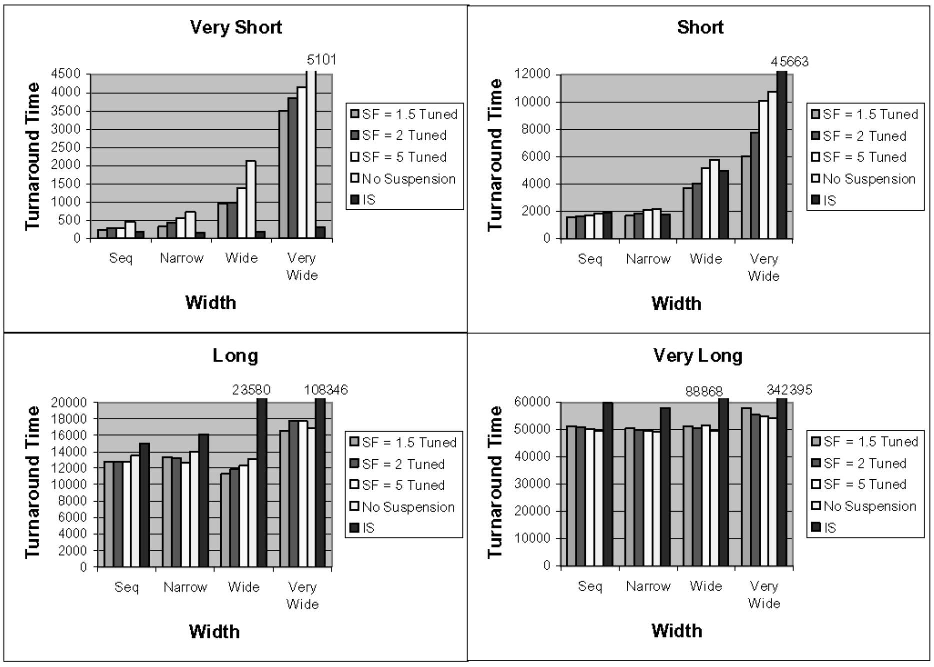

Figure 21 *Average turnaround time: Inaccurate estimates of run time; CTC trace. Compared to NS, SS improves the turnaround times for most of the categories with little deterioration to other categories. The performance of IS is bad for the long jobs*

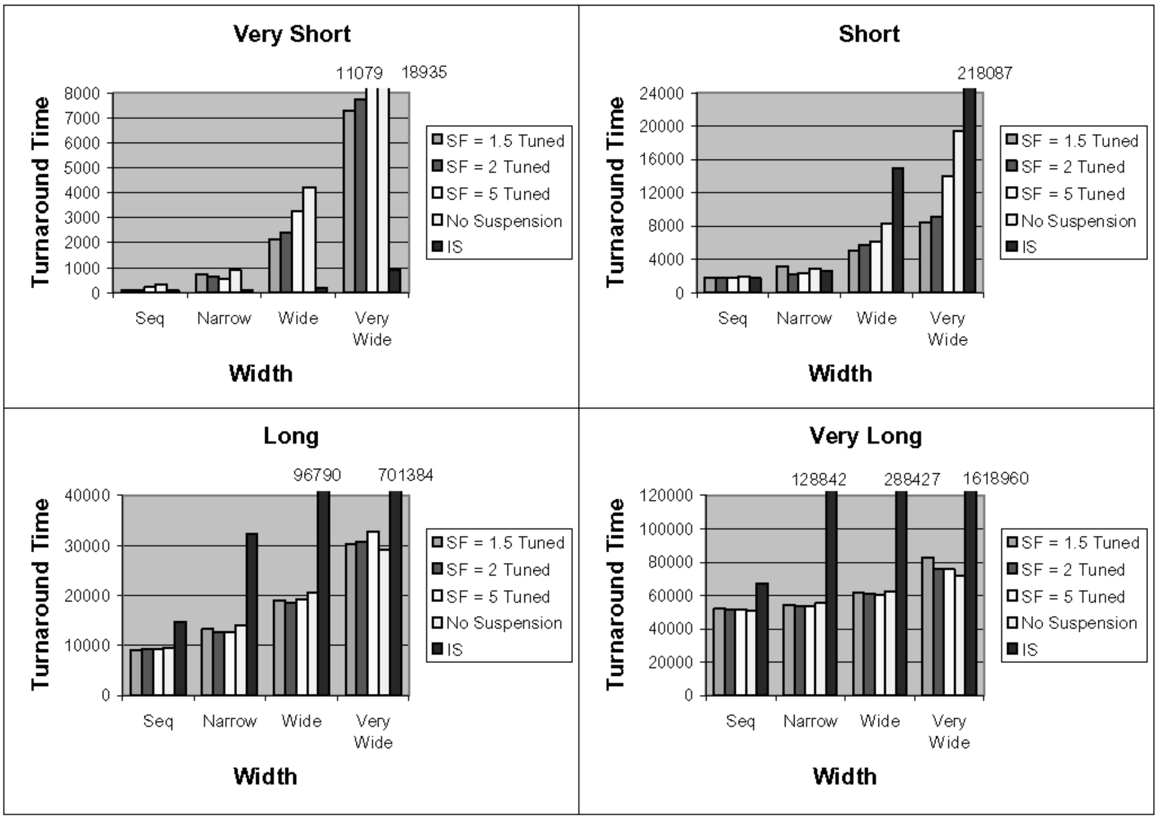

Figure 22 *Average turnaround time: Inaccurate estimates of run time; SDSC trace. Compared to NS, SS improves the turnaround times for most of the categories with little deterioration to other categories. The performance of IS is bad except for VS categories*

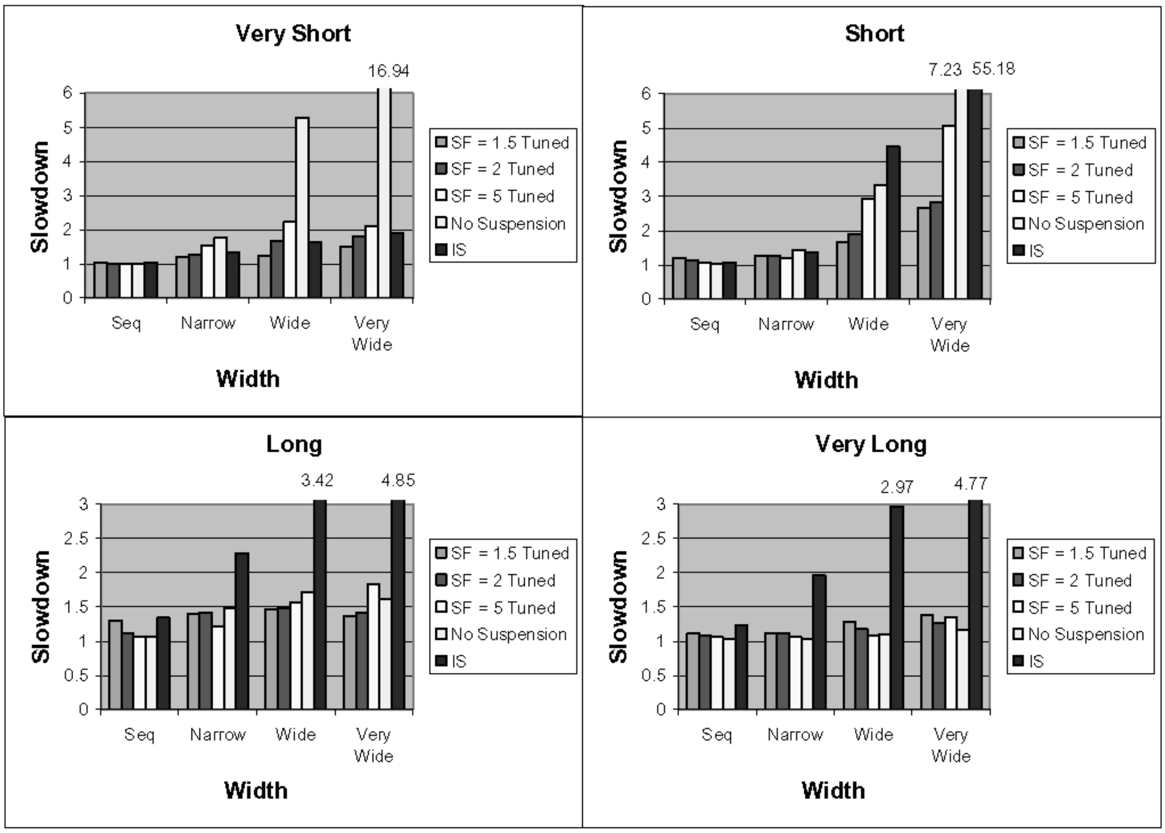

Figure 23 *Average slowdown of well estimated jobs: CTC trace. Compared to NS, SS significantly improves the slowdowns for most of the categories with little deterioration to other categories. The performance of SS is better than or comparable to IS for VS categories*

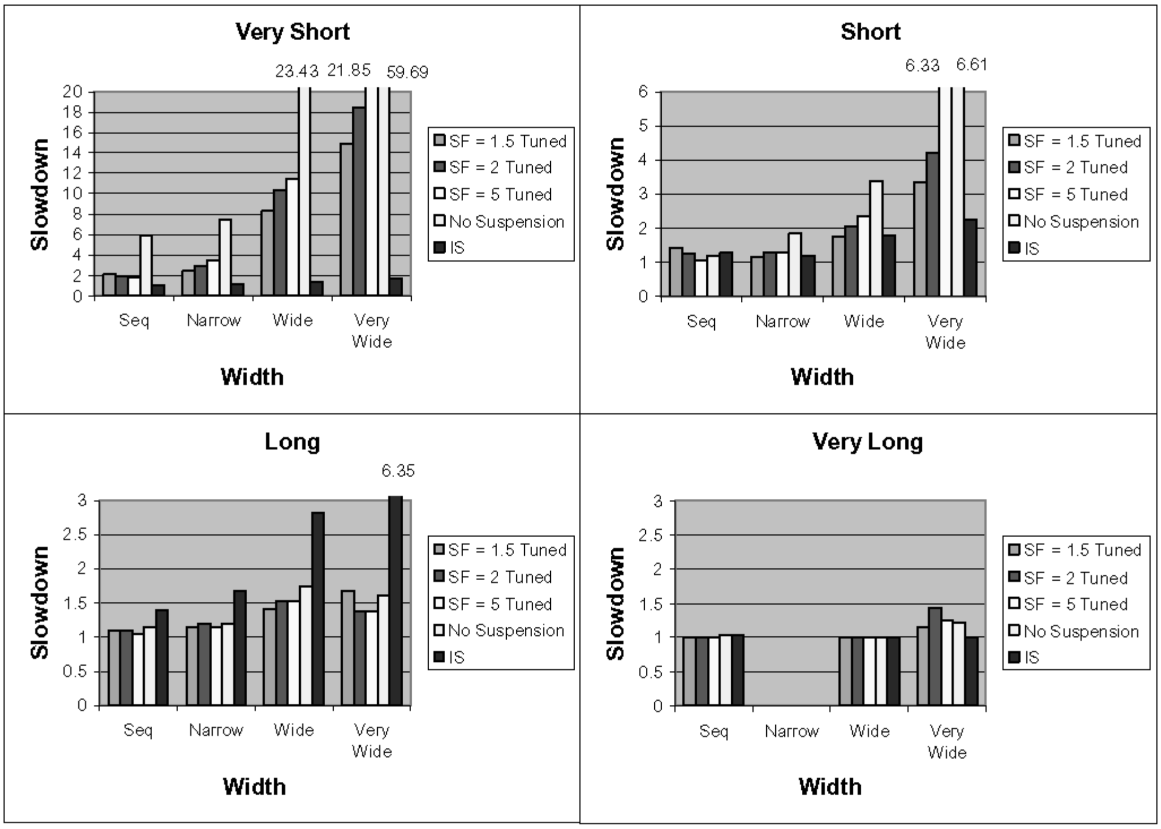

Figure 24 *Average slowdown of badly estimated jobs: CTC trace. Compared to NS, SS provides a slight improvement in slowdowns for many categories. SS tends to penalise the badly estimated jobs in VS categories. IS gives better performance for VS, S and VL categories*

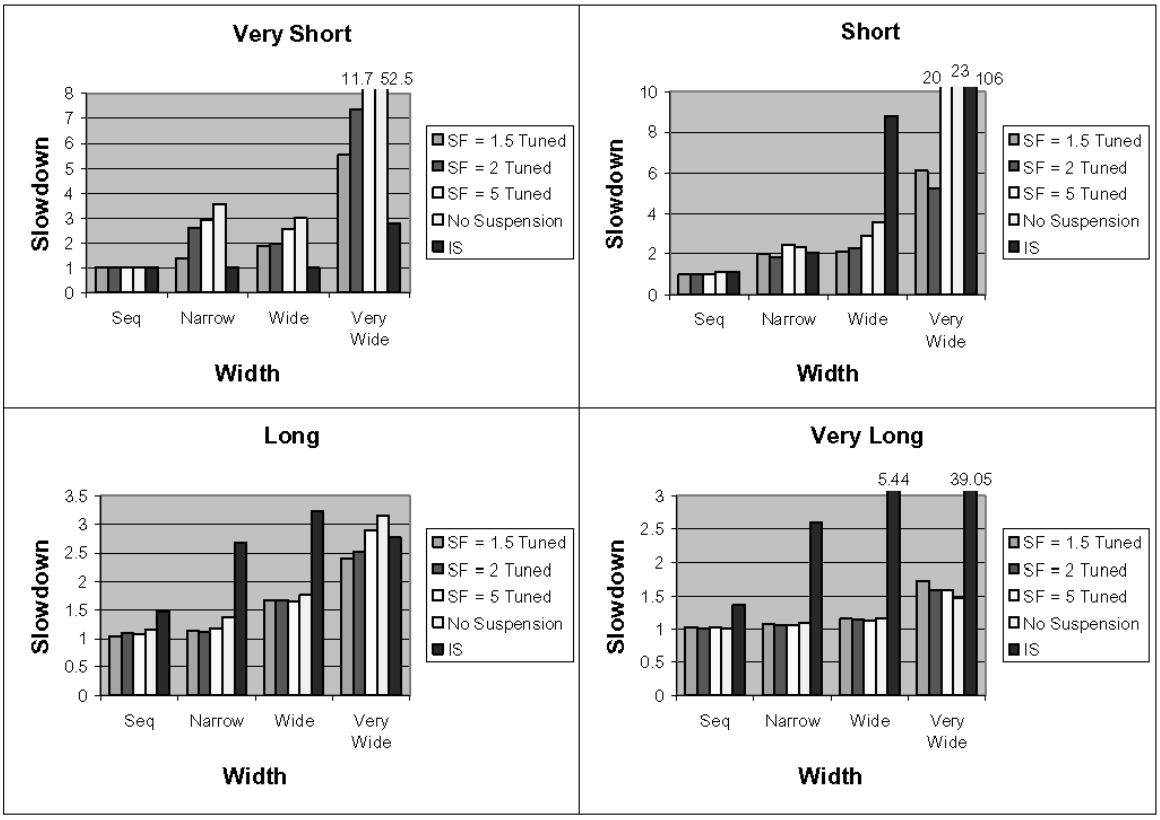

Figure 25 *Average slowdown of well estimated jobs: SDSC trace. Compared to NS, SS significantly improves the slowdowns for most of the categories with little deterioration to other categories. The performance of IS is bad except for VS categories*

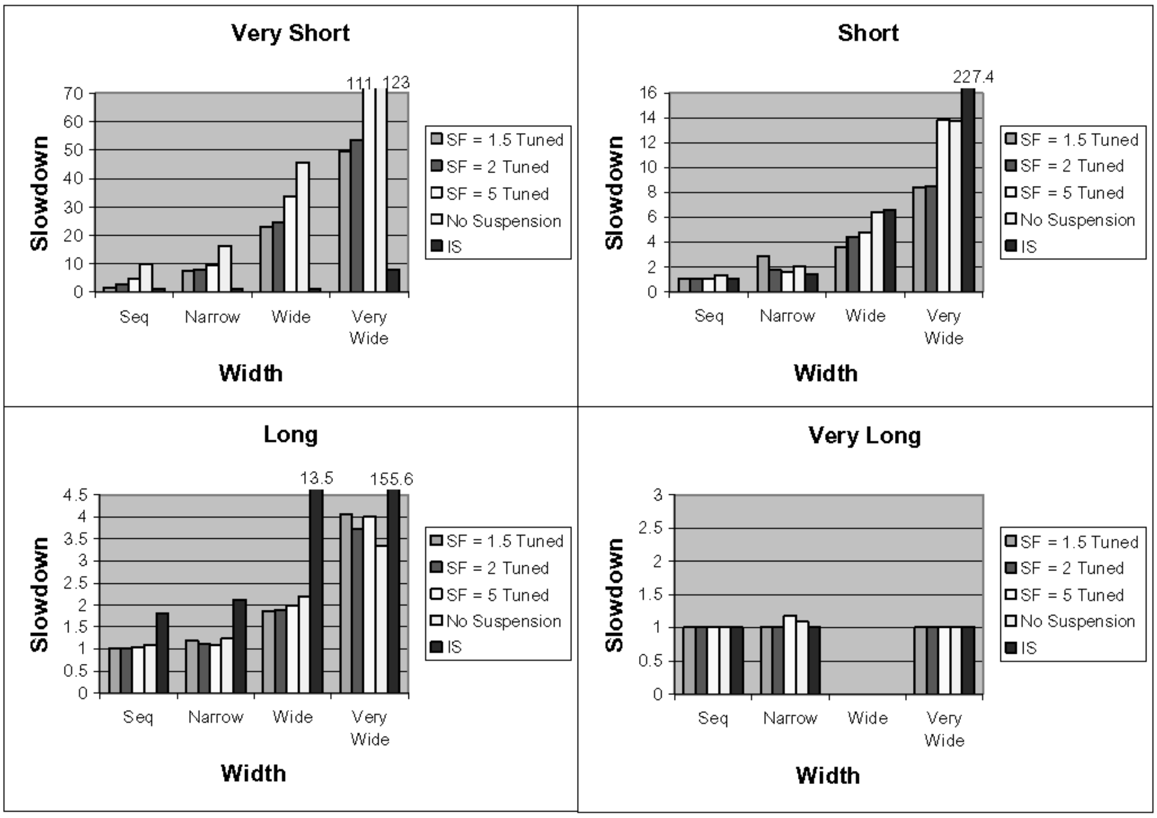

Figure 26 *Average slowdown of badly estimated jobs: SDSC trace. Compared to NS, SS provides a slight improvement in slowdowns for many categories. SS tends to penalise the badly estimated jobs in VS categories*

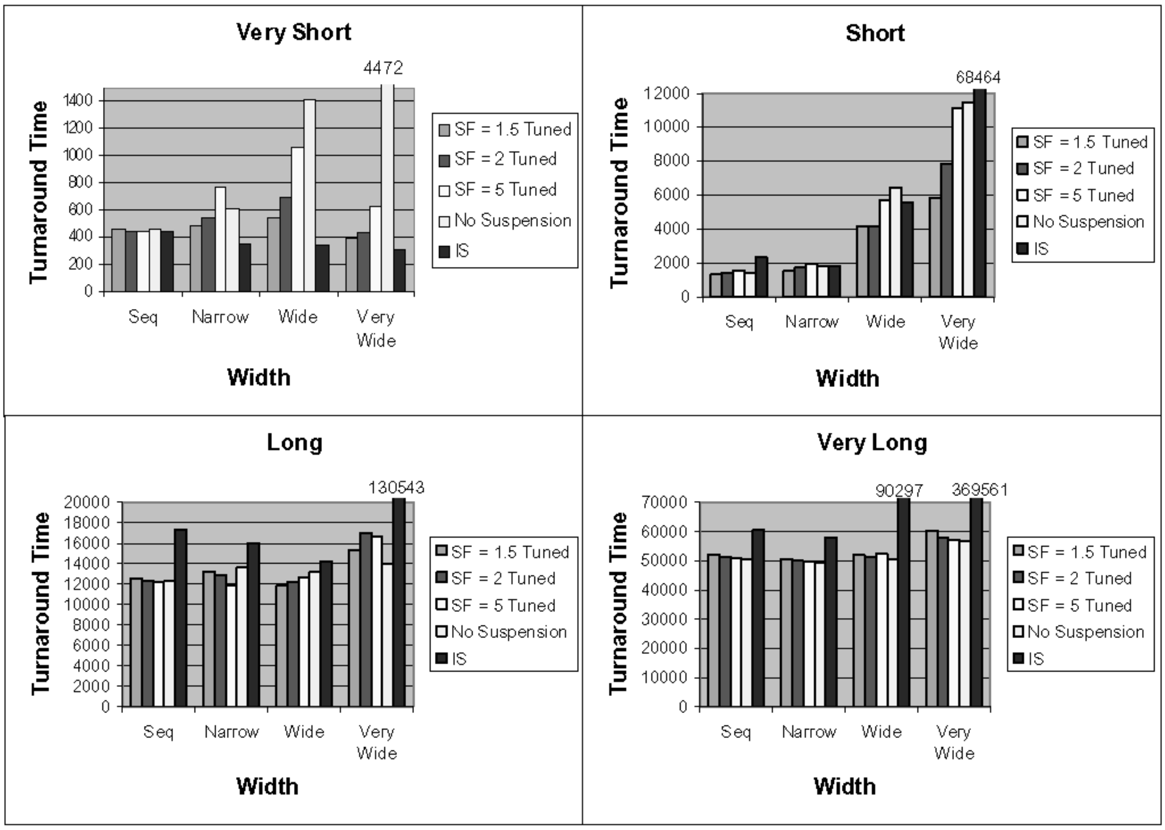

Figure 27 *Average turnaround time of well estimated jobs: CTC trace. Compared to NS, SS significantly improves the turnaround times for most of the categories with little deterioration to other categories. The performance of SS is comparable to IS for VS categories*

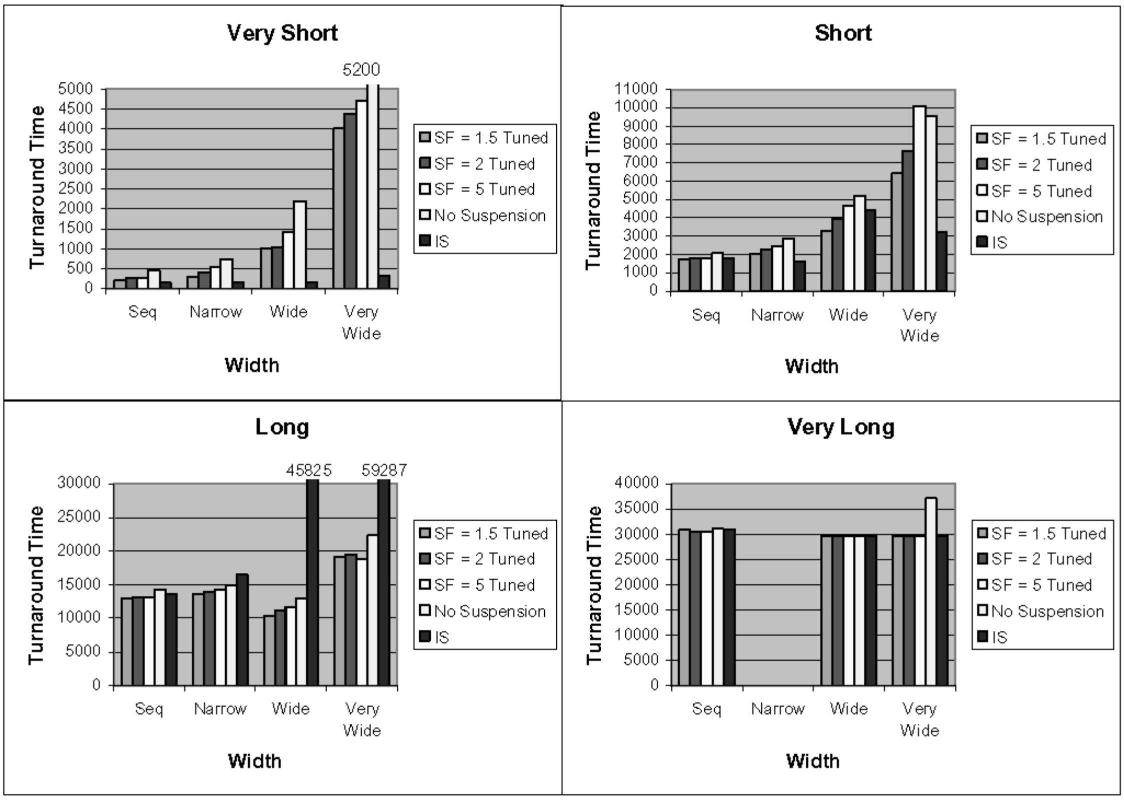

Figure 28 *Average turnaround time of badly estimated jobs: CTC trace. Compared to NS, SS provides a slight improvement in turnaround times for many categories. SS tends to penalise the badly estimated jobs in VS categories. IS gives better performance for VS, S, and VL categories*

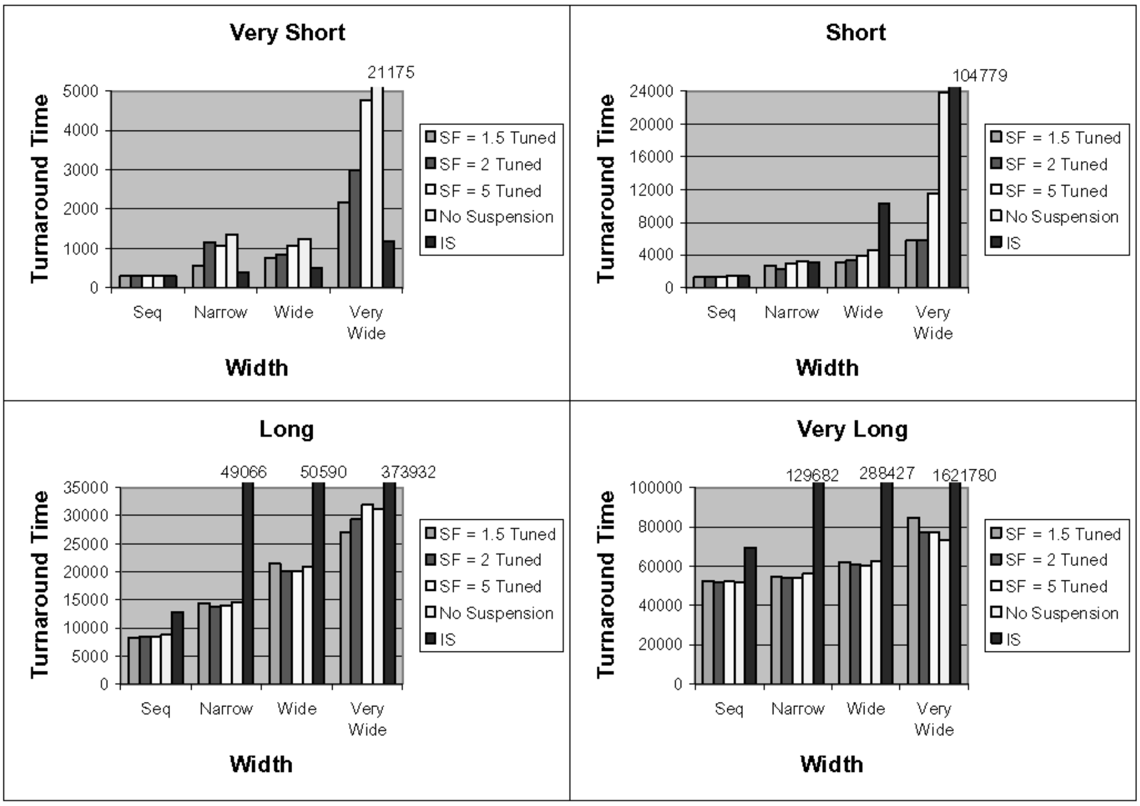

Figure 29 *Average turnaround time of well estimated jobs: SDSC trace. Compared to NS, SS significantly improves the turnaround times for most of the categories with little deterioration to other categories. The performance of IS is very bad for long jobs*

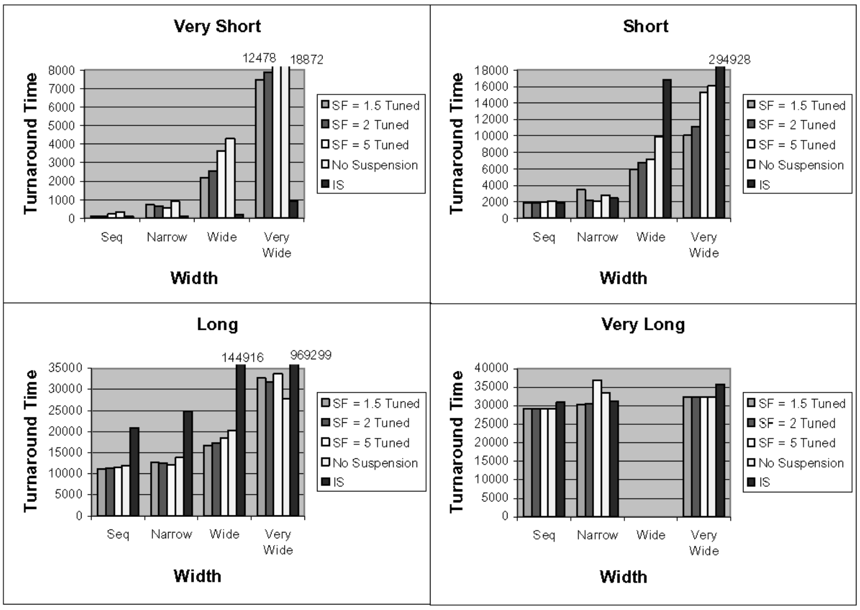

Figure 30 *Average turnaround time of badly estimated jobs: SDSC trace. Compared to NS, SS provides a slight improvement in turnaround times for many categories. SS tends to penalise the badly estimated jobs in VS categories. The performance of IS is bad except for VS categories*

5.1 Modelling of job suspension overhead

We have so far assumed no overhead for preemption of jobs. In this section, we present simulation results that incorporate overheads for job suspension. Since the job traces did not have information about job memory requirements, we considered the memory requirement of jobs to be random and uniformly distributed between 100 MB and 1 GB. The overhead for suspension is calculated as the time taken to write the main memory used by the job to the disk. The memory transfer rate that we considered is based on the following scenario: with a commodity local disk for every node, with each node being a quad, the transfer rate per processor was assumed to be 2 MB/s (corresponding to a disk bandwidth of 8 MB/s).

Figures 31 and 32 compare respectively the slowdowns and turnaround times of the proposed tuneable scheme with NS and IS in the presence of overhead for job suspension/restart for the CTC trace. Figures 33 and 34 compare respectively the slowdowns and turnaround times of the proposed tuneable scheme with NS and IS in the presence of overhead for job suspension/restart for the SDSC trace. One can observe that overhead does not significantly affect the performance of the SS scheme.

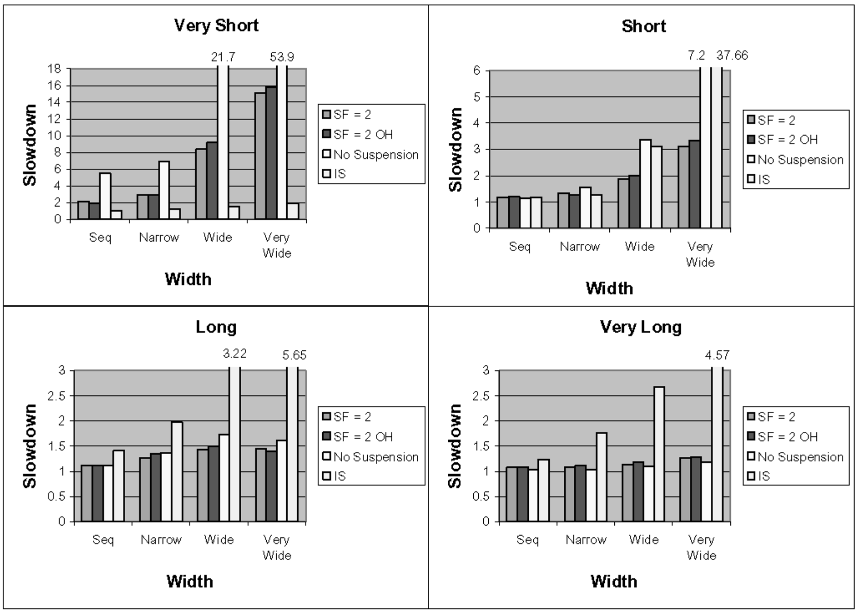

Figure 31 *Average slowdown with modelling of overhead for suspension/restart: CTC trace. The impact of overhead on the performance of SS scheme is minimal*

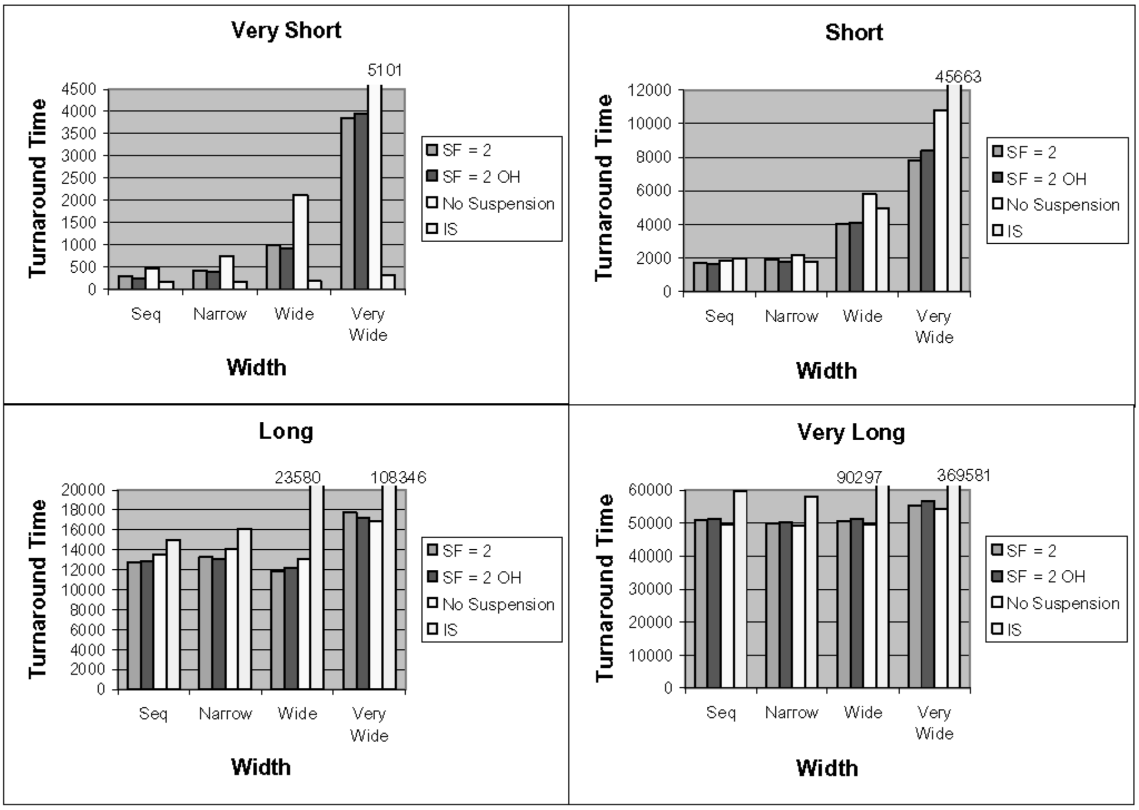

Figure 32 *Average turnaround time with modelling of overhead for suspension/restart: CTC trace. The impact of overhead on the performance of SS scheme is minimal*

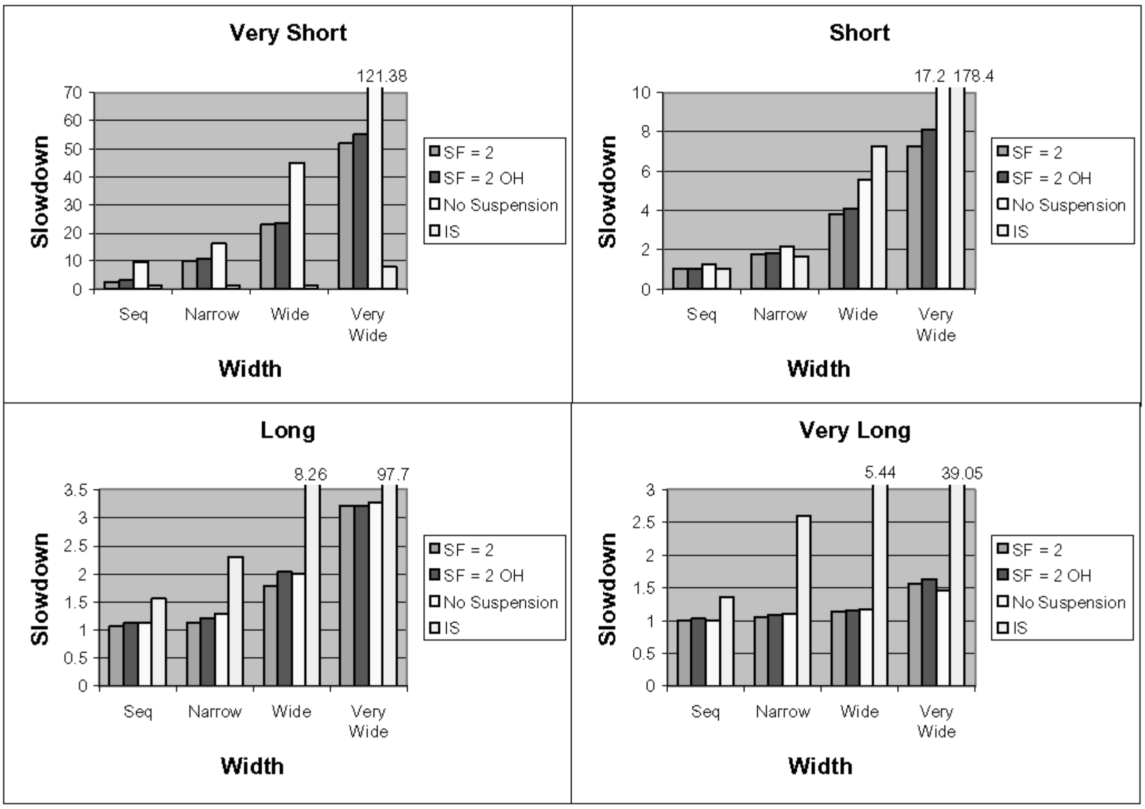

Figure 33 *Average slowdown with modelling of overhead for suspension/restart: SDSC trace. The impact of overhead on the performance of SS scheme is minimal*

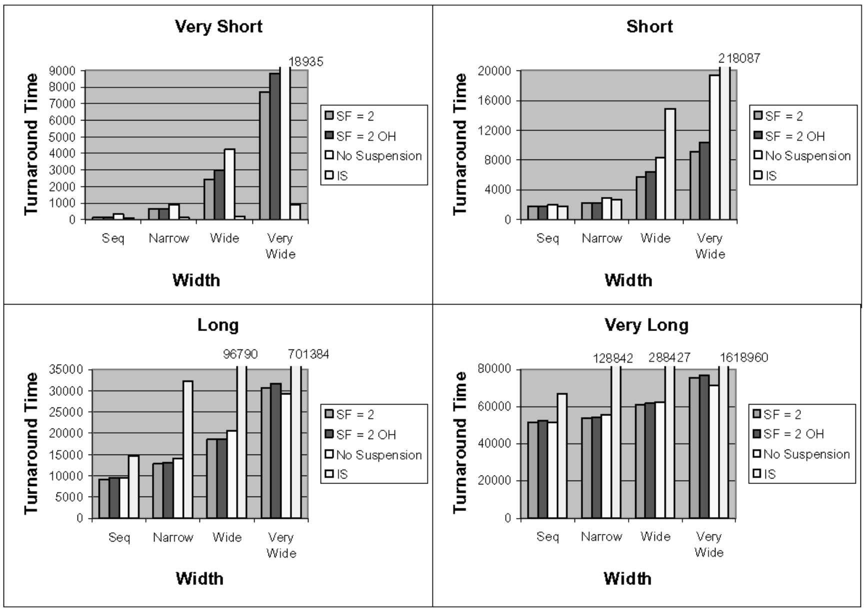

Figure 34 *Average turnaround time with modelling of overhead for suspension/restart: SDSC trace. The impact of overhead on the performance of SS scheme is minimal*

6 LOAD VARIATION

We have so far seen the performance of the Selective Suspension scheme under normal load. In this section, we present the performance of the SS scheme under different load conditions starting from the normal load (original trace) and increasing load until the system reaches saturation. The different loads correspond to modification of the traces by dividing the arrival times of the jobs by suitable constants, keeping their run time the same as in the original trace. For example, the job trace for a load factor of 1.1 is obtained by dividing the arrival times of the jobs in the original trace by 1.1.

For simplicity, we have reduced the number of job categories from 16 to 4 for the load variation studies: two categories based on their run time – Short (S) and Long (L) – and two categories based on the number of processors requested – Narrow (N) and Wide (W). The criteria used for job classifications are shown in Table 6. The distribution of jobs in the CTC and SDSC traces, corresponding to the four categories, is given in Tables 7 and 8 respectively. Figures 35 and 36 show the overall system utilisation for different schemes under different load conditions for the CTC and SDSC traces. One can observe that the SS scheme is able to achieve a better utilisation than the NS scheme at higher loads, whereas the overall system utilisation is very low under the IS scheme. Also, there is no significant increase in the overall system utilisation (for both the SS and NS schemes) when the load factor is increased beyond 1.6 (for CTC) and 1.3 (for SDSC). This result indicates that the system reaches saturation at a load factor of 1.6 (for CTC) and 1.3 (SDSC). We report the performance of the SS scheme for various load factors between 1.0 (normal) and 1.6 for the CTC trace and between 1.0 and 1.3 for the SDSC trace.

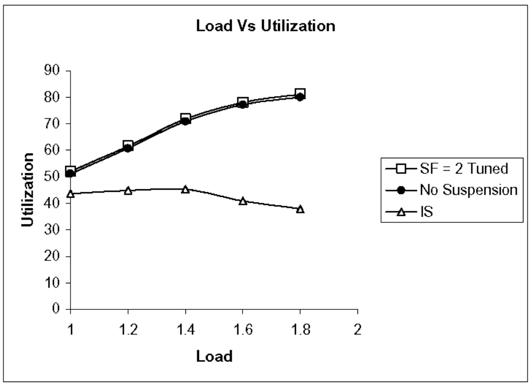

Figure 35 *Overall system utilisation under different load conditions: CTC trace. The overall system utilisation with the SS scheme is better than or comparable to the NS scheme. The performance of IS is much worse*

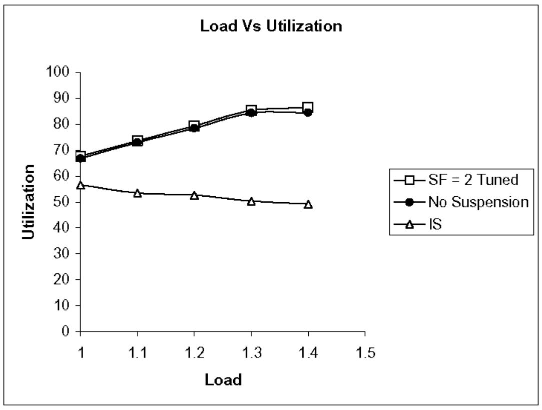

Figure 36 *Overall system utilisation under different load conditions: SDSC trace. The overall system utilisation with the SS scheme is better than or comparable to the NS scheme. The performance of IS is much worse*

Table 6 *Job categorisation criteria for load variation studies*

	≤8 Processors	*>8 Processors*
≤1 hr	SN	SW
>1 hr	LN	LW

Table 7 *Job distribution by category for load variation studies – CTC trace*

	≤8 Processors (%)	*>8 Processors* (%)
≤1 hr	44	13
>1 hr	30	13

Table 8 *Job distribution by category for load variation studies – SDSC trace*

	≤8 Processors (%)	*>8 Processors* (%)
≤1 hr	47	22
>1 hr	21	10

Figures 37–40 compare the performance of the SS scheme with the NS and IS schemes for different job categories under different load conditions for CTC and SDSC traces. It can be observed that the improvements obtained by the SS scheme are more pronounced under high load. The trends with respect to different categories under higher loads are similar to that observed under the normal load. It provides significant benefit to the short jobs without affecting the performance of long jobs. The IS scheme is better than the SS scheme only for the SN jobs in terms of average turnaround time, whereas it is better than SS for both SN and SW jobs in terms of average slowdown. It implies that the IS scheme improves the performance of only the relatively shorter jobs in the SW category by adversely affecting the performance of the relatively longer jobs. Also, the performance of the IS scheme is much worse for the long jobs, a very undesirable situation.

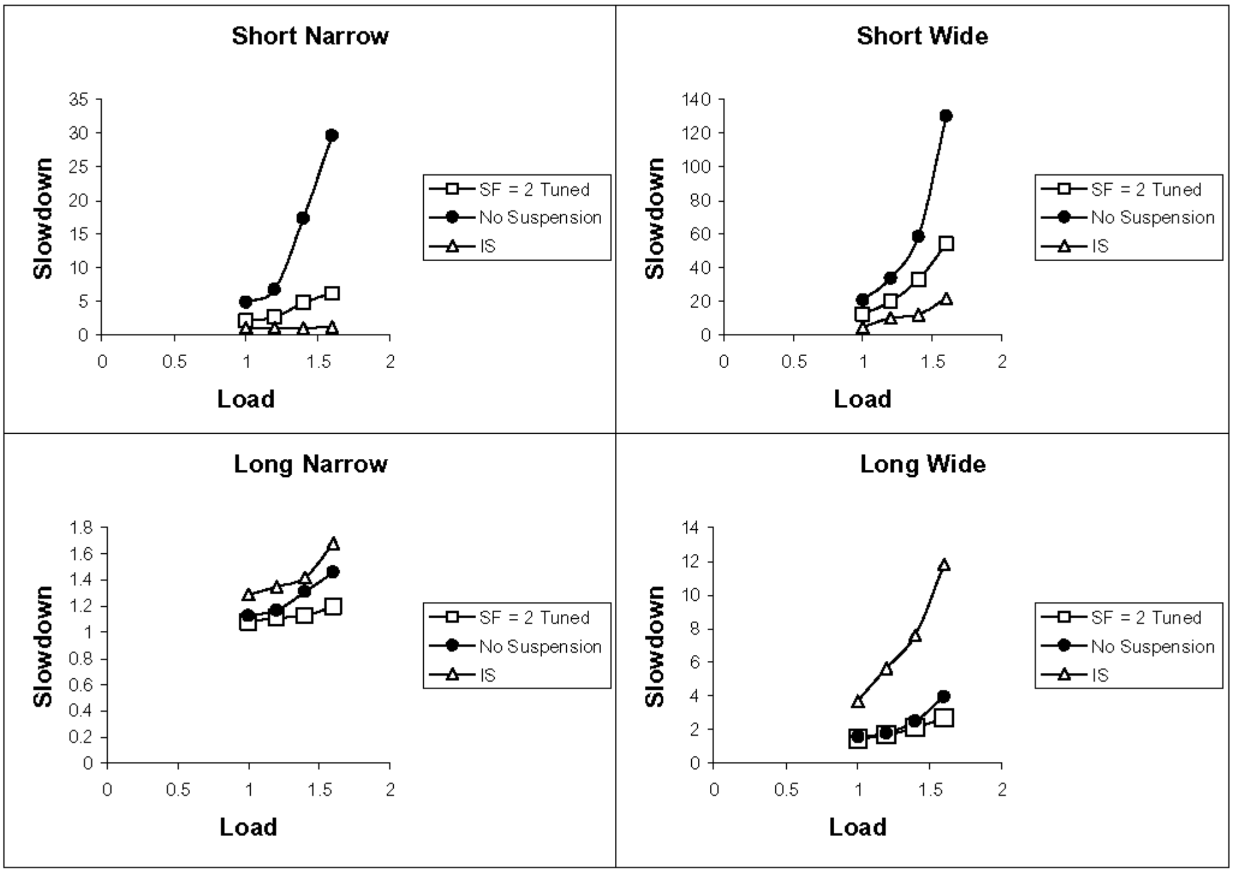

Figure 37 *Average slowdown: varying load; CTC trace. The improvements achieved by the SS scheme are more pronounced under high load*

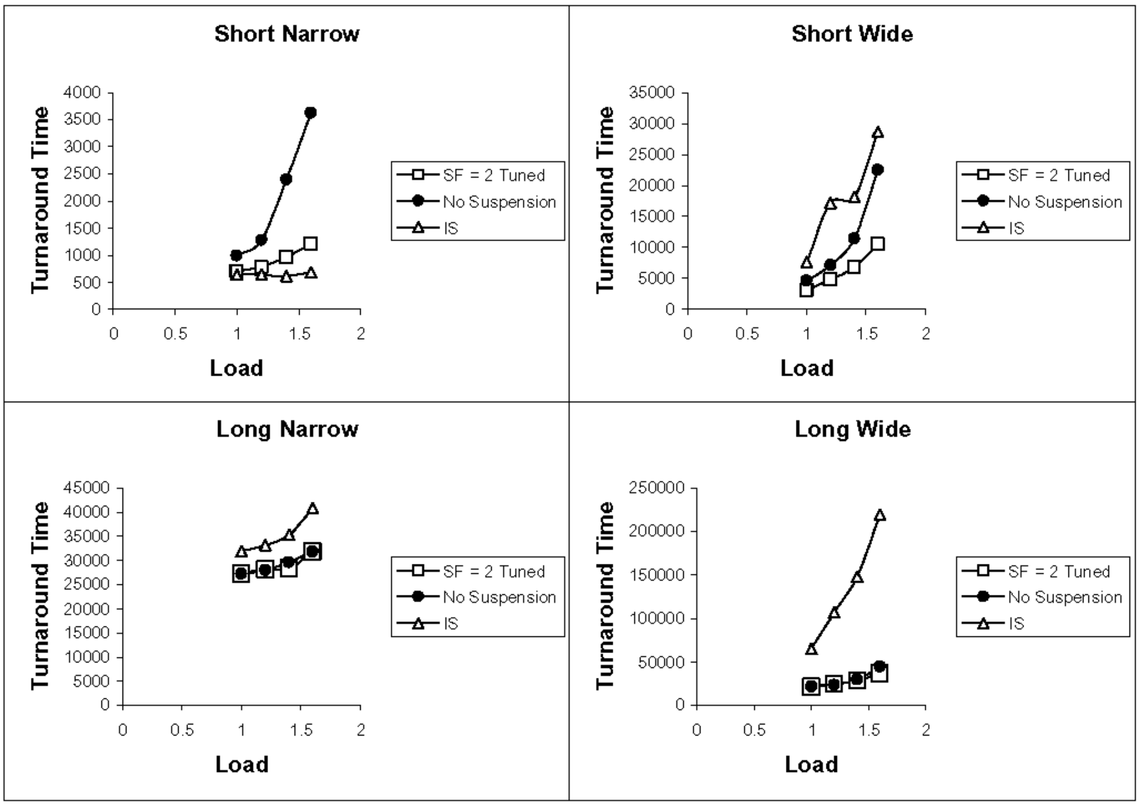

Figure 38 *Average turnaround time: varying load; CTC trace. The improvements achieved by the SS scheme are more pronounced under high load*

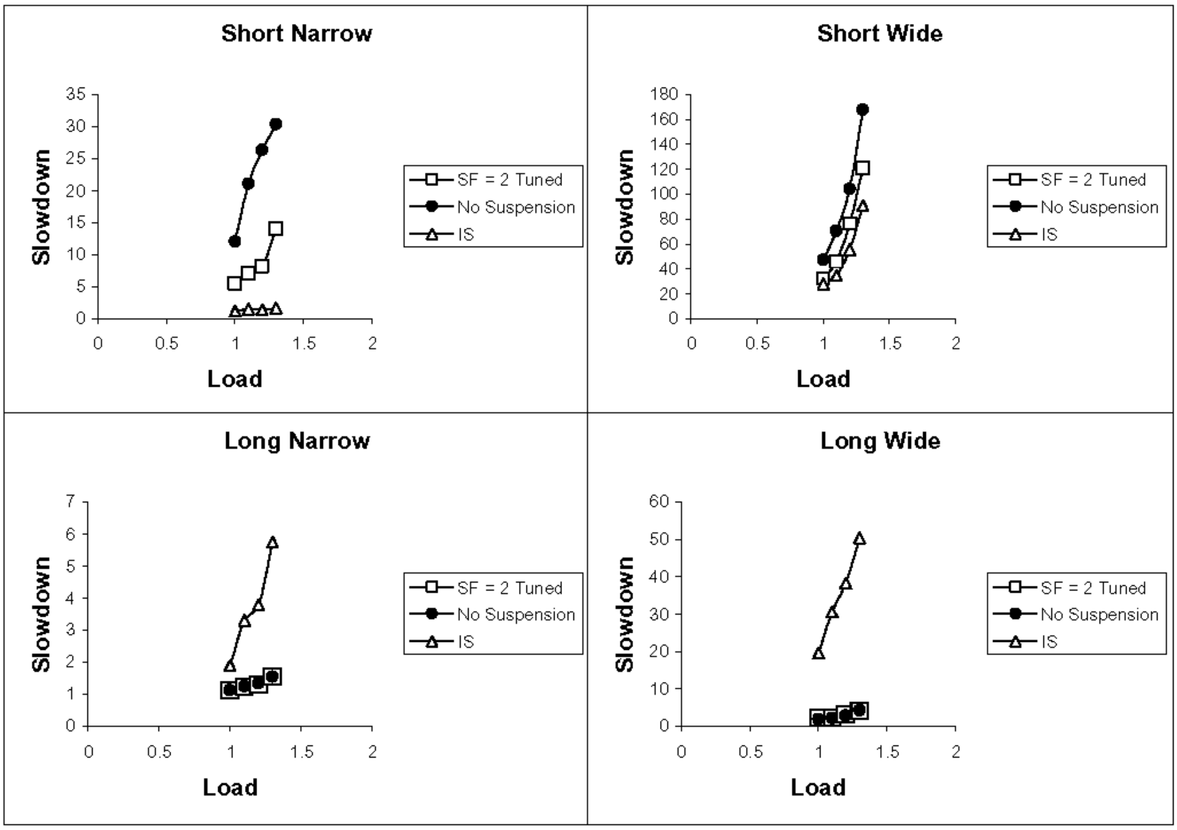

Figure 39 *Average slowdown: varying load; SDSC trace. The improvements achieved by the SS scheme are more pronounced under high load*

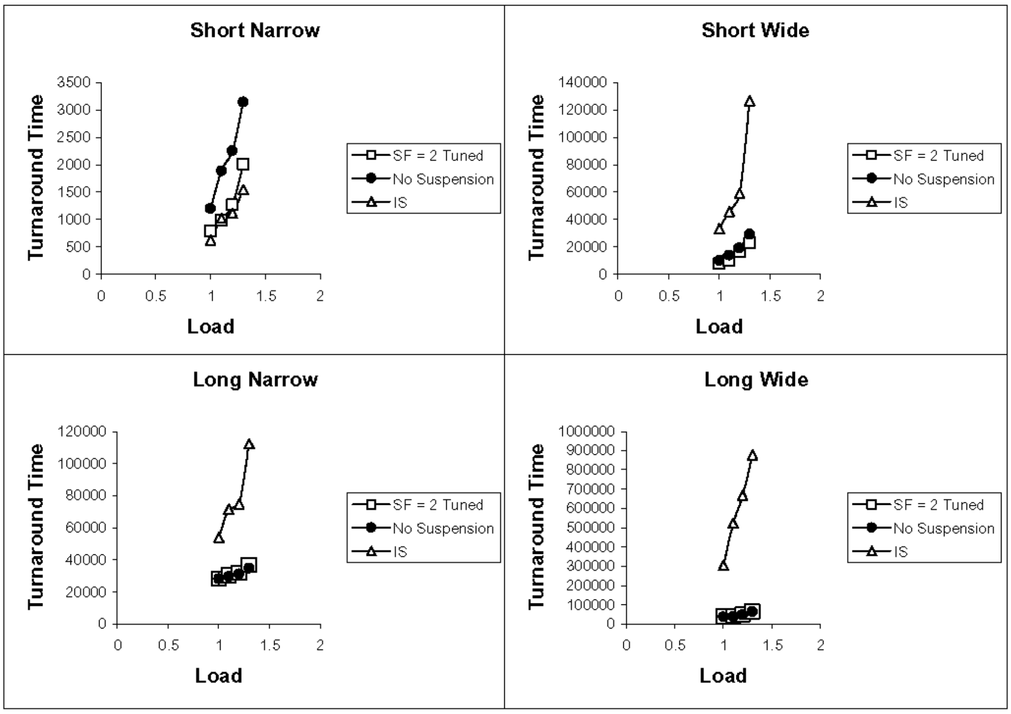

Figure 40 *Average turnaround time: varying load; SDSC trace. The improvements achieved by the SS scheme are more pronounced under high load*

Figures 41 and 42 compare respectively the average slowdowns and the average turnaround times of the jobs in the CTC trace against the overall system utilisation for various schemes. Figures 43 and 44 compare respectively the average slowdowns and the average turnaround times of the jobs in the SDSC trace against the overall system utilisation for various schemes. The SS scheme clearly is much better than both the IS and NS schemes. Even when the system is highly utilised, the SS scheme is able to provide much better response times for all categories of jobs. The IS scheme is not able to achieve high system utilisation.

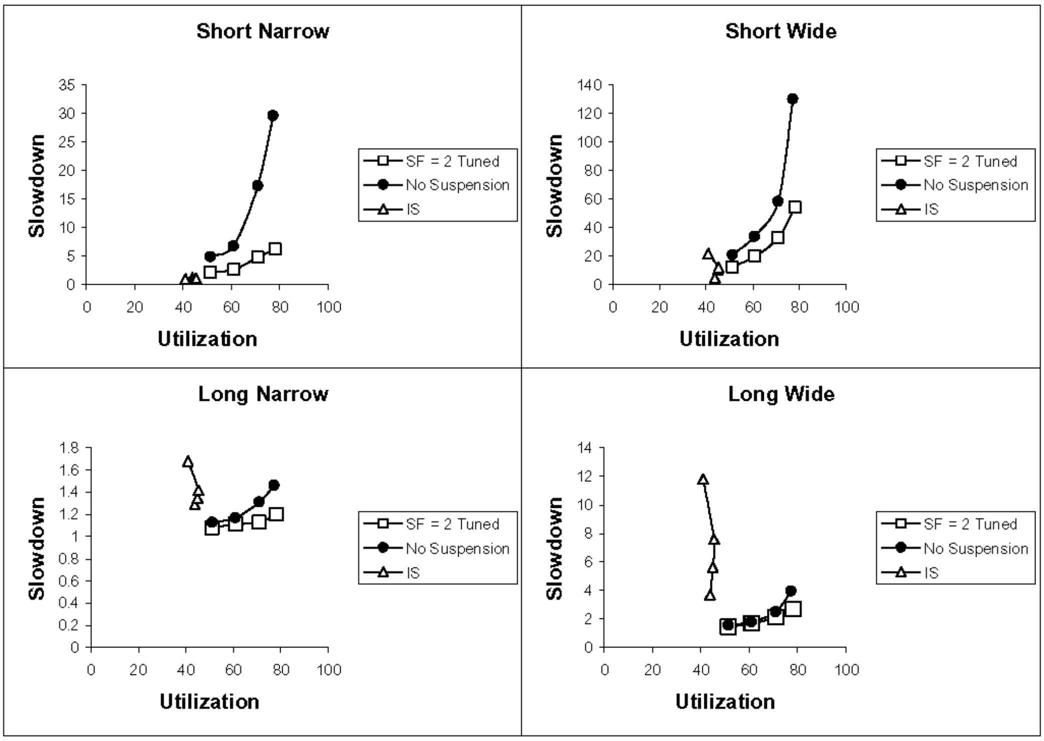

Figure 41 *Average slowdown vs. system utilisation: CTC trace. SS provides better performance even if the system is heavily utilised*

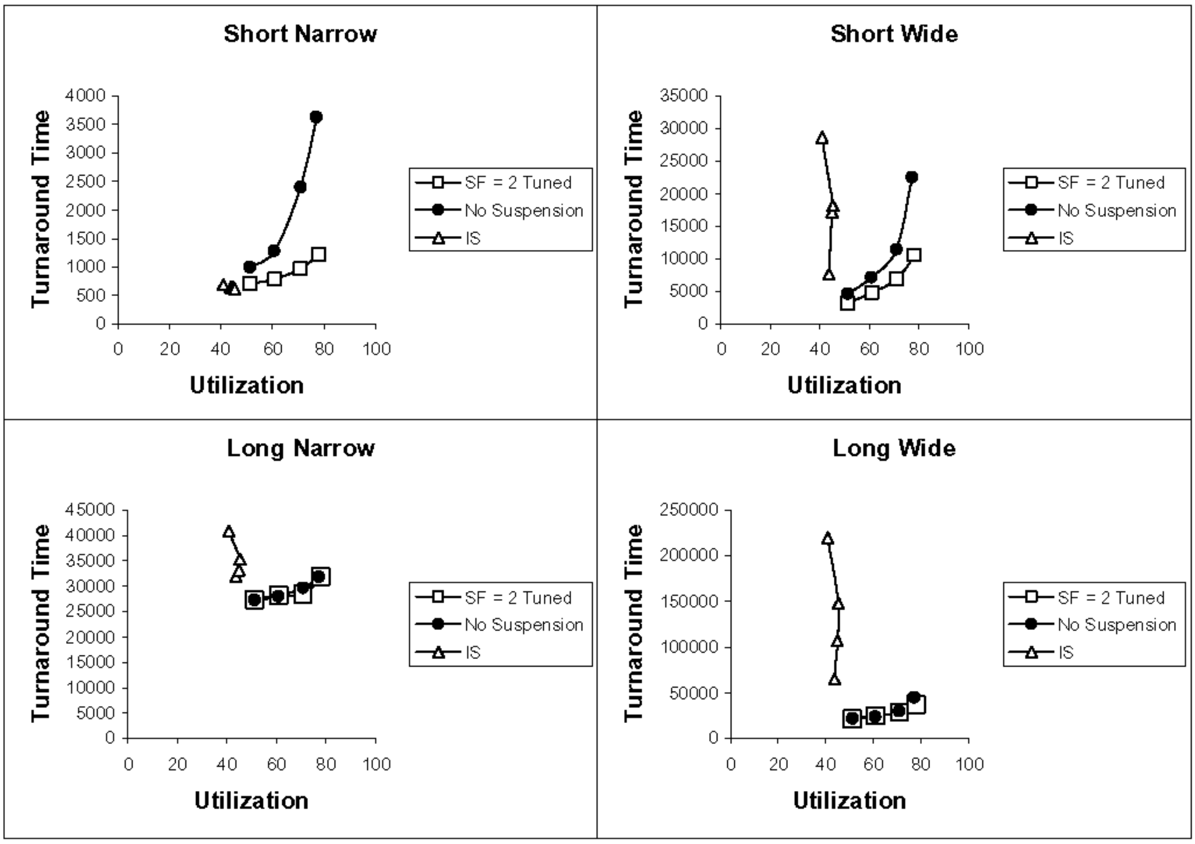

Figure 42 *Average turnaround time vs. system utilisation: CTC trace. SS provides better performance even if the system is heavily utilised*

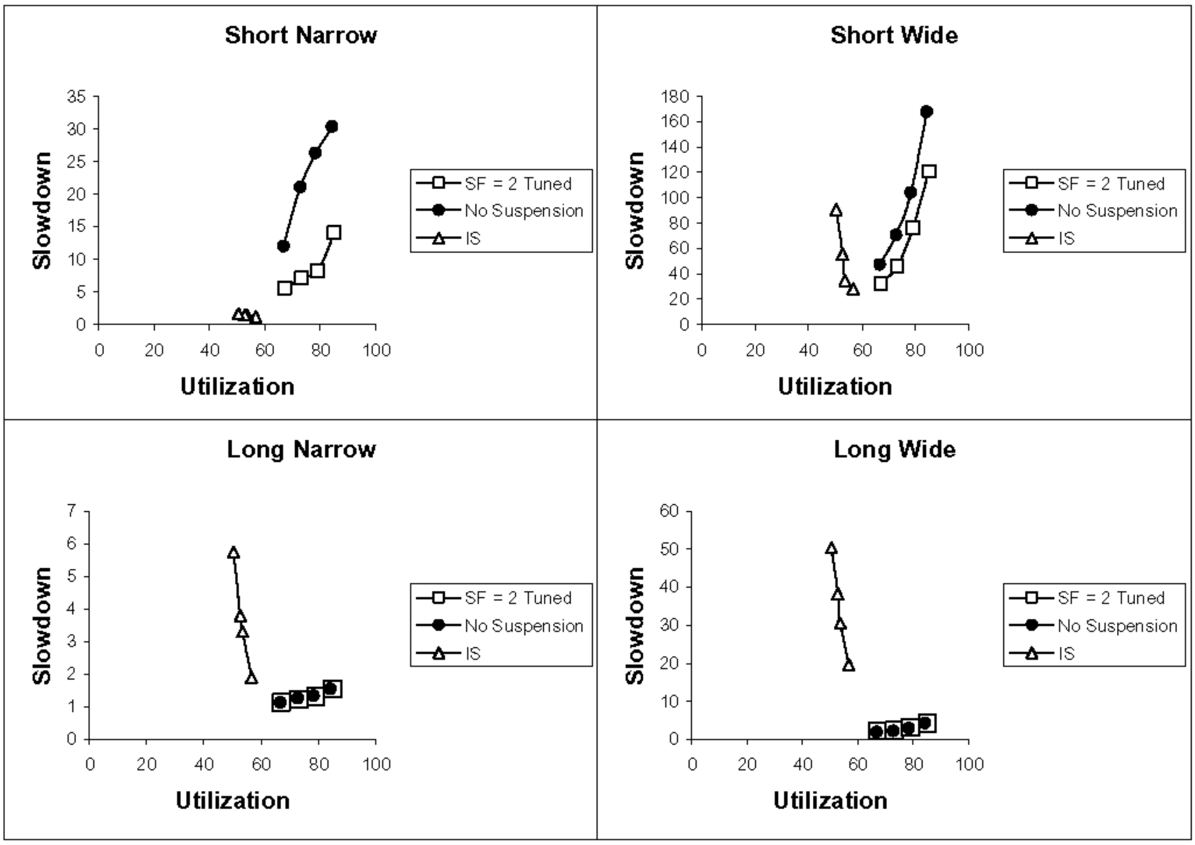

Figure 43 *Average slowdown vs. system utilisation: SDSC trace. SS provides better performance even if the system is heavily utilised*

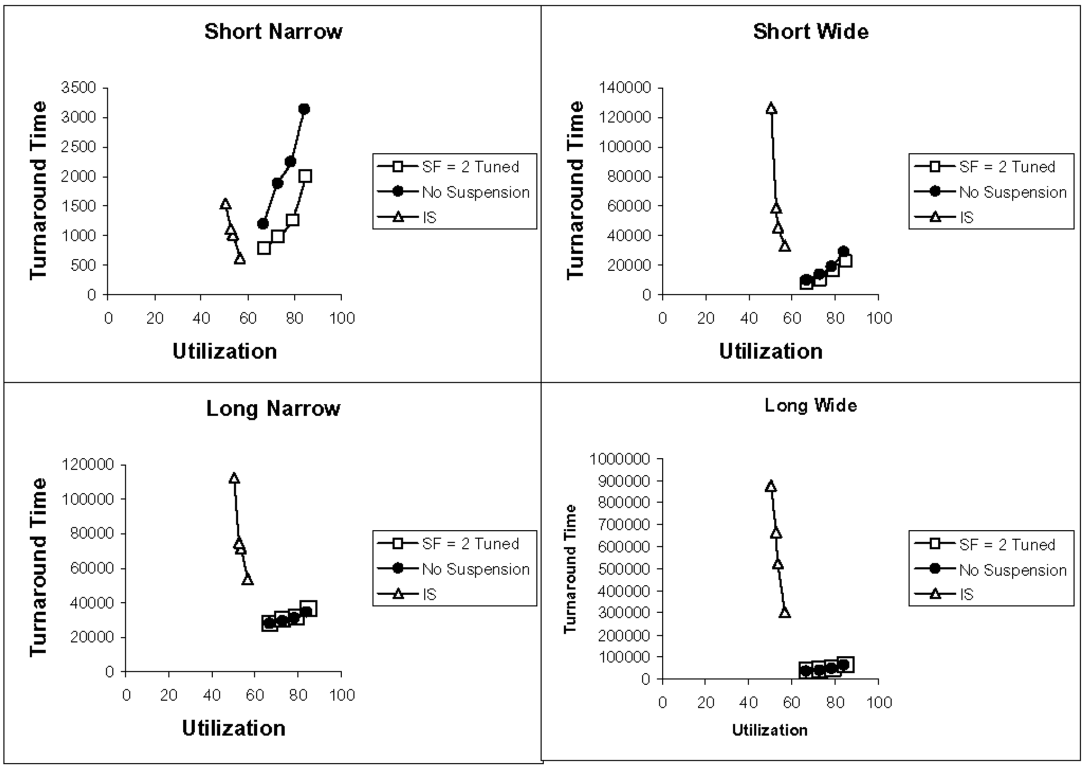

Figure 44 *Average turnaround time vs. system utilisation: SDSC trace. SS provides better performance even if the system is heavily utilised*

7 CONCLUSIONS

In this paper, we have explored the issue of preemptive scheduling of parallel jobs, using job traces from different supercomputer centres. We have proposed a tuneable, selective suspension scheme and demonstrated that it provides significant improvement in the average and the worst case slowdown of most job categories. It was also shown to provide better slowdown for most job categories over a previously proposed Immediate Service scheme. We also modelled the effect of overheads for job suspension, showing that even under stringent assumptions about available bandwidth to disk, the proposed scheme provides significant benefits over nonpreemptive scheduling and the Immediate Service strategy. We also evaluated the proposed schemes in the presence of inaccurate estimate of job run times and showed that the proposed scheme produced good results. Further, we showed that the Selective Suspension strategy provides greater benefits under high system loads compared to the other schemes.

ACKNOWLEDGEMENTS

We thank the anonymous referees for their helpful suggestions on improving the presentation of the paper. This work was supported in part by Sandia National Laboratories and in part by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, US Department of Energy, under Contract W-31-109-ENG-38.

REFERENCES

Aida, K. (2000) 'Effect of job size characteristics on job scheduling performance', in *Workshop on Job Scheduling Strategies for Parallel Processing*, Springer, Lecture Notes in Computer Science, Cancun, Mexico, Vol. 1911, pp.1–17, [Online] Available: citeseer.nj.nec.com/319169.html.

Anastasiadis, S.V. and Sevcik, K.C. (1997) 'Parallel application scheduling on networks of workstations', *Journal of Parallel and Distributed Computing*, Vol. 43, No. 2, pp.109–124, [Online] Available: citeseer.nj.nec.com/article/anastasiadis96parallel.html.

Arndt, O., Freisleben, B., Kielmann, T. and Thilo, F. (2000) 'A comparative study of online scheduling algorithms for networks of workstations', *Cluster Computing*, Saxony, Germany, Vol. 3, No. 2, pp.95–112, [Online] Available: citeseer.nj.nec.com/article/arndt98comparative.html.

Chiang, S.H. and Vernon, M.K. (2001) 'Production job scheduling for parallel shared memory systems', *Proceedings of International Parallel and Distributed Processing Symposium*, IEEE Computer Society, San Francisco, California, [Online] Available: citeseer.nj.nec.com/196999.html.

Chiang, S.H., Mansharamani, R.K. and Vernon, M.K. (1994) 'Use of application characteristics and limited preemption for run-to-completion parallel processor scheduling policies', *ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, ACM press, Nashville, Tennessee, USA, pp.33–44, [Online] Available: citeseer.nj.nec.com/chiang94use.html.

Cirne, W. (2003) 'When the herd is smart: aggregate behavior in the selection of job request', *IEEE Transactions on Parallel and Distributed Systems*, IEEE Press, Vol. 14, No. 2, February, pp.181–192 [Online] Available: citeseer.nj.nec.com/457615.html.

Cirne, W. and Berman, F. (2000) 'Adaptive selection of partition size for supercomputer requests', *Workshop on Job Scheduling Strategies for Parallel Processing*, Springer, Lecture Notes in Computer Science, Cancun, Mexico, Vol. 1911, pp.187–208, [Online] Available: citeseer.nj.nec.com/479768.html.

DasGupta, B. and Palis, M.A. (2000) 'Online real-time preemptive scheduling of jobs with deadlines', *Proceedings of the Third International Workshop on Approximation Algorithms for Combinatorial Optimization (Approx 2000)*, Springer-Verlag, Lecture Notes in Computer Science, Saarbrucken, Germany, pp.96–107 [Online] Available: citeseer.nj.nec.com/dasgupta00online.html.

Deng, X. and Dymond, P. (1996) 'On multiprocessor system scheduling', *Proceedings of the Eighth Annual ACM Symposium on Parallel Algorithms and Architectures*, ACM Press, Padua, Italy, pp.82–88.

Deng, X., Gu, N., Brecht, T. and Lu, K. (1996) 'Preemptive scheduling of parallel jobs on multiprocessors', *SODA: ACM-SIAM Symposium on Discrete Algorithms*, Society for Industrial and Applied Mathematics, Atlanta, Georgia, USA, pp.159–167 [Online] Available: citeseer.nj.nec.com/deng00preemptive.html.

Epstein, L. (2001) 'Optimal preemptive scheduling on uniform processors with non-decreasing speed ratios', *STACS 2001, Proceedings of 18th Annual Symposium on Theoretical Aspects of Computer Science*, Dresden, Germany, Vol. 2010, pp.230–248, [Online] Available: citeseer.nj.nec.com/epstein00optimal.html.

Feitelson, D.G. (2001) *Logs of Real Parallel Workloads from Production Systems*, http://www.cs.huji.ac.il/labs/parallel/workload/logs.html.

Feitelson, D.G. (2002) *Analyzing the Root Causes of Performance Evaluation Results*, Leibniz Center, Hebrew University, Tech. Rep.

Feitelson, D.G. and Jette, M.A. (1997) 'Improved utilization and responsiveness with gang scheduling', *Workshop on Job Scheduling Strategies for Parallel Processing*, Springer-Verlag, Lecture Notes in Computer Science, Geneva, Switzerland, Vol. 1291, pp.238–261.

Feitelson, D.G., Rudolph, L., Schwiegelshohn, U., Sevcik, K.C. and Wong, P. (1997) 'Theory and practice in parallel job scheduling', *Workshop on Job Scheduling Strategies for Parallel Processing*, Springer-Verlag, Lecture Notes in Computer Science, Geneva, Switzerland, Vol. 1291, pp.1–34.

Jackson, D., Snell, Q. and Clement, M.J. (2001) 'Core algorithms of the maui scheduler', *Workshop on Job Scheduling Strategies for Parallel Processing*, Springer, Lecture Notes in Computer Science, Cambridge, Massachussets, USA, Vol. 2221, pp.87–102, [Online] Available: citeseer.nj.nec.com/479768.html.

Jones, J.P. and Nitzberg, B. (1999) 'Scheduling for parallel supercomputing: a historical perspective of achievable utilization', *Workshop on Job Scheduling Strategies for Parallel Processing*, Springer, Lecture Notes in Computer Science, San Juan, Puerto Rico, Vol. 1291, pp.1–16, [Online] Available: citeseer.nj.nec.com/patton99scheduling.html.

Keleher, P.J., Zotkin, D. and Perkovic, D. (2000) 'Attacking the bottlenecks of backfilling schedulers', *Cluster Computing*, Saxony, Germany, Vol. 3, No. 4, pp.245–254, [Online] Available: citeseer.nj.nec.com/467800.html.

Krallmann, J., Schwiegelshohn, U. and Yahyapour, R. (1999) 'On the design and evaluation of job scheduling algorithms', *Workshop on Job Scheduling Strategies for Parallel Processing*, Springer, Lecture Notes in Computer Science, San Juan, Puerto Rico, Vol. 1291, pp.17–42, [Online] Available: citeseer.nj.nec.com/krallmann99design.html.

Lawson, B.G. and Smirni, E. (2002) 'Multiple-queue backfilling scheduling with priorities and reservations for parallel systems', *Workshop on Job Scheduling Strategies for Parallel Processing*, Springer, Lecture Notes in Computer Science, Edinburgh, Scotland, Vol. 2537, pp.72–87.

Lawson, B.G., Smirni, E. and Puiu, D. (2002) 'Self-adapting backfilling scheduling for parallel systems', *Proceedings of the International Conference on Parallel Processing*, IEEE Computer Society, Vancouver, Canada, pp.583–592.

Leutenneger, L.T. and Vernon, M.K. (1990) 'The performance of multiprogrammed multiprocessor scheduling policies', *ACM SIGMETRICS Conference on Measurement and Modelling of Computer Systems*, ACM Press, Boulder, Colorado, USA, May, pp.226–236, [Online] Available: citeseer.nj.nec.com/196999.html.

Lifka, D. (1995) 'The ANL/IBM SP scheduling system', *Workshop on Job Scheduling Strategies for Parallel Processing*, Springer, Lecture Notes In Computer Science, Santa Barbara, California, USA, Vol. 949, pp.295–303.

McCann, C., Vaswani, R. and Zahorjan, J. (1993) 'A dynamic processor allocation policy for multiprogrammed shared-memory multiprocessors', *ACM Transactions on Computer Systems*, Vol. 11, No. 2, pp.146–178.

Mu'alem, A.W. and Feitelson, D.G. (2001) 'Utilization, predictability, workloads, and user runtime estimates in scheduling the IBM SP2 with backfilling', *IEEE Transactions on Parallel and Distributed Systems*, Vol. 12, No. 6, pp.529–543.

Parsons, E.W. and Sevcik, K.C. (1997) 'Implementing multiprocessor scheduling disciplines', in Feitelson, D.G. and Rudolph, L. (Eds.): *Job Scheduling Strategies for Parallel Processing*, Lecture Notes in Computer Science, Springer-Verlag, San Juan, Puerto Rico Vol. 1291, pp.166–192.

Perkovic, D. and Keleher, P.J. (2000) 'Randomization, speculation, and adaptation in batch schedulers', *Proceedings of the 2000 ACM/IEEE conference on Supercomputing (CDROM)*, IEEE Computer Society, Dallas, Texas, USA, p.7.

Sabin, G., Kettimuthu, R., Rajan, A. and Sadayappan, P. (2003) 'Scheduling of parallel jobs in a heterogeneous multi-site environment', in *Workshop on Job Scheduling Strategies for Parallel Processing*, Springer, Lecture Notes in Computer Science, Seattle, Washington, USA, Vol. 2862, pp.87–104.

Schwiegelshohn, U. and Yahyapour, R. (2000) 'Fairness in parallel job scheduling', *Journal of Scheduling*, Vol. 3, No. 5, pp.297–320, [Online] Available: citeseer.ist.psu.edu/schwiegelshohn00fairness.html.

Sevcik, K.C. (1994) 'Application scheduling and processor allocation in multiprogrammed parallel processing systems', *Performance Evaluation*, Vol. 19, Nos. 2–3, pp.107–140, [Online] Available: citeseer.nj.nec.com/sevcik93application.html.

Skovira, J., Chan, W., Zhou, H. and Lifka, D. (1996) 'The easy – loadleveler api project', in *Workshop on Job Scheduling Strategies for Parallel Processing*, Springer, Lecture Notes in Computer Science, Honolalu, Hawaii, USA, Vol. 1162 pp.41–47, [Online] Available: citeseer.nj.nec.com/479768.html.

Srinivasan, S., Kettimuthu, R., Subramani, V. and Sadayappan, P. (2002) 'Selective reservation strategies for backfill job scheduling', in *Workshop on Job Scheduling Strategies for Parallel Processing*, Springer Lecture Notes in Computer Science, Edinburgh, Scotland, Vol. 2537, pp.55–71.

Srinivasan, S., Kettimuthu, R., Subramani, V. and Sadayappan, P. (2002) 'Characterization of backfilling strategies for parallel job scheduling', *Proceedings of the ICPP-2002 Workshops*, IEEE Computer Society, Vancouver, Canada, pp.514–519.

Srinivasan, S., Subramani, V., Kettimuthu, R., Holenarsipur, P. and Sadayappan, P. (2002) 'Effective selection of partition sizes for moldable scheduling of parallel jobs', *Proceedings of the 9th International Conference on High Performance Computing*, Springer, Lecture Notes In Computer Science, Bangalore, India, Vol. 2552, pp.174–183.

Streit, A. (2001) 'On job scheduling for HPC-clusters and the dynP scheduler,' *Proceedings of the 8th International Conference on High Performance Computing*, Springer-Verlag, Lecture Notes In Computer Science, Hyderabad, India, Vol. 2228, pp.58–67.

Subramani, V., Kettimuthu, R., Srinivasan, S. and Sadayappan, P. (2002) 'Distributed job scheduling on computational grids using multiple simultaneous requests', *Proceedings of the 11th IEEE International Symposium on High Performance Distributed Computing*, IEEE Computer Society, Edinburgh, Scotland, pp.359–366.

Subramani, V., Kettimuthu, R., Srinivasan, S., Johnston, J. and Sadayappan, P. (2002) 'Selective buddy allocation for scheduling parallel jobs on clusters', *Proceedings of the IEEE International Conference on Cluster Computing*, IEEE Computer Society, Chicago, Illinois, USA, pp.107–116.

Talby, D. and Feitelson, D.G. (1999) 'Supporting priorities and improving utilization of the ibm sp scheduler using slack-based backfilling', *Proceedings of the 13th International Parallel Processing Symposium*, IEEE Computer Society, San Juan, Puerto Rico, pp.513–517 [Online] Available: citeseer.nj.nec.com/talby99supporting.html.

Ward, W.A., Mahood, C.L. and West, J.E. (2002) 'Scheduling jobs on parallel systems using a relaxed backfill strategy', in *Workshop on Job Scheduling Strategies for Parallel Processing*, Springer, Lecture Notes in Computer Science, Edinburgh, Scotland, Vol. 2537, pp.88–102.

Zahorjan, J. and McCann, C. (1990) 'Processor scheduling in shared memory multiprocessors', in *ACM SIGMETRICS Conference on Measurement and Modelling of Computer Systems*, ACM Press, Boulder, Colorado, USA, May, pp.214–225, [Online] Available: citeseer.nj.nec.com/196999.html.

Zotkin, D. and Keleher, P. (1999) 'Job-length estimation and performance in backfilling schedulers', *Proceedings of the 8th High Performance Distributed Computing Conference*, IEEE Computer Society, Redondo Beach, California, USA, pp.236–243, [Online] Available: citeseer.nj.nec.com/196999.html.