Preprint
UCRL-136397

An Evaluation of Parallel Job Scheduling for ASCI Blue-Pacific

H. Franke, J. Jann, J. Moreira, P. Pattnaik, M. Jette

This article was submitted to
Supercomputing 1999, Portland, OR, November 14-19, 1999

November 9, 1999

An Evaluation of Parallel Job Scheduling for ASCI Blue-Pacific

H. Franke J. Jann J. Moreira P. Pattnaik

IBM T. J. Watson Research Center
Yorktown Heights NY 10598-0218
{frankeh,joefon,jmoreira,pratap}@us.ibm.com

M. Jette

Lawrence Livermore National Laboratory
Livermore CA 94550
jette@llnl.gov

Abstract

In this paper we analyze the behavior of a gang-scheduling strategy that we are developing for the ASCI Blue-Pacific machines. Using actual job logs for one of the ASCI machines we generate a statistical model of the current workload with hyper Erlang distributions. We then vary the parameters of those distributions to generate various workloads, representative of different operating points of the machine. Through simulation we obtain performance parameters for three different scheduling strategies: (i) first-come first-serve, (ii) gang-scheduling, and (iii) backfilling. Our results show that backfilling can be very effective for the common operating points in the 60-70% utilization range. However, for higher utilization rates, time-sharing techniques such as gang-scheduling offer much better performance.

1 Introduction

Parallel job scheduling in large systems with hundreds or thousands of processors can be very challenging. A good job scheduling system works to maximize such objective measures as average job response time and system utilization. It also attempts to maximize the more subjective measure of user happiness. It has been shown in the literature that the scheduling strategy can have significant impact on the performance characteristics of a large parallel system.

In this paper we analyze the behavior of a gang-scheduling strategy that we are developing for the ASCI Blue-Pacific machines, at the Lawrence Livermore National Laboratory (LLNL). The larger of these machines (the SST machine, used for classified research) has approximately 1500 nodes, or 6000 processors. (Each node is a 4-processor SMP.) Even the "smaller" machine (the CTR machine, open to general research) has 320 nodes and more than 1200 processors. Academic and industrial experience in job scheduling for such large systems is practically nonexistent.

The workload of the CTR machine is very diverse, with many small jobs that only use one or a few nodes but also some very large jobs that use up to 256 nodes. One of the main goals of job scheduling for the CTR machine is to provide good average response time while efficiently servicing these very large jobs. Typical scheduling strategies for large parallel systems include first-come first-serve (FCFS), backfilling, and gang-scheduling. (Gang-scheduling with multiprogramming level of 1 reverts to FCFS.) FCFS and backfilling are already provided through current implementations of LoadLeveler. We have extended LoadLeveler with gang-scheduling, in a system codenamed GangLL, to provide a more flexible scheduling strategy.

Our goal in this paper is to conduct a workload sensitivity analysis of the performance of three common scheduling strategies (FCFS, backfilling, and gang-scheduling) for the CTR machine. It is important to check the range of acceptable operability of a new scheduling system before it is deployed. Although we do have job logs of the current use of the CTR machine, an increase in demand is expected. (An increase in demand is both a cause for and an effect of a better scheduling system.) To evaluate the performance of the schedulers at increasing loads, we use the following approach: We start with a real workload obtained from job logs of this machine. We then generate a statistical model of this workload using hyper Erlang distributions. We vary the parameters of these distributions to generate several workloads with different characteristics. Finally, we obtain performance characteristics for each of these workloads through simulation of the various scheduling strategies.

Our study will shows good space-sharing strategies, in particular backfilling, can deliver good response time averaged over all jobs. We also show that gang-scheduling with modest multiprogramming levels likewise delivers good average response time, but with the following added benefits: (i) an estimate of the job termination time is unnecessary, (ii) the response time for large production jobs is reduced, and (iii) the availability of the system for large jobs is increased. It is important to emphasize the relevance of large jobs. One can easily reduce average job response time by always favoring small jobs at the expense of large jobs. However, large jobs, although not numerous, typically represent important production runs and improving their response time is an important step in achieving happiness of the user community.

The rest of this paper is organized as follows. Section 2 gives some details of the ASCI Blue-Pacific machines. Section 3 describes the scheduling strategies analyzed in this paper. Section 4 describes our methodology for evaluating the behavior of the different strategies. Section 5 presents and discusses the results of our evaluation. Finally, Section 6 presents our conclusions and discusses future work.

2 The ASCI Blue-Pacific Machines

The main machine in the ASCI Blue-Pacific program is the *Sustained Stewardship TeraOPS* (SST) "Hyper-Cluster". Figure 1 shows the high-level organization of the SST machine. It consists of three 488-node *sectors* that are connected via high performance gateway links. Each node is comprised of a 4-way SMP with 332 MHz PowerPC 604e processors and up to 2.5 GB of main memory. Each node executes its own operating system (AIX) image. The nodes in each sector are interconnected via a TBMX high-performance switch that delivers a bidirectional bandwidth of 150 MB/s per node.

While the SST machine is for use on classified applications, the ASCI Blue-Pacific program also provides computing resources to academic research through the ASCI Strategic Alliance Program. In this program the *Combined Technology Refresh* (CTR) machine provides 320 nodes of similar characteristic as the SST nodes, organized in a single sector. Features of both machines are summarized in Table 1. For the remaining of this paper we discuss job scheduling for the CTR machine. Techniques developed for the CTR machine may eventually be incorporated into the SST machine.

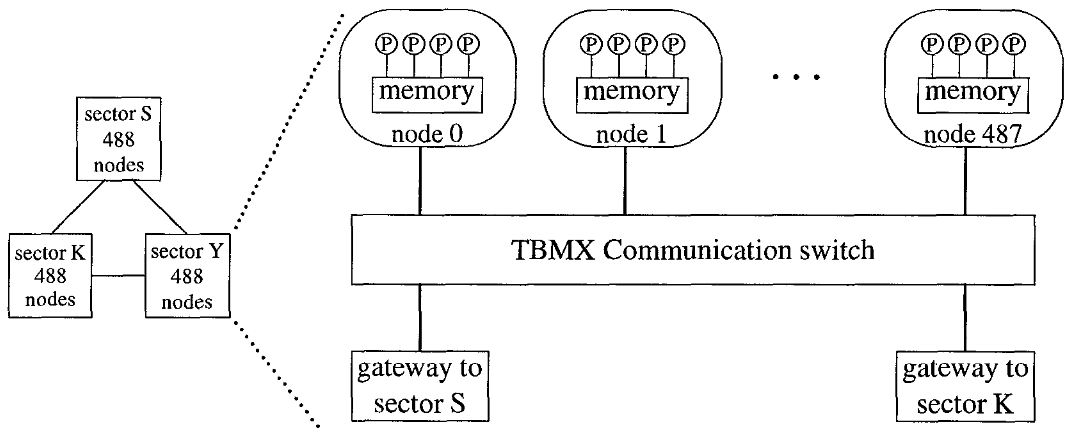


Figure 1: High-level organization of the ASCI Blue-Pacific machine.

Table 1: Characteristics of the ASCI Blue-Pacific machines.

characteristic	SST	CTR
Peak speed	3.9 teraOPS	850 gigaOPS
Memory	2.6 terabytes	480 gigabytes
Local Disk	17.3 terabytes	3.0 terabytes
Min/Max memory/node	1.5-2.5 gigabytes	1.5 gigabytes
Number of compute nodes (4-way SMPs)	1,464	320
Node-to-node bandwidth (TBMX switch, bidirectional)	150 MB/s	150 MB/s
Total number of processors	5,856	1,280
Processor-to-memory bandwidth	2.1 TB/s	2.1 TB/s
Compute node peak performance	2.656 gigaOPS	2.656 gigaOPS
Delivered RAID I/O bandwidth	6.4 GB/s	320 MB/s
Delivered I/O bandwidth to local disk	10.5 GB/s	4.7 GB/s
RAID storage	62.5 terabytes	10.0 terabytes

3 Scheduling strategies for ASCI Blue-Pacific

Despite the size of the CTR machine, we know from LLNL's experience [1, 4, 7] with this and other large systems that the demand for resources to execute jobs is often larger than the total number of processors available. The traditional solution to this problem in a space-sharing environment is to schedule jobs for execution strictly in the order they arrive (FCFS) and to queue the excess jobs: They wait to start execution until some running jobs have finished. It is well known in the literature that this can be highly detrimental to system and job performance. Simply queueing jobs until enough resources are available leads to low system utilization and high job response time. The problem can be alleviated in part by clever space-sharing scheduling techniques, such as backfilling [10]. With backfilling, jobs are also scheduled for execution in the order they arrive, but the scheduler looks past jobs that cannot be immediately started. Jobs are scheduled to run as soon as possible, so long as they do not delay a job that arrived earlier. This is also called *conservative backfilling*. Another variation of backfilling, called *aggressive backfilling*, schedules jobs to run as soon as possible, so long as they do not delay the *first* waiting job. To perform backfilling, the scheduler must have an estimate of the termination time for each job. The literature shows that both approaches result in comparable performance in real systems [2]. Therefore, in this paper we consider conservative backfilling only.

Backfilling can be very effective in increasing machine utilization and average job response time. However, it still presents some problems to users of large jobs. On one hand, execution of large jobs are postponed until enough processors are available. On the other hand, once large jobs start, they monopolize the resources for long periods of time, which in turn make the machine inaccessible to other users. Consequently, large computing centers typically restrict running large jobs to evenings and weekends.

A more flexible solution to the problem of executing a large number of jobs of varied characteristics is to share the machine's resources not only spatially but also temporally [1, 3, 5, 12]. The examples in the literature illustrate the benefit of adding this other axis of partitioning for parallel systems. It can result in better system utilization and reduced response time. In addition to time-slicing resources among several jobs, systems that support time-sharing typically also support the important feature of *preemption*. Preemption allows a lower priority job to be suspended so that a high priority job can run. Both time-slicing and preemption are supported in GangLL. When time-slicing multiple jobs in the same nodes, processes from all active jobs have to coexist in the same system image. This puts pressure on both the virtual and physical memory systems and effectively limits the degree of multiprogramming (MPL) that can be supported efficiently. For the CTR machine, an MPL of 2 or 3 is likely to be feasible. This degree of multiprogramming is expected to increase in future systems (*e.g.*, ASCI White) which will have larger memory.

Time-sharing on a large scale distributed parallel machine is complicated by two issues: (i) the native schedulers of the operating systems executing on each node are not coordinated, and (ii) tasks of distributed jobs interact tightly through the communication switch. The tasks of a parallel job must be coscheduled [11] (*i.e.*, must run concurrently on all nodes) or inefficient communication behavior results. Without coscheduling, receivers may not be ready when senders are and vice-versa. As a result, we adopt a coarse grain time-sharing strategy for GangLL that guarantees that tasks of the same job execute simultaneously despite the fact that the operating system images are not coordinated. We accomplish this by (i) partitioning the time axis into large time slices (order of seconds to minutes), (ii) populating the time slices with tasks from parallel jobs so that all tasks of a job occupy the same time-slice, and (iii) implementing the schedule at the individual nodes. The schedule is represented by an Ousterhout matrix [11], where the columns correspond to processors and the rows correspond to revolving time slices.

4 Evaluation methodology

When selecting and developing job schedulers for use in large MPP installations, it is important to understand their expected performance. The first step is to have a characterization and a procedure to synthetically generate the expected workloads. Our methodology for generating these workloads involves the following steps:

1. Fit a typical workload with mathematical models.
2. Generate synthetic workloads based on the derived mathematical models.
3. Determine the waiting time, response time and other parameters of interest for various types of jobs for various scheduling policies using these synthetic workloads and a simulation of the scheduler.

When fitting an MPP loadload, it is very useful to be able to find a compact mathematical representation that is expressible by a few parameters, is reasonably easy to use for the generation of synthetic workloads, and is also suitable for theoretical queuing analysis of scheduling algorithms.

MPP workloads often are over-dispersive. That is, job inter-arrival time distribution and job service time distribution each has a coefficient of variation that is greater than one. Distributions with coefficient of variation greater than one are also referred to as long-tailed distributions, and can be fitted adequately with Hyper Erlang Distributions of Common Order.

In an earlier work [6] we developed such a model, and demonstrated its efficacy by using it to fit a typical workload from the Cornell University Theory Center. Here we use this model to fit a typical workload from the ASCI Blue-Pacific System.

The Hyper Erlang Distribution of Common Order is a special form of a more general type of distribution named Phase Type Distribution. The Phase Type Distribution is a distribution (often used in queuing theory) which is capable of representing any stochastic process whose associated probability density function has a Laplace transform that is a rational function. Practically all the relevant systems one encounters in the stochastic modeling of computer workloads can be modeled by a Phase Type Distribution.

The Hyper Erlang Distribution of Common Order in turn is a generalization of 3 more commonly known distributions – namely exponential, hyper exponential, and Erlang distributions. The Hyper Erlang Distribution of Common Order distribution has a Laplace transform of the form

$$f^*(s) = \sum_{i=1}^{2} p_i \Big(\frac{\lambda_i}{s+\lambda_i}\Big)^n \tag{1}$$

where n, a positive integer, is called the order of the distribution, and $0 \leq p_i \leq 1$ with $p_1 + p_2 = 1$. The Erlang distribution is a special case of the Hyper Erlang Distribution of Common Order with one of the p_is equal to 1 (*e.g.*, $p_1 = 1$). The hyper exponential distribution is a Hyper Erlang Distribution of Common Order with $n = 1$. The exponential distribution is also a special case with $p_1 = 1$ and $n = 1$. The k^{th} non-central moment of a distribution, for all integers $k \geq 1$, can be obtained from the Laplace transform of the distribution by,

$$\mu_k = E[t^k] = (-1)^k \left[\frac{d^k f^*(s)}{ds^k}\right]_{s=0} \tag{2}$$

which, for Hyper Erlang distribution of Common Order, is

$$\mu_k = \sum_{i=1}^{2} p_i \frac{n(n+1)...(n+k-1)}{\lambda_i^k} \tag{3}$$

The moments for the Erlang distribution are obtained by setting $p_1 = 1$ and $p_2 = 0$ in equation 3, yielding

$$\mu_k = \frac{n(n+1)...(n+k-1)}{\lambda^k}. \tag{4}$$

The moments for the hyper exponential distribution are obtained by setting $n = 1$ in equation 3, yielding

$$\mu_k = \sum_{i=1}^{2} p_i \frac{k!}{\lambda_i^k} \tag{5}$$

The moments for the exponential distribution is obtain by letting $n = 1$ in equation 4, giving

$$\mu_k = \frac{k!}{\lambda^k} \tag{6}$$

There are a number of interrelationships among the moments to ensure physical situation. In particular, we must have nonnegative p_is and λ_is. The procedure for the extraction of the model parameters need to preserve these constraints. The algebra involved is detailed in [6]. Basically the procedure automatically selects the simplest distribution that is commensurate with the first three moments, and the related constraints in the observed data. More precisely, our modeling procedure involves the following steps:

1. First we group the jobs into classes, based on the number of processors they require to execute on. Each class is a bin in which the upper boundary is a power of 2.

2. Then we model the inter-arrival time distribution for each class, and the service time distribution for each class as follows:

3. From the job traces, we compute the first 3 moments, μ_1^o, μ_2^o, μ_3^o, of the observed inter-arrival time and those of the observed service time.

4. Then we select the Hyper Erlang Distribution of Common Order that fits these 3 observed moments. We chose to fit the moments of the model against those of the actual data, because the first 3 moments usually capture the generic features of the workload and are less susceptible to fluctuations. These three moments carry the information on the mean, variance, and skewness of the random variable respectively.

Next we generate various synthetic workloads from the observed workload by varying the inter-arrival rate, and cpu time used. The Hyper Erlang parameters for these synthetic workloads are obtained by multiplying the inter-arrival rate and the service time each by a separate multiplicative factor, and by specifying the number of jobs to generate. From these model parameters the actual job trace is obtained using the procedure described in [6]. Finally we simulate the effects of these synthetic workloads with a simulator and observe the results.

5 Results

In this section we discuss the performance characteristics of the three scheduling strategies available for the CTR machine: FCFS, backfilling, and gang-scheduling. We consider three different configurations of gang-scheduling, with multiprogramming levels of 2, 3, and 5. Note that these are the *maximum* multiprogramming levels allowed for that particular configuration. A smaller MPL is used if there are not enough jobs to fill the system. We also consider two different versions of backfilling: In one case, *perfect backfilling*, the estimate of the job execution time, as used by the scheduler, is exactly equal to the actual job execution time. In the other case, *realistic backfilling*, the estimate is obtained by multiplying the actual execution time by a uniformly distributed random number between 1.0 and 1.5. This corresponds to the user submitting the job requesting on the average 25% more than the actual execution time. When a job terminates before its estimated time, the freed up nodes are made available to newly arrived jobs.

Each job in the system is characterized by number of nodes it uses and its submission, start, and termination times. The submission time of a job is the time it was submitted for execution in the system. Its start time is the time it starts actually running on nodes, after possibly waiting in the queue. Finally, the termination time is the time the job completes execution and leaves the system. We describe the performance characteristics of our scheduling systems through plots of average response time and average wait time as functions of utilization. These system parameters are defined as follows:

1. *utilization:* This is a measure of how busy the nodes in the machine are during the simulated period. Total CPU time for one job is computed by multiplying its number of nodes by its execution time. We obtain utilization by summing CPU time for all jobs and dividing the total number of nodes times the length of simulated interval. The utilization is always a number between 0 and 1.

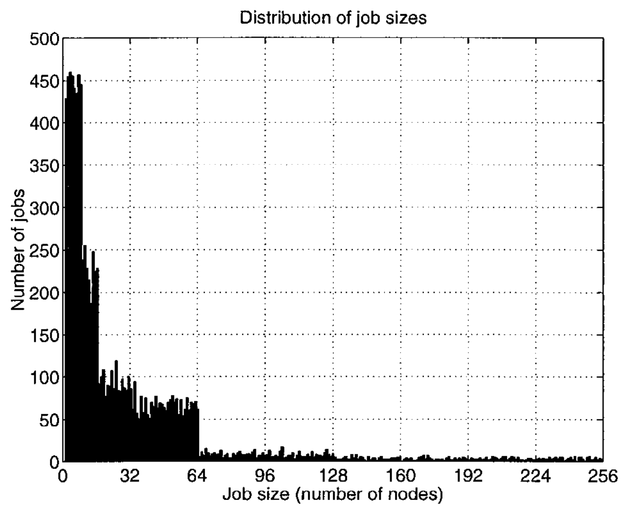
2. *average job response time:* The response time of a job is computed as its termination time minus its submission time. We then compute an average for all jobs as well as an average for *large* jobs. A large job is defined as any job that uses more than 32 nodes (or 10% of the CTR machine). This parameter is reported in seconds.

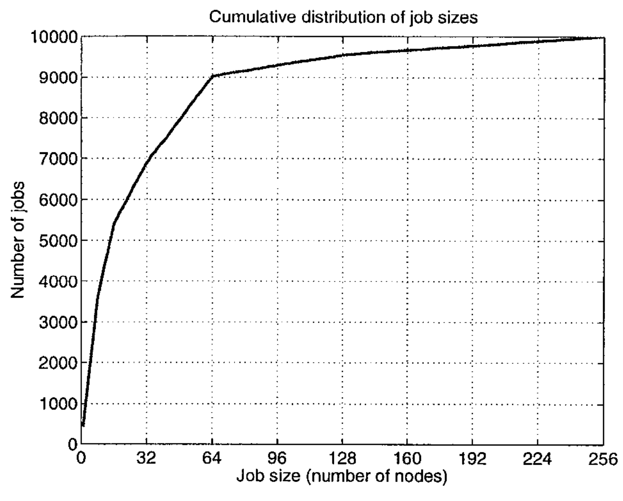
3. *average job wait time:* The wait time of a job is computed as its start time minus its submission time. We again compute an average for all jobs as well as an average for large jobs. This parameter is also reported in seconds.

The baseline workload consists of 10000 jobs generated by the hyper Erlang model for the actual LLNL workload. Some characteristics of this workload are shown in Figure 2 and Figure 3. Figure 2 reports the distribution of job sizes (number of nodes). For each job size, between 1 and 256, Figure 2(a) shows the number of jobs of that size, while Figure 2(b) plots the number of jobs with *at most* that size. Figure 3 reports the distribution of CPU time. For each job size, Figure 3(a) shows the sum of the CPU times for all jobs of that size, while Figure 3(b) is a plot of the sum of the CPU times for all jobs of that size. From Figure 2 and Figure 3 we observe that, although large jobs (those with more than 32 nodes), represent only 30% of the number of jobs, they constitute more than 80% of the total work performed in the system.

In addition to the baseline workload of Figure 2 and Figure 3 we generate 80 additional workloads, of 10000 jobs each, by varying the model parameters so as to increase average job execution time and to decrease average job interarrival time. For a fixed interarrival time, increasing job execution time typically increases utilization, until the system saturates. The same is true for decreasing job interarrival time for a fixed job execution time.

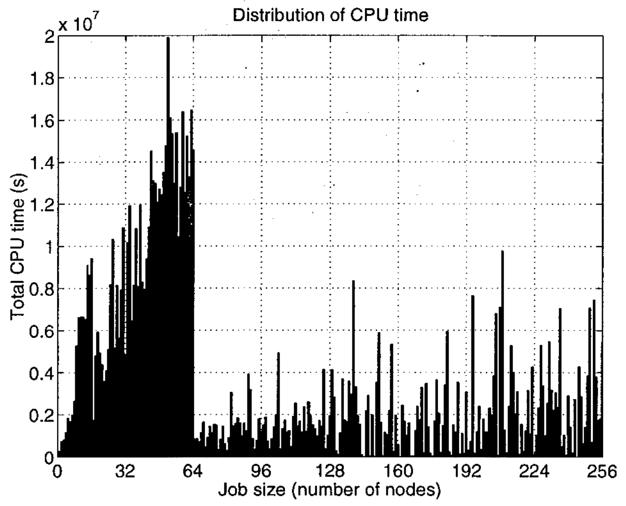
Results for response time and wait time as a function of utilization, averaged over all jobs, are shown in Figure 4 and Figure 5 respectively. Each plot is for a particular average job interarrival time and has results for each of the scheduling strategies considered. The same results, averaged only over large jobs, are shown in Figure 6 and Figure 7.

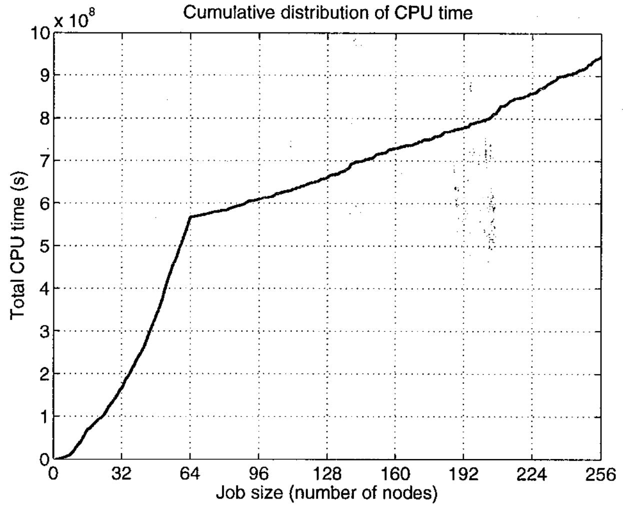




(a) (b)

Figure 2: Workload characteristics: distribution of job sizes.





(a) (b)

Figure 3: Workload characteristics: distribution of cpu time.

We first note that backfilling would be very effective in improving system performance if perfect estimation of job execution time were available. Over a wide range of utilizations, perfect backfilling performs just as good as gang-scheduling with MPL of 5 with respect to average job response time. However, in terms of average job wait time of average job response time for large jobs, perfect backfilling performs only as good as gang-scheduling with MPL of 3.

The realistic backfilling performs much better than FCFS over the important operation range between 0.60 and 0.70 utilization. This utilization rate is common in supercomputing centers such as NASA Ames and Cornell CTC [9, 13]. However, as we move towards the high utilization rates that LLNL has achieved in the past with various gang-scheduling systems [1, 8], realistic backfilling starts to degrade. Gang-scheduling

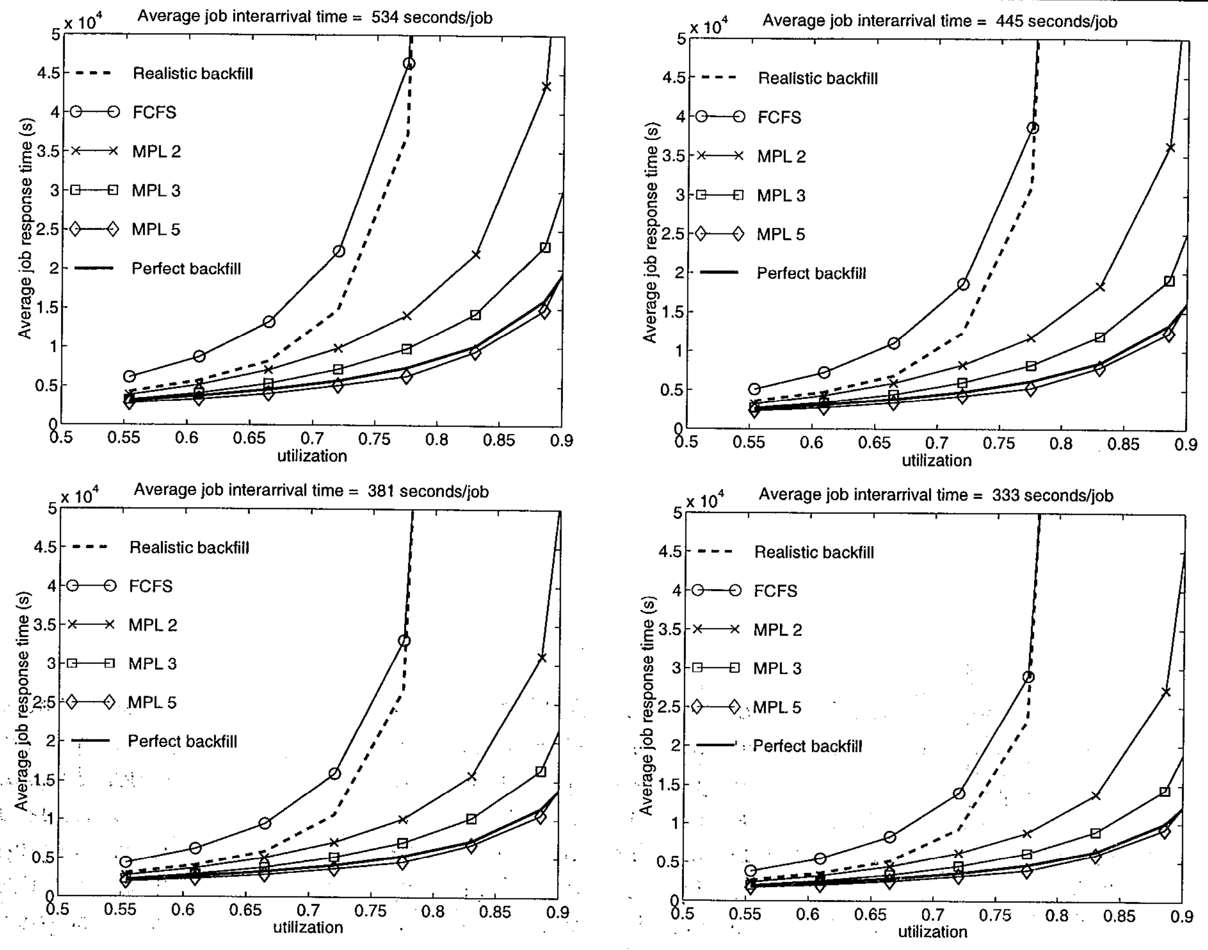


Figure 4: Response times as a function of utilization, averaged over all jobs.

with MPL of 2 always performs better than realistic backfilling with respect to average response time. It performs significantly better when average wait time or average response time for large jobs are considered.

6 Conclusions

The ASCI Blue-Pacific machines, because of their size and diverse workloads, present new challenges to parallel job scheduling systems. Strategies that rely exclusive on space-sharing, such as backfilling, can be effective for utilization rates of up to 70%. However, to achieve higher utilizations it is necessary to use time-sharing as well.

We have developed a gang-scheduling system, GangLL, that performs both space- and time-sharing of resources in a parallel system. Our performance analysis shows that, even at modest multiprogramming levels of 2 and 3, gang-scheduling can accommodate very high utilization rates of up to 80 or 90%.

References

[1] D. G. Feitelson and M. A. Jette. **Improved Utilization and Responsiveness with Gang Scheduling.** In *IPPS'97 Workshop on Job Scheduling Strategies for Parallel Processing*, volume 1291 of *Lecture*

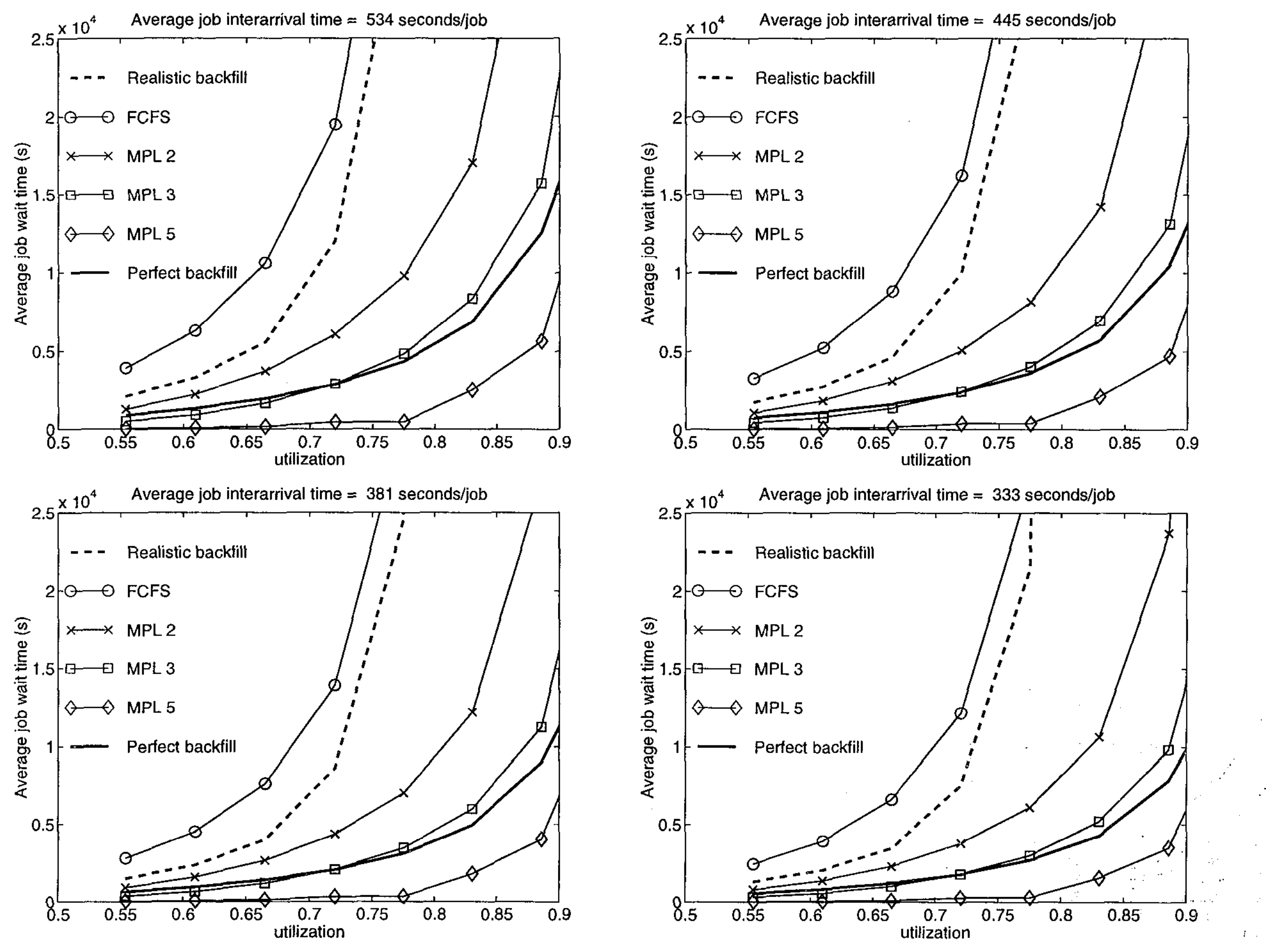


Figure 5: Wait times as a function of utilization, averaged over all jobs.

Notes in Computer Science, pages 238–261. Springer-Verlag, April 1997.

[2] D. G. Feitelson and A.M. Weil. **Utilization and predictability in scheduling the IBM SP2 with backfilling**. In *12th International Parallel Processing Symposium*, pages 542–546, April 1998.

[3] H. Franke, P. Pattnaik, and L. Rudolph. **Gang Scheduling for Highly Efficient Multiprocessors**. In *Sixth Symposium on the Frontiers of Massively Parallel Computation, Annapolis, Maryland*, 1996.

[4] B. Gorda and R. Wolski. **Time Sharing Massively Parallel Machines**. In *International Conference on Parallel Processing*, volume II, pages 214–217, August 1995.

[5] N. Islam, A. L. Prodromidis, M. S. Squillante, L. L. Fong, and A. S. Gopal. **Extensible Resource Management for Cluster Computing**. In *Proceedings of the 17th International Conference on Distributed Computing Systems*, pages 561–568, 1997.

[6] J. Jann, P. Pattnaik, H. Franke, F. Wang, J. Skovira, and J. Riordan. **Modeling of Workload in MPPs**. In *Proceedings of the 3rd Annual Workshop on Job Scheduling Strategies for Parallel Processing*, pages 95–116, April 1997. In Conjuction with IPPS'97, Geneva, Switzerland.

[7] M. Jette, D. Storch, and E. Yim. **Timesharing the Cray T3D**. *Cray User Group*, pages 247–252, March 1996.

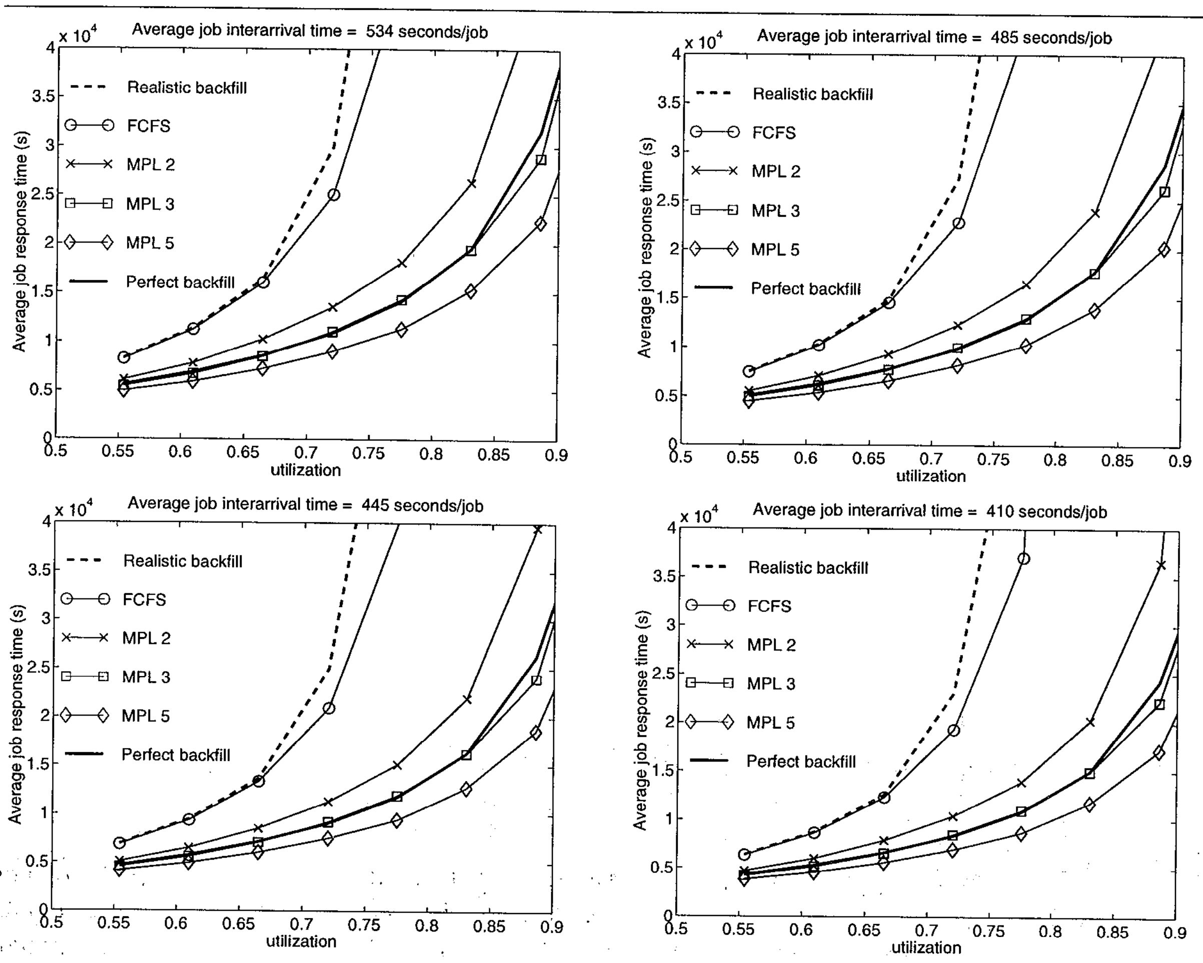


Figure 6: Response times as a function of utilization, averaged over large jobs (nodes > 32).

[8] M. A. Jette. **Expanding Symetric Multiprocessor Capability Through Gang Scheduling**. In *IPPS'98 Workshop on Job Scheduling Strategies for Parallel Processing*, March 1998.

[9] J. P. Jones and B. Nitzberg. **Scheduling for Parallel Supercomputing: A Historical Perspective of Achievable Utilization**. In *Proceedings of the 5th Annual Workshop on Job Scheduling Strategies for Parallel Processing*, April 1999. In conjunction with IPPS/SPDP'99, Condado Plaza Hotel & Casino, San Juan, Puerto Rico.

[10] D. Lifka. **The ANL/IBM SP scheduling system**. In *IPPS'95 Workshop on Job Scheduling Strategies for Parallel Processing*, volume 949 of *Lecture Notes in Computer Science*, pages 295–303. Springer-Verlag, April 1995.

[11] J. K. Ousterhout. **Scheduling Techniques for Concurrent Systems**. In *Third International Conference on Distributed Computing Systems*, pages 22–30, 1982.

[12] U. Schwiegelshohn and R. Yahyapour. **Improving First-Come-First-Serve Job Scheduling by Gang Scheduling**. In *IPPS'98 Workshop on Job Scheduling Strategies for Parallel Processing*, March 1998.

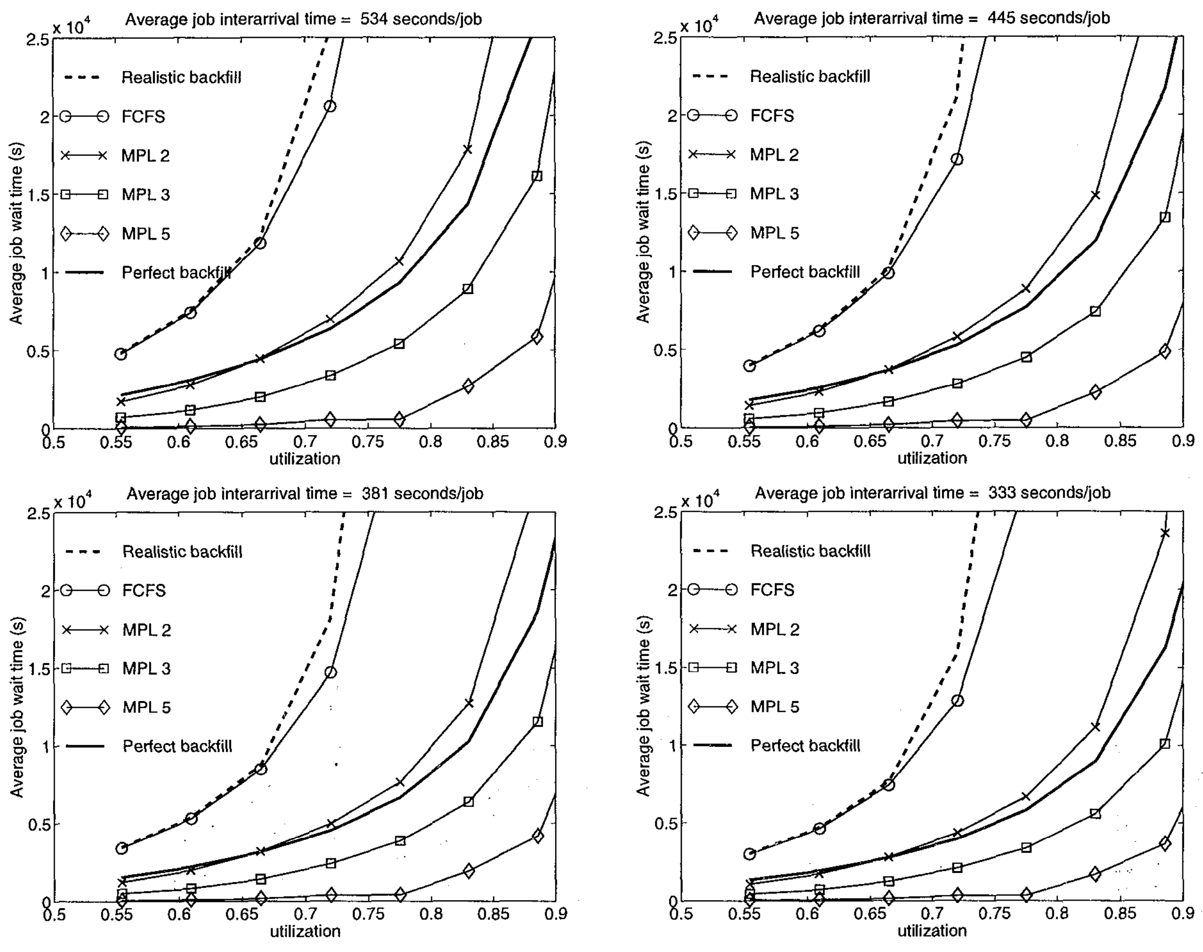


Figure 7: Wait times as a function of utilization, averaged over large jobs (nodes > 32).

[13] J. Skovira, W. Chan, H. Zhou, and D. Lifka. **The EASY-LoadLeveler API project**. In *IPPS'96 Workshop on Job Scheduling Strategies for Parallel Processing*, volume 1162 of *Lecture Notes in Computer Science*, pages 41–47. Springer-Verlag, April 1996.

*This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.