Workload Characteristics of a Multi-cluster Supercomputer

Hui Li* David Groep† Lex Wolters*

*Leiden Institute of Advanced Computer Science (LIACS), Leiden University
2333 CA, Leiden, The Netherlands

†National Institute for Nuclear and High Energy Physics (NIKHEF)
1098 SJ, Amsterdam, The Netherlands

E-mail: hli@liacs.nl davidg@nikhef.nl llexx@liacs.nl

Abstract

This paper presents a comprehensive characterization of a multi-cluster supercomputer[1] workload using twelve-month scientific research traces. Metrics that we characterize include system utilization, job arrival rate and interarrival time, job cancellation rate, job size (degree of parallelism), job run time, memory usage, and user/group behavior. Correlations between metrics (job runtime and memory usage, requested and actual runtime, etc) are identified and extensively studied. Differences with previously reported workloads are recognized and statistical distributions are fitted for generating synthetic workloads with the same characteristics. This study provides a realistic basis for experiments in resource management and evaluations of different scheduling strategies in a multi-cluster research environment.

1. Introduction

Workload characterization of parallel supercomputers is important to understand the system performance and develop workload models for evaluating different system designs and scheduling strategies [1, 2]. During the past several years, lots of workload data has been collected [3], analyzed [4, 5, 6], and modeled [7, 8, 9]. Benchmarks and standards are also proposed for job scheduling on parallel computers [10].

In previously studied workloads [4, 5, 6, 7], some characteristics are similar. For example, most of the workloads are collected from large custom-made production facilities (IBM SP2, SGI Origin, etc) in supercomputing centers. Jobs typically request "power-of-two" number of processors and have different arrival patterns in different periods (e.g. peak and none-peak hours in a daily cycle). Some characteristics, such as distributions and correlations, vary across different workloads [4, 5, 11]. Other characteristics are studied and reported separately, such as job cancellation rate [9] and conditional distributions (e.g. actual runtime distributions conditioned on requested runtime [4]). In this paper we compare our workload with previous reported ones on a per characteristics basis.

This paper presents a comprehensive workload characterization of the DAS-2 [12] supercomputer. The DAS-2 system is interesting in that it is built using the popular COTS (Commodity Off The Shelf) components (e.g. Intel Pentium processors and Ethernet networks) and consists of multiple distributed clusters serving the five participating universities. Not like other production machines, DAS-2 is dedicated to parallel and distributed computing research thus it has much lower system utilization. We analyze twelve-month workloads on DAS-2 clusters in year 2003. Characteristics include system utilization, job arrival rate and interarrival time, job cancellation rate, job size (degree of parallelism), job run time, memory usage, and user/group behavior. Correlations between metrics are also identified and studied.

The contributions of this paper reside in the following. Firstly, our study is based on cluster workloads. Cluster computing is a popular alternative in the HPC community and to our knowledge, not much work has been done in characterizing cluster workloads. Secondly, the system we study is a research facility. This provides an interesting comparison point to the well studied production workloads. Thirdly, we present a comprehensive characterization of the DAS-2 workloads. We not only analyze most of the metrics appeared in previous work, but also extensively study the correlations between different characteristics. Moreover, we fit the observed data with statistical distributions to facilitate synthetic workload generation. This research serves as a realistic basis in modeling cluster workloads, which contributes as input for evaluations of different scheduling strategies in a multi-cluster research environ-

1 Distributed ASCI Supercomputer-2 (DAS-2). ASCI stands for Advanced School for Computing and Imaging in the Netherlands.

cluster	location	#CPUs	period	#job entries
fs0	Vrije Univ. A'dam	144	01-12/2003	219618
fs1	Leiden Univ.	64	01-12/2003	39356
fs2	Univ. of A'dam	64	01-12/2003	65382
fs3	Delft Univ. of Tech.	64	01-12/2003	66112
fs4	Utrecht Univ.	64	02-12/2003	32953

Table 1. DAS-2 clusters and workload traces (A'dam - Amsterdam).

ment [13].

The rest of the paper is organized as follows. Section 2 provides an overview of the DAS-2 system and workload traces used in our study. Section 3 analyzes the overall system utilization. Section 4 describes the job arrival characteristics, including job arrival rate, job interarrival time and job cancellation rate. Distributions are fitted for job interarrival times and job cancellation lags. Section 5 describes job execution characteristics. This includes job size, job actual runtime, memory usage, and correlations between them. Distributions and/or conditional distributions are also provided. Section 6 describes user/group behavior and its implications in modeling and predictions. In section 7 conclusions are presented and future work is discussed.

2. The DAS-2 Supercomputer and Workload Traces

The DAS-2 supercomputer consists of five clusters located at five Dutch universities and is primarily used for computing and scientific research. The largest cluster (Vrije Universiteit) contains 72 nodes and the other four clusters have 32 nodes each. Every node contains two 1GHz Pentium III processors, 1GB RAM and 20GB local storage. The clusters are interconnected by the Dutch university internet backbone and the nodes within a local cluster are connected by high speed Myrinet as well as Fast Ethernet LANs. All clusters use openPBS [14] as local batch system (one and only one queue is configured for each cluster). Maui [15] (FCFS with backfilling) is used as the local scheduler. Jobs that require multi-clusters can be submitted using toolkits such as Globus [16]. DAS-2 runs RedHat Linux as the operating system.

We use job traces recorded in the PBS accounting logs for twelve months in year 2003 on the five clusters[2]. All jobs in the traces are *rigid* (jobs that do not change parallelism at runtime) batch jobs. An overview of the DAS-2 system and workload traces is provided in Table 1. As we can see, fs0 (VU) is the most active cluster, with more than two hundred thousand job entries. Next we have clusters at UvA (fs2) and Delft (fs3), each with more than sixty thousand entries. Leiden (fs1) and Utrecht (fs4) are relatively less active among the DAS-2 clusters. Next section gives a more detailed analysis on the overall system utilization.

3. System Utilization

Figure 1 shows the DAS-2 system utilization as function of time of day. Two plots are shown for every cluster. One is the average utilization of all days and the other is the average utilization of all active days in the year (excluding system down time and days without job arrivals[3]). In average, fs0 has the highest (22%) and fs3 has the lowest system utilization (7.3%) among DAS-2 clusters. The utilization (7.3% to 22%) is substantially lower than previously reported workloads (e.g. 50% in average excluding downtime [5]). This is because DAS-2 system is designed for scientific research and production jobs are precluded from it. The goal of DAS-2 is not on high utilization, but rather on provide fast response time and more available processors for university researchers. Moreover, DAS-2 schedulers define one special policy, which forbids jobs to be scheduled on nodes (SMP dual processor) of which one processor is already used by another job. This policy also has a certain negative impact on the overall system utilization.

We can see that the utilization approximately follows the daily job arrival rate (see Figure 2), although the differences between day and night are generally smaller. It is because nightly jobs often require more processors and run longer than daily jobs, despite substantially fewer job arrivals. This is particularly evident on cluster fs3 and fs4.

4. Job Arrival Characteristics

In this section we analyze the job arrival characteristics. We first describe the job arrival rate, focusing mainly on daily cycles. Daily peak and non-peak hours are identified. Secondly, we characterize the job interarrival times during

2 Logs of January on fs4 are not available.

3 Since we calculate the system utilization based on traces, we could not distinguish whether it is system down time or time without job arrivals.

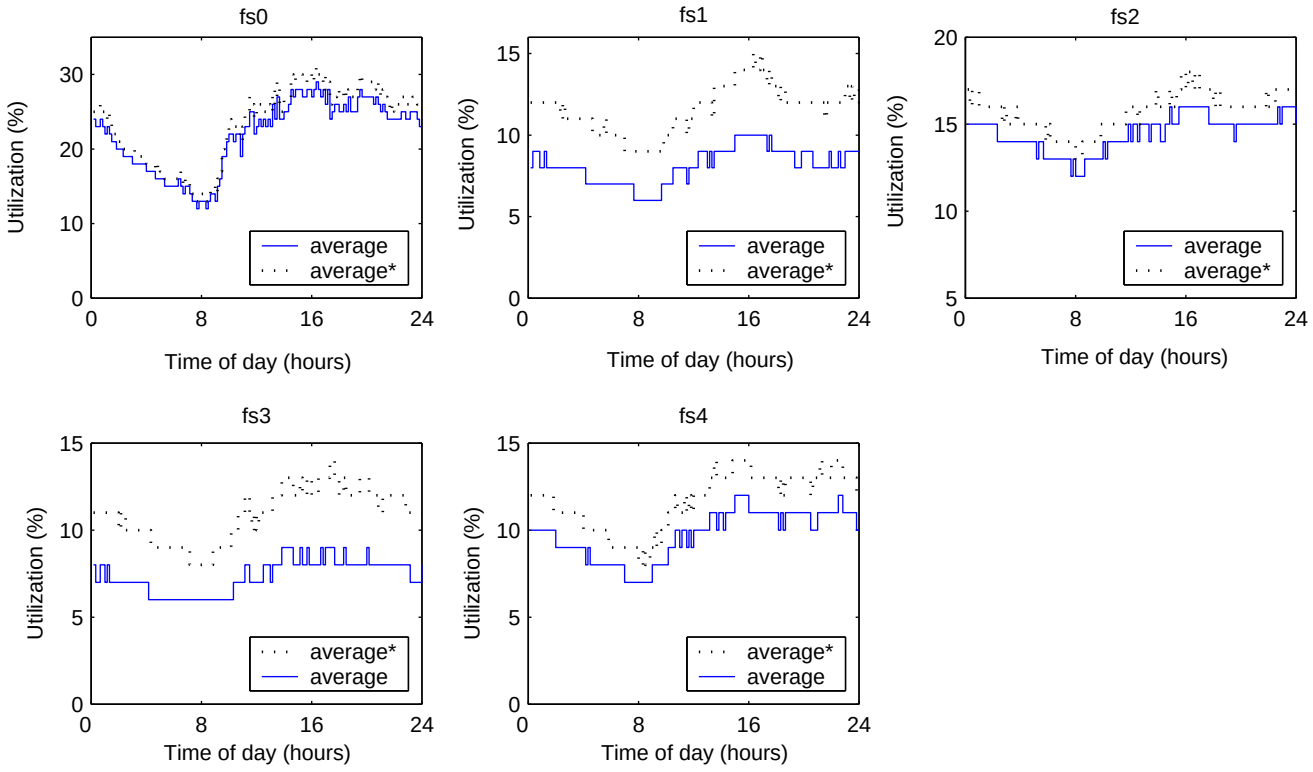


Figure 1. System utilization of DAS-2 clusters. "Average" stands for the average utilization of all days in the year. "Average*" stands for the average utilization of all active days in the year, excluding system downtime and days without job arrivals.

daily peak hours. Several statistical distributions are examined to fit the job interarrival times. Finally, job cancellation rate and cancellation lags are analyzed and modeled, since it may also affect the scheduling process.

4.1. Job Arrival Rate

As is studied in [7], job arrivals are expected to have cycles at three levels: daily, weekly, and yearly. In a yearly cycle, we find that workloads are not distributed evenly throughout the year. Instead, workloads concentrate on specific months and job entries in these months are around two or more times above average. We call them "job-intensive" months (October, November and December on fs0, August, November on fs1, November, December on fs2, May, December on fs3, and August, November on fs4). This is because of the different active users/groups on different clusters and they are active in specific periods during the year (see Section 6). In a weekly cycle, all clusters share similar characteristics. Wednesday has the highest average job arrival rate and decreases alongside, with Sunday and Saturday have the lowest arrival rate. This is natural since people generally work more during weekdays (Monday - Friday) than weekends (Saturday and Sunday).

The most important cycle is the daily cycle. As is shown in 2, clusters share similar daily workload distributions during weekdays. We identify the daily peak hours as from 9am to 7pm on all five clusters. This is in accordance with normal "working hours" at Dutch universities. Similar job arrival distributions are reported on other workloads with different peak hour periods (e.g. 8am to 6pm in [4], 8am to 7pm in [7]). Additionally, an intermediate period is reported from 6pm to 11pm in [4]. We observed similar characteristics on DAS-2 clusters, with an intermediate arrival period from 8pm to 1am and a low arrival period from 1am to 8am. The arrival rate per hour can be divided into three scales. The fs0 cluster has the highest one, with an average arrival rate of 108 jobs per hour and peak arrival rate exceeding 200 jobs per hour. In the middle there are fs2 and fs3, with average arrival rates of 31 and 32 jobs per hour each. Clusters fs1 and fs4 have average arrival rates of 19 and 15 jobs per hour, respectively.

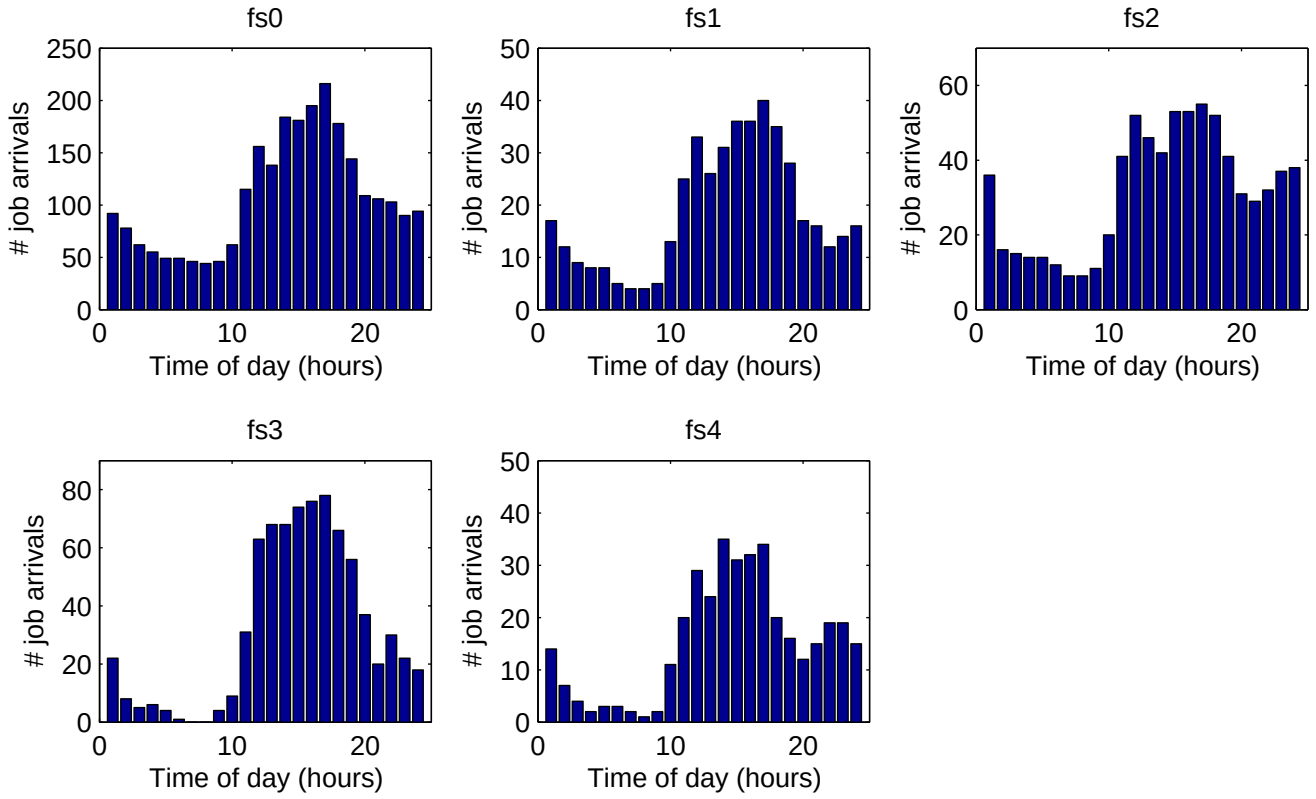


Figure 2. Daily cycle of job arrivals during weekdays on DAS-2 clusters.

cluster	period	M (s)	CV	best fitted distribution	KS
fs0	2003/12/02	17	1.6	gamma (a = 0.44, b = 39)	0.10
fs1	2003/11/25	26	2.4	gamma (a = 0.30, b = 86)	0.13
fs2	2003/12/29	14	1.3	hyperexp2 ($c1$=0.92, $\lambda1$=0.07, $c2$=0.08, $\lambda2$=100)	0.07
fs3	2003/05/26	10	1.8	hyperexp2 ($c1$=0.55, $\lambda1$=0.06, $c2$=0.45, $\lambda2$=0.42)	0.10
fs4	2003/08/13	62	3.0	hyperexp2 ($c1$=0.09, $\lambda1$=0.003, $c2$=0.91, $\lambda2$=0.03)	0.10

Table 2. High load distributions of job interarrival time during daily peak hours (M - Mean, CV - Coefficient of Variation, KS - maximal distance between the cumulative distribution function of the theoretical distribution and the sample's empirical distribution).

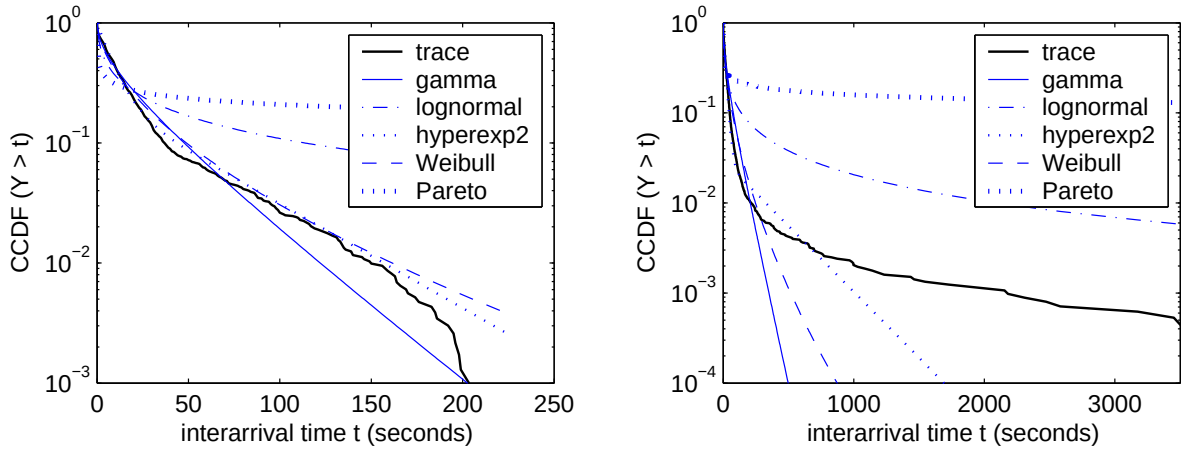
4.2. Job Interarrival Time

Based on the observed job interarrival patterns, we choose to characterize "representative" and "high load" period of job interarrival times. The representative period is defined as the peak hours during weekdays in job-intensive months. The high load period is the peak hours of the most heavily loaded days in the year. As is shown in Table 2, during high load period the *mean* ranges from 14 to 62 seconds and the *coefficient of variation* (CV) varies from 1.3 to 3.0 on DAS-2 clusters. The mean and CV are considerably larger in the representative period (see Table 3). Both small (1-2) and large CVs (3-6) have been reported in other workloads [4, 6].

We have selected several statistical models to fit the interarrival times of representative and high load period, including hyperexponential, gamma, Weibull, and heavy-tailed distributions like lognormal and Pareto [17]. We fit

cluster	period	M (s)	CV	best fitted distribution	KS
fs0	Dec	27	4.5	hyperexp2 ($c1$=0.04, $\lambda1$=0.003, $c2$=0.96, $\lambda2$=0.06)	0.15
fs1	Aug, Dec	66	3.6	Weibull (a = 22.6, b = 0.44)	0.10
fs2	Dec	44	5.0	Weibull (a = 26.1, b = 0.58)	0.08
fs3	May, Dec	23	6.0	Weibull (a = 11.6, b = 0.53)	0.14
fs4	Aug, Nov	86	5.1	Weibull (a = 33.2, b = 0.5)	0.09

Table 3. Representative distributions of job interarrival time during daily peak hours (M - Mean, CV - Coefficient of Variation, KS - maximal distance between the cumulative distribution function of the theoretical distribution and the sample's empirical distribution).



(a) High load distribution on fs0 on 12/02/2003 (mean = 17, CV = 1.6)

(b) Representative distribution on fs0 in December, 2003 (mean = 27, CV = 4.5)

Figure 3. Fitting distributions of interarrival time during peak hours on fs0.

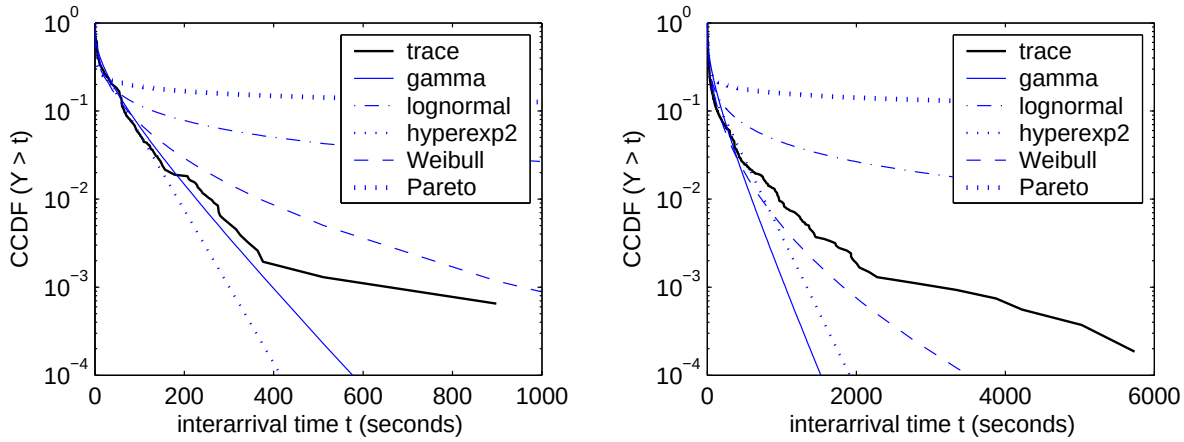
the above mentioned distributions (except hyperexponential) using *Maximum Likelihood Estimation* (MLE) method, and a two-phase hyperexponential distribution using *Expectation Maximization* (EM) algorithm[4] [18]. The goodness of fit is assessed using the Kolmogorov-Smirnov test.

Results of distribution fitting are shown in Table 2 and 3. Figure 3 and 4 further illustrate how well the different distributions fit the trace data on fs0 and fs1. Generally speaking, none of the chosen distributions pass the goodness of fit test. Some distributions, such as gamma and hyperexponential, fit the head of the sample distribution well but fail to fit the tail. Others like lognormal and Pareto, fit the tail but not the head. It seems not likely to find a model that fits all parts of the empirical distribution well. However, we provide the best fitted distributions for high load and representative period on DAS-2 clusters. For the high load period (see Table 2, gamma and two-phase hyperexponential give the best results among the distributions. One is slightly better than the other depending on the clusters. For the representative period where longer tails and larger CV are observed, Weibull distribution has the best Kolmogorov-Smirnov test results. The only exception occurs on fs0, where a two-phase hyperexponential distribution fits the sample tail better than Weibull. Parameters of fitted distributions are provided in Table 2 and 3.

4.3. Cancelled Jobs

Cancelled jobs may also affect the scheduling process and should be taken into account during workload modeling. In [9], reported job cancellation rates range from 12% to 23% and cancelled jobs are modeled separately. On DAS-

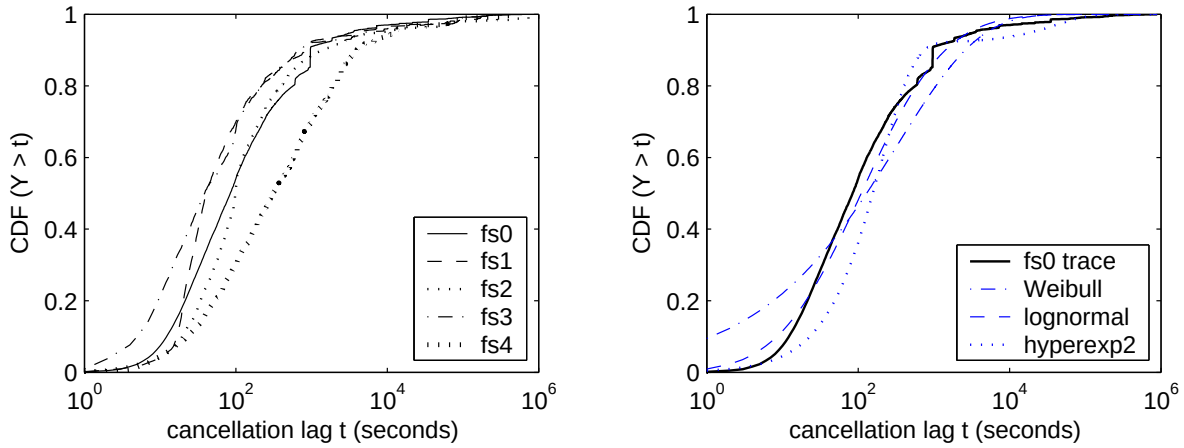
4 Matlab [19] and Dataplot [20] are used to calculate means, CVs, do MLE fitting and goodness of fit test. EMpht [21] is used to fit the hyperexponential distribution.



(a) High load distribution on fs1
on 11/25/2003 (mean = 26, CV = 2.4)

(b) Representative distribution on fs1
in Aug, Dec, 2003 (mean = 66, CV = 3.6)

Figure 4. Fitting distributions of interarrival time during peak hours on fs1.



(a) CDFs of cancellation lag
on DAS−2 clusters

(b) Fitting distributions of
cancellation lag on fs0

Figure 5. Distributions of cancellation lags on DAS-2 clusters.

2 clusters, as is shown in Table 4, lower cancellation rate are observed. The average percentage of cancelled jobs are 6.8% (range from 3.3% on fs3 to 10.6% on fs0).

The *cancellation lag* (CL) is defined as the time between job arrival and cancellation. On DAS-2 clusters, the average cancellation lag is 6429 seconds (Table 4). Plots of cancellation lag distributions (CDF) on a log scale are shown in Figure 5(a). In [9], log-uniform distribution is used to fit the cancellation lag. We examined three distributions (two-phase hyperexponential, lognormal and weibull). Figure 5(b) illustrates the fitting results on fs0. In general, lognormal provides the best fit for the observed data. However, only on fs4 it passes the goodness of fit test. Fitted lognormal parameters are provided in Table 4.

5. Job Execution Characteristics

In this section we describe the job execution characteristics. Firstly we characterize job size (number of processors requested), job actual runtime, and memory usage. Secondly the correlations between these metrics are extensively

cluster	cancelled jobs (%)	M (s)	CV	lognormal parameters	KS
fs0	10.6	3528	8.7	μ = 4.7, σ = 2.0	0.06
fs1	7.7	4749	6.4	μ = 4.4, σ = 2.0	0.16
fs2	3.6	13480	6.6	μ = 5.0, σ = 2.1	0.14
fs3	3.3	3931	6.5	μ = 4.0, σ = 2.3	0.09
fs4	8.6	6458	6.3	μ = 5.8, σ = 2.1	0.02
average	6.8	6429	6.9	μ = 4.8, σ = 2.1	0.09

Table 4. Job cancellation rates and cancellation lags (CL) on DAS-2 clusters (M - CL Mean, CV - CL Coefficient of Variation, KS - maximal distance between the cumulative distribution function of the theoretical distribution and the sample's empirical distribution).

cluster	serial(%)	two(%)	power-of-two(%)	others(%)	odd (except serial) (%)
fs0	2.8	59.4	78.1	19.1	4.2
fs1	2.4	42.8	60.5	37.1	0.2
fs2	4.7	39.6	61.9	33.4	0.4
fs3	1.4	73.6	96.1	2.5	0.03
fs4	0.9	85.3	97.6	1.5	0.05
average	2.4	60.1	78.8	18.7	1.0

Table 5. Job size characteristics on DAS-2 clusters.

studied and conditional distributions are defined for the job actual runtime.

5.1. Job Size

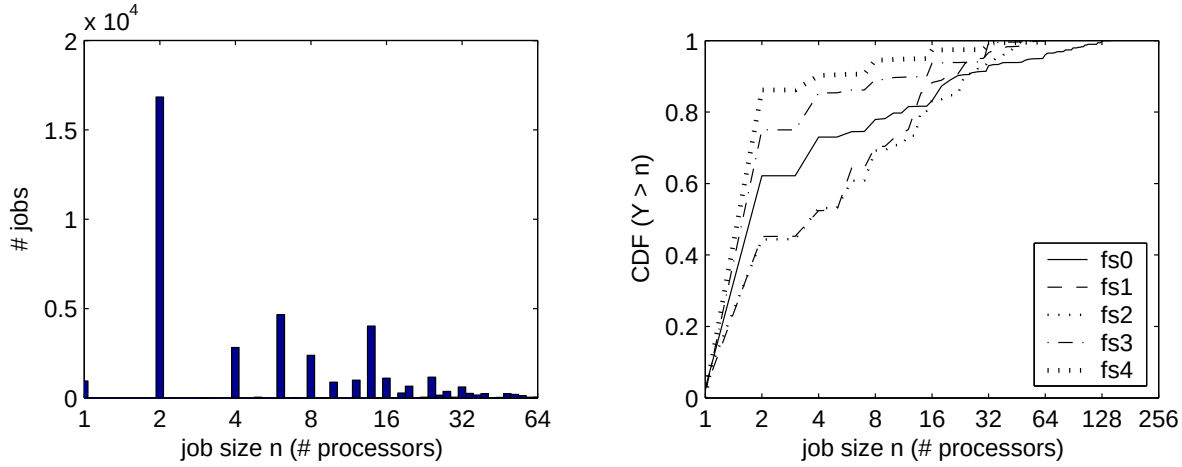
Table 5 shows the job size characteristics on DAS-2 clusters. The "power-of-two" phenomenon (78.8% in average) is clearly observed, as is found in many other workloads [4, 7, 9, 11]. However, the "power-of-two" sizes on cluster fs0, fs1, and fs2 are not as dominant as on fs3 and fs4. Instead, some multiple-2 sizes also contribute to a significant portion of the total number of jobs (e.g. 6 and 14 processors on fs1, shown in Figure 6(a)). The fractions of serial (0.9-4.7%) and odd numbers (1% in average) are significantly lower compared to previously reported workloads (30-40%). One possible explanation could be the special policy mentioned in Section 3 , which forbids jobs to be scheduled on nodes (SMP dual processor) with one processor busy. Researchers are not encouraged to submit multiprocessor jobs with odd numbers.

As we all noticed in Table 5, job size of *two* processors is surprisingly popular on DAS-2 clusters and it is chosen by a major fraction of jobs (range from 39.6% on fs2 to 85.3% on fs4). To find a proper explanation for this phenomenon, we analyze the internal structure of the workloads. On fs0, for instance, there are ten very active users (out of 130 users in total). The most active user submitted more than 40,000 jobs (18% of the total number of jobs on fs0) in consecutive seven weeks during October and November 2003, which is his/her only active period throughout the year. All of these jobs have the same name and request two processors. For the second most active user on fs0, around 90% of his/her jobs have a job size of two. On other DAS-2 clusters similar user behavior are observed, resulting in the popularity of job size two and power-of-two. We discuss more on user behavior and its impacts on workload modeling in Section 6.

In [7], the best results for fitting job sizes are obtained by gamma and two-stage uniform distributions. On DAS-2 clusters, we find that two-stage loguniform distribution provides the best fit for job sizes. Plots of the job size distributions on a log scale are shown in Figure 6(b).

5.2. Job Actual Runtime

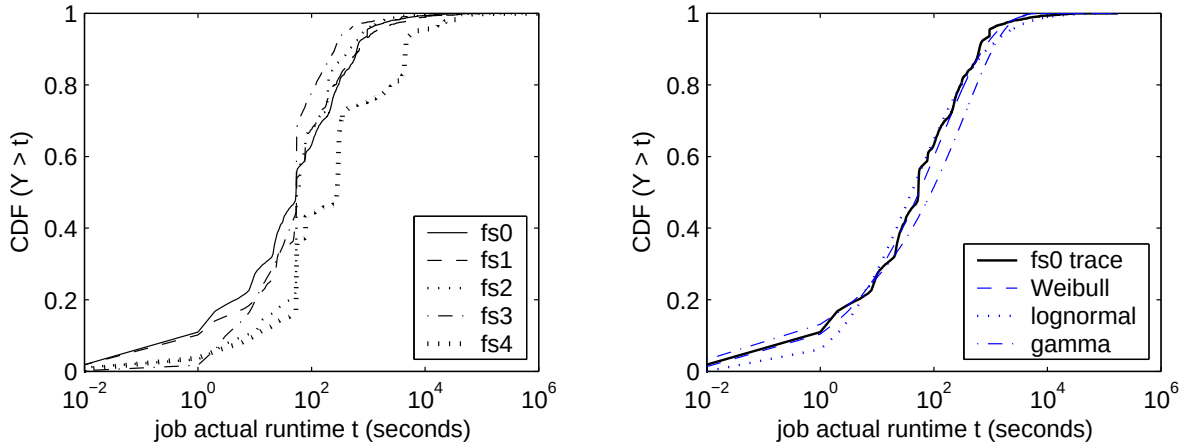
Job actual runtime has been extensively studied in previous reported workloads. Table 6 shows the characteristics of job actual runtimes on DAS-2 clusters. The actual runtimes range from 374 to 2427 seconds, which is lower then previously reported workloads (e.g. 3479 seconds on SDSC SP2 [6]). However, the CV (5.3 - 16) is substantially higher than other production systems (2 - 5) [4, 5, 6]. This is in accordance with the scientific and experimental nature of the DAS-2 usage: the majority of jobs have small execution



(a) Histogram of job size on fs1

(b) CDFs of job size on DAS−2 clusters

Figure 6. Distributions of job sizes on DAS-2 clusters.



(a) CDFs of job actual runtime on DAS−2 clusters

(b) Fitting distributions of job actual runtime on fs0

Figure 7. Distributions of job actual runtimes on DAS-2 clusters.

times and they vary a lot. Plots of the actual runtime distributions on a log scale are shown in Figure 7(a).

Different kinds of distributions have been used to model the actual runtime, for instance, loguniform in [22], hypergamma in [7] and Weibull in [4]. We evaluate gamma, lognormal and Weibull distributions for actual runtimes on DAS-2 clusters. Figure 7(b) shows the distribution fitting on fs0. Weibull and lognormal have similar goodness of fit test results, and they both fit better than gamma. Lognormal is a better model for samples that have a lower head and a longer tail (fs2, fs3, and fs4, see Figure 7(a)). Parameters of fitted distributions are listed in Table 6.

5.3. Memory Usage

The PBS [14] accounting logs record the maximum amount of physical memory used by the job. Hereafter we refer to memory usage as the maximum used physical memory. Memory usage per processor is defined as the maximum used memory divided by the number of processors requested.

Figure 8(a) shows the distributions of memory usage on DAS-2 clusters. It is clearly observed that three special values are chosen by a major fraction of jobs. These special values are 0KB, 324KB and 2600-3000KB (slightly differ-

cluster	mean (s)	CV	fitted distributions	KS
fs0	374	5.3	Weibull (a = 121.7, b = 0.46)	0.08
fs1	648	7.9	Weibull (a = 142.2, b = 0.45)	0.12
fs2	531	16	lognormal (μ = 4.2, σ = 1.8)	0.22
fs3	466	12	lognormal (μ = 3.7, σ = 1.7)	0.12
fs4	2427	6.4	lognormal (μ = 5.3, σ = 2.5)	0.13

Table 6. Job actual runtimes on DAS-2 clusters.

cluster	0KB (%)	324KB (%)	2600-3000KB (%)
fs0	32	19	34
fs1	29	20	16
fs2	25	18	21
fs3	40	17	34
fs4	24	6	62
average	30	16	33

Table 7. Three special memory usage values and their corresponding job percentages.

ent values in this range depending on the clusters), and their corresponding job percentages are listed in Table 7. We can see that a large fraction (30% in average) of jobs have very small memory usage[5]. 324KB and 2600-3000KB, on the other hand, contributes nearly one-sixth and one-third (in average) to the total number of jobs, respectively. The reason why memory usage concentrates on these special values might be that jobs typically have to load certain shared libraries (e.g. C, MPI, Globus), and these shared libraries normally require a fixed amount of memory. To verify this claim, we run MPI jobs (fractal computation) with different requested number of processors (4, 8, 16 and 32) on DAS-2 clusters. We found that memory usage for these jobs is almost the same (324KB, for job size 4, 8 and 16). The exception occurs for job size 32, of which memory usage jumps to 52,620KB. Other MPI programs also appears to use memory size of 324KB. Therefore, we might say that jobs which use 324KB memory most likely have to load certain libraries like MPI. Memory usage of 2600-3000KB could be explained by inclusion of other shared libraries or objects.

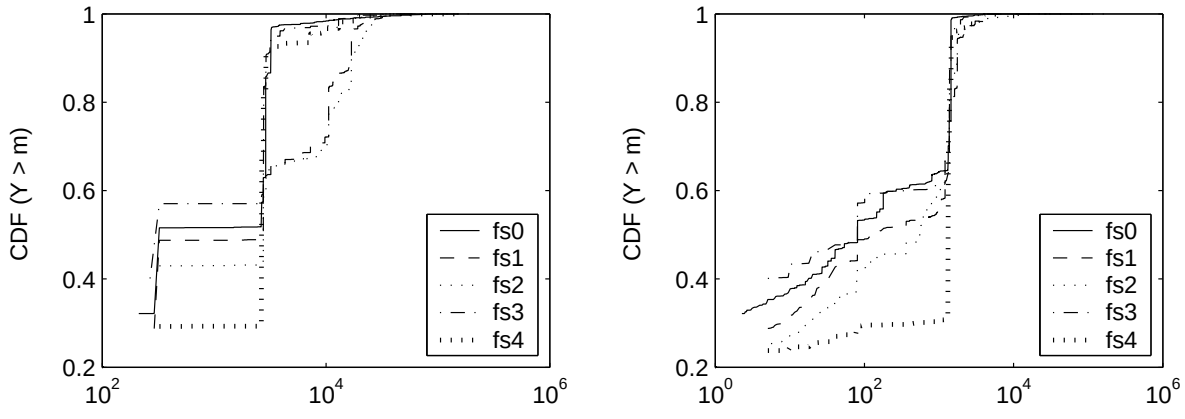
Distributions of memory usage per processor on a log scale are shown in Figure 8(b). As we can see, most of the jobs uses less than 10MB memory per processor (only 2% of the available amount). Correlations between memory usage and job sizes are discussed in next section.

5 0KB is recorded in the PBS accounting logs. It means that the job uses very small memory (rounded to zero) instead of saying that the job does not use memory at all.

5.4. Correlations Between Job Execution Characteristics

A simple way to check the correlations between job execution characteristics is to calculate the *Pearson's R correlation coefficients* between these variables. However, Pearson's R is very weak and misleading in our case since the variables we study are not normally distributed. Instead, we use *Spearman's rank correlation coefficients* to assess the relationship between job execution characteristics, as it makes no assumptions about the variable's distributions. Correlations that we studied are: memory usage versus job size, memory usage per processor versus job size, actual runtime versus job size, memory usage, and requested runtime. Spearman's r coefficients are listed in Table 8.

Firstly we examine the correlations between memory usage and job size. The Spearman's r coefficients show positive correlations. This indicates that larger size jobs (using more processors) tend to use more memory than smaller jobs. Similar characteristics are reported in [23]. Correlations between memory usage per processor and job size have two folds on DAS-2 clusters. On fs1-3 small positive correlations are observed, while on fs0 and fs4, weak inverse correlations are shown. We would expect that memory usage per processor would increase as the job size increases. However, as is discussed in Section 5.3, memory usage is concentrated on special values. Following the same example in Section 5.3, MPI programs with different job sizes (e.g. 4, 8, 16) use the same amount of memory (324KB). This will result an inverse correlation between memory usage per processor and job size. As the job size increases



(a) CDFs of memory usage

(b) CDFs of memory usage per processor

Figure 8. Distributions of memory usage and memory usage per processor on DAS-2 clusters.

cluster	memory versus job size	memory/proc versus job size	actual runtime versus job size	actual runtime versus memory	actual versus requested runtime
fs0	0.34	-0.02	0.01	0.72	0.44
fs1	0.59	0.22	0.27	0.71	0.61
fs2	0.64	0.13	0.46	0.68	0.45
fs3	0.25	0.08	-0.25	0.54	0.02
fs4	0.13	-0.08	-0.21	0.51	0.62

Table 8. Spearman's rank correlation coefficients between job execution characteristics.

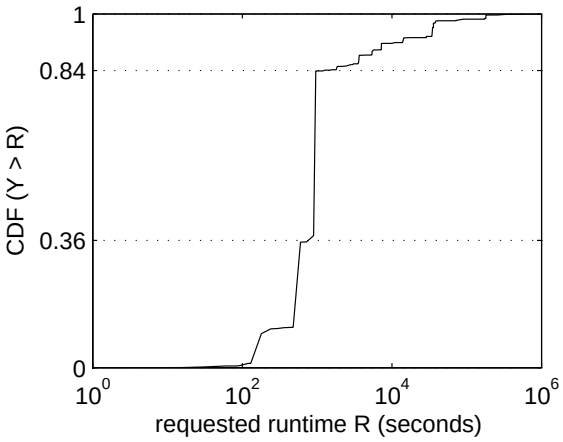
to a certain extent (e.g. 32), the maximum used memory jumps to another level (e.g. 52,620KB). Correspondingly the memory usage per processor grows rapidly and exceeds those of smaller job sizes. This explains why the correlations between memory usage per processor and job size are weak and two-fold.

Correlations between job actual runtime and other characteristics (e.g. job size, requested runtime, etc) are also extensively studied in previous workloads [4, 7, 9]. For job runtime and size, small positive correlation coefficients are reported in [7], meaning that in general larger jobs run longer than smaller jobs. On DAS-2 clusters, however, both positive and negative correlations are observed and it is hard to said in general how the actual runtime is related to size. The correlations between actual and requested runtime appear to be strong (except fs3). Naturally jobs with larger requested runtimes generally run longer. This is clearly observed in Figure 9, which illustrates the requested and actual runtime distributions on fs0. In Figure 9(a), we can see that requested runtimes can be divided into three ranges and each range contains a significant portion of jobs. Actual runtime distributions conditioned on these ranges are shown in Figure 9(b). Jobs with larger requested runtimes run longer is evident by the fact that their CDFs are below those of jobs with smaller requested runtimes.

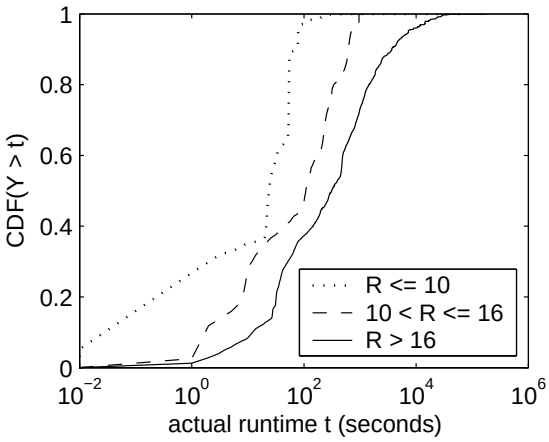
The most significant correlation is obtained between actual runtime and memory usage. This is also illustrated in Figure 10. However, as our observed memory usage is very special compared with other workloads [23], we choose to generate actual runtimes in a synthetic workload based on the requested runtimes. The fitted conditional actual runtime distributions for the five DAS-2 clusters are given is Table 9. Generally speaking, two-phase log-uniform, Weibull, and lognormal are the best fitted distributions for small, medium, and large requested runtimes, respectively. Exception occurs on fs3, where requested runtimes are only divided into medium and large ranges. Above all, distributions conditioned on requested runtimes are more realistic and accurate in modeling job actual runtimes.

6. User/group Behavior

User behavior has been discussed in [2, 11] as an important structure in the workloads. Workloads typically contain

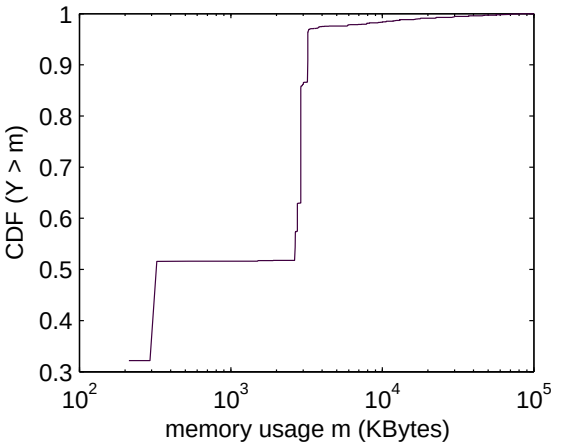


(a) CDF of requested runtime on fs0

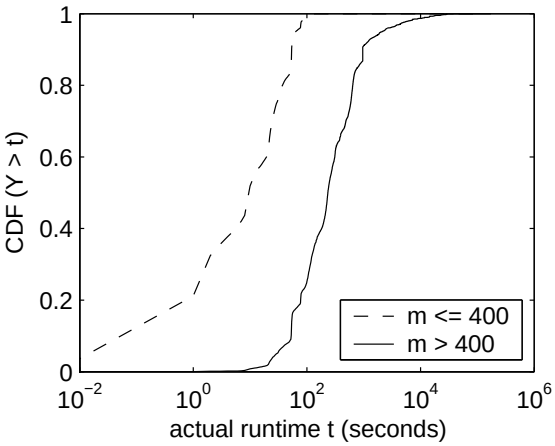


(b) CDFs of actual runtime t conditioned on requested runtime R (minutes) on fs0

Figure 9. CDF of requested runtime and conditional distributions of actual job runtime on fs0.



(a) CDF of memory usage on fs0



(b) CDFs of actual runtime t conditioned on memory usage m (KBytes) on fs0

Figure 10. CDF of memory usage and conditional distributions of actual job runtime on fs0.

a pool of users with different activity levels and periods. A few users and applications tend to dominate the workload. This special structure results in uniformity and predictability on short time scales, allowing better predictions to be made for improving the scheduler performance [11]. Similar structures are observed on the DAS-2 clusters. In Figure 11(a), we can see that there are twelve groups on fs0 in total. Six of them are dominant, contributing to the major fraction of the workload. Among the six groups two of them are the most active. They are local groups[6] at VU (CS staff/group 3 and student/group 7). On other clusters similar behavior is observed: local groups are the most active in their cluster workloads. Group Leiden and Delft are of special interest and they are active on most of the DAS-2 clusters. This is partially because Leiden students have to accomplish grid tasks utilizing more than one clusters, and Delft researchers are experimenting processor co-allocation on multi-clusters.

As to the users, 10 out of 130 are the most active on fs0 (see Figure 11(b)). We further analyze two users with the largest portion of jobs. User 7 submitted more than 40,000 jobs in consecutive seven weeks during October

6 The DAS-2 group and user accounts are mapped onto all five clusters.

cluster	small requested runtime (R - minutes)	middle requested runtime (R - minutes)	large requested runtime (R - minutes)
fs0	0<R≤10, m = 34s, CV = 1.2, loguniform-2 (l = -2.5, m = 1.2, h = 2.1, p = 0.1)	10<R≤16, m = 206s, CV = 1.2, Weibull (a = 150, b = 0.6)	R>16, m = 1624s, CV = 2.9, lognormal (μ = 5.4, σ = 2.2)
fs1	0<R≤10, m = 40s, CV = 0.9, loguniform-2 (l = -2.5, m = 1.2, h = 2, p = 0.08)	10<R≤60, m = 250s, CV = 1.5, Weibull (a = 184, b = 0.7)	R>60, m = 6022s, CV = 2.8, lognormal (μ = 6.4, σ = 2.9)
fs2	0<R≤10, m = 69s, CV = 0.8, loguniform-2 (l = -2.6, m = 1.6, h = 2.1, p = 0.03)	10<R≤60, m = 301s, CV = 1.5, Weibull (a = 229, b = 0.7)	R>60, m = 7473s, CV = 4.9, lognormal (μ = 6, σ = 2.7)
fs3	none	0<R≤61, m = 85s, CV = 1.8, Weibull (a = 71, b = 0.8)	R>61, m = 10060s,CV = 2.8, lognormal (μ = 6.9, σ = 2.6)
fs4	0<R≤16, m = 72s, CV = 1.5, loguniform-2 (l = -2.5, m = 1.7, h = 2.3, p = 0.04)	16<R≤600, m = 3131s, CV = 10.5, Weibull (a = 1369, b = 0.5)	R>600, m = 4270s,CV = 3.1, lognormal (μ = 6.6, σ = 2.1)

Table 9. Distributions of job actual runtimes conditioned on requested runtimes (loguniform-2 stands for two-stage log-uniform distribution, m - mean, CV - Coefficient of Variation).

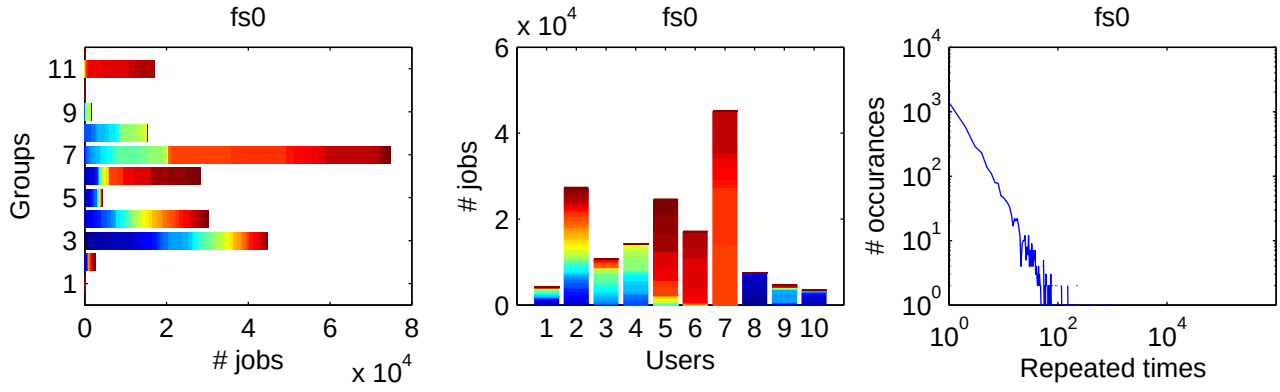
and November 2003, which is his/her only active period throughout the year. Moreover, these jobs all have the same name and request two processors. Jobs from user 2 are distributed evenly throughout the year, but 70% of them have the same name and 90% request two processors. This structure explains some of our main observations before - a majority of DAS-2 workloads have a job size of two processors, and certain applications appear many more times than others. Figure 11(c) shows the application repeated times and their number of occurrences on fs0. We can see that while lots of applications run only once or a small number of times, there are highly repeated applications that contribute to the heavy tail in the distribution. Similar phenomena are reported on other workloads [11]. Techniques and models have been proposed to capture the user behavior in the workloads [24].

7. Conclusions and Future Work

In this paper, we present a comprehensive characterization of a multi-cluster supercomputer (DAS-2) workload. We characterized system utilization, job arrival process (arrival rate, interarrival time, and cancellation rate), job execution characteristics (job size, runtime, and memory usage), correlations between different metrics, and user/group behavior. Differences of DAS-2 workloads compared with previously reported workloads include the following:

1. A substantially lower average system utilization (from 7.3% to 22%) is observed.
2. Lower job cancellation rates (3.3%-10.6%) are observed than in previously reported workloads (12%-23%).
3. Power-of-two phenomenon of job sizes is clearly observed, with an extreme popularity of job size *two*. The fraction of serial jobs (0.9%-4.7%) is much lower than other workloads (30%-40%).
4. The job actual runtimes are strongly correlated with memory usage as well as job requested runtimes. Conditional distributions based on requested runtime ranges are well fitted for actual runtimes.
5. A large portion of jobs has very small memory usage and several special values are used by a major fraction of jobs.

To facilitate generating synthetic workloads, we provide distributions and conditional distributions of the main characteristics. The distributions are summarized as follows:



(a) Jobs submitted by different groups in consecutive weeks

(b) Jobs submitted by 10 most active users in consecutive weeks

(c) Application repeat times and their number of occurances

Figure 11. Activity of groups, users and applications on the cluster fs0. From left to right the color changes (gray scale on a none-color printer) of bars symbolize the consecutive weeks in year 2003.

1. Interarrival time: in high load period, gamma or two phase hyperexponential are the most suitable distributions; in representative period, Weibull gives the best fit.

2. Cancellation lag: lognormal is the best fitted distribution.

3. Job size: two-stage loguniform is the most suitable distribution.

4. Actual runtime: Weibull or lognormal is the best fitted distribution.

5. Actual runtime conditioned on requested time ranges (R): for small R, two-stage loguniform is the most suitable distribution; for medium R, Weibull is the best fitted distribution; for large R, lognormal gives the best fit.

In future work, we plan to generate workload models based on the results in this paper and evaluate several scheduling strategies for DAS-2 clusters. Since the goal of DAS-2 system is to provide fast response time to researchers, load balancing techniques and higher level resource brokering are to be investigated. Another interesting point in a multi-cluster environment is co-allocation. Currently multi-cluster job information is not logged on the DAS-2 clusters. We plan to instrument the Globus gatekeeper to collect the necessary traces and identify the key characteristics for multi-cluster jobs.

8. Acknowledgments

The DAS-2 supercomputer is funded by NWO (Netherlands Organization for Scientific Research) and the participating universities. We thank Dr. Dick Epema (Delft University of Technology) and the referees for their many valuable suggestions that improved the quality of this paper.

References

[1] M. Calzarossa and G. Serazzi. Workload characterization: A survey. *Proc. IEEE*, 81(8): 1136–1150, 1993.

[2] D. G. Feitelson. Workload modeling for performance evaluation. *Lecture Notes in Computer Science*, 2459:114–141, 2002.

[3] Parallel Workload Archive. http://www.cs.huji.ac.il/labs/parallel/workload/.

[4] S.-H. Chiang and M. K. Vernon. Characteristics of a large shared memory production workload. *Lecture Notes in Computer Science*, 2221: 159–187, 2001.

[5] D. Feitelson and B. Nitzberg. Job characteristics of a production parallel scientific workload on the NASA ames iPSC/860. In D. G. Feitelson and L. Rudolph, editors, *Job Scheduling Strategies for Parallel Processing – IPPS'95 Workshop*, volume 949, pages 337–360. Springer, 1995.

[6] K. Windisch, V. Lo, R. Moore, D. Feitelson, and B. Nitzberg. A comparison of workload traces from two production parallel machines. In *6th Symp. Frontiers Massively Parallel Comput.*, pages 319–326, 1996.

[7] U. Lublin and D. G. Feitelson. The workload on parallel supercomputers: modeling the characteristics of rigid jobs. *J. Parallel and Distributed Comput.*, 63(11): 1105–1122, 2003.

[8] J. Jann, P. Pattnaik, H. Franke, F. Wang, J. Skovira, and J. Riodan. Modeling of workload in MPPs. In D. G. Feitelson and L. Rudolph, editors, *Job Scheduling Strategies for Parallel Processing*, pages 95–116. Springer Verlag, 1997.

[9] W. Cirne and F. Berman. A comprehensive model of the supercomputer workload. In *IEEE 4th Annual Workshop on Workload Characterization*, 2001.

[10] S. J. Chapin, W. Cirne, D. G. Feitelson, J. P. Jones, S. T. Leutenegger, U. Schwiegelshohn, W. Smith, and D. Talby. Benchmarks and standards for the evaluation of parallel job schedulers. In D. G. Feitelson and L. Rudolph, editors, *Job Scheduling Strategies for Parallel Processing*, pages 67–90. Springer-Verlag, 1999.

[11] A. B. Downey and D. G. Feitelson. The elusive goal of workload characterization. *Perf. Eval. Rev.*, 26(4): 14–29, 1999.

[12] The DAS-2 Supercomputer. http://www.cs.vu.nl/das2.

[13] S. Banen, A. Bucur and D. H. J. Epema. A Measurement-Based Simulation Study of Processor Co-Allocation in Multicluster Systems. In D. G. Feitelson and L. Rudolph, editors, *Job Scheduling Strategies for Parallel Processing*, pages 105–128. Springer-Verlag, 2003.

[14] Portable Batch System. http://www.openpbs.org.

[15] The Maui Scheduler. http://www.supercluster.org.

[16] The Globus project. http://www.globus.org.

[17] O. Allen. Probability, Statistics, and Queueing Theory with Computer Science Applications. Acdemic Press, 1978.

[18] R. E. A. Khayari, R. Sadre, B. R. Haverkort. Fitting world-wide web request traces with the EM-algorithm. Performance Evaluation 52, pp 175–191, Elsevier, 2003.

[19] Matlab. http://www.mathworks.com.

[20] Dataplot. http://www.itl.nist.gov/div898/software/dataplot/.

[21] The EMpht programme. http://www.maths.lth.se/matstat/staff/asmus/pspapers.html.

[22] Allen B. Downey. Using Queue Time Predictions for Processor Allocation. In D. G. Feitelson and L. Rudolph, editors, *Job Scheduling Strategies for Parallel Processing*, pages 35–57. Springer-Verlag, 1997.

[23] D. G. Feitelson Memory usage in the LANL CM-5 Workload. In D. G. Feitelson and L. Rudolph, editors, *Job Scheduling Strategies for Parallel Processing*, pages 78–94. Springer-Verlag, 1997.

[24] M. Calzarossa and G. Serazzi. Construction and use of multiclass workload models. *Performance Evaluation*, 19(4): 341–352, 1994.