Trace analysis report GWA-T-5

Trace analysis report NGS

This is the trace analysis report (generated by reportgen.py) for the NGS system. The trace data was taken from the filename ngs_jobs.gwf, which contains job data obtained from. Below is a summary of the contents of the trace data:

- Date first entry: Mon Dec 15 19:34:15 2003
- CPU time consumed by jobs: 269y 226d 3h 6m 38s
- Number of sites in the system: -
- Number of CPUs in the trace: -
- Number of jobs in the trace: 631737
- Number of users in the trace: 379
- Number of groups in the trace: 1

System-wide characteristics

System utilization

We define the overall system utilization as the ratio between the total CPU time consumed by users, and the total CPU time available to the users. We compute the total CPU time consumed by users as the sum of CPU time consumed by each job in the system; for failed jobs, only those that have effectively spent resource time are considered. We compute the total CPU time available as the number of CPUs multiplied by the duration of a fixed time interval, c.q. 10 minutes.

Below we show the statistical properties of both the overall system utilization and the overall system for non-zero values, that is, excluding all intervals that have system utilization equal to zero. This excludes values that may account for downtime of the system.

Figure 1 shows System utilization over time.

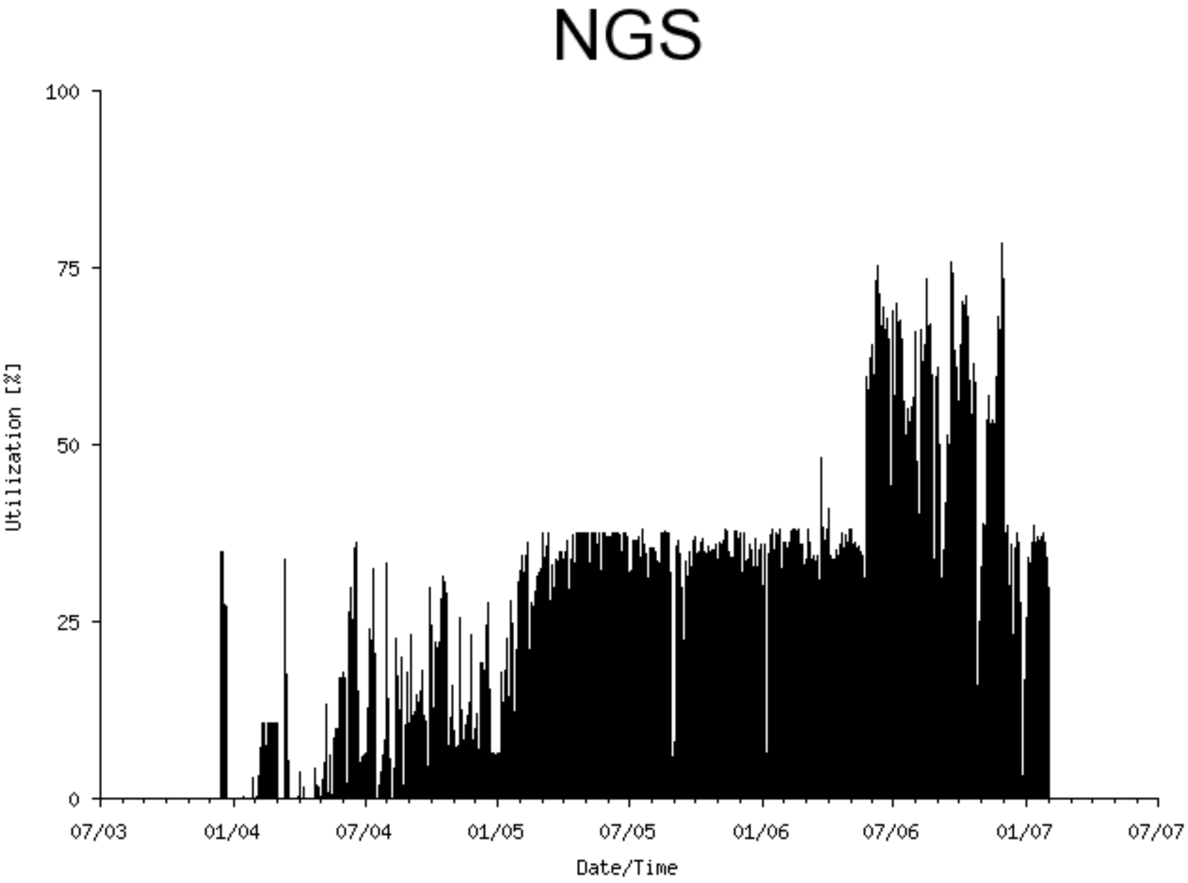

Figure 1: System utilization over time

Overall system utilization

- Minimum: 0.0 percent
- Maximum: 78.589 percent
- Average: 22.821 percent

Overall system utilization for non-zero values

- Minimum: 0.101 percent
- Maximum: 78.589 percent
- Average: 27.26 percent

Job arrival rate

We define the job arrival rate as the number of jobs that are submitted to the system in a fixed time interval. We compute the arrival rate for every hour by counting the all jobs that are recorded in the trace during that hour. This includes failed jobs and jobs that are cancelled before execution. Below we list the time periods in which the highest number of jobs were submitted to the system. We also summarize statistical properties for all job arrival rate values, and the statistical properties for arrival rate higher than zero. This excludes time periods that may account to downtime of the system.

Figure 2 shows Overall job arrival rate during hourly intervals.

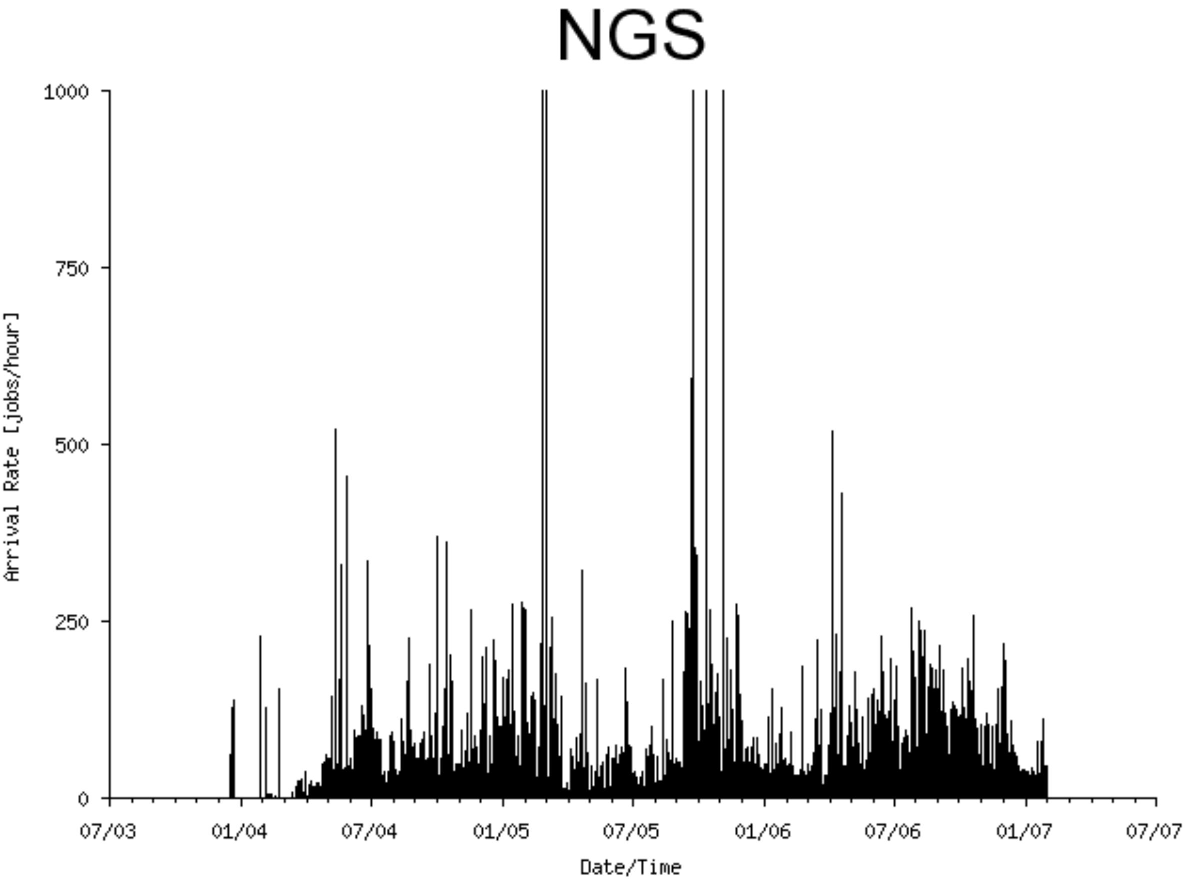

Figure 2: Overall job arrival rate during hourly intervals

Busiest time periods in terms of number of job submissions

- Busiest day: 2005-03-01
- Busiest week: 2005-38
- Busiest month: 2006-06

Overall job arrival metrics

- Minimum: 0.00 jobs/hour
- Maximum: 4994.00 jobs/hour
- Average: 23.02 jobs/hour

Overall job arrival metrics for non-zero values

- Minimum: 2.00 jobs/hour
- Maximum: 4994.00 jobs/hour
- Average: 28.57 jobs/hour

Job characteristics

We compute three important characteristics of jobs in the trace: number of CPUs used, the runtime of the job and the amount of memory used. Below we summarize the statistical properties for single jobs in the trace. We do not include jobs that were cancelled before execution, because those jobs did not consume resources from the system.

Figure 3 shows CDFs of the most important job characteristics.

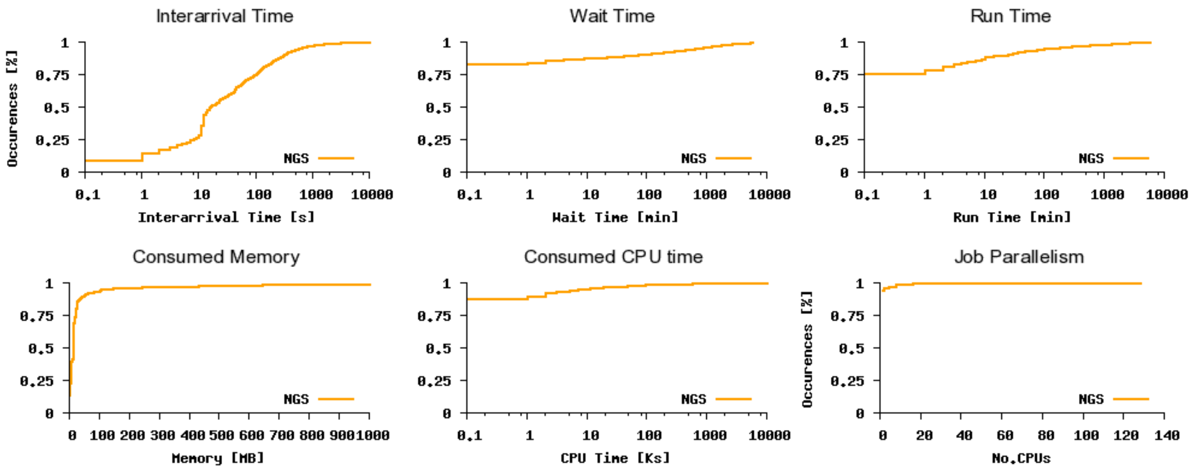

Figure 3: CDFs of the most important job characteristics

Number of CPUs used by a single job

- Minimum: 1 processors
- Maximum: 128 processors
- Average: 1.406 processors
- Standard deviation: 2.828
- Coefficient of variation: 2.012

Runtime of a single job

- Minimum: 0.00 seconds
- Maximum: 1206253.00 seconds
- Average: 2924.64 seconds
- Standard deviation: 17908.126
- Coefficient of variation: 6.123

Memory usage of a single job

- Minimum: 0.00 MB
- Maximum: 65993.98 MB
- Average: 38.91 MB
- Standard deviation: 226.215
- Coefficient of variation: 5.814

Sequential vs. Parallel jobs

Below we summarize the resource usage of all sequential and all parallel jobs, that is all jobs that use more than one processor. First we calculate the number of sequential jobs and the number of parallel jobs that are submitted to the system. Furthermore, we compute the consumed CPU time by multiplying the runtime of a job by the number of processors allocated to the job. Again, this is divided into parallel and sequential jobs. For the number of jobs and the consumed CPU time, the percentage of all jobs is displayed.

Number of jobs

- Sequential: 599090 jobs (94.83 percent)
- Parallel: 32647 jobs (5.17 percent)

Consumed CPU Time

- Sequential: 1313142197 seconds (15.44 percent)
- Parallel: 7189537079 seconds (84.56 percent)

User and group characteristics

User characteristics

Figure 4 shows The number of submitted jobs and the consumed CPU time by user.

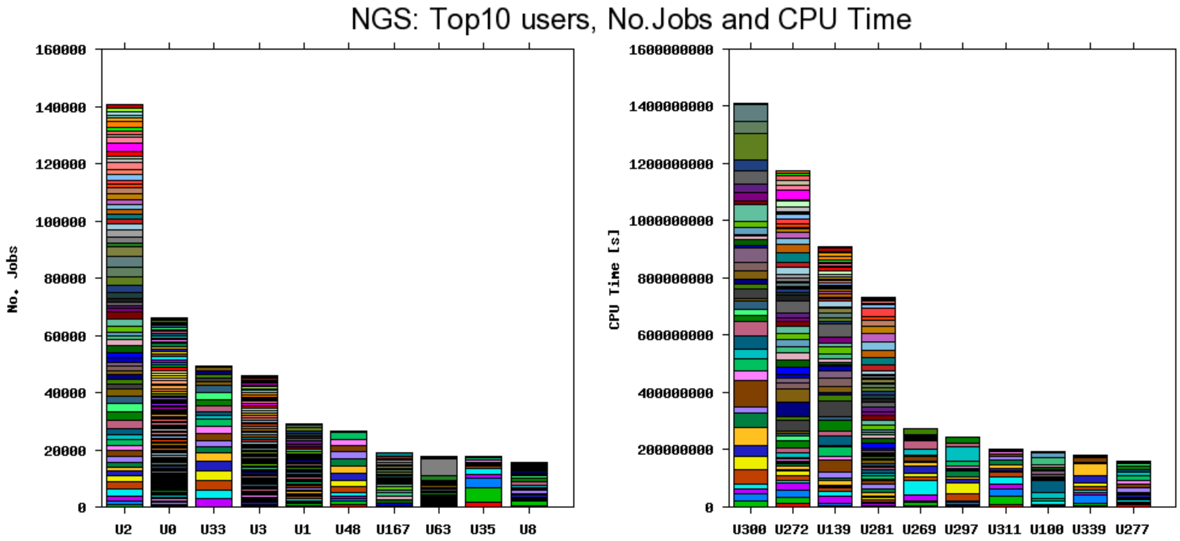

Figure 4: The number of submitted jobs (left) and consumed CPU time (right) by user. Only the top 10 users are displayed. The horizontal axis depicts the user's rank. The vertical axis shows the cumulated values, and the breakdown per week. Users have the same labels in the left and right sub-graphs

Top 10 users by number of job submitted to the system

Table 1 shows Top 10 users by number of jobs submitted to the system.

Table 1

Rank	UserID	Number of jobs	Percentage
1	U2	140449	22.23%
2	U0	66091	10.46%
3	U33	49460	7.83%
4	U3	45850	7.26%
5	U1	29114	4.61%
6	U48	26597	4.21%
7	U167	19058	3.02%
8	U63	17821	2.82%
9	U35	17575	2.78%
10	U8	15456	2.45%

Table 1

Rank	UserID	Number of jobs	Percentage
11	Other	204266	32.33%
12	Total	631737	100.00%

System utilization

- Minimum: 0.0 percent
- Maximum: 12.676 percent
- Average: 0.584 percent

Job arrival

- Minimum: 0.00 jobs/hour
- Maximum: 370.00 jobs/hour
- Average: 17.30 jobs/hour

Job characteristics

Number of CPUs used by a single job

- Minimum: 1 processors
- Maximum: 24 processors
- Average: 1.003 processors
- Standard deviation: 0.000
- Coefficient of variation: 0.000

Runtime of a single job

- Minimum: 0.00 seconds
- Maximum: 424859.00 seconds
- Average: 439.24 seconds
- Standard deviation: 5235.975
- Coefficient of variation: 11.921

Memory usage of a single job

- Minimum: 0.00 MB
- Maximum: 3033.47 MB
- Average: 15.44 MB
- Standard deviation: 58.936
- Coefficient of variation: 3.817

Top 10 users by consumed CPU time

Table 2 shows Top 10 users by consumed CPU time (in seconds).

Table 2

Rank	UserID	CPU seconds	Percentage
1	U300	1409246013	16.57%
2	U272	1173650589	13.80%
3	U139	907641122	10.67%
4	U281	733410806	8.63%
5	U269	271887472	3.20%
6	U297	242801528	2.86%

Table 2

Rank	UserID	CPU seconds	Percentage
7	U311	201431123	2.37%
8	U100	194939109	2.29%
9	U339	179104842	2.11%
10	U277	159954412	1.88%
11	Other	3028654582	35.62%
12	Total	8502721598	100.00%

System utilization

- Minimum: 0.0 percent
- Maximum: 65.608 percent
- Average: 18.614 percent

Job arrival

- Minimum: 0.00 jobs/hour
- Maximum: 2986.00 jobs/hour
- Average: 0.80 jobs/hour

Job characteristics

Number of CPUs used by a single job

- Minimum: 1 processors
- Maximum: 96 processors
- Average: 5.558 processors
- Standard deviation: 10.247
- Coefficient of variation: 1.844

Runtime of a single job

- Minimum: 0.00 seconds
- Maximum: 1194936.00 seconds
- Average: 29910.08 seconds
- Standard deviation: 59429.938
- Coefficient of variation: 1.987

Memory usage of a single job

- Minimum: 0.00 MB
- Maximum: 65993.98 MB
- Average: 74.44 MB
- Standard deviation: 1067.833
- Coefficient of variation: 14.345

Group characteristics

Figure 5 shows The number of submitted jobs and consumed CPU time by group.

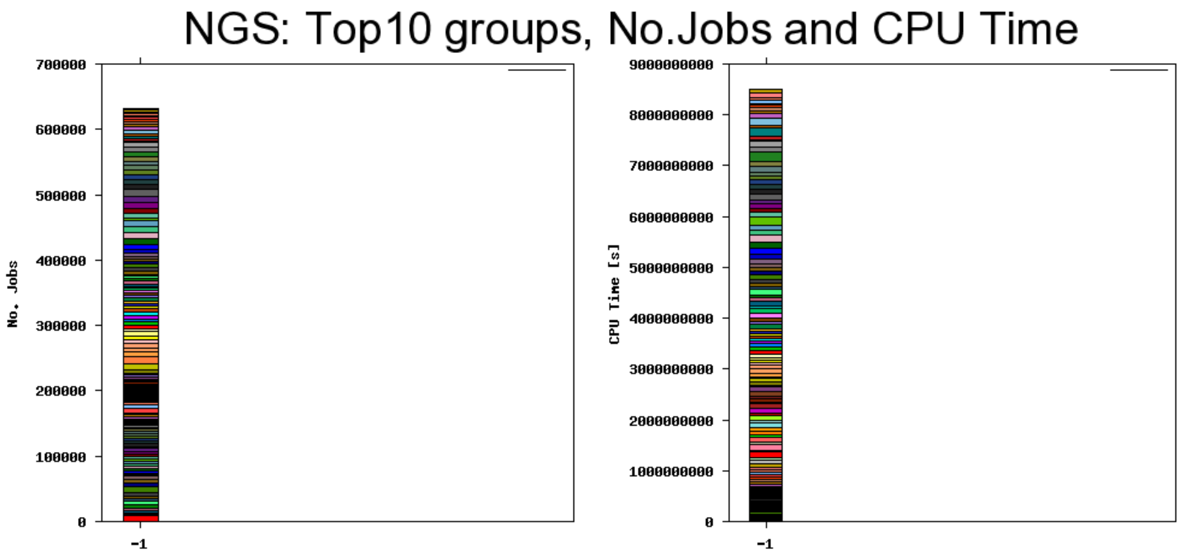

Figure 5: The number of submitted jobs (left) and consumed CPU time (right) by group. Only the top 10 groups are displayed. The horizontal axis depicts the groups rank. The vertical axis shows the cumulated values, and the breakdown per week. Groups have the same labels in the left and right sub-graphs

Table 3 shows Top 10 groups by number of jobs submitted to the system.

Table 3			
Rank	GroupID	Number of jobs	Percentage
1	1	631737	100.00%
2	Other	0	0.00%
3	Total	631737	100.00%

Table 4 shows Top 10 Groups by consumed CPU time (in seconds).

Table 4			
Rank	GroupID	CPU seconds	Percentage
1	1	8502721598	100.00%
2	Other	0	0.00%
3	Total	8502721598	100.00%

Performance analysis

Waiting and running jobs

Figure 6 shows The number of running and of waiting jobs during hourly intervals. The vertical axis is limited to 7500 for better visibility.

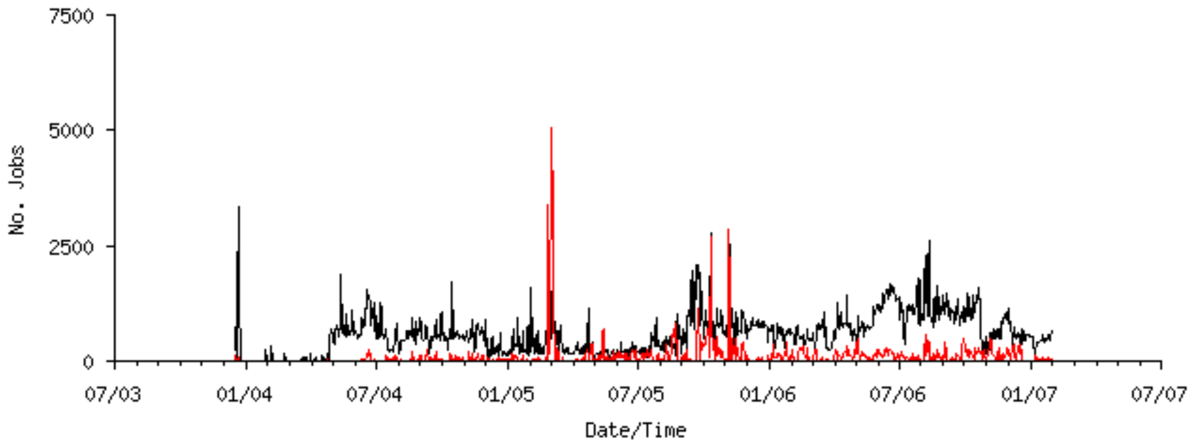

Figure 6: The number of running and of waiting jobs during hourly intervals. The vertical axis is limited to 7500 for better visibility

We compute the number of running and waiting jobs by considering a fixed time interval. In each time interval, we count in the trace the amount of jobs that have been submitted but not yet started, that is, waiting. We also count the number of jobs that have been submitted, and have started executing in the time interval, but did not finish executing, and thus are running. Below we show the values for an interval value of 3600 seconds, summarized in amounts per day. Also the summary for values higher than zero are displayed, which excludes the possible effect of downtime of the system.

Number of waiting jobs per day

- Minimum: 0 jobs
- Maximum: 5044 jobs
- Average: 111.79 jobs

Number of waiting jobs per day (non-zero values)

- Minimum: 1 jobs
- Maximum: 5044 jobs
- Average: 135.36 jobs

Number of running jobs per day

- Minimum: 0 jobs
- Maximum: 3350 jobs
- Average: 589.15 jobs

Number of running jobs per day (non-zero values)

- Minimum: 1 jobs
- Maximum: 3350 jobs
- Average: 622.94 jobs

Throughput

We compute the job throughput by considering a fixed time interval. In each time interval, we count in the trace the amount of jobs that have been submitted, started and finished executing. Below we show the values for an interval value of 3600 seconds, summarized in amounts per day. Also the summary for values higher than zero are displayed, which excludes the possible effect of downtime of the system.

Figure 7 shows Throughput during hourly intervals. The vertical axis of each individual site graph is limited to 7500 for better visibility.

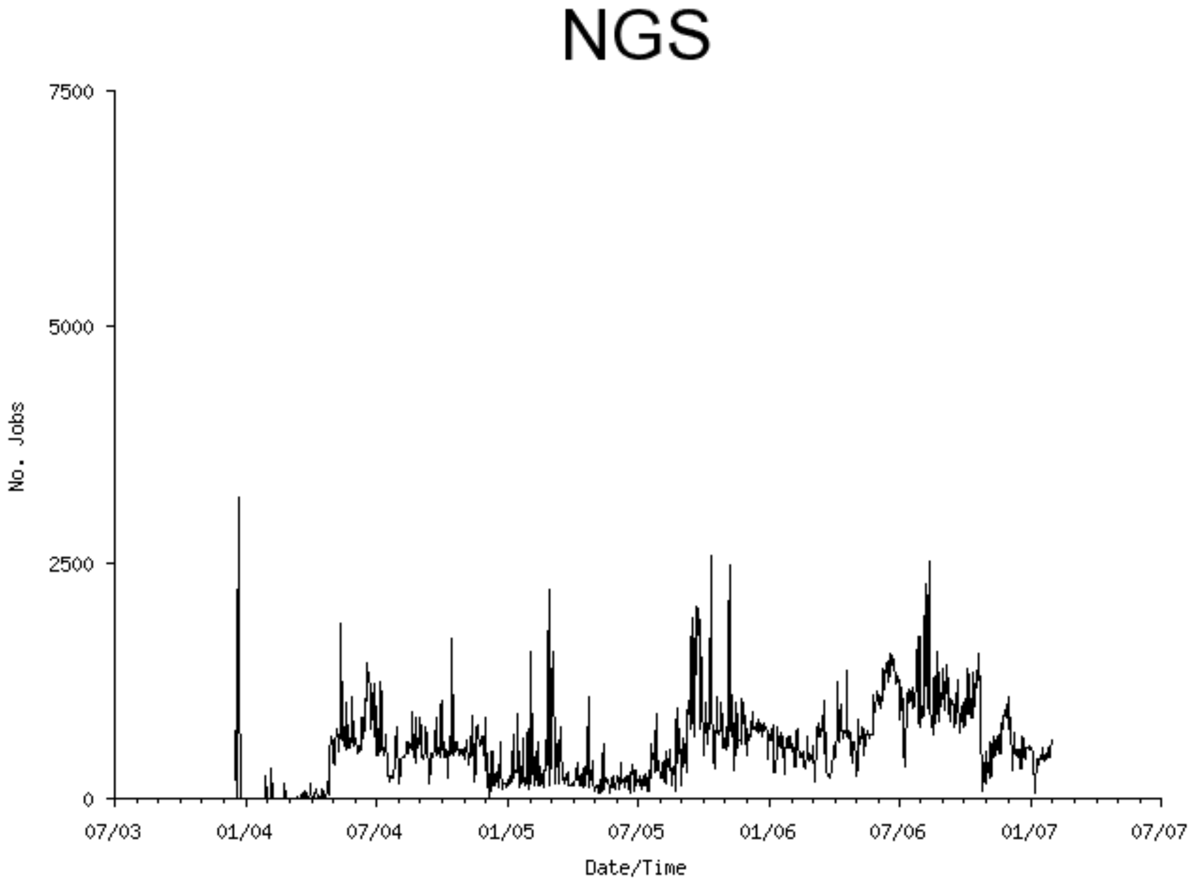

Figure 7: Throughput during hourly intervals. The vertical axis of each individual site graph is limited to 7500 for better visibility

Throughput per day

- Minimum: 0 jobs
- Maximum: 3201 jobs
- Average: 552.24 jobs

Throughput per day (non-zero values)

- Minimum: 1 jobs
- Maximum: 3201 jobs
- Average: 588.27 jobs

Completed jobs

Figure 8 shows The number of completed jobs during hourly intervals.

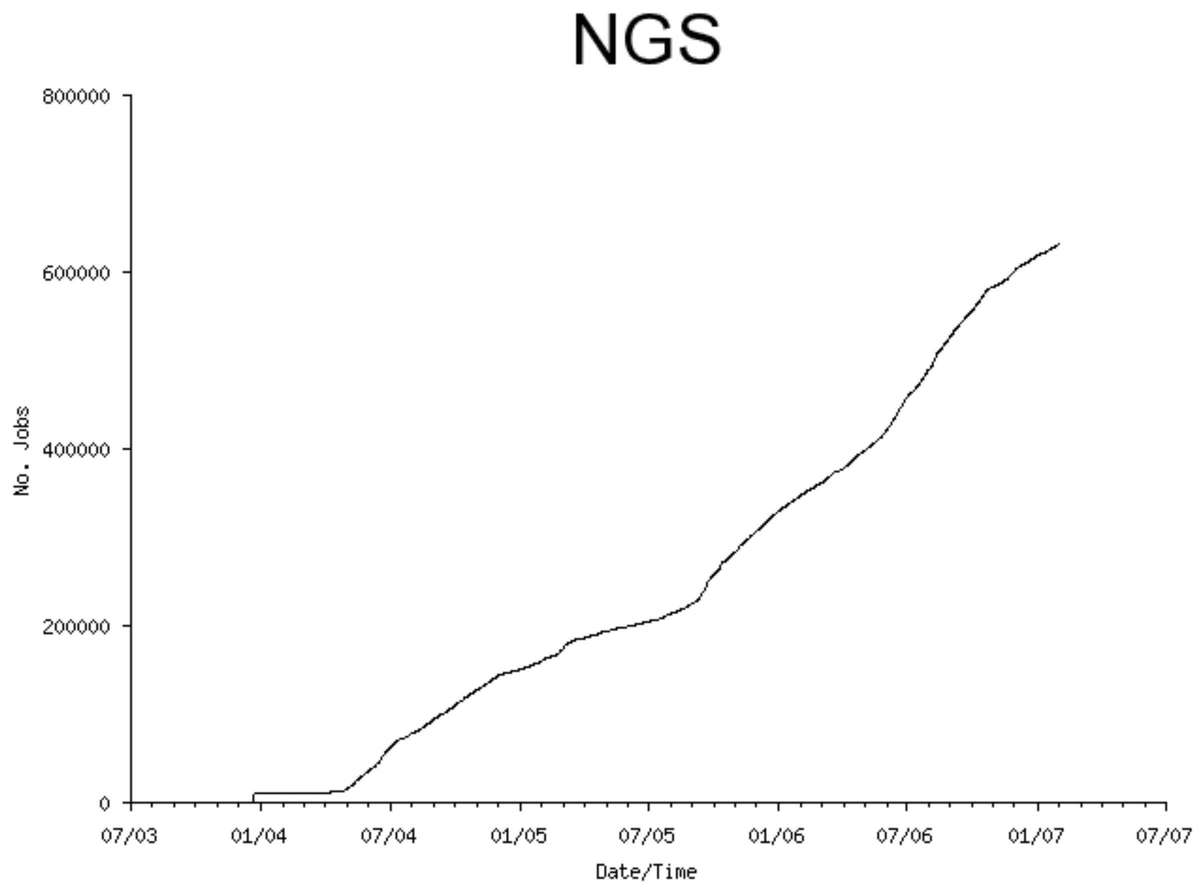

Figure 8: The number of completed jobs during hourly intervals

Workload model

This section contains the workload model for the analyzed trace. The workload model consists of several parameters: job size, job runtime, requested runtime and interrarivals of jobs. These parameters are modeled by fitting well-known distributions to the data obtained from the trace. In all cases, first a logarithmic transformation was performed on the dataset to diminish the effect of outliers and speed up the modelling process. The fitting was performed using the maximum likelihood estimation method, which tries to maximize the log-likelihood function of each distribution given a dataset.

Job size

Figure 9 shows Cumulative distribution function for the logarithm of the job sizes, with fitted distributions.

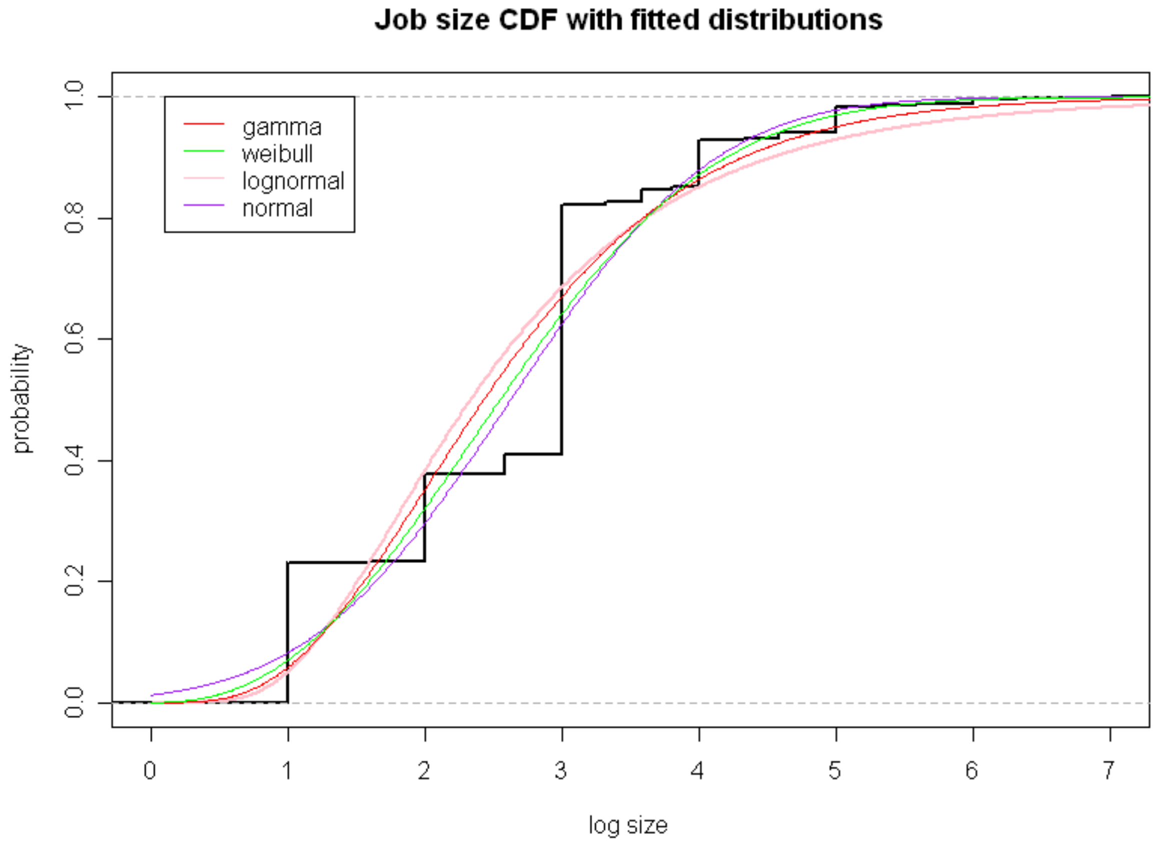

Figure 9: Cumulative distribution function for the logarithm of the job sizes, with fitted distributions

Job runtime

Figure 10 shows Cumulative distribution function for the logarithm of the job runtimes, with fitted distributions.

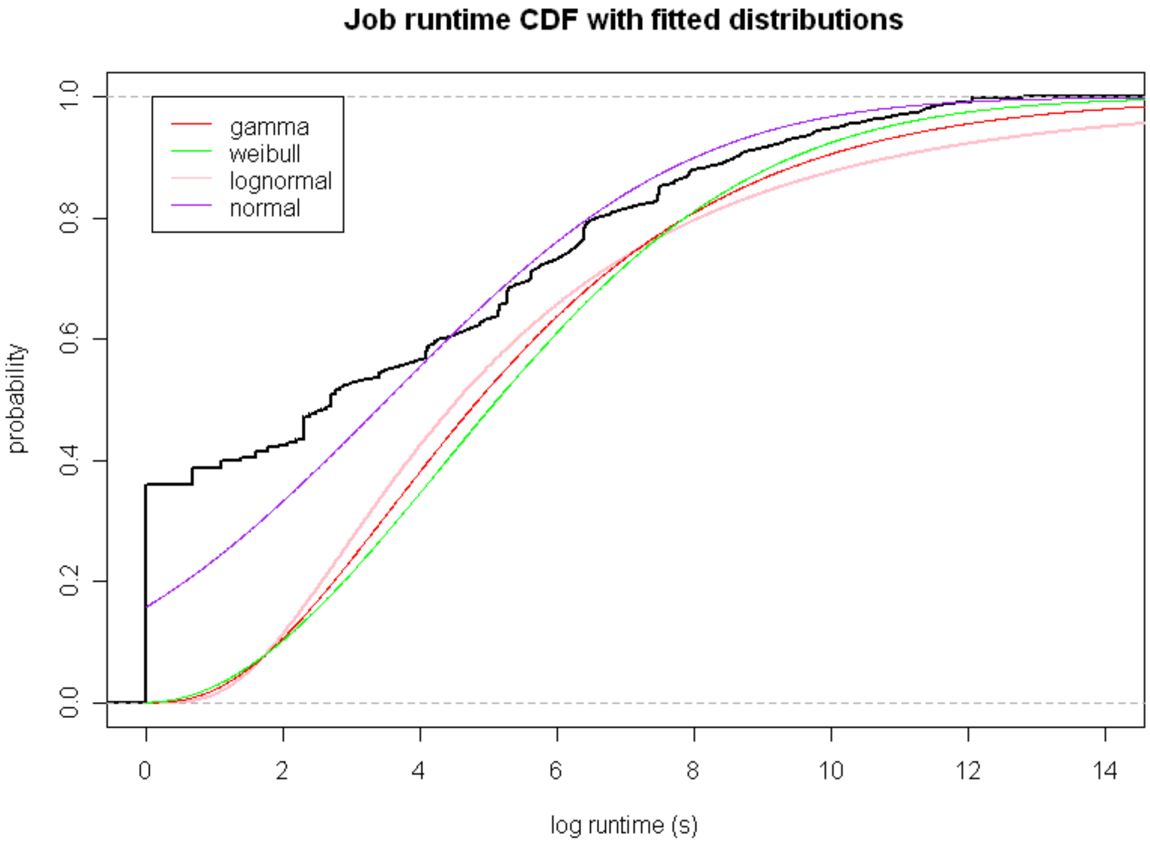

Figure 10: Cumulative distribution function for the logarithm of the job runtimes, with fitted distributions

Job requested runtime

Figure 11 shows Cumulative distribution function for the logarithm of the job requested runtimes, with fitted distributions.

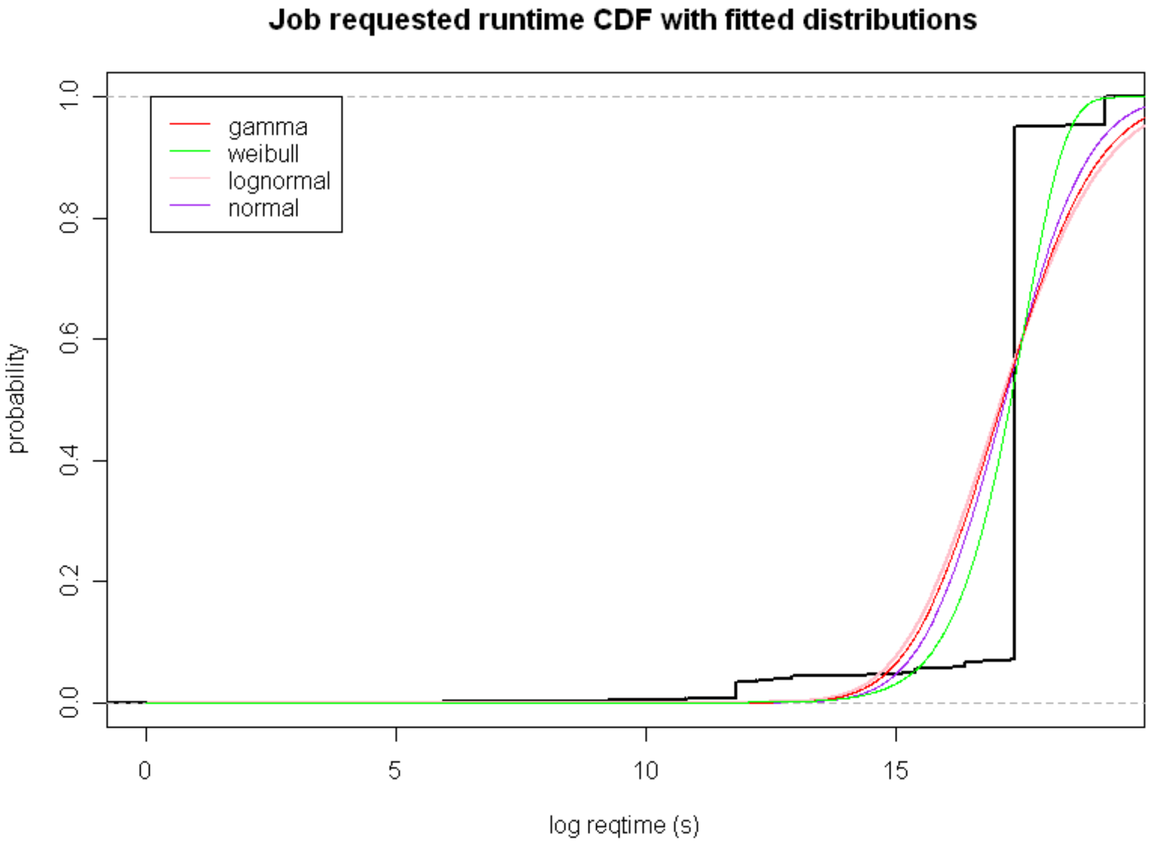

Figure 11: Cumulative distribution function for the logarithm of the job requested runtimes, with fitted distributions

Parameters of fitted distributions

Lognormal distribution

- meanlog: 2.8407907641311
- sdlog: 0.0921666208869822

Exponential distribution

- rate: 0.0581638692580985

Normal distribution

- sd: 1.30486789944264
- mean: 17.1928039306079

Weibull distribution

- shape: 21.6902532213033
- scale: 17.6050670698576

Hyperexponential distribution

- p: 5.82920054484107
- rate2: 0.0545599148050512
- rate1: 0.0858315857643764

Gamma distribution

- shape: 135.110805541538
- rate: 7.8585277397255

Goodness-of-fit (Kolmogorov-Smirnov test)

Table 7 shows for each distribution the results for the Kolmogorov-Smirnov test, which gives a measure for the distance of the distribution to the original dataset (lower distance => better fit).

Table 7

Distribution	Distance
Lognormal	0.494671704762857
Exponential	0.566179811045743
Normal	0.490082922539964
Weibull	0.46624466163712
Hyperexponential	0.54265998838151
Gamma	0.493841855045911

Job interarrival

Figure 12 shows Cumulative distribution function for the logarithm of the job interarrival, with fitted distributions.

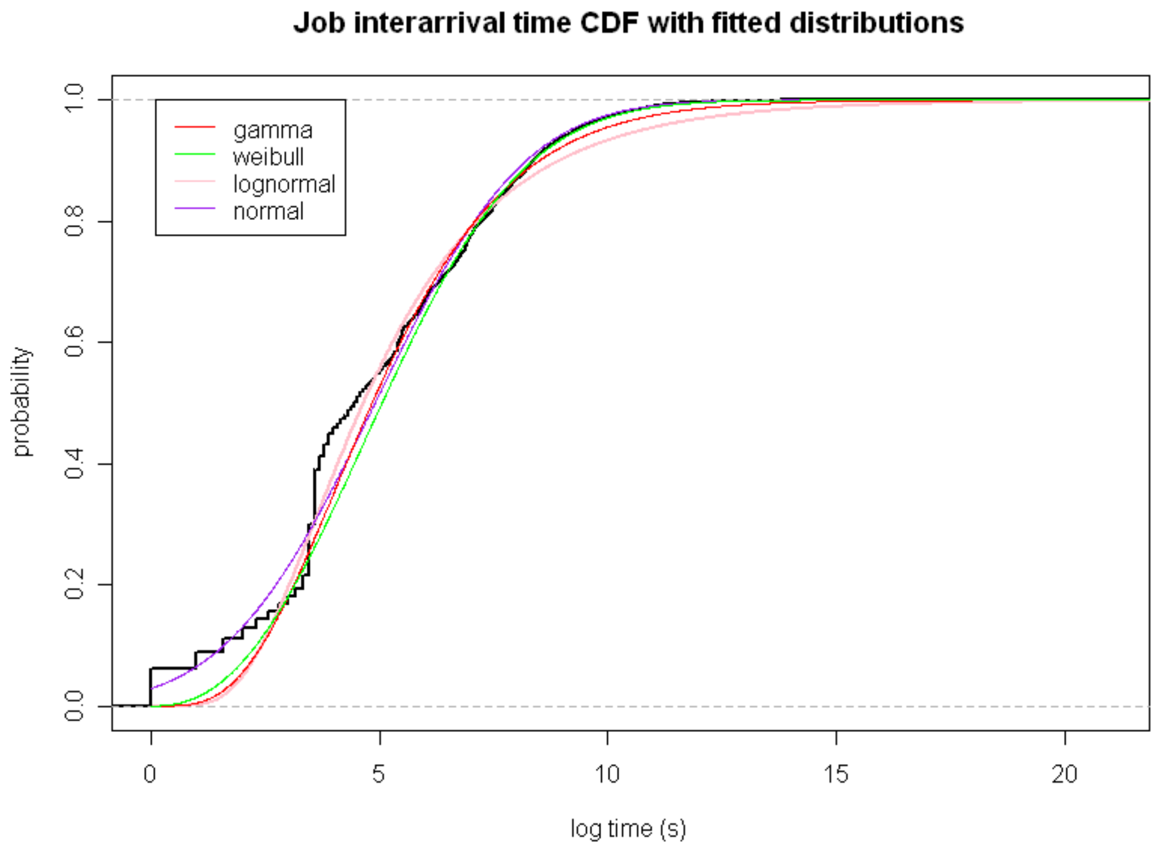

Figure 12: Cumulative distribution function for the logarithm of the job interarrival, with fitted distributions