Simulation results for JSQ Server Farms with Processor Sharing Servers

Varun Gupta[*] **Mor Harchol-Balter**[*]
Karl Sigman[†] **Ward Whitt**[†]

August 2007
CMU-CS-07-xxx

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[*]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[†]Department of Industrial Engineering and Operations Research, Columbia University, New York, NY, USA

Abstract

In [1] we present the first analysis of Join-the-Shortest-Queue (JSQ) routing in Web server farms, with Processor-Sharing (PS) servers. Please see that paper for details about the JSQ routing policy and its analysis, and for the relevant references. The analytical work in [1] was backed up by many simulation experiments, not all of which made it into the paper. This document is meant to serve as an online supplement to [1], providing the full set of simulation results and comparisons with analytical results. We hope that these will be useful for those who wish to extend our work on this problem.

Keywords: Shortest queue routing; JSQ; Processor sharing; Insensitivity; Single-queue approximation

1 The model

We will be concerned with JSQ/PS server farms, as depicted in Figure 1. We assume that jobs arrive according to a Poisson Process with rate λ. Each arrival is immediately dispatched by a front-end router to one of K servers, via the Join-the-Shortest-Queue (JSQ) routing policy. Under JSQ, a job goes to the server with the fewest current number of jobs. Ties are broken by flipping a fair coin. Each server employs Processor-Sharing (PS) scheduling of its jobs. That is, when there are n jobs at the server, each simultaneously receives $1/n$th of the total server capacity. The size of a job is defined as the time the job takes to run on a server in isolation. We assume that job sizes are sampled in an i.i.d. fashion from some distribution and the mean job size is denoted by μ^{-1}. Further, we define the load, ρ, of the system by

$$\rho = \frac{\lambda}{K\mu}.$$

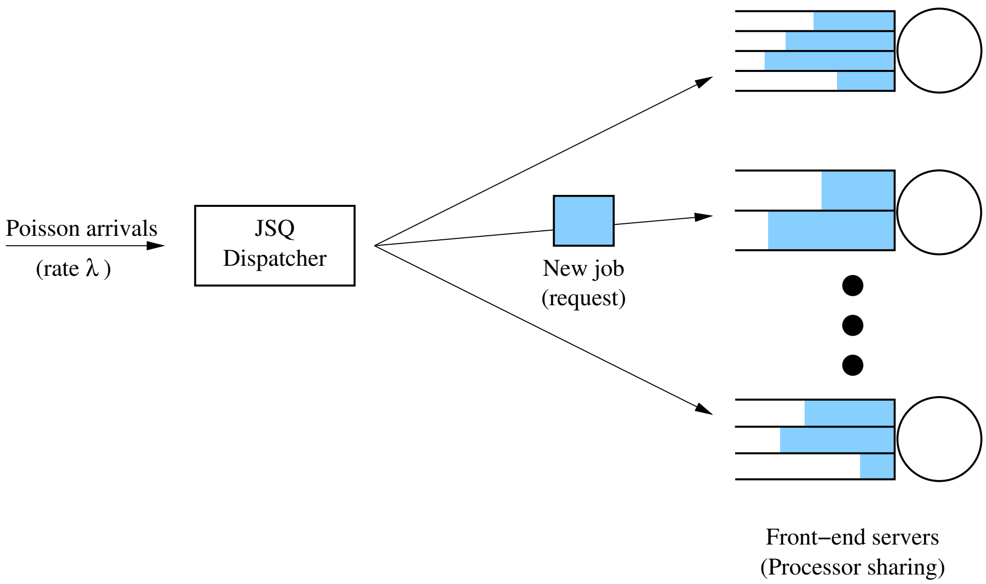


Figure 1: Server farm with front-end dispatcher and K identical processor sharing back-end servers.

Our performance metric throughout will be $E[N]$ and $E[N^2]$, where N, is a random variable denoting the *number of jobs at the first queue*. Without loss of generality, we pick the first queue to concentrate on, since all queues are statistically identical.

We are also interested in understanding the *rate of arrivals into the first queue* of the JSQ/PS server farm. The SQA (Single-Queue-Approximation) analysis technique introduced in [1] relies strongly on a quantity called $\lambda(n)$, which denotes the *conditional arrival rate into the first queue, given that there are n jobs at that queue*. Specifically, we define:

$$\lambda(n) = \lim_{t\to\infty} \frac{A_n(t)}{T_n(t)} \ ,$$

where $A_n(t)$ is the number of arrivals during the time interval $[0, t]$ finding n jobs in queue 1, while $T_n(t)$ is the total time during $[0, t]$ during which there are n jobs in queue 1. In this paper we will be making many detailed plots and tables concerning the quantity $\lambda(n)$, as a function of n.

We will consider several job size distributions throughout. Below we list a set of job size distributions all with mean 2, but with a wide range of values of variance. These distributions will be used often in our experiments, so we list them here:

1. `Deterministic`: point mass at 2 (variance $= 0$)

2. `Erlang2`: sum of two exponential random variables with mean 1 (variance $= 2$)

3. `Exponential`: exponential distribution with mean 2 (variance $= 4$)

4. `Bimodal-1`: (mean $= 2$, variance $= 9$)

$$X = \begin{cases} 1 & w.p.\ 0.9 \\ 11 & w.p.\ 0.1 \end{cases}$$

5. `Weibull-1`: Weibull with shape parameter $= 0.5$ and scale parameter $= 1$ (heavy-tailed, mean $= 2$, variance $= 20$)

6. `Weibull-2`: Weibull with shape parameter $= \frac{1}{3}$ and scale parameter $= \frac{1}{3}$ (heavy-tailed, mean $= 2$, variance $= 76$)

7. `Bimodal-2`: (mean $= 2$, variance $= 99$)

$$X = \begin{cases} 1 & w.p.\ 0.99 \\ 101 & w.p.\ 0.01 \end{cases}$$

2 Simulation Methodology

We simulate the queueing model shown in Figure 1 via a queueing simulator which we wrote in the C++ programming language.

Our first aim is to show that the first two moments of queue length at each server are nearly insensitive to the variability of the job size distribution. To demonstrate this claim, in Section 3, we consider two values of load, $\rho = 0.5$ (light load) and $\rho = 0.9$ (moderately heavy load) and a range of K values including $K = 2, 4, 8,$ and 16, and of course a range of job size distributions. For each load, each value of K, and each job size distribution, we run the simulation 50 times, where each run consists of $K \times 10^7$ departures (10^7 departures per server, on average).

These same simulations are also used to study the near-insensitivity of the conditional arrival rates, the $\lambda(n)$. See Section 4.

In Section 5, we look at the problem of obtaining empirical curves for $\lambda(n)$ ($n = 0, 1, 2$) under an exponential job-size distribution. Obtaining accurate empirical curve fits requires running our simulator on a much wider range of K and ρ values than in previous experiments. Each run consists of $2K \times 10^7$ departures, but due to the long simulation time, we only manage 2 runs per (K, ρ) combination. This appears sufficient because the results of the runs are consistent.

Confidence intervals throughout are derived using the following method. Let $X_1, X_2, \ldots, X_m$ denote the average queue length observed during m runs of the simulator. Let $\bar{X}$ denote the sample mean:

$$\bar{X} = \frac{1}{m} \sum_{i=1}^{m} X_m \tag{1}$$

Note that $\bar{X}$ is a random variable since each of X_i is an i.i.d. random variable. To find the 95% confidence interval for $\mathbf{E}[N]$, we make the assumption that $\bar{X}$ is distributed according to a normal distribution with mean $\mathbf{E}[N]$ and variance σ^2, where σ^2 is estimated by:

$$\sigma^2 = \frac{1}{m(m-1)} \sum_{i=1}^{m} (X_m - (\bar{X}))^2 \tag{2}$$

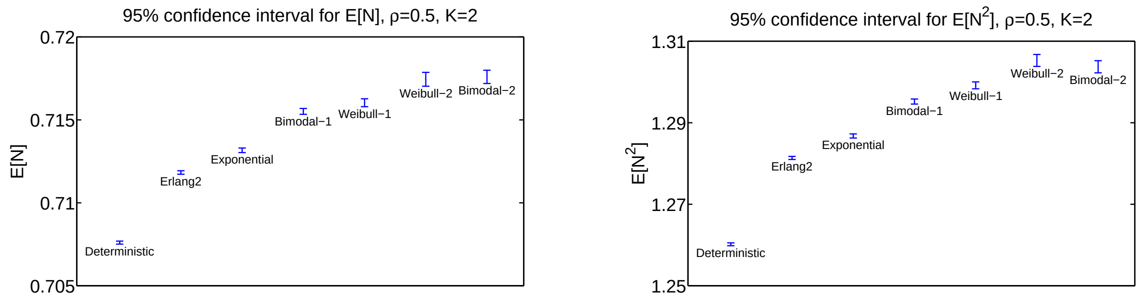
Finally, the 95% confidence interval for $\mathbf{E}[N]$ under the assumptions stated above is given by $[\bar{X} - 1.96\sigma, \bar{X} + 1.96\sigma]$.

3 Near-Insensitivity of $\mathbf{E}[N]$ and $\mathbf{E}\left[N^2\right]$ under general job size distributions

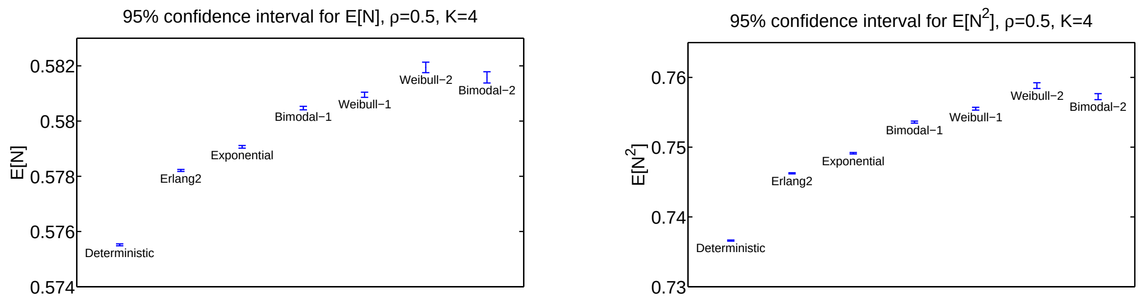
We simulate the $M/G/K/JSQ/PS$ model to measure $\mathbf{E}[N]$ and $\mathbf{E}[N^2]$ for job-size distributions listed in Section 1, where N denotes the number of jobs at the first queue. Figures 2 and 3 show the empirical mean and 95% confidence intervals for the first two moments of N, for $\rho = 0.5$ and $\rho = 0.9$ respectively. The numerical values are shown in Table 1.

We observe that $\mathbf{E}[N]$ for the distributions considered is within 1% of $\mathbf{E}[N]$ for the exponential job-size distribution, when $\rho = 0.5$, and within 1.5%, when $\rho = 0.9$. For $\mathbf{E}[N^2]$, the effect of variability is less than 2% for $\rho = 0.5$ and less than 3% for $\rho = 0.9$. We conclude that the first two moments of N seem to be *nearly insensitive* to the variability of job-size distribution.

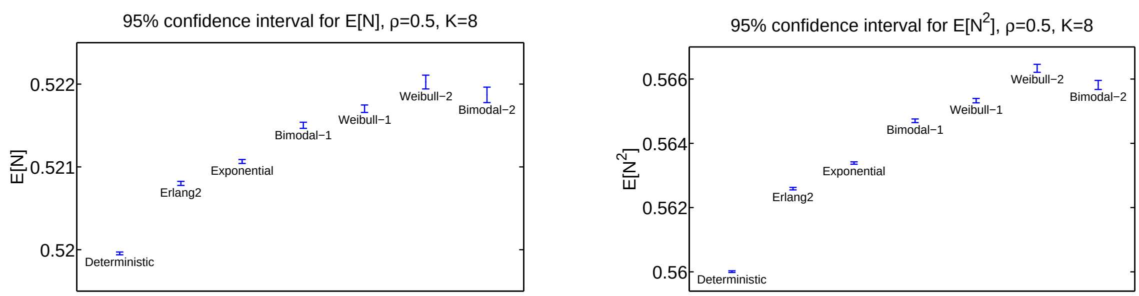
In addition, Table 2 shows the empirical estimates of $\mathbf{E}[N]$ under more values of K and ρ, for the case of an exponential job-size distribution with mean 1.



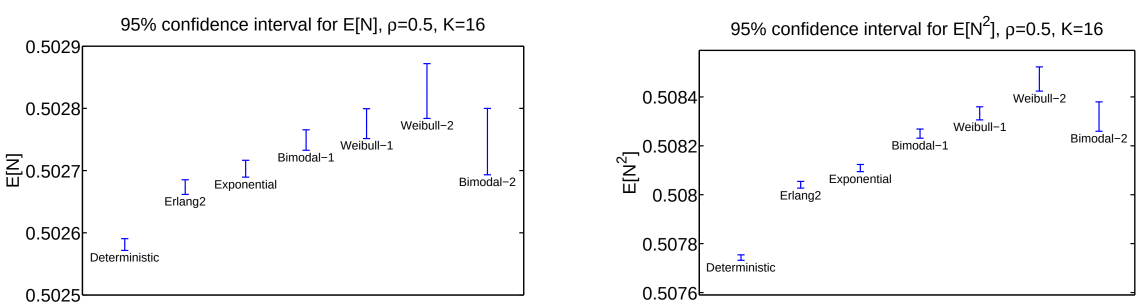
(a) $K = 2$



(b) $K = 4$

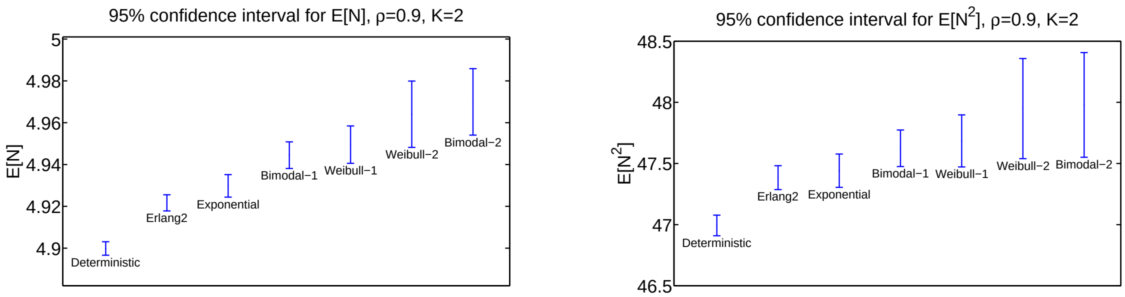


(c) $K = 8$

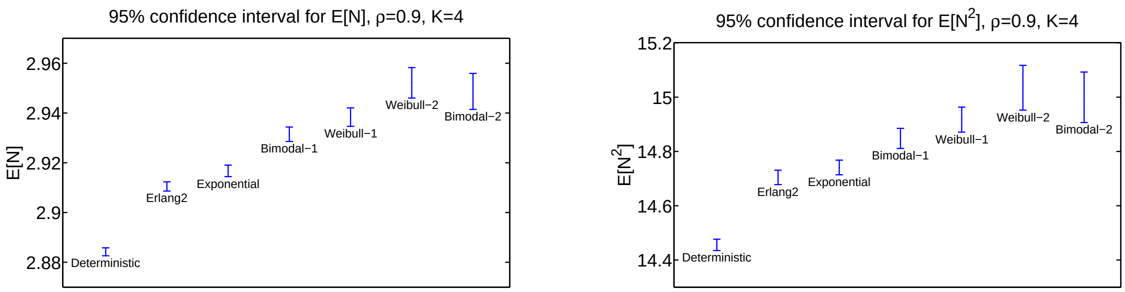


(d) $K = 16$

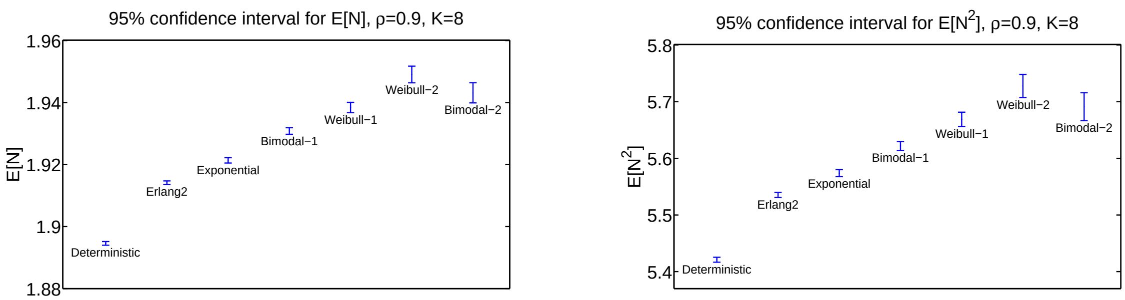
Figure 2: 95% Confidence intervals for mean (left column) and second moment (right column) of queue length of first queue in the $M/G/K/JSQ/PS$ model with $\rho = 0.5$ for the job-size distributions listed in Section 1, based on simulations.



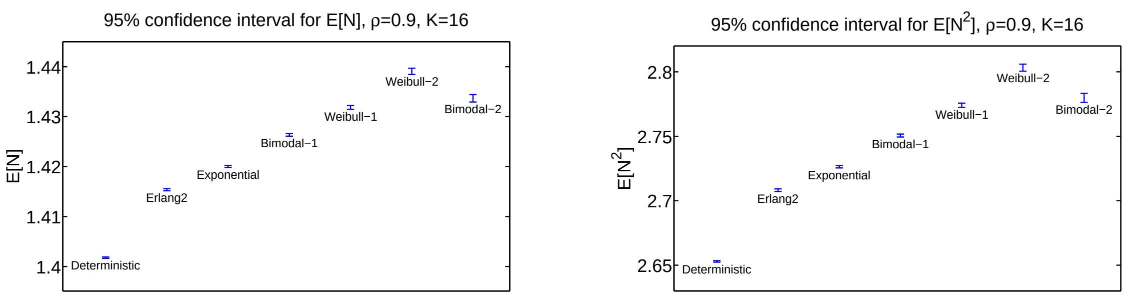
(a) $K = 2$



(b) $K = 4$



(c) $K = 8$



(d) $K = 16$

Figure 3: 95% Confidence intervals for mean (left column) and second moment (right column) of queue length of first queue in the $M/G/K/JSQ/PS$ model with $\rho = 0.9$ for the job-size distributions listed in Section 1, based on simulations.

		$\rho = 0.5$		$\rho = 0.9$	
		E$[N]$	**E**$[N^2]$	**E**$[N]$	**E**$[N^2]$
$K = 2$	`Deterministic`	0.7076	1.2602	4.8999	46.9934
	`Erlang2`	0.7118	1.2814	4.9217	47.3844
	`Exponential`	0.7132	1.2868	4.9298	47.4411
	`Bimodal-1`	0.7155	1.2952	4.9445	47.6245
	`Weibull-1`	0.7160	1.2992	4.9495	47.6847
	`Weibull-2`	0.7174	1.3053	4.9640	47.9490
	`Bimodal-2`	0.7176	1.3037	4.9700	47.9787
$K = 4$	`Deterministic`	0.5755	0.7366	2.8842	14.4559
	`Erlang2`	0.5782	0.7462	2.9105	14.7045
	`Exponential`	0.5791	0.7491	2.9168	14.7409
	`Bimodal-1`	0.5805	0.7536	2.9315	14.8479
	`Weibull-1`	0.5810	0.7555	2.9384	14.9171
	`Weibull-2`	0.5819	0.7588	2.9521	15.0346
	`Bimodal-2`	0.5816	0.7572	2.9487	14.9995
$K = 8$	`Deterministic`	0.5200	0.5600	1.8946	5.4210
	`Erlang2`	0.5208	0.5626	1.9142	5.5354
	`Exponential`	0.5211	0.5634	1.9214	5.5738
	`Bimodal-1`	0.5215	0.5647	1.9308	5.6217
	`Weibull-1`	0.5217	0.5653	1.9384	5.6689
	`Weibull-2`	0.5220	0.5663	1.9490	5.7277
	`Bimodal-2`	0.5219	0.5658	1.9431	5.6912
$K = 16$	`Deterministic`	0.5026	0.5077	1.4018	2.6529
	`Erlang2`	0.5027	0.5080	1.4154	2.7082
	`Exponential`	0.5027	0.5081	1.4200	2.7264
	`Bimodal-1`	0.5027	0.5082	1.4263	2.7506
	`Weibull-1`	0.5028	0.5083	1.4318	2.7740
	`Weibull-2`	0.5028	0.5085	1.4391	2.8032
	`Bimodal-2`	0.5027	0.5083	1.4337	2.7798

Table 1: First two moments of N for $K = 2, 4, 8$ and 16 servers, $\rho = 0.5$ and 0.9 for the job-size distributions listed in Section 1, based on simulations.

Table 2: $\mathbf{E}[N]$ for various ρ and K under an exponential job-size distribution with mean 1.

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2	0.1018	0.2132	0.3434	0.5044	0.7156	1.0141	1.4831	2.3754	4.9521
3	0.1003	0.2045	0.3196	0.4545	0.6245	0.8545	1.2015	1.8361	3.5955
4	0.1001	0.2017	0.3097	0.4321	0.5808	0.7762	1.0626	1.5692	2.9223
5	0.1000	0.2007	0.3052	0.4202	0.5559	0.7299	0.9796	1.4070	2.5252
6	0.1000	0.2003	0.3030	0.4133	0.5404	0.6998	0.9242	1.3011	2.2578
7	0.1000	0.2001	0.3018	0.4090	0.5298	0.6789	0.8843	1.2251	2.0704
8	0.1000	0.2001	0.3011	0.4063	0.5226	0.6633	0.8549	1.1678	1.9249
9	0.1000	0.2000	0.3007	0.4044	0.5173	0.6516	0.8320	1.1226	1.8124
10	0.1000	0.2000	0.3005	0.4031	0.5132	0.6423	0.8135	1.0866	1.7235
11	0.1000	0.2001	0.3003	0.4021	0.5101	0.6351	0.7988	1.0571	1.6507
12	0.1000	0.2000	0.3001	0.4015	0.5078	0.6292	0.7865	1.0320	1.5886
13	0.1000	0.2000	0.3000	0.4010	0.5061	0.6244	0.7760	1.0111	1.5384
14	0.1000	0.2000	0.3000	0.4007	0.5048	0.6205	0.7672	0.9925	1.4936
15	0.1000	0.2000	0.3000	0.4004	0.5036	0.6172	0.7596	0.9768	1.4558
16	0.1000	0.2000	0.3000	0.4003	0.5028	0.6146	0.7532	0.9628	1.4210
17	0.1000	0.2000	0.3000	0.4002	0.5022	0.6124	0.7475	0.9506	1.3910
18	0.1000	0.2000	0.3000	0.4001	0.5017	0.6106	0.7426	0.9396	1.3641
19	0.1000	0.2000	0.3000	0.4001	0.5014	0.6090	0.7384	0.9299	1.3401
20	0.1000	0.2000	0.3000	0.4001	0.5011	0.6077	0.7347	0.9209	1.3182
22	0.1000	0.2000	0.3000	0.4000	0.5007	0.6056	0.7285	0.9057	1.2799
24	0.1000	0.2000	0.3000	0.4000	0.5004	0.6042	0.7234	0.8932	1.2475
26	0.1000	0.2000	0.3000	0.4001	0.5003	0.6032	0.7194	0.8824	1.2201
28	0.1000	0.2000	0.3000	0.4000	0.5002	0.6024	0.7163	0.8734	1.1970
30	0.1000	0.2000	0.3000	0.4000	0.5001	0.6018	0.7137	0.8658	1.1765
32	0.1000	0.2000	0.3000	0.4000	0.5001	0.6014	0.7115	0.8591	1.1575
34	0.1000	0.2000	0.3000	0.4000	0.5000	0.6010	0.7099	0.8533	1.1411
36	0.1000	0.2000	0.3000	0.4000	0.5001	0.6008	0.7083	0.8483	1.1264
38	0.1000	0.2000	0.3000	0.4000	0.5000	0.6006	0.7070	0.8438	1.1132
40	0.1000	0.2000	0.3000	0.4000	0.5000	0.6005	0.7060	0.8397	1.1011
44	0.1000	0.2000	0.3000	0.4000	0.5000	0.6003	0.7045	0.8331	1.0806
48	0.1000	0.2000	0.3000	0.4000	0.5000	0.6001	0.7033	0.8278	1.0629
52	0.1000	0.2000	0.3000	0.4000	0.5000	0.6001	0.7025	0.8235	1.0476
56	0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7019	0.8200	1.0347
60	0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7014	0.8171	1.0235
64	0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7011	0.8147	1.0135
68	-	-	-	-	-	-	-	0.8126	1.0047
72	-	-	-	-	-	-	-	0.8109	0.9969
76	-	-	-	-	-	-	-	0.8095	0.9898
80	-	-	-	-	-	-	-	0.8082	0.9837
84	-	-	-	-	-	-	-	0.8072	0.9780
88	-	-	-	-	-	-	-	0.8063	0.9727
92	-	-	-	-	-	-	-	0.8055	0.9682
96	-	-	-	-	-	-	-	0.8048	0.9638
100	-	-	-	-	-	-	-	0.8043	0.9600
104	-	-	-	-	-	-	-	0.8038	0.9564
108	-	-	-	-	-	-	-	0.8034	0.9532

112	-	-	-	-	-	-	-	0.8029	0.9502
116	-	-	-	-	-	-	-	0.8026	0.9473
120	-	-	-	-	-	-	-	0.8023	0.9447
124	-	-	-	-	-	-	-	0.8021	0.9424
128	-	-	-	-	-	-	-	0.8018	0.9401
132	-	-	-	-	-	-	-	0.8016	0.9380
136	-	-	-	-	-	-	-	0.8015	0.9361
140	-	-	-	-	-	-	-	0.8013	0.9343
144	-	-	-	-	-	-	-	0.8012	0.9327
148	-	-	-	-	-	-	-	0.8010	0.9311

4 Near-Insensitivity of conditional arrival rates $\lambda(n)$ under general job size distributions

Tables 3 and 4 show the empirical estimates of the first seven conditional arrival rates normalized by the mean job size ($\lambda(n)/\mu : n = 0, 1, \dots, 6$) for $\rho = 0.5$ and $\rho = 0.9$, respectively, under a range of job size distributions and $K = 2, 4, 8$ and 16. We observe:

1. The conditional arrival rates $\lambda(n)$ for $n \geq 1$ are appreciably smaller than $\lambda(0)$ and this drop increases as the number of servers increases.

2. The conditional arrival rates are nearly insensitive to the job-size distribution.

		$\lambda(0)/\mu$	$\lambda(1)/\mu$	$\lambda(2)/\mu$	$\lambda(3)/\mu$	$\lambda(4)/\mu$	$\lambda(5)/\mu$	$\lambda(6)/\mu$	ρ^K
	`Deterministic`	0.6866	0.3392	0.2574	0.2525	0.2583	0.2626	0.2648	
	`Erlang2`	0.6845	0.3407	0.2636	0.2524	0.2513	0.2521	0.2536	
	`Exponential`	0.6840	0.3412	0.2648	0.2523	0.2504	0.2495	0.2500	
$K = 2$	`Bimodal-1`	0.6837	0.3416	0.2659	0.2524	0.2512	0.2522	0.2539	0.2500
	`Weibull-1`	0.6832	0.3419	0.2670	0.2533	0.2506	0.2497	0.2499	
	`Weibull-2`	0.6828	0.3421	0.2681	0.2543	0.2516	0.2511	0.2525	
	`Bimodal-2`	0.6834	0.3419	0.2669	0.2535	0.2525	0.2547	0.2553	
	`Deterministic`	0.8563	0.1565	0.0662	0.0648	0.0694	0.0724	0.0620	
	`Erlang2`	0.8537	0.1593	0.0704	0.0638	0.0638	0.0643	0.0644	
	`Exponential`	0.8529	0.1602	0.0715	0.0636	0.0627	0.0599	0.0468	
$K = 4$	`Bimodal-1`	0.8523	0.1609	0.0731	0.0633	0.0639	0.0634	0.0507	0.0625
	`Weibull-1`	0.8515	0.1618	0.0740	0.0645	0.0624	0.0615	0.0515	
	`Weibull-2`	0.8508	0.1625	0.0755	0.0651	0.0636	0.0630	0.0516	
	`Bimodal-2`	0.8517	0.1615	0.0744	0.0643	0.0641	0.0625	0.0538	
	`Deterministic`	0.9598	0.0417	0.0041	0.0040	0.0045	0.0000	0.0000	
	`Erlang2`	0.9585	0.0431	0.0049	0.0036	0.0023	0.0000	0.0000	
	`Exponential`	0.9581	0.0435	0.0050	0.0040	0.0020	0.0000	0.0000	
$K = 8$	`Bimodal-1`	0.9577	0.0439	0.0054	0.0036	0.0009	0.0000	0.0000	0.0039
	`Weibull-1`	0.9573	0.0444	0.0055	0.0042	0.0004	0.0000	0.0000	
	`Weibull-2`	0.9568	0.0449	0.0058	0.0043	0.0022	0.0000	0.0000	
	`Bimodal-2`	0.9574	0.0443	0.0058	0.0042	0.0045	0.0000	0.0000	
	`Deterministic`	0.9948	0.0052	0.0000	0.0000	0.0000	0.0000	0.0000	
	`Erlang2`	0.9946	0.0054	0.0000	0.0000	0.0000	0.0000	0.0000	
	`Exponential`	0.9946	0.0054	0.0000	0.0000	0.0000	0.0000	0.0000	
$K = 16$	`Bimodal-1`	0.9945	0.0055	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	`Weibull-1`	0.9945	0.0056	0.0000	0.0000	0.0000	0.0000	0.0000	
	`Weibull-2`	0.9944	0.0056	0.0000	0.0000	0.0000	0.0000	0.0000	
	`Bimodal-2`	0.9944	0.0055	0.0000	0.0000	0.0000	0.0000	0.0000	

Table 3: First few conditional arrival rates normalized by the mean job size, $\lambda(n)/\mu$, for $K = 2, 4, 8$ and 16 servers, $\rho = 0.5$, for the job-size distributions listed in Section 1.

		$\lambda(0)/\mu$	$\lambda(1)/\mu$	$\lambda(2)/\mu$	$\lambda(3)/\mu$	$\lambda(4)/\mu$	$\lambda(5)/\mu$	$\lambda(6)/\mu$	ρ^K
$K = 2$	`Deterministic`	1.4297	0.9729	0.8334	0.8081	0.8074	0.8100	0.8122	0.8100
	`Erlang2`	1.4214	0.9722	0.8431	0.8167	0.8116	0.8108	0.8105	
	`Exponential`	1.4196	0.9727	0.8452	0.8175	0.8117	0.8105	0.8101	
	`Bimodal-1`	1.4191	0.9728	0.8455	0.8168	0.8103	0.8097	0.8099	
	`Weibull-1`	1.4173	0.9743	0.8485	0.8189	0.8118	0.8099	0.8094	
	`Weibull-2`	1.4164	0.9754	0.8501	0.8194	0.8120	0.8098	0.8091	
	`Bimodal-2`	1.4187	0.9734	0.8467	0.8175	0.8107	0.8096	0.8099	
$K = 4$	`Deterministic`	2.2374	0.9864	0.6932	0.6578	0.6608	0.6650	0.6682	0.6561
	`Erlang2`	2.2170	0.9939	0.7070	0.6632	0.6579	0.6573	0.6578	
	`Exponential`	2.2116	0.9955	0.7104	0.6640	0.6571	0.6561	0.6558	
	`Bimodal-1`	2.2074	0.9954	0.7139	0.6644	0.6573	0.6574	0.6590	
	`Weibull-1`	2.2017	0.9991	0.7173	0.6659	0.6565	0.6549	0.6545	
	`Weibull-2`	2.1972	1.0011	0.7207	0.6670	0.6572	0.6548	0.6535	
	`Bimodal-2`	2.2029	0.9962	0.7173	0.6662	0.6577	0.6576	0.6591	
$K = 8$	`Deterministic`	3.4100	0.8849	0.4720	0.4383	0.4423	0.4464	0.4487	0.4305
	`Erlang2`	3.3753	0.8999	0.4807	0.4352	0.4320	0.4317	0.4318	
	`Exponential`	3.3630	0.9040	0.4842	0.4352	0.4312	0.4311	0.4308	
	`Bimodal-1`	3.3516	0.9061	0.4895	0.4362	0.4334	0.4349	0.4365	
	`Weibull-1`	3.3385	0.9117	0.4919	0.4368	0.4307	0.4295	0.4294	
	`Weibull-2`	3.3257	0.9157	0.4967	0.4382	0.4302	0.4282	0.4279	
	`Bimodal-2`	3.3394	0.9086	0.4949	0.4387	0.4346	0.4363	0.4390	
$K = 16$	`Deterministic`	4.9526	0.6407	0.2165	0.1923	0.1947	0.1966	0.2004	0.1853
	`Erlang2`	4.9027	0.6571	0.2195	0.1876	0.1862	0.1863	0.1846	
	`Exponential`	4.8823	0.6627	0.2215	0.1870	0.1852	0.1861	0.1869	
	`Bimodal-1`	4.8628	0.6679	0.2251	0.1881	0.1882	0.1899	0.1907	
	`Weibull-1`	4.8379	0.6746	0.2272	0.1878	0.1850	0.1846	0.1851	
	`Weibull-2`	4.8138	0.6813	0.2308	0.1885	0.1850	0.1845	0.1804	
	`Bimodal-2`	4.8409	0.6737	0.2292	0.1890	0.1895	0.1897	0.1817	

Table 4: First few conditional arrival rates normalized by the mean job size, $\lambda(n)/\mu$, for $K = 2, 4, 8$ and 16 servers, $\rho = 0.9$, for the job-size distributions listed in Section 1.

5 Obtaining empirical curves for $\lambda(n)$

The analysis in [1] hinges on being able to approximate $\lambda(n)$ for all values of n, in the case where the job size distribution is exponential.

We claim in that paper that $\lambda(n)$ is approximately constant when $n \geq 3$. Specifically we claim that:

$$\lambda(n) \approx \mu\rho^K \text{ for all } n \geq 3 , \tag{3}$$

for $\rho \leq 0.95$. Simulations of the $M/M/K/JSQ/FCFS$ model show this approximation to be consistently within 2% of the actual values (provided that ρ is not too extreme, i.e., for $0.3 \leq \rho \leq 0.95$). Some evidence for this was presented in [1]. Further, the last column of Tables 3 and 4 show that ρ^K is a very good approximation for $\lambda(n)/\mu$, $n \geq 3$, particularly when the job size distribution is exponential.

The fact that $\lambda(n)$ is relatively constant for $n \geq 3$ reduces the problem of approximating all $\lambda(n)$ to determining just $\lambda(0)$, $\lambda(1)$, and $\lambda(2)$. The second of these, $\lambda(1)$ is derived analytically in [1].

The approximations for $\lambda(2)$ and $\lambda(0)$ were obtained empirically using MATLAB's curve fitting toolbox (version 1.1.5), which uses a trust-region method for a nonlinear least-squares fit.

Tables 5, 6 and 7 show $\lambda(0)$, $\lambda(1)$ and $\lambda(2)$ respectively for many values of load ρ in the range $\rho = 0.1$ to $\rho = 0.9$ and number of servers K from 2 to 148, under an exponential job size distribution with mean 1. We then use the data in these tables to find a functional closed-form approximation for $\lambda(0)$ and $\lambda(2)$.

For each value of load, ρ, we approximate $\lambda(2)$ as a function of K by a simple exponential function of the form

$$\lambda(2) \approx \mu\left(u_\rho v_\rho^K\right) \tag{4}$$

Empirical fit yields the following functions of ρ:

$$u_\rho = c_3\rho^3 + c_2\rho^2 + c_1\rho + c_0 \text{ and } v_\rho = c_2'\rho^2 + c_1'\rho + c_0' ,$$

where $c_3 = -0.29$, $c_2 = 0.8822$, $c_1 = -0.5349$, and $c_0 = 1.0112$, while $c_2' = -0.1864$, $c_1' = 1.195$, and $c_0' = -0.016$.

For $\lambda(0)$, we use a function with two exponential terms, namely,

$$\lambda(0) \approx \mu\left(a_\rho - b_\rho c_\rho^K - d_\rho e_\rho^K\right) \tag{5}$$

where $c_\rho, e_\rho < 1$. The constant a_ρ in (5) is clearly the limit of $\frac{\lambda(0)}{\mu}$ as $K \to \infty$. Lemma 5.1 gives the value of this limit. From Table 5, one can see that $\lambda(0)$ converges for high K (depending on ρ), in accordance with Lemma 5.1 below.

Lemma 5.1

$$\lim_{K\to\infty} \frac{\lambda(0)}{\mu} = \frac{\rho}{1-\rho} \tag{6}$$

Proof: For any value of $\rho < 1$, as the number of servers becomes large enough, any arrival will find at least one server idle with high probability. Therefore, $\lambda(i) \approx 0$ for $i \geq 1$. Equating the expressions for time average arrival rates into any queue,

$$(1-\rho)\lambda(0) = \mu\rho \quad \text{or} \quad \frac{\lambda(0)}{\mu} = \frac{\rho}{1-\rho}\ .$$

■

The remaining functions b_ρ, c_ρ, d_ρ, and e_ρ are determined empirically for $0.3 \leq \rho \leq 0.95$; we do not have accurate enough simulations outside this range. The final functions are

$$\begin{aligned}
b_\rho &= \frac{-0.0263\rho^2 + 0.0054\rho + 0.1155}{\rho^2 - 1.939\rho + 0.9534} \\
c_\rho &= -6.2973\rho^4 + 14.3382\rho^3 - 12.3532\rho^2 + 6.2557\rho - 1.005 \\
d_\rho &= \frac{-226.1839\rho^2 + 342.3814\rho + 10.2851}{\rho^3 - 146.2751\rho^2 - 481.1256\rho + 599.9166} \\
e_\rho &= 0.4462\rho^3 - 1.8317\rho^2 + 2.4376\rho - 0.0512
\end{aligned}$$

Table 5: $\lambda(0)$ for various ρ and K under an exponential job size distribution with mean 1.

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2	0.1092	0.2342	0.3726	0.5229	0.6841	0.8555	1.0371	1.2296	1.4324
3	0.1107	0.2444	0.4018	0.5827	0.7877	1.0170	1.2713	1.5520	1.8583
4	0.1110	0.2479	0.4149	0.6152	0.8519	1.1283	1.4476	1.8126	2.2255
5	0.1111	0.2491	0.4212	0.6339	0.8937	1.2081	1.5842	2.0298	2.5538
6	0.1111	0.2496	0.4244	0.6449	0.9219	1.2668	1.6936	2.2160	2.8494
7	0.1111	0.2498	0.4262	0.6518	0.9411	1.3119	1.7829	2.3776	3.1214
8	0.1111	0.2499	0.4271	0.6565	0.9556	1.3466	1.8568	2.5200	3.3719
9	0.1111	0.2499	0.4276	0.6596	0.9662	1.3745	1.9184	2.6448	3.6059
10	0.1111	0.2500	0.4281	0.6618	0.9740	1.3959	1.9705	2.7553	3.8238
11	0.1111	0.2500	0.4283	0.6633	0.9800	1.4142	2.0157	2.8554	4.0267
12	0.1111	0.2500	0.4284	0.6645	0.9847	1.4287	2.0534	2.9444	4.2201
13	0.1111	0.2500	0.4284	0.6652	0.9883	1.4404	2.0858	3.0238	4.4008
14	0.1111	0.2500	0.4285	0.6656	0.9911	1.4500	2.1129	3.0957	4.5734
15	0.1111	0.2500	0.4286	0.6659	0.9930	1.4577	2.1375	3.1610	4.7346
16	0.1111	0.2500	0.4286	0.6662	0.9944	1.4642	2.1587	3.2197	4.8881
17	0.1111	0.2500	0.4285	0.6663	0.9955	1.4698	2.1766	3.2740	5.0330
18	0.1111	0.2500	0.4285	0.6664	0.9966	1.4743	2.1925	3.3217	5.1714
19	0.1111	0.2500	0.4286	0.6665	0.9974	1.4780	2.2068	3.3668	5.3031
20	0.1111	0.2500	0.4285	0.6665	0.9979	1.4810	2.2194	3.4062	5.4252
22	0.1111	0.2500	0.4286	0.6666	0.9987	1.4857	2.2406	3.4775	5.6600
24	0.1111	0.2500	0.4286	0.6666	0.9991	1.4894	2.2561	3.5390	5.8715
26	0.1111	0.2500	0.4286	0.6668	0.9996	1.4922	2.2685	3.5895	6.0596
28	0.1111	0.2500	0.4285	0.6666	0.9997	1.4942	2.2798	3.6335	6.2384
30	0.1111	0.2500	0.4285	0.6666	0.9998	1.4955	2.2882	3.6727	6.4010
32	0.1111	0.2500	0.4286	0.6667	0.9998	1.4966	2.2953	3.7056	6.5464
34	0.1111	0.2500	0.4286	0.6667	1.0000	1.4974	2.3020	3.7341	6.6833
36	0.1111	0.2500	0.4286	0.6666	1.0001	1.4980	2.3060	3.7602	6.8078
38	0.1111	0.2500	0.4286	0.6667	0.9999	1.4983	2.3091	3.7830	6.9242
40	0.1111	0.2500	0.4286	0.6666	1.0000	1.4989	2.3132	3.8013	7.0295
44	0.1111	0.2500	0.4286	0.6667	0.9999	1.4993	2.3188	3.8346	7.2248
48	0.1111	0.2500	0.4286	0.6667	1.0000	1.4994	2.3226	3.8611	7.3897
52	0.1111	0.2500	0.4286	0.6666	1.0000	1.4997	2.3253	3.8830	7.5331
56	0.1111	0.2500	0.4286	0.6666	1.0000	1.4996	2.3275	3.9006	7.6578
60	0.1111	0.2500	0.4286	0.6666	1.0001	1.4998	2.3286	3.9147	7.7720
64	0.1111	0.2500	0.4286	0.6667	1.0000	1.5000	2.3298	3.9266	7.8694
68	-	-	-	-	-	-	-	3.9370	7.9576
72	-	-	-	-	-	-	-	3.9451	8.0335
76	-	-	-	-	-	-	-	3.9530	8.1013
80	-	-	-	-	-	-	-	3.9581	8.1677
84	-	-	-	-	-	-	-	3.9632	8.2250
88	-	-	-	-	-	-	-	3.9685	8.2732

92	-	-	-	-	-	-	-	3.9725	8.3203
96	-	-	-	-	-	-	-	3.9756	8.3620
100	-	-	-	-	-	-	-	3.9786	8.3990
104	-	-	-	-	-	-	-	3.9810	8.4374
108	-	-	-	-	-	-	-	3.9839	8.4725
112	-	-	-	-	-	-	-	3.9849	8.5004
116	-	-	-	-	-	-	-	3.9869	8.5266
120	-	-	-	-	-	-	-	3.9884	8.5503
124	-	-	-	-	-	-	-	3.9896	8.5788
128	-	-	-	-	-	-	-	3.9912	8.5984
132	-	-	-	-	-	-	-	3.9914	8.6180
136	-	-	-	-	-	-	-	3.9928	8.6392
140	-	-	-	-	-	-	-	3.9936	8.6563
144	-	-	-	-	-	-	-	3.9947	8.6743
148	-	-	-	-	-	-	-	3.9950	8.6901

Table 6: $\lambda(1)$ for various ρ and K under an exponential job size distribution with mean 1.

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2	0.0177	0.0646	0.1356	0.2279	0.3396	0.4702	0.6190	0.7865	0.9717
3	0.0035	0.0227	0.0646	0.1326	0.2302	0.3605	0.5276	0.7363	0.9894
4	0.0007	0.0086	0.0326	0.0806	0.1597	0.2785	0.4473	0.6787	0.9865
5	0.0002	0.0034	0.0175	0.0510	0.1137	0.2179	0.3801	0.6219	0.9730
6	0.0000	0.0015	0.0099	0.0336	0.0830	0.1725	0.3245	0.5688	0.9526
7	0.0000	0.0007	0.0058	0.0227	0.0616	0.1381	0.2771	0.5193	0.9273
8	0.0000	0.0004	0.0036	0.0156	0.0462	0.1112	0.2377	0.4732	0.8984
9	0.0000	0.0002	0.0022	0.0108	0.0350	0.0902	0.2047	0.4313	0.8679
10	0.0000	0.0001	0.0013	0.0074	0.0266	0.0737	0.1768	0.3931	0.8371
11	0.0000	0.0000	0.0008	0.0051	0.0203	0.0606	0.1536	0.3591	0.8053
12	0.0000	0.0000	0.0005	0.0035	0.0156	0.0500	0.1340	0.3279	0.7744
13	0.0000	0.0000	0.0003	0.0024	0.0120	0.0415	0.1172	0.3006	0.7446
14	0.0000	0.0000	0.0001	0.0016	0.0093	0.0346	0.1031	0.2753	0.7160
15	0.0000	0.0000	0.0001	0.0011	0.0071	0.0291	0.0911	0.2532	0.6875
16	0.0000	0.0000	0.0000	0.0008	0.0056	0.0245	0.0806	0.2331	0.6600
17	0.0000	0.0000	0.0000	0.0005	0.0043	0.0207	0.0717	0.2151	0.6340
18	0.0000	0.0000	0.0000	0.0004	0.0034	0.0176	0.0640	0.1990	0.6088
19	0.0000	0.0000	0.0000	0.0003	0.0027	0.0150	0.0573	0.1843	0.5844
20	0.0000	0.0000	0.0000	0.0002	0.0021	0.0128	0.0515	0.1710	0.5616
22	0.0000	0.0000	0.0000	0.0001	0.0013	0.0094	0.0419	0.1482	0.5189
24	0.0000	0.0000	0.0000	0.0000	0.0008	0.0070	0.0343	0.1293	0.4797
26	0.0000	0.0000	0.0000	0.0000	0.0005	0.0052	0.0283	0.1135	0.4442
28	0.0000	0.0000	0.0000	0.0000	0.0003	0.0040	0.0236	0.1001	0.4127
30	0.0000	0.0000	0.0000	0.0000	0.0002	0.0030	0.0197	0.0889	0.3839
32	0.0000	0.0000	0.0000	0.0000	0.0001	0.0023	0.0166	0.0793	0.3571
34	0.0000	0.0000	0.0000	0.0000	0.0001	0.0017	0.0141	0.0711	0.3332
36	0.0000	0.0000	0.0000	0.0000	0.0001	0.0013	0.0119	0.0639	0.3112
38	0.0000	0.0000	0.0000	0.0000	0.0000	0.0010	0.0101	0.0576	0.2915
40	0.0000	0.0000	0.0000	0.0000	0.0000	0.0008	0.0086	0.0522	0.2735
44	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0064	0.0431	0.2423
48	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0047	0.0359	0.2158
52	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0035	0.0302	0.1931
56	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0027	0.0255	0.1741
60	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0020	0.0217	0.1578
64	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0015	0.0186	0.1436
68	-	-	-	-	-	-	-	0.0160	0.1312
72	-	-	-	-	-	-	-	0.0138	0.1204
76	-	-	-	-	-	-	-	0.0120	0.1106
80	-	-	-	-	-	-	-	0.0104	0.1023
84	-	-	-	-	-	-	-	0.0091	0.0948
88	-	-	-	-	-	-	-	0.0079	0.0878

92	-	-	-	-	-	-	-	0.0069	0.0819
96	-	-	-	-	-	-	-	0.0061	0.0763
100	-	-	-	-	-	-	-	0.0054	0.0713
104	-	-	-	-	-	-	-	0.0047	0.0668
108	-	-	-	-	-	-	-	0.0042	0.0628
112	-	-	-	-	-	-	-	0.0037	0.0590
116	-	-	-	-	-	-	-	0.0033	0.0554
120	-	-	-	-	-	-	-	0.0029	0.0523
124	-	-	-	-	-	-	-	0.0026	0.0494
128	-	-	-	-	-	-	-	0.0023	0.0467
132	-	-	-	-	-	-	-	0.0020	0.0441
136	-	-	-	-	-	-	-	0.0018	0.0418
140	-	-	-	-	-	-	-	0.0016	0.0396
144	-	-	-	-	-	-	-	0.0014	0.0376
148	-	-	-	-	-	-	-	0.0013	0.0357

Table 7: $\lambda(2)$ for various ρ and K under an exponential job size distribution with mean 1.

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2	0.0104	0.0423	0.0962	0.1705	0.2640	0.3767	0.5086	0.6612	0.8348
3	0.0011	0.0098	0.0332	0.0767	0.1447	0.2416	0.3735	0.5472	0.7692
4	0.0003	0.0030	0.0136	0.0375	0.0820	0.1571	0.2754	0.4526	0.7070
5	0.0001	0.0014	0.0058	0.0179	0.0454	0.1002	0.2010	0.3714	0.6455
6	0.0000	0.0005	0.0021	0.0079	0.0241	0.0629	0.1449	0.3031	0.5885
7	0.0000	0.0001	0.0006	0.0033	0.0124	0.0388	0.1037	0.2469	0.5345
8	0.0000	0.0000	0.0002	0.0013	0.0063	0.0239	0.0746	0.2010	0.4849
9	0.0000	0.0000	0.0001	0.0004	0.0032	0.0146	0.0536	0.1632	0.4394
10	0.0000	0.0000	0.0000	0.0002	0.0015	0.0092	0.0383	0.1333	0.3991
11	0.0000	0.0000	0.0000	0.0001	0.0008	0.0055	0.0273	0.1083	0.3618
12	0.0000	0.0000	0.0000	0.0000	0.0004	0.0035	0.0199	0.0884	0.3278
13	0.0000	0.0000	0.0000	0.0000	0.0002	0.0021	0.0142	0.0720	0.2979
14	0.0000	0.0000	0.0000	0.0000	0.0001	0.0013	0.0102	0.0588	0.2700
15	0.0000	0.0000	0.0000	0.0000	0.0001	0.0008	0.0072	0.0477	0.2456
16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0052	0.0390	0.2226
17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0037	0.0318	0.2019
18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0027	0.0258	0.1834
19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0019	0.0210	0.1671
20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0014	0.0172	0.1515
22	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0007	0.0115	0.1248
24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0075	0.1026
26	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0050	0.0850
28	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0033	0.0704
30	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0022	0.0582
32	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0014	0.0477
34	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0010	0.0394
36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0006	0.0327
38	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0269
40	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0222
44	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0153
48	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0104
52	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0070
56	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0047
60	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0033
64	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0022
68	-	-	-	-	-	-	-	0.0000	0.0015
72	-	-	-	-	-	-	-	0.0000	0.0010
76	-	-	-	-	-	-	-	0.0000	0.0007
80	-	-	-	-	-	-	-	0.0000	0.0005
84	-	-	-	-	-	-	-	0.0000	0.0003
88	-	-	-	-	-	-	-	0.0000	0.0002

92	-	-	-	-	-	-	-	0.0000	0.0001
96	-	-	-	-	-	-	-	0.0000	0.0001
100	-	-	-	-	-	-	-	0.0000	0.0001
104	-	-	-	-	-	-	-	0.0000	0.0000
108	-	-	-	-	-	-	-	0.0000	0.0000
112	-	-	-	-	-	-	-	0.0000	0.0000
116	-	-	-	-	-	-	-	0.0000	0.0000
120	-	-	-	-	-	-	-	0.0000	0.0000
124	-	-	-	-	-	-	-	0.0000	0.0000
128	-	-	-	-	-	-	-	0.0000	0.0000
132	-	-	-	-	-	-	-	0.0000	0.0000
136	-	-	-	-	-	-	-	0.0000	0.0000
140	-	-	-	-	-	-	-	0.0000	0.0000
144	-	-	-	-	-	-	-	0.0000	0.0000
148	-	-	-	-	-	-	-	0.0000	0.0000

6 Conclusion

This short paper provides simulation results on JSQ/PS server farms. We study the first and second moments of queue length and the conditional arrival rates into the first queue, under a wide range of loads, number of servers, and job size distributions.

7 Acknowledgements

Varun Gupta and Mor Harchol-Balter were supported by NSF Division of Computer and Network System, Grant CNS-0615262. Ward Whitt was supported in part by NSF Grant DMI-0457095.

References

[1] Varun Gupta, Mor Harchol-Balter, Karl Sigman, and Ward Whitt. Analysis of join-the-shortest-queue routing for web server farms. In *PERFORMANCE 2007. IFIP WG 7.3 International Symposium on Computer Modeling, Measurement and Evaluation*, Cologne, Germany, October 2007.