Approximate performance analysis of generalized join the shortest queue routing

Jori Selen
Eindhoven University of Technology, Netherlands
j.selen@tue.nl

Ivo Adan
Eindhoven University of Technology, Netherlands
i.j.b.f.adan@tue.nl

Stella Kapodistria
Eindhoven University of Technology, Netherlands
s.kapodistria@tue.nl

ABSTRACT

In this paper we propose a highly accurate approximate performance analysis of a heterogeneous server system with a processor sharing (PS) service discipline and a general job-size distribution under a *generalized join the shortest queue* (GJSQ) routing protocol. The GJSQ routing protocol is a natural extension of the well-known join the shortest queue (JSQ) routing policy that takes into account the non-identical service rates in addition to the number of jobs at each server. The performance metrics that are of interest here are the equilibrium distribution and the mean and standard deviation of the number of jobs at each server. We show that the latter metrics are near-insensitive to the job-size distribution using simulation experiments. By applying a *single queue approximation* (SQA) we model each server as a single server queue with a state-dependent arrival process, independent of other servers in the system, and derive the distribution of the number of jobs at the server. These state-dependent arrival rates are intended to capture the inherent correlation between servers in the original system and behave in a rather atypical way.

CCS Concepts

•Mathematics of computing → Queueing theory; Markov processes; •General and reference → Performance;

Keywords

Heterogeneous servers; routing policy; approximations

1. INTRODUCTION

1.1 Motivation

This work is motivated by web server farms. Server farms have gained popularity for providing scalable and reliable computing and web services. Most commonly the objective in analyzing such a system lies in the determination of an optimal or near-optimal load balancing routing protocol so as to maximize the performance of the system, see, e.g., [1, 6, 8], where the performance of interest is usually the mean response time for an arbitrary job. In this paper the objective is to report some interesting properties of the arrival flow to each server and suggest an approximation approach for the GJSQ routing protocol. We consider farms with heterogeneous servers, which is motivated by the different hardware and the wide variety of computing capacities regarding processing power and memory access performance seen in practice in server farms [13]. We assume that service requests arrive to the system according to a Poisson process. Upon arrival, a front-end dispatcher routes the request to one of the servers. After the request has been routed to the server, we assume that it cannot balk or jokey. All requests routed to a server are sharing the provided service (think of bandwidth, CPU, or RAM). We assume a PS service discipline at each server since it closely approximates the scheduling policies [7, 10] employed by most commodity operating systems (e.g., Linux CPU time-sharing) and is a popular policy in computing centers (e.g., Cisco Local Director, IBM Network Dispatcher and Microsoft Sharepoint, see [3] for a survey).

In [5] the authors consider a server farm consisting of homogeneous servers, where upon arrival jobs are routed according to the JSQ routing protocol. This protocol in case of homogeneous servers, due to the PS service discipline, is performing near-optimal in terms of the mean response time. However, as indicated by Whitt in [15], the JSQ policy is far from optimal in case of heterogeneous servers. In [2] the authors comment on the performance of various systems under different routing protocols and conclude that the shortest expected delay (SED) routing protocol is near-optimal in terms of mean response time. The SED policy is a policy that routes jobs upon arrival to the queue promising the minimum expected delay (which also includes the processing time). In case of exponential job-size distributions, the GJSQ and SED routing protocols are identical and in case of homogeneous servers GJSQ and JSQ are the same. However, in case of general job-size distributions and heterogeneous servers we assume that the only available information are the service rates and the number of jobs at each server, i.e. we do not keep track of residual processing times.

Due to the complexity and the various challenges that the model at hand presents, we restrict our analysis to the case

VALUETOOLS 2015, December 14-16, Berlin, Germany

DOI 10.4108/eai.14-12-2015.2262695

of two heterogeneous servers with a general job-size distribution under the GJSQ routing protocol. From here onwards we refer to this model as the $M/G(1,s)/2/GJSQ/PS$ system, abbreviated as the GJSQ model, where G is the job-size distribution and 1 and s are the service rates at servers 1 and 2, respectively. The approach described in this paper can be seen as a first stepping stone towards the analysis of heterogeneous server farms with PS servers; a very broad area, full of interesting problems. Moreover, the ideas presented here extend the work of Gupta et al. [5] on the analysis of the JSQ routing for homogeneous web server farms.

1.2 Related work

To the best of our knowledge there is no previous mathematical analysis of the GJSQ system. In [14], Selen et al. derive the joint equilibrium distribution of the number of jobs at each server in the $M/M(1,s)/2/SED/FCFS$ model. They prove that this distribution can be expressed as an infinite series of product forms using the compensation approach. The benefit of that approach is that it produces, by truncating the series expression, numerical results with an a priori set accuracy level. Unfortunately, the compensation approach is not appropriate (in its current setting) for multiple servers, nor for general job-size distributions. Before [14], very little was known regarding the mathematical analysis of the SED policy. In [11], the authors suggest two models that act as upper and lower bounds to the SED system. However, they do not provide closed form expressions for the equilibrium distribution of these two bounding models, but only an algorithmic approach based on matrix analytic methods. Furthermore, in [4, 9], the authors show that the SED routing policy is asymptotically optimal in terms of the mean response time and results in complete resource pooling in the heavy traffic limit. This heavy traffic limit result may be used in a similar manner as in [12]. However, after a few numerical experiments, we concluded that this approximation in our case results in poor estimates and for this reason we did not proceed in this direction. On the contrary, the approach developed by Gupta et al. [5] on approximating the distribution of the number of jobs at each server, as we show in this paper, is appropriate for the GJSQ setting with heterogeneous servers. More concretely, in [5], the authors develop the SQA method that accurately determines the distribution of the number of jobs at each server by modeling each queue as an $M_n/M/1/PS$ system with state-dependent arrival rates. These state-dependent arrival rates are referred to as the *conditional arrival rates* and are constructed in such a way that they capture the inherent correlated behavior of the complete server farm.

1.3 Contributions

We believe that we provide the first approximate analysis of the equilibrium distribution and moments of the number of jobs at each server in the GJSQ system (and by Little's law also the mean response time for an arbitrary job). Moreover, the approximation is highly accurate: we encounter a maximum relative difference between the approximation and simulations of 2.2%. In deriving these approximations, we provide three key contributions:

1. The mean and standard deviation of the number of jobs at each server and the conditional arrival rates are near-insensitive to the job-size distribution. This allows us to study the more tractable model with an exponential job-size distribution.

2. In case of an exponential job-size distribution, the SQA method yields the same equilibrium distribution for the number of jobs at each server as in the original GJSQ model.

3. For the application of the SQA method we present an approach for the derivation of the conditional arrival rates. In particular, we show that the conditional arrival rates, say $\lambda_i(n),\ i=1,2,\ n\in\mathbb{N}_0$, to server 1 satisfy
$$\lambda_1(n)\to\rho^{1+s} \text{ as } n\to\infty, \tag{1}$$
where ρ is the load on the system, see Section 2, and the conditional rates to server 2 for large n oscillate between s different points. Note that the former result is similar to the result obtained in [5] for the case $s=1$, however the latter result is very atypical and is discussed in greater detail in Section 2.3.

1.4 Outline

The rest of the paper is organized as follows. In Section 2 we give a detailed model description and formally define and investigate the time-average and conditional arrival rates. Section 3 is devoted to showing that the performance metrics of interest are near-insensitive to the job-size distribution. We describe the SQA and determine the conditional arrival rates in Section 4. The approximations are evaluated in Section 5. In Section 6 we present some conclusions.

2. MODEL DESCRIPTION

2.1 Heterogeneous servers

We consider a system of two heterogeneous servers and a single dispatcher. The servers employ a PS service discipline and can have different service rates, i.e. server 1 has service rate 1 and server 2 has service rate s. Jobs arrive to the dispatcher according to a Poisson process with rate λ and are routed immediately to one of the servers. Jobs cannot switch servers after being routed. We detail the routing policy in Section 2.2. The size of a job is drawn from a general distribution G. Without loss of generality we assume that the mean job size is 1. Note that, for example, the (residual) processing time of a (residual) size G job that runs on server 2 that is currently serving q_2 jobs is given by Gq_2/s.

In what follows we assume that s is a positive integer number. In the general case $s\in\mathbb{R}_+$ we can bound the corresponding system by two systems with service rates given by the closest two integers to s.

2.2 Routing policy

The routing policy employed by the dispatcher is a state-aware policy, i.e. the dispatcher is aware of the number of jobs at each server just before an arrival instant, q_1 and q_2, and the service rates. The GJSQ routing policy routes an arriving job to the server with the smallest index $(q_i+1)/s_i$, where s_i is the service rate at server i. In case of a tie, the job is randomly routed to one of the servers. These indexes may be interpreted as an estimate of the expected processing time for the arriving job, made by the dispatcher who is unaware of the job-size distribution and the remaining

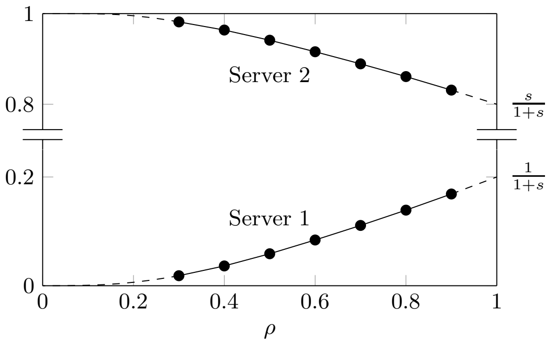


Figure 1: Simulated long-term fraction of jobs routed to server i as a function of the load ρ, where $s = 4$ and the job-size distribution is exponential. Dashed lines represent expected behavior.

processing times of the jobs currently in service, and furthermore ignores future arrivals.

Under this routing policy, we define the load on this system as

$$\rho := \lambda/(1+s). \tag{2}$$

Throughout the rest of this paper we assume that $\rho < 1$.

Although not necessarily optimal, GJSQ routing outperforms JSQ routing when servers are non-identical. GJSQ routing attempts to balance the load on the servers by taking into account the different service rates in addition to the information on the current number of jobs at each server. In Figure 1 we show that the long-term fraction of jobs routed to the two servers is a function of the load ρ. In light traffic GJSQ assigns all jobs to the fast server and in heavy traffic the load is divided proportionally according to the service rates. This is in contrast with JSQ routing, which assigns a long-term fraction of the jobs to server 1 that decreases from $1/2$ to $1/(1+s)$ for increasing load ρ (verified through simulation).

2.3 Arrival rates

We briefly introduce two important concepts related to the (time-average) arrival rates to each server. These concepts will be used throughout the paper.

Definition 1. In the GJSQ model, the *time-average arrival rate* to server i is defined as

$$\overline{\lambda}_i := \lim_{t\to\infty} \frac{A_i(t)}{t}, \tag{3}$$

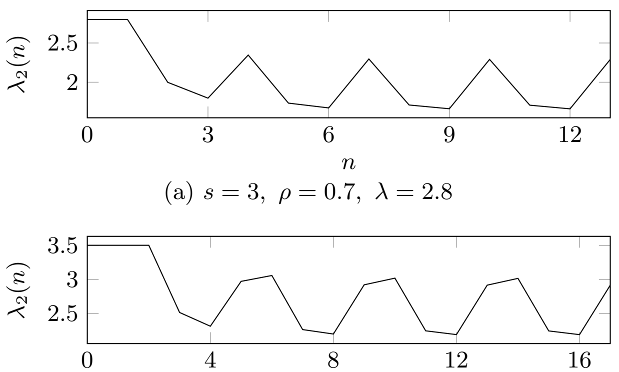
where $A_i(t)$ is the number of arrivals at server i during the time interval $[0, t]$.

Definition 2. In the GJSQ model, the *conditional arrival rate* to server i, given that server i has n jobs, is defined as

$$\lambda_i(n) := \lim_{t\to\infty} \frac{A_{i,n}(t)}{T_{i,n}(t)}, \tag{4}$$

where $A_{i,n}(t)$ is the number of arrivals at server i during the time interval $[0, t]$ that see n jobs at server i on arrival (excluding themselves), and $T_{i,n}(t)$ is the total time spent by server i with n jobs during the time interval $[0, t]$.

The two definitions above are related. Assuming it exists, let $\pi_i(n)$ be the equilibrium probability that there are n jobs at server i, then $\overline{\lambda}_i = \sum_{n=0}^{\infty} \lambda_i(n)\pi_i(n)$.



(a) $s = 3,\ \rho = 0.7,\ \lambda = 2.8$

(b) $s = 4,\ \rho = 0.7,\ \lambda = 3.5$

Figure 2: The conditional arrival rates to server 2 oscillate between s points.

Figure 2 depicts the conditional arrival rates to server 2 for varying s. Intuitively it makes sense that if a server has many jobs, the other server will probably have few jobs and thus it is less likely that the dispatcher routes the job to that server. However, what we see here is a peculiar repeating pattern that has s different points and does not align with this intuition. We see that if server 2 has a multiple of s jobs (or one less), fewer jobs are routed to server 2. This pattern is difficult to explain, but it is definitely related to the probability that server 1 has a lower index than server 2, given that server 2 currently has n jobs. We expect and indeed verify that this probability also follows such a repeating pattern. Additionally, states in server 2 are somewhat similar if they differ by a multiple of s jobs, which can be derived from the equilibrium distribution in [14].

3. NEAR-INSENSITIVITY

In [5] the authors establish a near-sensitivity property in the setting of a homogeneous server farm with JSQ routing. In particular, the first and second moment of the number of jobs at server i, Q_i, and the conditional arrival rates are near-insensitive to the job-size distribution. The near-insensitivity of these two metrics seems related to the insensitivity of the equilibrium distribution to the job-size distribution in PS servers, see, e.g., Theorem 4.2 in [5]; and the fact that the routing policy only uses the number of jobs at each server when making a decision, as opposed to, e.g., using residual processing times. The GJSQ routing decisions are based on the dynamically changing number of jobs at each server as well as the service rates. Indeed, one expects the near-insensitivity properties to extend also to the case of heterogeneous servers and GJSQ routing. Establishing this near-insensitivity property is important, since it allows us to limit our attention to the more tractable GJSQ system with an exponential job-size distribution.

3.1 Simulation settings

To support our claims, we simulate the GJSQ model. A simulation consists of $2 \cdot 10^6$ job departures and each simulation is repeated 50 times. Statistics are only computed for departed jobs, i.e. data of jobs that are still in service at the end of the simulation are discarded. In Table 1 we list the four job-size distributions considered in this paper.

Name	Distribution	Support	Variance
uni	Uniform	$[0,2]$	1/3
exp	Exponential	$[0,\infty)$	1
weib	Weibull	$(0,\infty)$	5
logn	Log-normal	$(0,\infty)$	10

Table 1: Job-size distributions used in simulations.

3.2 Near-insensitivity results

In Table 2 we show simulated statistics on the mean and standard deviation $\sigma(\cdot)$ of Q_i for the GJSQ model with various job-size distributions. For the settings considered in Table 2, the mean number of jobs at server i deviates by no more than 3.6% from the exponential case, while the standard deviation deviates by at most 4.4%. The largest deviations from the exponential case occur for the log-normal job-size distribution. This is as expected, since this job-size distribution has a variance that is 10 times higher than the variance of the exponential job-size distribution. Although the results are not as strong as those shown in Figure 3 of [5], we conclude that the more volatile environment of heterogeneous servers and GJSQ routing also has the near-insensitivity property for $\mathbb{E}[Q_i]$ and $\sigma(Q_i)$. Moreover, the performance in terms of the mean response time is also near-insensitive to the job-size distributions by Little's law.

Concerning the conditional arrival rates, we see in Figure 3 that the simulated values for the job-size distributions of Table 1 match the results of the algorithm for the exponential case [14]. Simulated values for states where the sample standard deviation is not too high differ by at most 5% from the results for the exponential case. So, also the conditional arrival rates are near-insensitive to the job-size distribution.

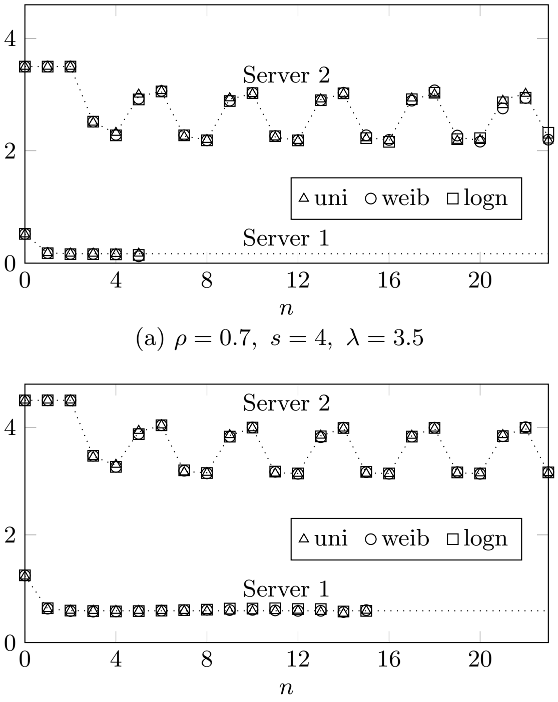
4. SINGLE QUEUE APPROXIMATION

We have established near-insensitivity of $\mathbb{E}[Q_i]$, $\sigma(Q_i)$ and the conditional arrival rates to the job-size distributions. Thus, we may limit our attention to systems with an exponential job-size distribution. In this section we derive an approximation for the distribution of the number of jobs at each server using the SQA, which models server i as an $M_n/M_i/1/PS$ queue with state-dependent arrival rates $\lambda_i(n)$, see also Section 3 in [5]. SQA is exact when the job-size distribution is exponential and the routing belongs to a specific class of routing policies; the following theorem is a version of Theorem 3.1 in [5] that is applicable to the GJSQ model.

Definition 3. A *stationary state-aware routing policy* is a time-stationary routing policy that only uses information about the number of jobs at the servers and the service rates at the instant of an arrival. The decisions may be made probabilistically, possibly biased in favor of certain servers.

THEOREM 1. *Consider the $M/M(1,s)/2/\mathcal{R}/\mathcal{S}$ queueing model, where $\mathcal{R}$ is any stationary state-aware routing policy, e.g. GJSQ, and $\mathcal{S}$ is any stationary, size-independent, work-conserving service discipline, e.g. PS. Consider server i in this model. Then SQA with the exact conditional arrival rates $\lambda_i(\cdot)$ yields the same equilibrium distribution for the number of jobs at each server as in the original model.*

It remains to specify the conditional arrival rates $\lambda_i(n)$ for both servers. We combine exact limiting results for $n \geq N_i$ and approximation results for $n < N_i$, where $N_1 = 3$ and $N_2 = 2s$. These choices for N_i result in accurate approximations.



(a) $\rho = 0.7,\ s = 4,\ \lambda = 3.5$

(b) $\rho = 0.9,\ s = 4,\ \lambda = 4.5$

Figure 3: Simulated conditional arrival rates in the GJSQ system with various job-size distributions. The dotted curves represent values determined by the algorithm in [14] for the exponential job-size distribution.

We note that Theorem 1 implies that in order to determine the conditional arrival rates, we may assume a FCFS service discipline.

In Figure 2 we have seen that the conditional arrival rates $\lambda_i(n)$ exhibit a repeating pattern from some n and onwards. We rigorously characterize this limiting repeating pattern in the next theorem.

THEOREM 2. *For the $M/M(1,s)/2/GJSQ/PS$ queueing model with $s \in \mathbb{N}$,*

$$\lim_{n\to\infty} \lambda_1(n) = \rho^{1+s}, \tag{5}$$

$$\lim_{n\to\infty} \lambda_2(sn + r) = \begin{cases} s\frac{A(r+1)}{A(r)}, & r = 0, 1, \dots, s-2, \\ s\rho^{1+s}\frac{A(0)}{A(s-1)}, & r = s-1, \end{cases} \tag{6}$$

where

$$A(r) = \sum_{i=1}^{s} \eta_i \frac{\beta_i}{\rho^{1+s} - \beta_i} i_+(\rho^{1+s}, \beta_i, r) + h(r) + \frac{\beta_{s+1}}{1 - \beta_{s+1}} i_-(\rho^{1+s}, \beta_{s+1}, r), \tag{7}$$

and the variables $\beta_1, \dots, \beta_{s+1}$, $\eta_1, \dots, \eta_s$, and the functions $h(\cdot)$, $i_+(\cdot)$, $i_-(\cdot)$ are defined in Lemma 5.11 in [14].

PROOF. See Appendix A. □

For the rates $\lambda_1(n)$, $n < 3$ and $\lambda_2(n)$, $n < 2s$ we provide approximations that are functions of s and ρ. For server

			Job-size distribution				SQA	
s	ρ	Metric	uni	exp	weib	logn	Value	Diff.
2	0.7	$\mathbb{E}[Q_1]$	0.9139 (0.0030)	0.9232 (0.0030)	0.9223 (0.0029)	0.9361 (0.0038)	0.9077	1.7%
		$\sigma(Q_1)$	1.0404 (0.0049)	1.0505 (0.0050)	1.0560 (0.0056)	1.0704 (0.0067)	1.0462	0.4%
		$\mathbb{E}[Q_2]$	2.0111 (0.0059)	2.0289 (0.0061)	2.0222 (0.0057)	2.0519 (0.0074)	2.0329	−0.2%
		$\sigma(Q_2)$	2.0302 (0.0099)	2.0465 (0.0106)	2.0506 (0.0114)	2.0813 (0.0141)	2.0484	−0.1%
	0.9	$\mathbb{E}[Q_1]$	3.2244 (0.0336)	3.2797 (0.0336)	3.2316 (0.0298)	3.2396 (0.0266)	3.2188	1.9%
		$\sigma(Q_1)$	3.2208 (0.0590)	3.2716 (0.0723)	3.2186 (0.0575)	3.2002 (0.0505)	3.2161	1.7%
		$\mathbb{E}[Q_2]$	6.6841 (0.0676)	6.7915 (0.0674)	6.6834 (0.0587)	6.6988 (0.0524)	6.6424	2.2%
		$\sigma(Q_2)$	6.4289 (0.1185)	6.5288 (0.1453)	6.4186 (0.1141)	6.3828 (0.1016)	6.4091	1.9%
4	0.7	$\mathbb{E}[Q_1]$	0.4688 (0.0016)	0.4747 (0.0017)	0.4705 (0.0017)	0.4667 (0.0022)	0.4741	0.1%
		$\sigma(Q_1)$	0.6685 (0.0024)	0.6730 (0.0026)	0.6700 (0.0029)	0.6652 (0.0031)	0.6655	1.1%
		$\mathbb{E}[Q_2]$	2.5386 (0.0063)	2.5507 (0.0069)	2.5177 (0.0070)	2.4997 (0.0067)	2.5866	−1.4%
		$\sigma(Q_2)$	2.5082 (0.0102)	2.5179 (0.0115)	2.4936 (0.0133)	2.4744 (0.0122)	2.5457	−1.1%
	0.9	$\mathbb{E}[Q_1]$	1.8662 (0.0191)	1.8793 (0.0145)	1.8830 (0.0196)	1.9400 (0.0223)	1.8813	−0.1%
		$\sigma(Q_1)$	1.9404 (0.0338)	1.9539 (0.0314)	1.9801 (0.0444)	2.0394 (0.0408)	1.9566	−0.1%
		$\mathbb{E}[Q_2]$	8.2405 (0.0769)	8.2773 (0.0597)	8.2631 (0.0783)	8.4863 (0.0861)	8.3642	−1.0%
		$\sigma(Q_2)$	7.6982 (0.1374)	7.7507 (0.1264)	7.8485 (0.1815)	8.0912 (0.1652)	7.7692	−0.2%

Table 2: Simulated mean and standard deviation of Q_i, for the GJSQ system with various s, ρ and job-size distributions. Sample standard deviation is shown in parentheses. Last two columns show the value obtained by the SQA and the relative difference with respect to the exponential case.

1 we use a multiple linear regression model to fit an approximate function for the conditional arrival rates on data obtained from the algorithm in [14] for $s = 1, 2, 3, 4$ and ρ from 0.3 to 0.99. Obviously, one can also use conditional arrival rates obtained by simulation for these fitting purposes. We carefully select a set of 5 independent variables for each conditional arrival rate. This leads to the following approximate conditional arrival rates for server 1:

$$\frac{\lambda_1(0)}{\rho^{1+s}} \approx \left[\begin{array}{ccccc} s\rho & s & \frac{s}{\rho} & 1 & \frac{\rho^2}{s^2} \end{array}\right]\beta_0, \quad (8)$$

$$\frac{\lambda_1(1)}{\rho^{1+s}} \approx \left[\begin{array}{ccccc} s\rho^2 & 1 & \frac{1}{\rho} & \frac{1}{s\rho} & \rho^{1/s} \end{array}\right]\beta_1, \quad (9)$$

$$\frac{\lambda_1(2)}{\rho^{1+s}} \approx 1 + \left[\begin{array}{ccccc} s\rho & \frac{1}{s\rho} & \frac{\rho}{s^2} & \frac{1}{s^2} & \rho^{1/s} \end{array}\right]\beta_2, \quad (10)$$

where

$$\beta_0 = \left[\begin{array}{ccccc} 0.669 & -1.90 & 1.23 & 1.86 & -0.192 \end{array}\right]^T,$$

$$\beta_1 = \left[\begin{array}{ccccc} -0.00856 & 1.37 & -0.0578 & 0.123 & -0.254 \end{array}\right]^T,$$

$$\beta_2 = \frac{1}{100}\left[\begin{array}{ccccc} -0.131 & -0.820 & -6.48 & 10.4 & 0.893 \end{array}\right]^T.$$

For $s = 1$, one should consider $\lambda_1(\cdot) = \lambda_2(\cdot)$ and use the approximations presented in (8)-(10).

For server 2, let us note that $\lambda_2(n) = \lambda,\ n = 0, \ldots, s-2$ due to the GJSQ routing. Using a multiple regression model in this case is more difficult, since the number of states for which we need to obtain a fit increases with s. To circumvent a possibly complex fitting procedure, we establish a relation between the conditional arrival rates for the states $n = s-1, s, \ldots, 2s-1$ and the limiting conditional arrival rates determined in Theorem 2. Namely,

$$\lambda_2(n) \approx \Big(1 + \big(\frac{1}{s} - \frac{\rho}{2s-1}\big)\frac{1}{2^{n-(s-1)}}\Big)\lambda_2^{\lim}(n-s), \quad (11)$$

where for convenience $\lambda_2^{\lim}(r)$ is defined as the right-hand side of (6) and $\lambda_2^{\lim}(-1) = \lambda_2^{\lim}(s-1)$. The approximations (8)-(11) behave in various limiting regimes as expected:

Proposition 3.

1. *For $s \to \infty$, we have that $\lambda_1(n) \downarrow 0$ and $\lambda_2(n) = \lambda$ for all $n \in \mathbb{N}_0$. No job will join server 1, since the processing times in server 2 are instantaneous.*

2. *In the light-traffic regime, i.e. $\rho \downarrow 0$, we find that $\lambda_1(n) \downarrow 0,\ n \in \mathbb{N}_0$ and $\lambda_2(n) \downarrow 0,\ n \geq s-1$.*

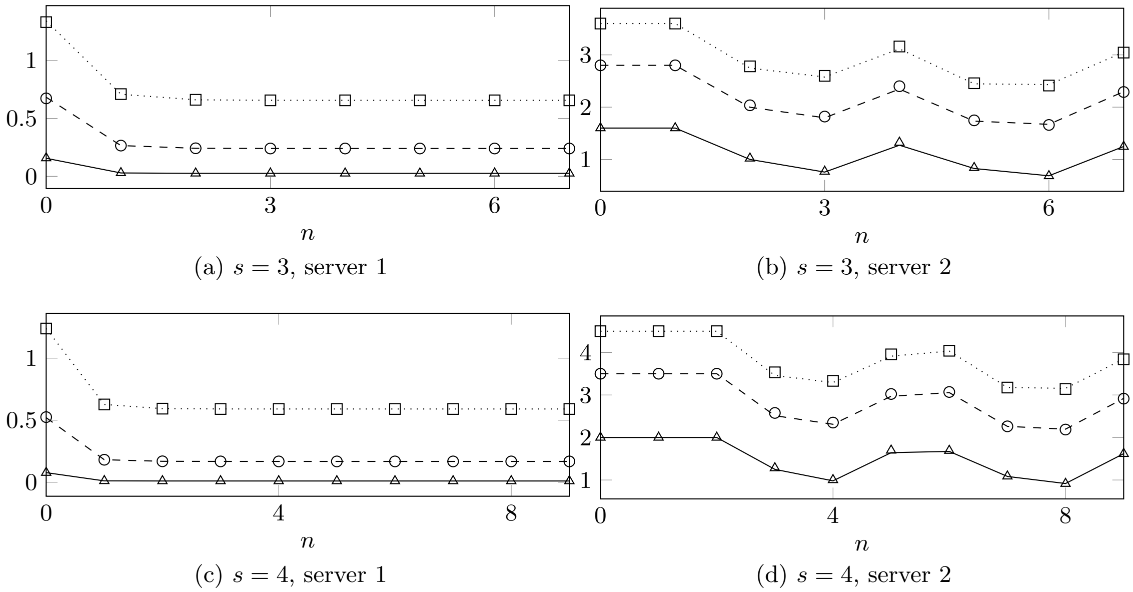
3. *In the heavy-traffic regime, i.e. $\rho \uparrow 1$, we establish that $\overline{\lambda}_1/\lambda = 1/(1+s)$ and $\overline{\lambda}_2/\lambda = s/(1+s)$ which is consistent with the findings in Figure 1.*

Proof. 1. Follows straightforwardly by taking the limit $s \to \infty$ in (8)-(10) while taking into account that $\rho = \lambda/(1+s)$. Furthermore, observe that $\lambda_2(n) = \lambda,\ n = 0, \ldots, s-2$, so that $\lim_{s\to\infty} \lambda_2(n) = \lambda,\ n \in \mathbb{N}_0$.
2. See Appendix B.
3. From the approximate conditional arrival rates $\lambda_1(\cdot)$ one can derive (approximate) equilibrium probabilities $\pi_1(\cdot)$. Then, $\overline{\lambda}_1 = \sum_{n=0}^{\infty} \lambda_1(n)\pi_1(n) = \sum_{n=0}^{\infty} \pi_1(n+1) = 1 - \pi_1(0)$ by exploiting the balance equations. For $\rho \uparrow 1$ it can be verified that $\pi_1(0) \downarrow 0$, so that $\lim_{\rho\uparrow 1} \overline{\lambda}_1/\lambda = 1/(1+s)$. The result for server 2 follows analogously. □

5. EVALUATING THE APPROXIMATION

We are now in a position to evaluate the proposed approximations. First, we show that the approximations for the conditional arrival rates follow closely the exact values, which were determined using the algorithm in [14]. Second, we establish that the mean and standard deviation of the number of jobs at each server is also well approximated.

Figure 4 compares the conditional arrival rates obtained from the algorithm in [14] and the approximations derived in the previous section. For the cases considered in the figure, the maximum relative difference of the approximation with respect to the values determined by the algorithm is 1.5% for $\lambda_1(\cdot)$ and 4.1% for $\lambda_2(\cdot)$. Since both methods consider exponential job-size distributions, the difference is due to the fitting error introduced in the approximations of the conditional arrival rates in Section 4 and the truncation error in the algorithm in [14]. However, since the truncation



(a) $s = 3$, server 1

(b) $s = 3$, server 2

(c) $s = 4$, server 1

(d) $s = 4$, server 2

Figure 4: Comparison of conditional arrival rates determined by the algorithm in [14] (lines) and our approximations (marks) for $\rho = 0.4$ (—,△), $\rho = 0.7$ (- -,○), and $\rho = 0.9$ (····,□).

error has been chosen to be of the order 10^{-5}, it has little influence.

In the two rightmost columns of Table 2 we provide the mean and standard deviation of the number of jobs at both servers determined using the SQA. We report highly accurate approximations that deviate less than 2.2% from the case with an exponential job-size distribution for the listed values of s and ρ. Although our approximations are not aimed at the case $s = 1$, we report accurate approximations also in this setting with maximum relative differences of the same order as in Section 6.1 of [5].

6. CONCLUSION

In this paper, we provide an approximate performance analysis of a queueing system consisting of two heterogeneous PS servers with service rates 1 and $s \in \mathbb{N}$, respectively, a general job-size distribution and GJSQ routing. More concretely, we derived the approximate equilibrium distribution of the number of jobs at each server using the SQA method. In order to apply SQA, we established that the GJSQ system is near-insensitive to the job-size distribution and thus we approximated the system at hand with exponentially distributed job-sizes. We then approximated the conditional arrival rates for the exponential case, by combining exact limiting results for large number of jobs and approximation results, which were obtained using a multiple linear regression model, for small number of jobs. Ultimately, the aforementioned approach resulted in approximations that are highly accurate; we reported a maximum relative difference with respect to exact or simulation results of 4.1% for the conditional arrival rates and 2.2% for the mean and standard deviation of the number of jobs at each server.

In this paper we set the groundwork for the analysis of server farms with heterogeneous servers under the GJSQ routing policy by analyzing the case of two servers. Of course, server farms consist of multiple servers so it is in our future plans to extend the analysis presented in this paper to more than two servers. The most difficult aspect of this task would be the derivation of the conditional arrival rates, which possibly has to rely on simulation data, since the approach in [14] is in its current setting restricted to two servers. In Figure 5 we present an example of the simulated conditional arrival rates in case of three servers with service rates 1, 2 and 5. Note that the structure of the conditional arrival rates is as expected, i.e. the number of points in the repeating pattern is directly related to the rate of service, but the exact values of these points differ from the values obtained by formulas (5) and (6).

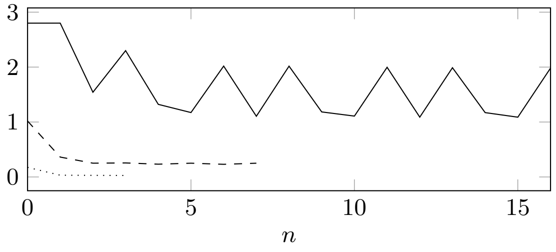


Figure 5: Simulated conditional arrival rates for a system with three servers with service rates 1 (····), 2 (- -), and 5 (—), with $\rho = 0.7$.

7. ACKNOWLEDGMENTS

This work was supported by an NWO free competition grant and the NWO Gravitation Project NETWORKS.

8. REFERENCES

[1] E. Altman, U. Ayesta, and B. Prabhu. Load balancing in processor sharing systems. *Telecommunication Systems*, 47(1-2):35–48, 2011.

[2] S.A. Banawan and J. Zahorjan. Load sharing in heterogeneous queueing systems. In *IEEE INFOCOM '89*, pages 731–739, 1989.

[3] V. Cardellini, E. Casalicchio, M. Colajanni, and P.S. Yu. The state of the art in locally distributed web-server systems. *ACM Computing Surveys*, 34(2):263–311, 2002.

[4] G.J. Foschini and J. Salz. A basic dynamic routing problem and diffusion. *IEEE Transactions on Communications*, 26(3):320–327, 1978.

[5] V. Gupta, M. Harchol-Balter, K. Sigman, and W. Whitt. Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation*, 64(9):1062–1081, 2007.

[6] M. Harchol-Balter. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action.* Cambridge University Press, 2013.

[7] M. Harchol-Balter, B. Schroeder, N. Bansal, and M. Agrawal. Size-based scheduling to improve web performance. *ACM Transactions on Computer Systems*, 21(2):207–233, 2003.

[8] E. Hyytiä, A. Penttinen, and S. Aalto. Size- and state-aware dispatching problem with queue-specific job sizes. *European Journal of Operational Research*, 217(2):357–370, 2012.

[9] C.N. Laws. Resource pooling in queueing networks with dynamic routing. *Advances in Applied Probability*, 24(3):699–726, 1992.

[10] Z. Liu, M.S. Squillante, and J.L. Wolf. On maximizing service-level-agreement profits. In *ACM conference on Electronic Commerce*, pages 213–223. ACM, 2001.

[11] J.C.S. Lui, R.R. Muntz, and D. Towsley. Bounding the mean response time of the minimum expected delay routing policy: An algorithmic approach. *IEEE Transactions on Computers*, 44(12):1371–1382, 1995.

[12] R.D. Nelson and T.K. Philips. An approximation to the response time for shortest queue routing. In *SIGMETRICS '89*, pages 181–189, 1989.

[13] M.C. Ortiz. Building the dynamic data center. *Dell Power Solutions*, 03:50–53, 2010.

[14] J. Selen, I.J.B.F. Adan, S. Kapodistria, and J.S.H. van Leeuwaarden. Steady-state analysis of shortest expected delay routing. arXiv:1509.03535v2, 2015.

[15] W. Whitt. Deciding which queue to join: Some counterexamples. *Operations Research*, 34(1):55–62, 1986.

APPENDIX

A. PROOF OF THEOREM 2

The proof is based on the exact results of the related $M/M(1,s)/2/SED/FCFS$ system, with $s \in \mathbb{N}$, presented in [14]. Although we obtain similar results for the limiting conditional arrival rates for server 1 as in [5], we use here a completely different approach in deriving the limits.

In [14], the state space $\{(q_1,q_2) \mid (q_1,q_2) \in \mathbb{N}_0^2\}$ of the Markov process is transformed to the state space $\{(m,n,r) \mid m \in \mathbb{N}_0,\ n \in \mathbb{Z},\ r = 0,1,\ldots,s-1\}$ where $m = \min(q_1, \lfloor \frac{q_2}{s} \rfloor)$, $n = \lfloor \frac{q_2}{s} \rfloor - q_1$ and $r = \mathrm{mod}(q_2, s)$. Let us denote the equilibrium probabilities for the three-dimensional state space as $p(m,n,r)$. The equilibrium probability $p(m,n,r)$ has a series expression, i.e. $p(m,n,r) = \sum_{l=0}^{\infty} x(l,m,n,r)$, namely, for $m \geq 0,\ n \geq 1$,

$$p(m,n,r) = C\sum_{l=0}^{\infty}\sum_{i=1}^{(s+1)^l}\sum_{j=1}^{s}\beta_{l,d(i)+j}^{n}\big(\eta_{l,d(i)+j}\alpha_{l,i}^{m} + \nu_{l+1,d(i)+j}\alpha_{l+1,d(i)+j}^{m}\big)i_+(\alpha_{l,i},\beta_{l,d(i)+j},r). \qquad (12)$$

For $m \geq 0$,

$$p(m,0,r) = C\sum_{l=0}^{\infty}\sum_{i=1}^{(s+1)^l}\alpha_{l,i}^{m}h_{l,i}(r). \qquad (13)$$

For $m \geq 0,\ n \leq -1$,

$$\begin{aligned} p(m,n,r) = {}& \\ & C\sum_{l=0}^{\infty}\sum_{i=1}^{(s+1)^l}\eta_{l,i(s+1)}\alpha_{l,i}^{m}\beta_{l,i(s+1)}^{-n}i_-(\alpha_{l,i},\beta_{l,i(s+1)},r) \\ & +C\sum_{l=0}^{\infty}\sum_{i=1}^{(s+1)^l}\nu_{l+1,i(s+1)}\alpha_{l+1,i(s+1)}^{m}\beta_{l,i(s+1)}^{-n}\times \\ & \qquad i_-(\alpha_{l+1,i(s+1)},\beta_{l,i(s+1)},r). \end{aligned} \qquad (14)$$

For the exact interpretation of each variable we refer the reader to [14]. In [14] the authors establish the following properties:

1. There exists a positive integer N such that $p(m,n,r)$ converges absolutely for all $m \geq 0,\ |n| \geq 1$ with $m + |n| > N$.
2. For $m + |n| > N$, we have $|x(l,m,n,r)| < u(l)$ and $\sum_{l=0}^{\infty} u(l) < \infty$.
3. The series $\sum_{m+|n|>N} p(m,n,r),\ r = 0,1,\ldots,s-1$ converges absolutely.
4. $|\alpha_{l,i}| > |\beta_{l,d(i)+j}|$ and $|\beta_{l,i}| > |\alpha_{l+1,i}|$ with $\alpha_{0,1} = \rho^{1+s} < 1$.

In this proof we make use of the dominated convergence theorem for complex-valued functions.

A.1 Server 1

The limiting conditional arrival rate to server 1 can be determined from

$$\lim_{n\to\infty}\lambda_1(n) = \lim_{n\to\infty}\frac{\pi_1(n+1)/\alpha_{0,1}^{n+1}}{\pi_1(n)/\alpha_{0,1}^{n}}\alpha_{0,1}. \qquad (15)$$

The marginal distribution for server 1 is given by, where $m = \lfloor \frac{q_2}{s} \rfloor$ and $r = \mathrm{mod}(q_2, s)$,

$$\begin{aligned} \pi_1(n) &= \sum_{m=0}^{\infty}\sum_{r=0}^{s-1}p(\min(n,m),m-n,r) \\ &= \sum_{m=0}^{n-1}\sum_{r=0}^{s-1}p(m,m-n,r) + \sum_{r=0}^{s-1}p(n,0,r) \\ &\quad + \sum_{m=1}^{\infty}\sum_{r=0}^{s-1}p(n,m,r). \end{aligned} \qquad (16)$$

Furthermore,

$$\lim_{n\to\infty}\frac{\pi_1(n)}{\alpha_{0,1}^{n}} = \lim_{n\to\infty}\sum_{m=0}^{n-1}\sum_{r=0}^{s-1}\frac{p(m,m-n,r)}{\alpha_{0,1}^{n}}$$

$$+\sum_{r=0}^{s-1} \lim_{n\to\infty} \frac{p(n,0,r)}{\alpha_{0,1}^n} + \sum_{m=1}^{\infty}\sum_{r=0}^{s-1} \lim_{n\to\infty} \frac{p(n,m,r)}{\alpha_{0,1}^n}, \qquad (17)$$

where the interchange of the limit and the series for the third term on the right-hand side of (17) is allowed by the dominated convergence theorem, because one can bound $p(n,m,r)$ from above by $p(0,m,r)$ and $\sum_{m=0}^{\infty} p(0,m,r) < \infty$ since it is a subseries of $\sum_{m+|n|>N} p(m,n,r)$, which converges absolutely by property 3. Furthermore, we know that $\lim_{m\to\infty} p(m,n,r) = \lim_{m\to\infty} \sum_{l=0}^{\infty} x(l,m,n,r)$ which is equal to $\sum_{l=0}^{\infty} \lim_{m\to\infty} x(l,m,n,r)$ by the dominated convergence theorem for complex-valued functions in combination with property 2. This allows us to compute the second and third term on the right-hand side of (17). The first term on the right-hand side of (17) can be determined as follows

$$\lim_{n\to\infty} \sum_{m=0}^{n-1}\sum_{r=0}^{s-1} \frac{p(m,m-n,r)}{\alpha_{0,1}^n}$$

$$= C\Big(\lim_{n\to\infty} \sum_{l=0}^{\infty} \sum_{i=1}^{(s+1)^l} \eta_{l,i(s+1)} \frac{\left(\frac{\alpha_{l,i}}{\alpha_{0,1}}\right)^n - \left(\frac{\beta_{l,i(s+1)}}{\alpha_{0,1}}\right)^n}{\frac{\alpha_{l,i}}{\beta_{l,i(s+1)}} - 1} \times$$

$$\sum_{r=0}^{s-1} i_-(\alpha_{l,i}, \beta_{l,i(s+1)}, r)$$

$$+ \lim_{n\to\infty} \sum_{l=0}^{\infty} \sum_{i=1}^{(s+1)^l} \nu_{l+1,i(s+1)} \frac{\left(\frac{\beta_{l,i(s+1)}}{\alpha_{0,1}}\right)^n - \left(\frac{\alpha_{l+1,i(s+1)}}{\alpha_{0,1}}\right)^n}{1 - \frac{\alpha_{l+1,i(s+1)}}{\beta_{l,i(s+1)}}} \times$$

$$\sum_{r=0}^{s-1} i_-(\alpha_{l+1,i(s+1)}, \beta_{l,i(s+1)}, r)\Big)$$

$$= C\frac{\eta_{0,s+1}}{\frac{\alpha_{0,1}}{\beta_{0,s+1}} - 1} \sum_{r=0}^{s-1} i_-(\alpha_{0,1}, \beta_{0,s+1}, r). \qquad (18)$$

Interchange of the limit and series is again allowed here since one can bound the absolute value of the summands from above by $v(l)$ for sufficiently large n and $\sum_{l=0}^{\infty} v(l) < \infty$. One can finally establish that

$$\lim_{n\to\infty} \frac{\pi_1(n)}{\alpha_{0,1}^n} = C\Big(\frac{\eta_{0,s+1}}{\frac{\alpha_{0,1}}{\beta_{0,s+1}} - 1} \sum_{r=0}^{s-1} i_-(\alpha_{0,1}, \beta_{0,s+1}, r)$$

$$+ \sum_{r=0}^{s-1} h_{0,1}(r) + \sum_{j=1}^{s} \frac{\eta_{0,j}\beta_{0,j}}{1-\beta_{0,j}} \sum_{r=0}^{s-1} i_+(\alpha_{0,1}, \beta_{0,j}, r)\Big). \qquad (19)$$

Thus, by (15) and (19), $\lim_{n\to\infty} \lambda_1(n) = \alpha_{0,1} = \rho^{1+s}$.

A.2 Server 2

The limiting conditional arrival rate to server 2 can be determined from $\lim_{n\to\infty} \lambda_2(sn+r)$, where, for $r = 0, 1, \ldots, s-2$,

$$\lim_{n\to\infty} \lambda_2(sn+r) = \lim_{n\to\infty} s\frac{\pi_2(sn+r+1)/\alpha_{0,1}^n}{\pi_2(sn+r)/\alpha_{0,1}^n} \qquad (20)$$

and for $r = s-1$,

$$\lim_{n\to\infty} \lambda_2(sn+r) = \lim_{n\to\infty} s\frac{\pi_2(sn+r+1)/\alpha_{0,1}^{n+1}}{\pi_2(sn+r)/\alpha_{0,1}^n}\alpha_{0,1}. \qquad (21)$$

The marginal distribution for server 2 is given by, for $r = 0, 1, \ldots, s-1$,

$$\pi_2(sn+r) = \sum_{q_1=0}^{\infty} p(\min(q_1,n), n-q_1, r)$$

$$= \sum_{q_1=0}^{n-1} p(q_1, n-q_1, r) + p(n,0,r) + \sum_{q_1=1}^{\infty} p(n,-q_1,r). \qquad (22)$$

For $\pi_2(sn+r+1)$, $r = s-1$ we should replace n by $n+1$ and r by 0 in (22). Furthermore, for $r = 0, 1, \ldots, s-1$,

$$\lim_{n\to\infty} \frac{\pi_2(sn+r)}{\alpha_{0,1}^n} = \lim_{n\to\infty} \sum_{q_1=0}^{n-1} \frac{p(q_1, n-q_1, r)}{\alpha_{0,1}^n}$$

$$+ \lim_{n\to\infty} \frac{p(n,0,r)}{\alpha_{0,1}^n} + \lim_{n\to\infty} \sum_{q_1=1}^{\infty} \frac{p(n,-q_1,r)}{\alpha_{0,1}^n}. \qquad (23)$$

Using identical arguments as for the limiting conditional arrival rate for server 1, we establish for $r = 0, 1, \ldots, s-1$,

$$\lim_{n\to\infty} \frac{\pi_2(sn+r)}{\alpha_{0,1}^n} = A(r), \qquad (24)$$

where $A(r)$ is given in (7). Finally, combining (20)-(21) and (24) proves (6).

B. PROOF OF PROPOSITION 3, POINT 2

By letting $\rho \downarrow 0$ in (8)-(10) and $\lambda_1(n) \approx \rho^{1+s}$, $n \geq 3$ we immediately find that $\lambda_1(n) \downarrow 0$, $n \in \mathbb{N}_0$.

We note that in (11) the factors on the right-hand side in front of $\lambda_2^{\lim}(n-s)$ go to a constant for $\rho \downarrow 0$. So, what remains is that we establish that $\lim_{\rho\downarrow 0} \lambda_2^{\lim}(r) = 0$, $r = 0, 1, \ldots, s-1$. This part of the proof relies heavily on the asymptotic results of [14]. We denote $\alpha = \rho^{1+s}$ and investigate for $r = 0, 1, \ldots, s-1$,

$$\frac{A(\alpha,r)}{\alpha^{r/s}} = \sum_{i=1}^{s} \eta_i \frac{\beta_i/\alpha}{1-\beta_i/\alpha} u_i \big(\frac{\beta_i}{\alpha}\big)^{r/s} + \alpha^{1/s}\frac{h(r)}{\alpha^{(r+1)/s}}$$

$$+ \alpha^{1-r/s}\frac{\beta_{s+1}/\alpha}{1-\beta_{s+1}} i_-(\alpha, \beta_{s+1}, r), \qquad (25)$$

where we used that $i_+(\alpha, \beta_i, r) = u_i \beta_i^{r/s}$ with u_i the i-th unit root of $u^s = 1$, which is established in Lemma 5.6 of [14]. Now,

$$\lim_{\alpha\downarrow 0} \frac{A(\alpha,r)}{\alpha^{r/s}} = c(r), \qquad (26)$$

where $c(r)$ is some constant. In the following we denote c_i as some constant that can be a function of r. Equation (26) follows from the fact that for $\alpha \downarrow 0$ we have that $\beta_i/\alpha \to c_1 < 1$, $i = 1, 2, \ldots, s$, $\beta_{s+1}/\alpha \to c_2$ (Lemma 5.15(i)(a) and (i)(c) of [14]); $h(r)/\alpha^{(r+1)/s} \to c_3(r)$ (Appendix B, part (c) of [14]); $i_-(\alpha, \beta_{s+1}, r) \to c_4(r)$ (Lemma 5.15(i)(d) of [14]); $\beta_{s+1} \to 0$ (Corollary 5.14 in [14]); and $\sum_{i=1}^{s} \eta_i u_i \to c_5$ ($\alpha \downarrow 0$ in (5.46) of [14]).

Finally, for $r = 0, 1, \ldots, s-2$,

$$\lim_{\alpha\downarrow 0} \frac{A(\alpha,r+1)}{A(\alpha,r)} = \lim_{\alpha\downarrow 0} \frac{\alpha^{(r+1)/s}}{\alpha^{r/s}} \frac{A(\alpha,r+1)/\alpha^{(r+1)/s}}{A(\alpha,r)/\alpha^{r/s}}$$

$$= \lim_{\alpha\downarrow 0} \alpha^{1/s}\frac{c(r+1)}{c(r)} = 0 \qquad (27)$$

and for $r = s-1$,

$$\lim_{\alpha\downarrow 0} \alpha\frac{A(\alpha,0)}{A(\alpha,s-1)} = \lim_{\alpha\downarrow 0} \alpha^{1/s}\frac{c(0)}{c(s-1)} = 0. \qquad (28)$$