QUEUES WITH RESEQUENCING, PART III:
LIGHT TRAFFIC LIMITS AND INTERPOLATION APPROXIMATIONS

by

Subir Varma[1] and Armand M. Makowski[2]

Electrical Engineering Department and Systems Research Center
University of Maryland, College Park, Maryland 20742

ABSTRACT

In this paper we combine the heavy traffic limits from [14], with light traffic limits, to obtain polynomial approximations for several resequencing models. Among the models analyzed thus, include the following. (1): The Baccelli, Gelenbe and Plateau [1] model of resequencing, which was analyzed by the above authors. They gave a rather complicated expression for the transforms of the performance measures of interest. However, by using our methods, we are able to obtain simple explicit expressions for these measures. (2): Resequencing due to parallel single server queues. This model was analyzed by Gün and Jean-Marie [4]. However the expression that they obtained for the response time of the system, was amenable to computation only in Markovian case. Using our methods, We show how to obtain approximations for the case of non-Markovian service times.

[1] The work of this author was supported by a fellowship from the Systems Research Center, University of Maryland

[2] The work of this author was supported partially through NSF Grant ECS-83-51836 and partially through NSF Engineering Research Centers Program Grant CDR-88-03012-04

1. Introduction

In [14] we developed heavy traffic limits for a variety of queueing systems with resequencing. In the present paper our objective is to obtain polynomial approximations for some of those models. To that end we calculate the light traffic limits using the Reiman–Simon theory [11] and combine them with the heavy traffic limits of [14]. For an introduction to the literature concerning resequencing systems, the reader may consult our companion paper [13] or the survey [3].

In Section 2 we obtain polynomial approximations for the Baccelli, Plateau, Gelenbe (BGP) model. An earlier analysis of this model using complex analytic methods [1] yielded a complicated expression for the Laplace transform of the end–to–end delay, from which it was very difficult to obtain explicit formulas. However using our methods, we obtain a quadratic approximation to the average waiting time which agrees extremely well with experimental results. In Sections 3 and 4 we obtain polynomial approximations for the resequencing model in which the disordering is due to K single server queues operating in parallel. This model was analyzed by Gün and Jean–Marie [4], who gave an expression for the average end–to–end delay for the case when the arrivals are Poisson. However, this expression is difficult to evaluate except in the case when the services are exponential. Using, heavy and light traffic theory, we give simple but good approximation for non–exponential service times, such as deterministic service times (in Section 3) and r^{th} order Erlangian service times (in Section 4). Lastly in Section 5 we obtain polynomial approximations for a generalized BGP model, in which the disordering system, which is composed of two single server queues operating in parallel, is followed by a single server queue.

2. Polynomial approximations for the BGP model

The reader may recall that the BGP model operates as follows (Fig 1): customers arrive into an infinite server disordering system after receiving service from which, they are resequenced and sent into a single server queue. In this section we develop light traffic approximations for the BGP model in the special case when the arrival process is Poisson with rate λ, the disordering process is exponential with rate ν and the service process is also exponential with rate μ. We combine the heavy traffic limit of [14] with the light traffic limits of this section, to obtain an approximation that provides good estimates for the entire range of λ. Also note that even though we restrict out attention to exponential service and disordering distributions, this methodology can be applied to an arbitrary service and disordering distributions.

We now proceed to obtain formulae for $\overline{W}(0)$ and $\overline{W}'(0)$ with the help of [14] It is trivial to see that

$$\overline{W}(0) = 0 \tag{2.1}$$

since if only one customer arrives over the entire time interval, then the only delay it encounters before getting served is the disordering delay.

We now proceed to calculate $\overline{W}'(0)$. Let $W(t, d_0, d_1, s_1)$ be the waiting time of the customer that arrives at time zero with disordering delay d_0, given that another customer arrives at time t with disordering delay d_1 and service time s_1 at the single server queue. It is clear that

$$W(t, d_0, d_1, s_1) = \begin{cases} 0, & \text{if } t > 0 \\ \max(0, t + d_1 - d_0 + s_1), & \text{if } t \leq 0, \end{cases} \tag{2.2}$$

Define the RVs X and Y as follows,

$$X = d_1 + s_1, \tag{2.3}$$

$$Y = X - d_0. \tag{2.4}$$

Then it can be easily shown that X has the density function f_X given by

$$f_X(x) = \frac{\mu\nu}{\nu - \mu}(e^{\mu x} - e^{-\nu x}), \quad x \geq 0. \tag{2.5}$$

and Y has the distribution function F_Y given by

$$F_Y(x) = 1 + \frac{\mu e^{-\nu x}}{2(\nu - \mu)} - \frac{\nu^2 e^{-\mu x}}{(\nu^2 - \mu^2)}, \quad x \in I\!R. \tag{2.6}$$

Note that

$$W := W(t, d_0, d_1, s_1) = \max(0, t + Y), \quad \text{if } t \leq 0$$

so that the RV W has distribution F_W given by

$$\begin{aligned} F_W(x) &= I\!\!P(Y+t \le x), \\ &= 1 + \frac{\mu e^{-\nu x} e^{\nu t}}{2(\nu-\mu)} - \frac{\nu^2 e^{-\mu x} e^{\mu t}}{(\nu^2-\mu^2)}, \quad x \ge 0. \end{aligned} \tag{2.7}$$

Using the fact that

$$\overline{\psi}(\{t\}) = \int_0^\infty (1 - F_W(x))dx, \quad \text{if } t < 0$$

it follows that

$$\overline{\psi}(\{t\}) = \frac{\nu^2 e^{\mu t}}{\mu(\nu^2-\mu^2)} - \frac{\mu e^{\nu t}}{2\nu(\nu-\mu)}, \quad \text{if } t < 0 \tag{2.8}$$

Finally combining (2.8) with (VI.2.7), we obtain

$$\overline{W}'(0) = \frac{\nu^2}{\mu^2(\nu^2-\mu^2)} - \frac{\mu}{2\nu^2(\nu-\mu)}. \tag{2.9}$$

For the case $\nu = \mu$, if we use L'Hospitals rule and take the limit as $\nu \to \mu$ in (3.8), we obtain,

$$\overline{W}'(0) = \frac{7}{4\mu^2}. \tag{2.10}$$

We now combine the light traffic estimates with the heavy traffic estimates in Appendix A to obtain a first order approximation to the average waiting time of the resequencing model. If we specialize the heavy traffic result of equation (A1) to the case when all RV's all exponentially distributed, we obtain

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda)\overline{W}(\lambda) = 1. \tag{2.11}$$

Finally from (2.1), (2.9) and (2.11), we obtain as the first order approximation to the average waiting time in steady state as

$$\begin{aligned} \hat{W}(\lambda) = & \frac{\lambda\nu^2}{\mu(\nu^2-\mu^2)(\mu-\lambda)} - \frac{\lambda\mu^2}{2\nu^2(\nu-\mu)(\mu-\lambda)} + \frac{\lambda^2}{\mu^2(\mu-\lambda)} \\ & - \frac{\lambda^2\nu^2}{\mu^2(\nu^2-\mu^2)(\mu-\lambda)} + \frac{\lambda^2\mu}{2\nu^2(\nu-\mu)(\mu-\lambda)}, \quad \mu \ne \nu, \\ & 0 \le \lambda < \mu \end{aligned} \tag{2.12}$$

and

$$\hat{W}(\lambda) = \frac{\lambda}{\mu - \lambda} + \frac{3}{4(\mu - \lambda)}[\frac{\lambda}{\mu} - (\frac{\lambda}{\mu})^2], \quad \mu = \nu, \quad 0 \leq \lambda < \mu. \tag{2.13}$$

This approximation agrees extremely well with simulation results (see Section 2.1).

Note that in (2.12)

$$\lim_{\nu \uparrow \infty} \hat{W}(\lambda) = \frac{\lambda}{\mu} \frac{1}{\mu - \lambda}, \quad 0 \leq \lambda < \mu$$

which is the average waiting time in a $M/M/1$ queue, as one would expect, because as $\nu \uparrow \infty$ the disordering delay goes to zero and the system like an ordinary $M/M/1$ queue.

2.1 Simulation results

The approximation (2.12) is compared with simulation results in the case $\nu = 2$ and $\mu = 1$. Substituting these values into (3.12) we obtain

$$\hat{W}(\lambda) = \frac{\lambda}{24(1 - \lambda)}(29 - 5\lambda), \quad 0 \leq \lambda < 1.$$

The 95% confidence levels have been obtained in all cases.

λ	$\overline{W}(\lambda)$	$\hat{W}(\lambda)$	% Error
0.1	0.133 ± 0.003	0.13	2.25
0.2	0.29 ± 0.006	0.29	1.69
0.3	0.49 ± 0.011	0.49	1.80
0.4	0.76 ± 0.017	0.75	1.83
0.5	1.12 ± 0.029	1.10	1.78
0.6	1.65 ± 0.049	1.62	1.82
0.7	2.51 ± 0.097	2.48	1.19
0.8	4.21 ± 0.24	4.16	1.19
0.9	9.50 ± 0.31	9.19	3.26

3. Approximations for the parallel queue resequencing model: The case of deterministic services

Our objective in this section is to obtain interpolation approximations for the average response time $\overline{T}(\lambda)$, of the resequencing model in which the disordering is due to K single server queues operating in parallel (Fig 2). We further assume that the input into the system is a Poisson process and the the customers are switched to the various queues by a Bernoulli switch with equi–probable switching probability. This system was analyzed by Gün and Jean–Marie [4] who gave an expression for the average response time of this model in terms of the virtual waiting process for the single server queues. However except for the case in which the service times are exponential, the virtual waiting time is hard to obtain.

In this section we assume that the service times are deterministic and equal to $\frac{1}{\mu}$. In the next section we treat the case in which the service times possess the Erlangian distribution. The heavy traffic limit for this system is given in equation ($A2$), which in the case of deterministic service times and $n = 1$, reduces to

$$\lim_{\lambda \to K\mu} (K\mu - \lambda)\overline{T}_K(\lambda) = \frac{KH_K}{2}. \tag{3.1}$$

We now proceed to obtain the light traffic limits for this system. It is easy to that

$$\overline{T}_K(0) = \frac{1}{\mu} \tag{3.2}$$

since if only one customer arrives over the entire time interval, then it does not encounter any queueing or resequencing delay in the system.

We now proceed with the calculation of $\overline{T}'_K(0)$. Let $T(t)$ be the response time of a customer that arrives at time zero, given that another customer arrived at time t. Then it easy to see that

$$T(t) = \begin{cases} \frac{1}{\mu} & \text{if } t \geq 0 \text{ or if } t < 0 \text{ and the two go to} \\ & \text{different queues,} \\ \frac{1}{\mu} + \max(0, t + \frac{1}{\mu}) & \text{if } t < 0 \text{ and both go to the same queue.} \end{cases} \tag{3.3}$$

It is plain from (3.3) that for the case $t \geq 0$

$$T(t) = \overline{\psi}(\{t\}) = \frac{1}{\mu} = \overline{\psi}(\emptyset)$$

so that (II.2.7) now reduces to

$$\overline{T}'(0) = \int_{-\infty}^{0} (\overline{\psi}(\{t\}) - \overline{\psi}(\emptyset))dt. \tag{3.4}$$

Taking note of the fact that both customers go to the same queue with probability $\frac{1}{K}$, while they go to different queues with probability $\frac{K-1}{K}$, and substituting (3.3) into (3.4), we obtain

$$\begin{aligned} \overline{T}'_K(0) &= \frac{1}{K} \int_{t=-\frac{1}{\mu}}^{0} (t + \frac{1}{\mu})dt \\ &= \frac{1}{2K\mu^2}. \end{aligned} \tag{3.5}$$

Combining (3.1), (3.2) and (3.5) we obtain the following approximation $\hat{T}_K(\lambda)$ for the average response time of the system,

$$\hat{T}_K(\lambda) = \frac{K}{(K\mu - \lambda)} - \frac{\lambda}{2\mu(K\mu - \lambda)} + \frac{H_K - 1}{2K(K\mu - \lambda)}(\frac{\lambda}{\mu})^2, \quad 0 \leq \lambda < K\mu. \tag{3.6}$$

This approximation agrees extremely well with simulation results (see Section 8.3.1).

3.1 Simulation results

Approximation (3.5) is compared with simulation for the case when $\mu = 1$, while $K = 2, 5$ and 10.

λ	$\overline{T}_2(\lambda)$	$\hat{T}_2(\lambda)$	%Error
0.2	1.05 ± 0.001	1.05	0.01
0.4	1.13 ± 0.002	1.14	0.88
0.6	1.24 ± 0.004	1.25	0.81
0.8	1.38 ± 0.007	1.40	1.45
1.0	1.60 ± 0.012	1.62	1.25
1.2	1.95 ± 0.024	1.97	1.02
1.4	2.53 ± 0.048	2.57	1.58
1.6	3.70 ± 0.111	3.80	2.70
1.8	7.53 ± 0.16	7.52	0.13

λ	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error
0.5	1.06 ± 0.001	1.06	0.11
1.0	1.16 ± 0.003	1.16	0.44
1.5	1.31 ± 0.006	1.30	1.42
2.0	1.52 ± 0.011	1.50	1.31
2.5	1.85 ± 0.020	1.81	2.16
3.0	2.36 ± 0.039	2.31	2.16
3.5	3.23 ± 0.072	3.18	1.55
4.0	5.01 ± 0.167	4.98	0.60
4.5	10.68 ± 0.21	10.51	1.59

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error
1	1.07 ± 0.002	1.06	0.65
2	1.20 ± 0.004	1.17	2.50
3	1.41 ± 0.009	1.34	4.96
4	1.73 ± 0.016	1.59	8.06
5	2.19 ± 0.028	1.98	9.58
6	2.90 ± 0.053	2.62	9.65
7	4.08 ± 0.104	3.74	8.33
8	6.43 ± 0.257	6.08	5.44
9	13.69 ± 0.22	13.31	2.77

4. Approximations for the parallel queue resequencing model: The case of Erlang services

The model to be analysed is the same as in the last section, except for the fact that now the parallel queues are assumed to possess a r^{th} order Erlang service distribution with rate μ. The heavy traffic limit ($A2$) for the case $n = 1$ reduces to

$$\lim_{\lambda \to K\mu} (K\mu - \lambda)\overline{T}_K(\lambda) = \frac{r+1}{2r} K H_K. \tag{4.1}$$

As in the last section, we have

$$\overline{T}_K(0) = \frac{1}{\mu} \tag{4.2}$$

and we now proceed to calculate $\overline{T}'_K(0)$.

Let $T(t, s_0, s_1)$ be the response time of a customer that arrives at time zero with service time s_0, given that another customer arrived at time t with service time s_1. We see that

$$T(t, s_0, s_1) = \begin{cases} s_0, & \text{if } t \geq 0 \\ s_0 + \max(0, t + s_1), & \text{if } t < 0 \text{ and the two customers} \\ & \text{join the same queue,} \\ \max(s_0, t + s_1), & \text{if } t < 0 \text{ and the two customers} \\ & \text{join different queues.} \end{cases} \tag{4.3}$$

As before, the two customers join the same queue with probability $\frac{1}{K}$, while they join different queues with probability $\frac{K-1}{K}$. Let $\overline{T}'_{K,I}(0)$ be the first derivative of the average response time which is obtained under the assumption that the two customers join the same queue, and let $\overline{T}'_{K,II}(0)$ be this derivative obtained under the assumption that the two customers join different queues. Then it is clear that

$$\overline{T}'_K(0) = \frac{1}{K}\overline{T}'_{K,I}(0) + \frac{K-1}{K}\overline{T}'_{K,II}(0)$$

When both customers join the same queue, the calculation of $\overline{T}'_{K,I}(0)$ reduces to the calculation of the corresponding quantity in a single server queue, since resequencing does not play any role. From the light traffic limits for the single server queue obtained in [9], we conclude that

$$\overline{T}'_{K,I}(0) = \frac{r+1}{2r\mu^2}. \tag{4.4}$$

We now treat the case where the two customers join different queues. Note that

$$\overline{T}'_{K,II}(0) = \int_{t=0}^{\infty}\int_{s_0=0}^{\infty}\int_{s_1}^{\infty} \max(0, s_1 - s_0 - t)dt \ h_r(s_1)h_r(s_0)ds_1 \ ds_0 \tag{4.5}$$

where h_r, is the density function of a r^{th} order Erlang distribution, given by

$$h_r(x) = \frac{r\mu(r\mu x)^{r-1}e^{-r\mu x}}{(r-1)!}, \quad x \geq 0. \qquad r = 1, 2 \ldots (4.6)$$

Interchanging the order of integration in (4.5), we then get

$$\overline{T}'_{K,II}(0) = \int_{s_1=0}^{\infty}\int_{s_0=0}^{s_1}\int_{t=0}^{s_1-s_0} (s_1 - s_0 - t)dt \ h_r(s_0)h_r(s_1)ds_0 ds_1. \qquad r = 1, 2 \ldots$$

Carrying out the integration with respect to t and simplifying, we conclude that

$$\overline{T}'_{K,II}(0) = \frac{r+1}{2r\mu^2} - \int_{s_1=0}^{\infty}\int_{s_0=0}^{s_1} s_0 s_1 h_r(s_0) h_r(s_1) ds_0 ds_1. \qquad r = 1, 2 \ldots (4.7)$$

Substituting the expression (4.6) for h_r into (4.7) and making some further simplifications, we obtain

$$\overline{T}'_{K,II}(0) = \frac{1}{\mu^2}(\frac{r+1}{2r} - Q), \qquad r = 1, 2 \ldots (4.8)$$

where

$$Q = \frac{1}{(r!)^2}\int_{v=0}^{\infty} v^r e^{-v} dv \int_{u=v}^{\infty} u^r e^{-u} du = \frac{1}{2}. \qquad r = 1, 2 \ldots (4.9)$$

Combining (4.4) and (4.8), we obtain

$$\overline{T}'_{K}(0) = \frac{1}{\mu^2}(\frac{r+1}{2r} - \frac{K-1}{2K}). \qquad r = 1, 2 \ldots (4.10)$$

Finally combining (4.1), (4.2) and (4.10) we obtain an approximation $\hat{T}_K(\lambda)$ for the average response time of the system in the form

$$\begin{aligned}\hat{T}_K(\lambda) = &\frac{K}{K\mu - \lambda} + \left[K(\frac{r+1}{2r} - \frac{K-1}{2K}) - 1\right]\frac{\lambda}{\mu(K\mu - \lambda)} \\ &+ \left[\frac{r+1}{2r}\frac{H_K}{K} - (\frac{r+1}{2r} - \frac{K-1}{2K})\right](\frac{\lambda}{\mu})^2\frac{1}{(K\mu - \lambda)}, \\ &\qquad 0 \le \lambda < K\mu,\ r = 1, 2 \ldots (4.11)\end{aligned}$$

This approximation agrees extremely well with simulation results (see Section 4.1).

4.1 Simulation results

Approximation (4.11) is compared with simulation for the case when $r = 2, \mu = 1$, while $K = 2, 5$ and 10.

λ	$\overline{T}_2(\lambda)$	$\hat{T}_2(\lambda)$	% Error
0.2	1.11 ± 0.005	1.11	0.09
0.4	1.25 ± 0.007	1.25	0.08
0.6	1.43 ± 0.010	1.44	0.70
0.8	1.69 ± 0.016	1.74	2.96
1.0	2.05 ± 0.028	2.06	0.49
1.2	2.59 ± 0.048	2.61	0.77
1.4	3.49 ± 0.089	3.54	1.43
1.6	5.31 ± 0.232	5.40	1.69
1.8	10.90 ± 0.28	11.01	1.03

λ	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error
0.5	1.19 ± 0.005	1.19	0.25
1.0	1.43 ± 0.007	1.43	0.35
1.5	1.74 ± 0.017	1.74	0.28
2.0	2.15 ± 0.028	2.15	0.30
2.5	2.72 ± 0.047	2.73	0.37
3.0	3.57 ± 0.083	3.59	0.56
3.5	5.08 ± 0.174	5.02	1.18
4.0	7.87 ± 0.121	7.88	0.13
4.5	16.14 ± 0.372	16.44	1.86

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error
1	1.31 ± 0.007	1.32	0.76
2	1.67 ± 0.014	1.71	2.39
3	2.11 ± 0.023	2.18	3.32
4	2.67 ± 0.038	2.79	4.49
5	3.43 ± 0.062	3.60	4.95
6	4.56 ± 0.110	4.78	4.82
7	6.50 ± 0.225	6.69	2.92
8	10.49 ± 0.574	10.44	0.47
9	20.97 ± 0.414	21.52	2.62

5. Light traffic approximations for a generalized BGP model

The model to be analyzed in section operates as follows (Fig 3): customers arriving according to a Poisson process with rate λ are Bernoulli switched with equal probability to K single server queues operating in parallel with identical exponential service rates ν. After they leave this system they are resequenced in a resequencing buffer and sent to the buffer of single server queue with exponential service rate μ. After getting served in this queue they leave the system. Note that the parallel $M/M/1$ queues can also be in heavy traffic, in addition to the single server queue. Hence we shall assume that $K\nu > \mu$, so that as the the arrival rate λ increases from zero, the single server queue goes into heavy traffic earlier than the K $M/M/1$ queues.

We now proceed to find the light traffic limits for the average waiting time in this system $\overline{W}_K(\lambda)$, which is defined as the total time that a customer spends in the resequencing buffer plus the buffer of the single server queue.

It is trivial to see that

$$\overline{W}_K(0) = 0 \tag{5.1}$$

since if only one customer arrives over the entire time interval, then it does not encounter any resequencing or queueing delay.

We now proceed to calculate $\overline{W}'_K(0)$. Let $W(t, c_0, c_1, s_1)$ be the waiting time of the customer that arrives at time zero with service time c_0 at one of the $M/M/1$ queues, given that another customer arrives at time t with service time at one of the $M/M/1$ queues equal to c_1 and service time at the single server queue equal to s_1. Then it clear that

$$W(t, c_0, c_1, s_1) = \begin{cases} 0, & \text{if } t > 0; \\ \max(0, t + c_1 + s_1 - c_0), & \text{if } t \leq 0 \text{ and the} \\ & \text{customers are} \\ & \text{routed to} \\ & \text{different queues.} \\ \max(0, t + c_1 + s_1 - c_0 - \max(0, t + c_1)), & \text{if } t \leq 0 \text{ and the} \\ & \text{customers are} \\ & \text{routed to} \\ & \text{same queue.} \end{cases} \tag{5.2}$$

Note that the customers are routed to the same queue with probability $\frac{1}{K}$, while they are routed to different queues with probability $\frac{K-1}{K}$. In the case when the customers are to different queues, the waiting time is exactly the same as for the case when the disordering system is an infinite server queue. Hence

$$\overline{W}'_K(0) = \frac{K-1}{K} \int_{t=-\infty}^{0} \int_{c_0} \int_{c_1} \int_{s_1} \max(0, t + c_1 + s_1 - c_0) H_1(dc_0) H_1(dc_1) H_2(ds_1)$$

$$+\frac{1}{K} \int_{t=-\infty}^{0} \int_{c_0} \int_{c_1} \int_{s_1} \max(0, t+c_1+s_1-c_0-\max(0, t+c_1)) H_1(dc_0) H_1(dc_1) H_2(ds_1) \tag{5.3}$$

where H_1 and H_2 are distribution functions of exponential distributions with rate ν and μ respectively.

The first of these integrals was already calculated in Section 8.2 so that

$$\overline{W}'_K(0) = \frac{(K-1)\nu^2}{K\mu^2(\nu^2-\mu^2)} - \frac{(K-1)\mu}{2K\nu^2(\nu-\mu)} + \frac{1}{K}\int_{t=-\infty}^{0}\int_{c_0}\int_{c_1}\int_{s_1} \max(0, t+c_1+s_1-c_0-\max(0,t+c_1))H_1(dc_0)H_1(dc_1)H_2(ds_1). \tag{5.4}$$

We now proceed to calculate the second integral which we denote as I. Define the RV X as follows,

$$X = c_1 + t. \tag{5.5}$$

Then it can be easily shown that a has the following density function,

$$f_X(x) = \nu e^{\nu t} e^{-\nu x}, \quad x \geq t. \tag{5.6}$$

Also define the RV Y as

$$Y = s_1 - c_0, \tag{5.7}$$

then after some calculation we come to the conclusion that b has the following distribution function,

$$F_Y(x) = \begin{cases} 1 - \frac{\nu}{\nu+\mu}e^{-\mu x}, & \text{if } x \geq 0 \\ e^{\nu x} - \frac{\nu}{\nu+\mu}e^{(\nu+2\mu)x}, & \text{if } x < 0\ . \end{cases} \tag{5.8}$$

With the help of (5.5) and (5.7), equation (5.2) simplifies to

$$W(X,Y) = \max(0, X+Y-\max(0,X)) \tag{5.9}$$

for the case $t \geq 0$. Our next objective is to find the distribution function of the RV $W(X,Y)$. Note that for $z \geq 0$

$$\begin{aligned} I\!P(W(X,Y) \leq z) &= I\!P(X+Y-\max(0,X) \leq z) \\ &= \int_{x=t}^{\infty} I\!P(x+Y-\max(0,x) \leq z \mid X=x) f_X(x)dx \end{aligned} \tag{5.10}$$

Taking not of the fact that the RV's X and Y are independent, the above expression simplifies to

$$\begin{aligned}\mathbb{P}(W(X,Y) \leq z) = &\int_{x=t}^{0} \mathbb{P}(x+Y \leq z) f_X(x)dx + \int_{x=0}^{\infty} \mathbb{P}(Y \leq z) f_X(x)dx \\ =&(1 - \frac{\nu}{\nu+\mu} e^{-\mu z}) \int_{x=0}^{\infty} \nu \exp^{\nu t} e^{-\nu x} dx \\ &+ \int_{x=t}^{0} (1 - \frac{\nu}{\nu+\mu} e^{-\mu(z-x)}) \nu \exp^{\nu t} e^{-\nu x} dx \\ =&1 - \frac{\nu}{\nu+\mu} e^{\nu t} e^{-\mu z} - \frac{\nu^2}{\nu^2-\mu^2} e^{-\mu z}(e^{\mu t} - e^{\nu t}), \quad z \geq 0. \qquad (5.11)\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{E}W(X,Y) = &\int_{z=0}^{\infty} [1 - \mathbb{P}(W(X,Y) \leq z)]dz \\ &= \frac{\nu}{\mu(\nu+\mu)} e^{\nu t} + \frac{\nu^2}{\mu(\nu^2-\mu^2)}(e^{\mu t} - e^{\nu t}). \qquad (5.12)\end{aligned}$$

Integrating over t we finally come to the conclusion that

$$I = \frac{1}{\mu^2}. \qquad (5.13)$$

so that combining equations (5.4) and (5.13),

$$\overline{W}'_K(0) = \frac{(K-1)\nu^2}{K\mu^2(\nu^2-\mu^2)} - \frac{(K-1)\mu}{2K\nu^2(\nu-\mu)} + \frac{1}{K\mu^2}. \qquad (5.14)$$

Note that if we specialize the heavy traffic result of [14, Theorem 3.1, Part (a)] to the case when all RV's all exponentially distributed, we obtain

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{W}_K(\lambda) = 1. \qquad (5.15)$$

Combining (5.1), (5.14) and (5.15) we obtain as the first order approximation $\hat{W}_K(\lambda)$ to the average waiting time for the case when $K\nu > \mu$ and $\nu \neq \mu$, as

$$\begin{aligned}\hat{W}_K(\lambda) =&\frac{\lambda}{K\mu(\mu-\lambda)} + \frac{(K-1)\lambda\nu^2}{K\mu(\nu^2-\mu^2)(\mu-\lambda)} - \frac{(K-1)\lambda\mu^2}{2K\nu^2(\nu-\mu)(\mu-\lambda)} \\ &+ \frac{(K-1)\lambda^2}{K\mu^2(\mu-\lambda)} - \frac{(K-1)\lambda^2\nu^2}{K\mu^2(\nu^2-\mu^2)(\mu-\lambda)} + \frac{(K-1)\lambda^2\mu}{2K\nu^2(\nu-\mu)(\mu-\lambda)}, \\ &\qquad 0 \leq \lambda < \mu. \quad (5.16)\end{aligned}$$

Note that

$$\lim_{\nu\uparrow\infty} \hat{W}_K(\lambda) = \frac{\lambda}{\mu}\frac{1}{\mu-\lambda}, \quad 0 \le \lambda < \mu$$

which is the average waiting time in a $M/M/1$ queue, as one would expect, because as $\nu \uparrow \infty$ the disordering delay goes to zero.

In the case $\nu = \mu$, a similar calculation leads us to the conclusion that

$$\hat{W}_K(\lambda) = \frac{7(K-3)}{4K}\frac{\lambda}{\mu}\frac{1}{\mu-\lambda} - \frac{(3-3K)}{4K}(\frac{\lambda}{\mu})^2\frac{1}{\mu-\lambda}, \quad 0 \le \lambda < \mu. \tag{5.17}$$

Even though approximation (6.16) is valid in the range $K\nu > \mu$, simulation results suggest that it performs quite poorly when $K\nu$ is close to μ. It performs best when $K\nu \gg \mu$ (see Section 5.1), and we suggest that the reader who is interested in applying (5.16), choose a ν such that at least $\nu \ge \mu$.

5.1 Simulation results

Approximation (5.16) is compared with simulation for the case when $K = 2, \nu = 2$ and $\mu = 1$. Substituting these values into (5.16) we obtain

$$\hat{W}_2(\lambda) = \frac{\lambda}{1-\lambda}(1.104 - 0.104\lambda), \quad 0 \le \lambda < 1.$$

λ	$\overline{W}_2(\lambda)$	$\hat{W}_2(\lambda)$	% Error
0.1	0.124 ± 0.004	0.121	2.42
0.2	0.27 ± 0.006	0.271	0.37
0.3	0.46 ± 0.010	0.456	0.22
0.4	0.72 ± 0.016	0.71	1.39
0.5	1.07 ± 0.028	1.05	1.86
0.6	1.60 ± 0.050	1.56	2.50
0.7	2.46 ± 0.104	2.41	2.03
0.8	4.12 ± 0.07	4.08	0.97
0.9	8.91 ± 0.25	9.09	2.02

APPENDIX A

In this appendix we give the heavy traffic formulae from [14] that have been used in this paper.

(i): Let $\overline{W}(\lambda)^{(n)}$ be the n^{th} moment of the waiting time in the BGP model. Assume that the arrival process has rate λ and variance σ_0^2 while the service time process has rate μ and variance σ^2. Then the following heavy traffic limit holds.

$$\lim_{\lambda\uparrow\mu}(\mu-\lambda)\overline{W}^{(n)}(\lambda) = n!(\frac{\sigma_0^2+\sigma^2}{2}\mu^2)^n \qquad\qquad n=1,2\ldots(A1)$$

(ii): Let $\overline{T}_K^{(n)}(\lambda)$ be the n^{th} moment of the response time, in the resequencing model in which the disordering is due to K parallel queues. Assume that the arrival process is Poisson with rate λ while the service process has rate μ and variance σ^2. Then the following heavy traffic limit holds.

$$\lim_{\lambda\uparrow K\mu}(K\mu-\lambda)^n\overline{T}_K^{(n)}(\lambda) = n!\left[(\sigma^2+\frac{1}{\mu^2})\frac{K\mu^2}{2}\right]^n\sum_{k=1}^{K}\binom{K}{k}\frac{(-1)^{k+1}}{k^n}. \qquad (A2)$$

REFERENCES

[1] F. Baccelli, E Gelenbe and B. Plateau,"An end-to-end approach to the resequencing problem," *JACM*, Vol. 31, No. 3, pp. 474-485 (1984).

[2] F. Baccelli,"A queueing model for timestamp ordering in a distributed system," *Performance '87*, Brussels, pp. 413-431 (1987).

[3] F. Baccelli and A.M. Makowski,"Queueing systems with synchronization constraints," *Proceeding of the IEEE*, (1989).

[4] L.Gün and A. Jean-Marie,"Parallel queues with resequencing," Manuscript, University of Maryland, (1990).

[5] J.M. Harrison, *Brownian motion and stochastic flow systems*, J. Wiley and Sons, New York (1985).

[6] E. Horlatt and D. Mailles,"Etude du resequencement dans un reseau de files d'attente" *Technical Report No. 125, Universitie P. et M. Curie*, (1986).

[7] J. Kollerstrom,"Heavy traffic theory for queues with several servers. I," *J. Appl. Prob.*, Vol. 11, pp. 544-552 (1974).

[8] A.M. Law and W.D. Kelton, *Simulation modeling and analysis*, Mc-Graw Hill, New York (1982).

[9] M.I. Reiman and B. Simon, "An interpolation approximation for queueing systems with Poisson input," *Oper. Res.*, Vol. 36, No. 3, pp. 454-469 (1988).

[10] M.I. Reiman and B. Simon, "Light traffic limits of sojourn time distributions in Markovian queuing networks," *Commun. Statist.-Stochastic Models*, Vol. 4, No. 2, pp. 191-233 (1988).

[11] M.I. Reiman and B. Simon, "Open queueing systems in light traffic," *Maths. of Oper. Res.*, Vol. 14, No. 1, pp. 26-59 (1989).

[12] S. Varma, "Some problems in queueing systems with resequencing," MS Thesis, University of Maryland; also available as SRC Technical Report No. TR-87-192 (1987).

[13] S. Varma and A.M. Makowski," Resequencing systems, Part I: Structural Properties," In preparation.

[14] S. Varma and A.M. Makowski," Resequencing systems, Part II: Heavy traffic limits," In preparation.

[15] S. Varma,"Performance evaluation of the time–stamp ordering algorithm in a distributed database," Manuscript (1990).

[16] S. Varma and A.M. Makowski, "Heavy traffic limits for fork–join queues," In preparation, (1990)

[17] S. Varma and A.M. Makowski, "Interpolation approximations for fork–join queues," In preparation, (1990).

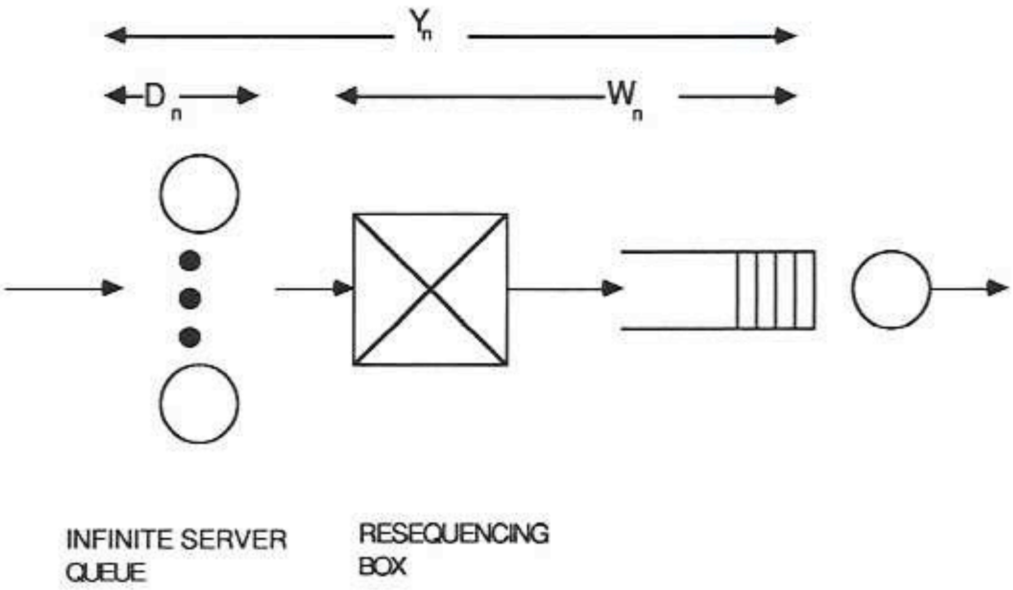

Fig. 1. The BGP model

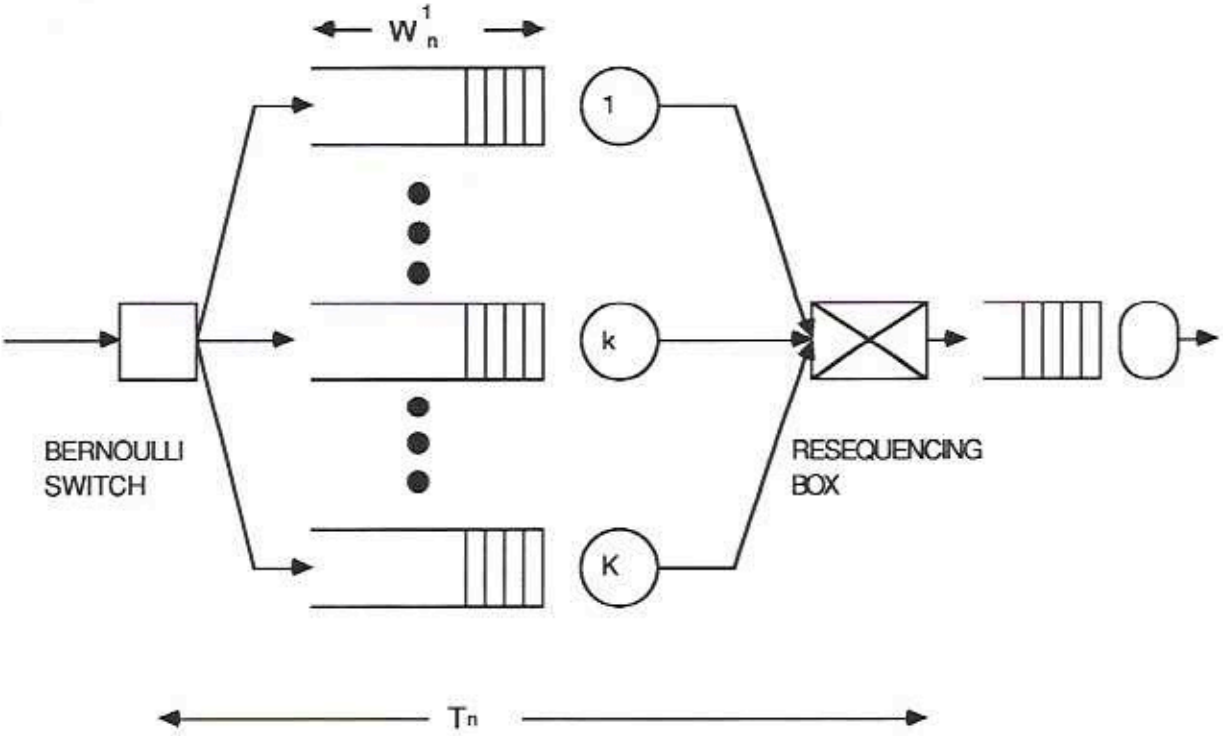

Fig. 2. A generalized BGP model