A Brownian control problem for a simple queueing system in the Halfin-Whitt regime

Rami Atar[*]
Technion - I.I.T.
Haifa 32000, Israel
atar@ee.technion.ac.il

Avi Mandelbaum[†]
Technion - I.I.T.
Haifa 32000, Israel
avim@ie.technion.ac.il

Marty Reiman
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974, USA
marty@research.bell-labs.com

May 21, 2002

Abstract

We consider a formal diffusion limit for a control problem of a multi-type multi-server queueing system, in the regime proposed by Halfin and Whitt, in the form of a control problem where the dynamics are driven by a Brownian motion. In one dimension, a pathwise minimum is obtained and is characterized as the solution to a SDE. The pathwise solution to a special multi-dimensional problem (corresponding to a multi-type system) follows.

1 Introduction

Brownian control problems (BCPs) were proposed by Harrison [8] as formal diffusion limits for queueing network control problems, to provide a basis for identifying and analyzing "good" or nearly optimal control policies. Since then, several authors have studied methods for providing optimal solutions to the BCPs, as well as suboptimal policies for the queueing networks which asymptotically achieve these optima (see [13] and references therein). The formal limit is obtained under the so called heavy traffic scaling (which we refer to here as the *classical* heavy traffic scaling), in which time is speeded up by a factor of N, and queue lengths are normalized by a factor of $\sqrt{N}$. In the classical heavy trafic regime, a multi-server model with a fixed number of servers gives rise to a diffusion limit identical to that obtained for a single server with accelerated service. In systems where the number of servers is large (e.g., in models for call centers [5]), it is reasonable to consider an alternative heavy traffic asymptotic regime, namely the one that was proposed by Halfin and Whitt [6]. Under this regime, the number of servers is scaled up by a factor of N, the number of customers in queue and the number of idle servers are scaled down by a factor of $\sqrt{N}$, and time

[*]Research supported in part by the US-Israel Binational Science Foundation and the fund for the promotion of research at the Technion

[†]Research supported in part by the fund for the promotion of research at the Technion, by Technion V.P.R. fund for the promotion of sponsored research, and by the Israel Science Foundation (grant no. 388/99).

is not scaled (for recent results on these diffusion limits under fixed policies, see [10], [11], [12]). Typically, the diffusion limits obtained under the classical heavy traffic scaling give rise to reflected diffusions, while the scaling of Halfin and Whitt gives rise to diffusions with nonlinear (but piecewise linear) drift. In the current work we consider a BCP obtained as a formal limit under the scaling of Halfin and Whitt. Rather than formulating a general framework, we consider in this short paper only the most simple example of a queueing network service control problem, as depicted in Figure 1. A related work is [9], where the HJB equation for the Brownian control problem under study is proved to have a unique solution. Another control problem in the Halfin-Whitt regime is studied in [1], although the objective function there is different.

Before analyzing the control problem referred to above, we formulate a BCP in dimension one, for which we show that a pathwise solution exists. This solution is otherwise characterized as the solution to a SDE. We point out the analogy with the classical BCP [7], where the pathwise minimum agrees with the solution to the Skorohod equation.

In many cases, it has been shown that BCPs (in the classical setting) which correspond to networks with several customer classes or service stations, and are therefore multi-dimensional, have a reduction to a one dimensional problem, and as a result, a cost such as the weighted average queue length possesses a pathwise minimum. The BCPs discussed in the current paper turn out to be more complicated in that pathwise minimal solutions do not exist even in very simple two-dimensional problems. Consider a network consisting of two classes of customers 1 and 2, served by a pool of statistically identical servers, where class i customers are served at rate μ_i and the number of class i customers in the system at time t is $Q_i(t)$, $i = 1, 2$. The quantity which corresponds in the BCP to the weighted average queue length $Q^c(t) = \sum_i c_i Q_i(t)$ does not in general have a pathwise minimum, and in particular, minimizing different (monotone) functionals of $Q^c(t)$ may give rise to different optimizing policies. However, in the special case where $\mu_1 = \mu_2$ (but $c_1 \neq c_2$), we show that the quantity corresponding to in the BCP to $Q^c(t)$ does have a pathwise minimum. This is done by showing that the dimensionality of the problem can be reduced, and by using the one-dimensional solution. Our argument applies to an arbitrary number of classes, but we consider only two classes, to keep the notation simple. In the model that we consider, we also allow for customer abandonments from the queues. Heuristically, the solution to the BCP suggests priority to the class i for which c_i is greater. However, as is known in the classical scaling (e.g., [3]), an actual asymptotically optimal policy for the queueing network may have to be more involved than what is reflected by solutions to the limit problem.

Although the BCPs in the context considered here may fail to have the especially convenient form of solution that the classical ones have, they still provide an obvious simplification of the underlying queueing network control problems, and may help identifying asymptotically optimal policies for particular costs. We pursue this direction in the paper [2].

In Section 2 we formally derive a BCP for a two-dimensional network. In Section 3 we consider pathwise minimum results for a coressponding one-dimensional problem. Finally, in Section 4 we identify a two-dimensional BCP that has a pathwise minimum, by showing that it can be reduced to a one dimensional problem.

2 Formal derivation of a Brownian control problem

The configuration of the queueing system under study is depicted in Figure 1. The arrival rate to queue i is λ_i, $i = 1, 2$. Abandonments from queue i occur at rate θ_i per customer per unit time. There are N statistically identical servers, and service to class i is performed at rate μ_i. A controller dynamically schedules the services.

Let Q_{i0}, Q_{i1} denote the number of class-i customers waiting in the queue, and, respectively, being served. The total number of customers of class i in the system is then $Q_i = Q_{i0} + Q_{i1}$. To ease the exposition, we consider a Markovian network (Poisson arrivals and exponential services), although more general networks give rise to the same BCP. The state of the system will be given by the collection of the four variables Q_{ij}, $i = 1, 2$, $j = 0, 1$. Note that if we assumed the policy is a non-idling one, we would have a three dimensional problem, e.g. with variables $Q_{10} + Q_{11}$, Q_{20} and Q_{21}, since then $Q_{10} = ((Q_{10} + Q_{11}) + Q_{21} - N)^+$. However, at least in the prelimit problem, it makes sense to allow for idling policies. Let A_i denotes the arrival process of class i customers, and S_i the potential number of service completions in class i up to time t by a single server, namely, a Poisson process of rate μ_i. Similarly, $R_i(t)$ denotes a process used to count abandonments and is a Poisson process of rate θ_i. A_i, S_i and R_i, $i = 1, 2$ are independent.

Following Bell and Williams [3] and Williams [13], the control policy will be associated with a process $T = (T_1, T_2)$, where $T_i(t)$ denotes the accumulated time devoted to class i up to time t, summed over all servers. Note that $T_i(t)$ is also the integral up to time t of the number of servers serving class i customers. The composition $S_i(T_i(t))$, $i = 1, 2$ which gives the number of class-i customers served by one server up to time $T_i(t)$, is equal in law to the number of class-i customers that are atually served up to time t. In the same spirit, if $U_i(t)$ denotes the waiting time before service, accumulated up to time t, summed over all class i customers, then it is equal to the integral up to time t of the queue length Q_{i0}, and $R_i(U_i(t))$ then gives the number of abandonments from queue i until time t.

The constraints that the processes above must satisfy are as follows. For $i = 1, 2$ and $j = 0, 1$ and $t \geq 0$, $Q_{ij}(t) \geq 0$. Moreover, $Q_{11}(t) + Q_{21}(t) \leq N$, $t \geq 0$. Finally, the two components of T are nondecreasing processes.

We introduce two more quantities. Although they do not carry additional information, it will be convenient to use them to express the constraints. The total number of class i customers in the system at time t will be denoted by $Q_i(t) = \sum_j Q_{ij}(t)$. $I(t)$ denotes the idle time until time t, summed over all servers. The time derivatives of T, U and I satisfy $\dot{T}_i = Q_{i1}$, $\dot{U}_i = Q_{i0}$ and $\dot{I} = N - \sum_i Q_{i1}$.

The equations satisfied by the above quantities are

$$\begin{cases} Q_i(t) &= Q_i(0) + A_i(t) - S_i(T_i(t)) - R_i(U_i(t)) \\ U_i(t) &= \int_0^t Q_i(s)ds - T_i(t) \\ I(t) &= Nt - T_1(t) - T_2(t). \end{cases} \tag{1}$$

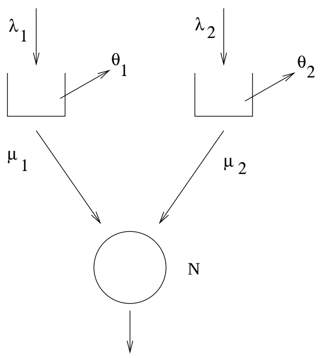

Figure 1: A queueing model

The constraints discussed before are now fully described by:

$$T_i, U_i, I \text{ are nondecreasing.} \tag{2}$$

Note, in particular, that $Q_i \geq 0$ follows from (1) and (2).

We adopt the notation of [3, 13] for the renormalized processes: $\bar{X}$ and $\hat{X}$ correspond to fluid and, respectively, diffusion scaling. The superscript N will be used to denote the parameters and variables corresponding to the Nth system. The parameters undergo the following scaling: $\lambda_i^N/N \to \lambda_i$, $\theta_i^N \to \theta_i$ and $\mu_i^N \to \mu_i$, but for simplicity we shall consider only the case where $\lambda_i^N = N\lambda_i$, $\theta_i^N = \theta_i$ and $\mu_i^N = \mu_i$. Due to this simplification, the heavy traffic assumption $\lambda_1^N/(N\mu_1^N) + \lambda_2^N/(N\mu_2^N) \to 1$ as $N \to \infty$ takes the form

$$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} = 1. \tag{3}$$

The initial conditions will correspond to the steady state fluid approximation solutions, namely, $Q_i^N(0) = \frac{\lambda_i}{\mu_i}N$. The following equations define the scaled processes involved.

$$\left\{ \begin{array}{lcl} \bar{T}_i^N(t) & = & N^{-1}T^N(t) \\ \bar{U}_i^N(t) & = & N^{-1}U^N(t) \\ \hat{A}_i^N(t) & = & N^{-1/2}(A_i^N(t) - N\lambda_i t) \\ \hat{S}_i^N(t) & = & N^{-1/2}(S_i^N(Nt) - N\mu_i t) \\ \hat{R}_i^N(t) & = & N^{-1/2}(R_i^N(Nt) - N\theta_i t) \\ \hat{U}_i^N(t) & = & N^{-1/2}U_i^N(t) \\ \hat{I}^N(t) & = & N^{-1/2}I^N(t) \\ \hat{Q}_i^N(t) & = & N^{-1/2}(Q_i^N(t) - Q_i^N(0)) \end{array} \right.$$

Letting

$$\bar{T}^*(t) = \left(\frac{\lambda_1}{\mu_1}t,\ \frac{\lambda_2}{\mu_2}t\right),$$

and introducing the processes

$$\begin{aligned}
\hat{Y}_i^N(t) &= N^{1/2}(\bar{T}_i^*(t) - \bar{T}_i^N(t)), \quad i = 1, 2, \\
\hat{X}_i^N(t) &= \hat{A}_i^N(t) - \hat{S}_i^N(\bar{T}_i^N(t)) - \hat{R}_i^N(\bar{U}_i^N(t)),
\end{aligned}$$

we obtain the following equations for the normalized quantities

$$\begin{cases}
\hat{Q}_i^N(t) &= \hat{X}_i^N(t) + \mu_i \hat{Y}_i^N(t) - \theta_i \hat{U}_i^N(t), \\
\hat{U}_i^N(t) &= \int_0^t \hat{Q}_i^N(s)ds + \hat{Y}_i^N(t), \\
\hat{I}^N(t) &= \hat{Y}_1^N(t) + \hat{Y}_2^N(t).
\end{cases}$$

Note that by assumption $\hat{X}_i^N(0) = 0$. Since we would like the limit control problem to correspond to the family of queueing network control problems for which the fluid asymptotics of T^N is given by $\bar{T}^*$, we impose the assumption that $\bar{T}^N \to \bar{T}^*$. The processes $\hat{A}_i^N$, $\hat{S}_i^N \circ \bar{T}_i^N$ and, respectively, $\hat{R}_i^N \circ \bar{U}_i^N$ then formally converge to Brownian motions with mean zero and variances λ_i, λ_i and, respectively, 0.

We can now state the BCP for the system. The costs that we consider are somewhat arbitrary in view of the fact that we will only be interested here with pathwise solutions. Let $\tilde{X}_i$ be independent Brownian motions with variances $2\lambda_i$, $i = 1, 2$. One is required to minimize either

$$\lim_{t\to\infty} t^{-1}(c_1 Q_1(t) + c_2 Q_2(t)),$$

or

$$E\int_0^\infty e^{-\gamma t}(c_1 Q_1(t) + c_2 Q_2(t))dt,$$

using a control process (Y_1, Y_2, U_1, U_2) such that the processes (Q, U, I) satisfy

$$\begin{cases}
Q_i(t) &= X_i(t) + \mu_i Y_i(t) - \theta_i U_i(t), \\
U_i(t) &= \int_0^t Q_i(s)ds + Y_i(t), \\
I(t) &= Y_1(t) + Y_2(t). \\
U_i \text{ and } I \text{ are nondecreasing.}
\end{cases} \tag{4}$$

3 On a one dimensional control problem

In [7] a one dimensional BCP is defined which corresponds to the classical heavy traffic scaling, and it is shown that it has a unique pathwise minimum. The minimum is otherwise given as the solution to the one dimensional Skorohod equation. We consider here a control problem that is analogous to it in both respects: It has a unique pathwise minimizer; and its solution can be characterized as the unique solution to a certain differential equation. The equation for the minimum is

$$dQ(t) = dX(t) + \mu Q^-(t)dt - \theta Q^+dt,$$

$$Q(0) = X(0),$$

where we denote $x^+ = \max(0, x)$ and $x^- = \max(0, -x)$, and where X is the driving Brownian motion.

The one dimensional BCP is to minimize (pathwise) the cost $c_1Q_1 + c_2Q_2$ using controls Y and U such that

$$\begin{cases} Q(t) = X(t) + \mu Y(t) - \theta U(t), \\ \int_0^t Q(s)ds = U(t) - Y(t) \\ Y, U \text{ are non-decreasing}, \\ Y(0) = U(0) = 0. \end{cases} \tag{5}$$

Proposition 1 *Let $X \in C$ be given and consider the relations (5). Assume $\theta \neq \mu$. Then there is a unique solution (Q^*, Y^*, U^*, I^*) to (5) in C, for which Y^* and U^* are minimal in the following sense: For any solution (Q, Y, U, I) to (5) one has*

$$U(t) \geq U^*(t) \quad t \geq 0, \tag{6}$$

and

$$Y(t) \geq Y^*(t) \quad t \geq 0. \tag{7}$$

Moreover, Q^* is given by the unique solution q to

$$q(t) = X(t) + \mu \int_0^t q^-(s)ds - \theta \int_0^t q^+(s)ds, \tag{8}$$

and U^* and Y^* are given by

$$U^*(t) = \int_0^t (Q^*(s))^+ds, \tag{9}$$

$$Y^*(t) = \int_0^t (Q^*(s))^-ds. \tag{10}$$

Remarks: (a) In case that $\theta = \mu$ there are multiple solutions.

(b) Minimality or maximality of Q^* also holds, depending on the relation between θ and μ. In case that $\theta < \mu$, one has

$$Q(t) \geq Q^*(t), \quad t \geq 0,$$

and in case $\theta > \mu$,

$$Q(t) \leq Q^*(t), \quad t \geq 0$$

holds. This follows from the proof.

(c) Equation (8) was obtained by Halfin and Whitt [6] in the case $\theta = 0$ as the weak limit of a queueing system undergoing the above scaling. Garnett, Mandelbaum and Reiman generalized the result of [6] to accommodate abandonment.

(d) Equations (9) and (10) merely express the fact that under the optimal policy, the cummulative idle time and the cummulative waiting time are minimal. They also indicate that under the optimal policy, when $Q \geq 0$ one has $dY = 0$ and when $Q < 0$ one has $dU = 0$. This, in fact, together with (5) characterizes the solution (Q^*, Y^*, U^*) (see Proposition 2).

(e) In fact, a statement stronger than (6) holds (for $\theta < \mu$): $U - U^*$ is non-decreasing. On the other hand, as can be shown by some simple examples, $Y - Y^*$ is not necessarily non-decreasing.

Proof: Since x^- is Lipschitz in x it is classical that (8) has a unique solution. Therefore the functions Q^*, Y^*, U^* and I^* are well defined. The relations between Y^*, U^* and Q^* expressed in (9) and (10) are immediate consequences of (5) and (8). We will first treat the case $\theta < \mu$. It will be shown that

$$Q(t) \geq Q^*(t), \quad t \geq 0, \tag{11}$$

and (6) and (7) hold for an arbitrary solution (Q, Y, U, I) to (5). We claim that

$$\eta(t) \equiv U(t) - \int_0^t Q^+(s)ds \quad \text{is non-decreasing.} \tag{12}$$

Indeed, (5) imposes that both $U(t)$ and $Y(t) = U(t) - \int_0^t Q(s)ds$ are non-decreasing. Hence for $0 \leq s < t$ one has

$$\begin{aligned} U(t) - U(s) - \int_s^t Q^+(\theta)d\theta &= \int_s^t 1_{Q>0}d\eta + \int_s^t 1_{Q\leq 0}d\eta \\ &= \int_s^t 1_{Q>0}dY + \int_s^t 1_{Q\leq 0}dU \\ &\geq 0, \end{aligned} \tag{13}$$

where the last line follows by monotonicity of the integrators and non-negativity of the integrands. Since $s < t$ are arbitrary, (12) holds. From the second line in (5), we have that

$$\eta = U - \int_0^\cdot Q^+ds = Y - \int_0^\cdot Q^-ds.$$

Now, from the first line in (5) we have

$$Q(t) = X(t) + \mu \int_0^t Q^-(s)ds - \theta \int_0^t Q^+(s)ds + (\mu - \theta)\eta(t). \tag{14}$$

The solution to this equation is monotone in η in the following sense: If $\tilde{\eta} - \eta$ is nondecreasing with $\eta(0) = \tilde{\eta}(0)$ and if Q $[\tilde{Q}]$ denotes the solution corresponding to η [respectively, $\tilde{\eta}$] then $\tilde{Q} \geq Q$ (see [4]). Since $\eta \geq 0$ and Q^* corresponds to $\eta = 0$, (11) follows.

Next, since $Q \geq Q^*$, we have that $Q^+ \geq (Q^*)^+$. We therefore obtain from (9) that

$$U \geq \int_0^\cdot Q^+ds \geq \int_0^\cdot (Q^*)^+ds = U^*,$$

and (6) follows. Now (7) follows from the first line in (5), (6) and (11). This completes the proof in the case $\theta < \mu$

In case that $\theta > \mu$ one can transform the problem as follows: Replace Q by $-Q$ and X by $-X$; interchange Y with U and μ with θ. The proposition is then valid for the transformed problem, and therefore asserts about the original problem that (6) and (7) are valid, and that (11) is valid with an inverted inequality. $\square$

We next show that the solution to the control problem can be characterized as follows.

Proposition 2 *The solution (Q^*, Y^*, U^*) of Proposition 1 uniquely solves (5) and*

$$\begin{cases} \int_0^{\cdot} 1_{Q \geq 0} dY = 0, \\ \\ \int_0^{\cdot} 1_{Q \leq 0} dU = 0, \end{cases} \tag{15}$$

given that $\theta \neq \mu$ and $X \in C$.

Proof: It follows from Proposition 1 that (Q^*, Y^*, U^*) solves (5) and (15). Let (Q, Y, U) satisfy both (5) and (15). Then it follows from (13) that for $t > s$, $\eta(t) - \eta(s) = -\int_s^t 1_{Q=0} dY$ and also that $\eta(t) - \eta(s) \geq 0$. Therefore $\eta = 0$ and Q must satisfy equation (8). As discussed before, this equation has a unique solution, hence $Q = Q^*$. Having $\theta \neq \mu$, U and Y are now uniquely determined by the first two lines of (5) as $U = (\mu - \theta)^{-1}(Q - X + \mu \int_0^{\cdot} Q)$ and $Y = (\mu - \theta)^{-1}(Q - X + \theta \int_0^{\cdot} Q)$. $\square$

4 Reduction of the control problem to one dimension

We show that under special assumptions on the parameters it is possible to reduce the dimensionality of the problem, and obtain pathwise minimum for $Q^c = c_1 Q_1 + c_2 Q_2$. We assume

$$\mu \equiv \mu_1 = \mu_2 > \theta \equiv \theta_1 = \theta_2.$$

Assume without loss that $c_1 > c_2$. Consider the processes $\tilde{Q} = Q_1 + Q_2$, $\tilde{X} = X_1 + X_2$ and $\tilde{U} = U_1 + U_2$. Write

$$Q^c(t) = (c_1 - c_2) Q_1(t) + c_2 \tilde{Q}(t).$$

Pathwise minimality for Q^c will be obtained by a control that achieves simultaneously pathwise minimality for Q_1 and for $\tilde{Q}$. From the statement of the BCP (4) it follows that the following relations must be satisfied

$$\begin{cases} \tilde{Q}(t) & = & \tilde{X}(t) + \mu I(t) - \theta \tilde{U}(t), \\ \\ \tilde{U}(t) & = & \int_0^t \tilde{Q}(s) ds + I(t), \\ \\ \tilde{U},\ I \text{ are nondecreasing.} \end{cases} \tag{16}$$

Proposition 1 (see also remark (b)) shows that a minimal pathwise $\tilde{Q}$ exists, under the constraints specified in (16). It is given as the unique solution to

$$\tilde{Q}(t) = \tilde{X}(t) + \mu \int_0^t \tilde{Q}^-(s)ds - \theta \int_0^t \tilde{Q}^+(s)ds, \tag{17}$$

while $\tilde{U}$ and I are given by

$$\tilde{U}(t) = \int_0^t (\tilde{Q}(s))^+ ds, \quad I = \int_0^t (\tilde{Q}(s))^- ds. \tag{18}$$

Note that the set of constraints specified in (16) is a subset of that in (4). Hence, if we can find U_1, U_2, Y_1, Y_2 satisfying (4), and at the same time $U_1 + U_2 = \tilde{U}$, $Y_1 + Y_2 = I$, where $\tilde{U}$ and I are as in (18), then (17) will also serve as a pathwise minimal $\tilde{Q}$ for (4). The choice that we make is to let $U_1(t) = 0$. With this, U_1 and $U_2 = \tilde{U}$ and I automatically are nondecreasing. Hence the constraints of (4) are all satisfied, and $\tilde{Q}$ of (17) is minimal for (4). To see that Q_1 is minimal as well, note that by (4), Q_1 is given by

$$Q_1(t) = X_1(t) - \mu \int_0^t Q_1(s)ds + \eta(t),$$

where $\eta(t) = (\mu - \theta) \int_0^t U_1(s)ds \geq 0$ for all t. By monotonicity of the solution to the equation in the last display with respect to η, Q_1 is minimized by $\eta = 0$. This is achieved by $U_1 = 0$. As a result, Q_1 is minimal, and since also $\tilde{Q}$ is minimal, so is Q^c.

References

[1] Armony M. and Maglaras, C., *Customer contact centers with multiple service channels,* preprint

[2] Atar, R., Mandelbaum A. and Reiman, M. *Scheduling a multi-class queue with many i.i.d. servers: asymptotic optimality in heavy-traffic.* In preperation.

[3] Bell, S. L. and Williams, R. J., *Dynamic Scheduling of a System with Two Parallel Servers in Heavy Traffic with Complete Resource Pooling: Asymptotic Optimality of a Continuous Review Threshold Policy*, preprint

[4] Birkhoff, Garrett; Rota, Gian-Carlo, *Ordinary differential equations.* Fourth edition. Wiley, New York, 1989.

[5] Garnett, O., Mandelbaum, A and Reiman, M., *Designing a call center with impatient customers*, preprint

[6] Halfin, Shlomo and Whitt, Ward., *Heavy-traffic limits for queues with many exponential servers.* Oper. Res. 29 (1981), no. 3, 567–588.

[7] Harrison, J. Michael, *Brownian motion and stochastic flow systems.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Wiley, Inc., New York, 1985.

[8] Harrison, J. Michael, *Brownian models of queueing networks with heterogeneous customer populations.* Stochastic differential systems, stochastic control theory and applications (Minneapolis, Minn., 1986), 147–186, IMA Vol. Math. Appl., 10, Springer, New York, 1988

[9] J. M. Harrison and A. Zeevi. Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. preprint.

[10] Mandelbaum, A., Massey, W.A., Reiman, M., Rider, B. and Stolyar, A. *Queue length and waiting times for multiserver queues with abandonment and retrials*, Proceedings of the Fifth INFORMS Telecommunications Conference, 2000.

[11] Mandelbaum, A., Massey, W.A., Reiman, M. and Stolyar, A. *Waiting time asymptotics for the time varying multiserver queue with abandonment and retrials*, Allerton Conference Proceedings, 1999.

[12] Puhalskii, A. A.; Reiman, M. I. *The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime.* Adv. in Appl. Probab. 32 (2000), no. 2, 564–595

[13] Williams, R. J., *On dynamic scheduling of a parallel server system with complete resource pooling.* Analysis of communication networks: call centres, traffic and performance (Toronto, ON, 1998), 49–71, Fields Inst. Commun., 28, Amer. Math. Soc., Providence, RI, 2000.