

e - c o m p a n i o n

ONLY AVAILABLE IN ELECTRONIC FORM

Electronic Companion—"Service-Level Differentiation in Call Centers with Fully Flexible Servers" by Itay Gurvich, Mor Armony, and Avishai Mandelbaum, *Management Science*, doi 10.1287/mnsc.1070.0825.

Service Level Differentiation in Call Centers with Fully Flexible Servers:

Technical Appendix

Itay Gurvich¹

Mor Armony²

Avishai Mandelbaum³

In this technical appendix we provide proofs for the various results stated in the manuscript titled: “Service Level Differentiation in Call Centers with Fully Flexible Servers”.

The notational convention is as follows: for any stochastic process $B(\cdot)$, $B(t)$ corresponds to the value of the process at time t . $B(\infty)$ and B denote interchangeably the process in steady-state, and $B(\cdot)$ denote the entire process. To relate this appendix to the main body of the paper one should note that Proposition 5.1 is a summary of Propositions 10.4 and 10.5. Proposition 5.2 is a summary of Propositions 11.1 and 11.2.

We start with the proof of Theorem 6.1 which assumes that all of the propositions that were stated in the body of the paper hold. These propositions, in turn, will be established in the subsequent sections.

9. Asymptotic Optimality

Proof of Theorem 6.1

Recall that for every fixed r , the problem formulation (16) is of the form

$$\begin{aligned} & \text{minimize } N \\ & \text{subject to } E[W^r] \leq T^r \\ & \quad P\{W_i^r > T_i^r\} \leq \alpha_i, i = 1, \dots, J-1, \\ & \quad N \in \mathbb{Z}_+, \pi \in \Pi \end{aligned} \tag{A1}$$

And consider the relaxed problem,

¹ Graduate School of Business, Columbia University, ig2126@columbia.edu

² Stern School of Business, New York University, marmony@stern.nyu.edu

³ Faculty of Industrial Engineering and Management, Technion Institute of Technology, avim@ie.technion.ac.il

$$\begin{aligned} & \text{minimize } N \\ & \text{subject to } E[W^r] \leq T^r \\ & \quad N \in \mathbb{Z}_+, \pi \in \Pi \end{aligned} \tag{A2}$$

Let $\underline{N}^r$ be the optimal solution for (A2) in the r^{th} system. Letting N^{*r} be the optimal solution for (A1) in the r^{th} system, we clearly have that for all $r \geq 0$, $\underline{N}^r \leq N^{*r}$. Now, consider a FCFS $M/M/N$ system with the same values of λ^r and μ , let $W_{\lambda^r, \mu}^{FCFS}(N)$ be defined as before and consider the single class optimization problem

$$\begin{aligned} & \text{minimize } N \\ & \text{subject to } E[W_{\lambda^r, \mu}^{FCFS}(N)] \leq T^r \\ & \quad N \in \mathbb{Z}_+ \end{aligned} \tag{A3}$$

Let $\underline{\underline{N}}^r$ be the solution to (A3). Then we claim that for all $r \geq 0$ $\underline{\underline{N}}^r = \underline{N}^r$. To see that this is indeed the case it suffices to note that by Little's Law the problem given by equation (A2) is equivalent to

$$\begin{aligned} & \text{minimize } N \\ & \text{subject to } E[Q^r] \leq \lambda^r T^r \\ & \quad N \in \mathbb{Z}_+ \end{aligned} \tag{A4}$$

and by the same argument the problem given in equation (A3) is equivalent to

$$\begin{aligned} & \text{minimize } N \\ & \text{subject to } E[Q_{\lambda^r, \mu}^{FCFS}(N)] \leq \lambda^r T^r \\ & \quad N \in \mathbb{Z}_+ \end{aligned} \tag{A5}$$

Now, since we have a common service rate μ it is straightforward to show that the overall queue length is minimized by any work conserving policy and in particular by FCFS, so that the two problems (A4) and (A5) are equivalent and so are, in turn, the problems (A2) and (A3).

Overall, we have shown that $\underline{\underline{N}}^r \leq N^{*r}$, so that the $M/M/N$ staffing problem constitutes a lower bound on the optimal staffing level in (A1). The argument would be complete if we can show that $\underline{\underline{N}}^r$ is asymptotically feasible, and thus, since it is also a lower bound, it is necessarily asymptotically optimal. To establish asymptotic feasibility, note that by Proposition 5.2

$$\frac{E[W^r]}{E[W_{\lambda^r, \mu}^{FCFS}(\underline{\underline{N}}^r)]} \rightarrow 1, \tag{A6}$$

so that the global constraint is asymptotically satisfied using ITP and SCS. This however, does not guarantee the asymptotic feasibility of $\underline{\underline{N}}^r$ and ITP since we have to show that the individual

SL constraints for classes $i = 1, \dots, J - 1$ are satisfied. Note, however, that by Propositions 5.1 and 5.2

$$P\{W_i^r > T_i^r\} \approx P\{W_{\lambda^r, \mu}^{FCFS}(\underline{N}^r) > 0\} \prod_{j=1}^{J-1} (\sigma_j^r)^{K_{j+1}^r - K_j^r} \bar{F}(\underline{N}^r T_i^r; \sigma_i^r, \sigma_{i-1}^r) \quad (\text{A7})$$

In particular, by choosing the thresholds through (18) the constraints are asymptotically satisfied. To see that the solution suggested in Remark 2.1 is also asymptotically feasible, it suffices to note that by Markov's inequality

$$P\{W_i^r > T_i^r\} \leq P\{W_i^r > 0\} \frac{E[W_i^r | W_i^r > 0]}{T_i^r}. \quad (\text{A8})$$

Using the convergence of $N^{*r} E[W_i^r | W_i^r > 0]$ that is given in Proposition 5.2 then shows that the thresholds determined by 2.1 are asymptotically feasible using the lower bound staffing level and in turn asymptotically optimal. ■

10. Performance Analysis of ITP and SCS

For simplicity of presentation, we chose to prove most of the results for $\gamma = 1/2$. The proofs for $\gamma \in (0, 1/2)$ are given in Section 12. The results for arbitrary $\gamma > 1/2$ are substantially simpler to establish and the proofs are omitted. The analysis consists of several steps. The limits of the steady state performance measures for the ITP scheduling policy are obtained by first examining the diffusion limits for the entire stochastic process. Then, using tightness arguments we deduce the convergence of the steady state distributions. Consequently, the analysis is presented in two subsections: Subsection 10.1 below establishes a functional central limit theorem (FCLT) for the overall number of customers in system under the ITP scheduling policy. As corollaries we obtain convergence for the queue lengths and waiting times of the different classes.

Section 10.2 focuses on steady state analysis. In subsection 10.2.1 we give a simple set of necessary conditions and a set of sufficient conditions for existence of steady state under the ITP policy. These conditions are less tight than the conditions of [15] but they are much simpler to check and provide insight on the system behavior under the ITP policy. Then, based on the diffusion limits and using some tightness arguments we prove the convergence of the steady state overall number

of customers in system. As corollaries we obtain convergence of the probability of delay for the lowest priority class J . Proposition 10.5 then gives asymptotic expressions for the probabilities of delay of the higher priority classes $1, \dots, J-1$.

Before we proceed with the proof of our main results we need the following lemma which is a summary of some of the results in Section 9 of Borst et. al. [3]. To this end, let

$$\rho^r = \frac{\lambda^r}{N^r \mu}.$$

Lemma 10.1 *Consider the sequence λ^r and the sequence of staffing levels N^r determined through SCS. Then, $N^r - R \approx \beta^r \sqrt{R}$, where β^r is the unique solution to*

$$\alpha_\gamma(\beta^r) \frac{1}{\beta^r \mu \sqrt{R}} = T^r, \quad (\text{A9})$$

with $\alpha_\gamma(\cdot)$ as given in (34). Hence, one may write

$$\sqrt{N^r}(1 - \rho^r) \approx \beta^r > 0, \quad (\text{A10})$$

and in particular, if $\gamma = 1/2$, we have that

$$\sqrt{N^r}(1 - \rho^r) \rightarrow \beta > 0. \quad (\text{A11})$$

In addition to ρ^r we define $\rho_C^r := \frac{\lambda^r}{\mu(N^r - K_j^r)}$. Note that whenever $K_j^r \ll \sqrt{r}$, $\sqrt{N^r}(1 - \rho^r) \rightarrow \beta > 0$ if and only if $\sqrt{N^r}(1 - \rho_C^r) \rightarrow \beta > 0$.

10.1. Diffusion Limits

Consider a sequence of $M/M/N^r/\{K_i^r\}$ systems indexed by $r = R$. Let $A_j^r(t) : j = 1, \dots, J$ be the total number of arrivals into class j up to time t (i.e. a $\text{Poisson}(\lambda_j)$ process). Due to Functional Strong Law of Large Numbers (FSLLN) and FCLT we have

$$\frac{1}{N^r} A_j^r(t) \Rightarrow \hat{\lambda}_j t, \quad (\text{A12})$$

where $\hat{\lambda}_j = \lim_{r \rightarrow \infty} \frac{\lambda_j^r}{N^r}$, $j = 1, \dots, J$, and

$$\frac{1}{\sqrt{N^r}}(A_j^r(t) - \lambda_j^r t) \Rightarrow BM(0, \hat{\lambda}_j), \quad (\text{A13})$$

where $BM(0, \hat{\lambda}_j)$ is Brownian motion with drift 0 and infinitesimal variance $\hat{\lambda}_j$. Also, let $Z^r(t)$ and $Q_i^r(t)$, $i = 1, \dots, J$ be, respectively, the overall number of busy agents and the number of class i ($i = 1, \dots, J$) customers in queue at time t in the r^{th} system. Then,

$$Y^r(t) = Z^r(t) + \sum_{i=1}^J Q_i^r(t) \quad (\text{A14})$$

is the total number of customers in the r^{th} system at time t . Finally, for $r = 1, 2, \dots$ define the centered and scaled process

$$X^r(t) = \frac{Y^r(t) - (N^r - K^r)}{\sqrt{N^r}}. \quad (\text{A15})$$

Proposition 10.1 Assume (15), $\sqrt{N^r}(1 - \rho_C^r) \rightarrow \beta > 0$, and $X^r(0) \Rightarrow X(0)$, where $\Rightarrow$ stands for weak convergence. Then

$$X^r(\cdot) \Rightarrow X(\cdot), \quad (\text{A16})$$

where X is a diffusion process with infinitesimal drift given by

$$m(x) = \begin{cases} -\beta\mu & x \geq 0 \\ -(\beta + x)\mu & x \leq 0 \end{cases} \quad (\text{A17})$$

and state independent infinitesimal variance $\sigma^2 = 2\mu$.

Remark 10.1 By [7] the same limit is obtained for a sequence of $M/M/N^r - K_J^r$ systems with $N^r = R + \beta\sqrt{R}$. Hence, the process of the overall number of customers in the $M/M/N^r/\{K^r\}$ system is approximately equal in law to the number of customers in the associated $M/M/N^r - K_J^r$ system. In particular, if $K_J^r \ll \sqrt{r}$, the overall number of customers in system is approximately equal in probability law to the number of customers in the associated $M/M/N^r$ system.

Proof: For simplicity we prove the proposition for a system with $J = 2$. The proof is similar for arbitrary number of classes as will be explained at the end. The proof consists of two steps: In the

first step we introduce another system (denoted by (B)) which is equivalent in law to the original $M/M/N/\{K_i\}$ system (denoted by (A)). In the second step we use a coupling argument and the convergence together theorem (Theorem 11.4.7 in [16]) to conclude the proof.

Definition of systems B and C:

Consider the original server pool of N servers. Split the server pool into two distinct pools: one with $N^r - K^r$ servers and the other with K^r servers, where $K^r = K_J^r$. Throughout the proof we will denote these two pools by “the $N - K$ Pool” and “the K pool” respectively.

In system B the following routing policy is used: as long as the total number in system is below $N - K$ route all customers to the $N - K$ pool. When there are more than $N - K$ busy servers route any arriving high priority customer to the K pool. Upon a service completion, if there are any customers in service in the K pool, preempt one of these customers and assign him/her to the server that was just released in the $N - K$ pool. Since we have a common μ for all priority classes, systems (A) and (B) can be coupled so that the total number in system process will have the same sample paths and the same probability law. Thus, proving the weak convergence of (B) will result in the desired weak convergence for (A).

Finally, let us further introduce a System C which is an $M/M/m$ queue with the same arrival and service rates as System B and with $m = N - K$ servers.

Denote by $Y_B^r(t)$ the total number in system process for system (B) and by $Y_C^r(t)$ the total number in system for system C. Also, denote by $Z_{K,B}^r(t)$ the number of busy servers from the K pool in system B. As before, define

$$X_B^r(t) = \frac{Y_B^r(t) - (N^r - K^r)}{\sqrt{N^r}} \quad (\text{A18})$$

and

$$X_C^r(t) = \frac{Y_C^r(t) - (N^r - K^r)}{\sqrt{N^r}} \quad (\text{A19})$$

By our assumption that $\lim_{r \rightarrow \infty} \sqrt{N^r}(1 - \rho_C^r) = \beta$, $0 < \beta < \infty$ we have from [7] that $X_C^r \Rightarrow X$.

Coupling:

Next we discuss the coupling of systems

(B) and (C). We will show that these two systems can be coupled so that the distance (in the *sup* norm) between them is bounded by an expression that converges to zero as $r \rightarrow \infty$. Having that, the proposition will follow by the convergence together theorem. In the following paragraphs we fix $r > 0$ and eliminate the superscript from the notation.

The coupled sample paths are described as follows: We use the same sample path of arrivals for both systems. For simplicity let us assume that both systems are initiated with $N - K$ busy servers and an arrival of a customer. As long as $Y_B(t) > N - K$ and $Y_C(t) > N - K$ we can generate the departures for system C and for the $N - K$ pool of system B from a common Poisson process with rate $(N - K)\mu$. System B will also have departures from the K pool generated by an independent Poisson process. During the time that both system are above $N - K$ the difference between them can be at most as the number of departures due to service completions (and not preemption) from the K pool.

Now, assume that system B goes below $N - K$. We will continue to generate the departures for system C and for the $N - K$ pool from the same Poisson process but with a thinning (as in [19]). i.e. If system B has a customer count of j at a departure epoch and system C has l customers, than the candidate departure event generated from the Poisson process with rate $l\mu$, is an actual departure for system B with probability j/l (recall that $j \leq l$). During the epoch in which system B is below $N - K$ the distance between the two systems in consideration can only decrease. If the two systems meet they will proceed together until they hit $N - K$ for the first time.

Denote by $D_K^r(T)$ the departures from the K pool up to time T . Then, we can write (see for example [9])

$$D_k(T) = \mathcal{N} \left(\int_0^T Z_{K^r, B}^r(\tau) \mu d\tau \right) \quad (\text{A20})$$

Where, $\mathcal{N}$ is a Poisson process with rate 1.

By the construction of the sample paths we have that for all $T \geq 0$ the distance between the two systems can be bounded by the number of departures from the K pool up to that time. More formally, for the r^{th} system we have

$$\sup_{0 \leq t \leq T} \|Y_B^r(t) - Y_C^r(t)\| \leq \mathcal{N} \left(\int_0^T Z_{K^r, B}^r(\tau) \mu d\tau \right) \quad (\text{A21})$$

or,

$$\sup_{0 \leq t \leq T} \|X_B^r(t) - X_C^r(t)\| \leq \frac{1}{\sqrt{N^r}} \mathcal{N} \left(\int_0^T Z_{K^r, B}^r(\tau) \mu d\tau \right) \quad (\text{A22})$$

Provided that

$$\frac{1}{\sqrt{N^r}} \mathcal{N} \left(\int_0^T Z_{K^r, B}^r(\tau) \mu d\tau \right) \Rightarrow 0, \quad (\text{A23})$$

and applying the convergence together theorem the proposition follows.

To establish (A23) it is enough to show that for each r , $Z_{K^r, B}(t) + Q_1^r(t)$ can be path wise bounded by an $M/M/1$ queue with arrival rate $\lambda^r = \lambda_1^r$ and with service rate $(N^r - K^r)\mu$. This is shown in the following way: Assume we initiate both systems by zero. Every jump up in $Z_{K^r}^r(t) + Q_1^r(t)$ is necessarily a jump up in the associated $M/M/1$. The opposite is not correct since if more then K^r servers are idle an arrival of high priority will not result in an increase in $Z_{K^r, B}^r(t) + Q_1^r(t)$. Assume that at time $t \geq 0$ both systems are not empty (in particular assume that $Z_{K^r, B}(t) + Q_1^r(t) = j > 0$). In particular, the time until the next departure is exponential with rate $(N^r - K^r + j)\mu$. Then, as before, we will use thinning - every service completion in $Z_{K^r, B}^r(t) + Q_1^r(t)$ will result in a service completion in the $M/M/1$ with probability $\frac{N^r - K^r}{N^r - K^r + j}$. Thus we have proved that for all $t \geq 0$, $Z_{K^r}^r(t) + Q_1^r(t)$ can be path wise bounded by the associated $M/M/1$.

By (15) this $M/M/1$ is under-loaded and by Theorems 4.1 and 4.2 of [9]

$$\frac{1}{\sqrt{N^r}} \int_0^T Z_{K^r, B}^r(\tau) \mu d\tau \Rightarrow 0. \quad (\text{A24})$$

Since the Poisson process $\mathcal{N} \left(\int_0^T Z_{K^r, B}^r(\tau) \mu d\tau \right)$ admits the decomposition (see for example [14])

$$\mathcal{N} \left(\int_0^T Z_{K^r, B}^r(\tau) \mu d\tau \right) = \int_0^T Z_{K^r, B}^r(\tau) \mu d\tau + M^r(T) \quad (\text{A25})$$

where M^r is a martingale with quadratic variation function that is bounded by $K^r t$, we have the desired result. Thus, we have established the convergence (A16). To prove the result for a general number of classes one would proceed in a similar way to the two class case. Particularly, the K pool will only serve higher priority customers (this time with thresholds). Finally, one would need

to show that $Z_{K^r, B}^r(t) + \sum_{i=1}^{J-1} Q_i^r(t)$ can be bounded by an under-loaded $M/M/1$ queue and hence the proof follows. ■

Corollary 10.1 *Let $X(\cdot)$ be the diffusion process described in Proposition 10.1. Then the steady-state distribution of $X(\infty)$ has a density $f(\cdot)$ which satisfies:*

$$f(x) = \begin{cases} \exp\{-\beta x\} \alpha(\beta), & x \geq 0 \\ \frac{\phi(\beta+x)}{\Phi(\beta)} (1 - \alpha(\beta)), & x < 0 \end{cases} \quad (\text{A26})$$

where $P\{X(\infty) > 0\} = \alpha(\beta)$.

Proof: This result follows directly from [7]. ■

A consequence of Proposition is that $X^r(t)$ (the scaled and normalized process of the overall number of customers in system) becomes approximately sufficient to describe the asymptotic behavior of the J dimensional process $(Z(\cdot) + Q_1(\cdot), Q_2(\cdot), \dots, Q_J(\cdot))$. This state space collapse property of the $M/M/N/\{K_i\}$ model is summarized by the following corollary where we set $K^r \equiv K_J^r$.

Corollary 10.2 (State Space Collapse) *Denote by $\mathcal{E}^r(t)$ the number of busy servers above the level of $N^r - K^r$, i.e. $\mathcal{E}^r(t) = [Z^r(t) - (N^r - K^r)]^+$. Then*

$$\begin{aligned} \frac{1}{\sqrt{N^r}} \mathcal{E}^r(\cdot) &\Rightarrow 0 \\ \frac{1}{\sqrt{N^r}} Q_i^r(\cdot) &\Rightarrow 0, \forall i \leq J-1 \\ \frac{1}{\sqrt{N^r}} Q_J^r(\cdot) &\Rightarrow X^+(\cdot) \end{aligned} \quad (\text{A27})$$

Proof: Note that $\mathcal{E}^r(t) + Q_1^r(t)$ is just $Z_K^r(t) + \sum_{i=1}^{J-1} Q_i^r(t)$, hence the result follows from the proof of Proposition 10.1. ■

The next corollary shows how to obtain the limit of the virtual waiting time for class J as a function of the limit queue length process X , where ξ_J is as defined in (15).

Corollary 10.3 *Let $W_i^r(\cdot)$ be the virtual waiting time process for class i . Then, if there exists $-\infty < c < \infty$, such that*

$$\sqrt{N} \left(\frac{\lambda_J^r}{N^r} - \xi_J \mu \right) \rightarrow c, \quad (\text{A28})$$

then

$$\sqrt{N^r} W_J^r(\cdot) \Rightarrow \frac{1}{\xi_J \mu} [X]^+(\cdot). \quad (\text{A29})$$

Proof: By the FCLT for the arrivals and by (A28) we have the convergence

$$V^r(t) = \sqrt{N^r} \left(\frac{A_J^r(t)}{N^r} - \xi_J \mu t \right) \Rightarrow V(t), \quad (\text{A30})$$

where $V(t) = \hat{A}(t) + ct$ and $\hat{A}$ is a $BM(0, \hat{\lambda}_J)$. Define $\hat{Q}^r(\cdot) = \frac{1}{\sqrt{N^r}} Q_J^r(\cdot)$. Then, by corollary 10.2 we have that $\hat{Q}^r(\cdot) \Rightarrow [X]^+(\cdot)$.

The convergence of $V^r(\cdot)$ and $\hat{Q}^r(\cdot)$ does not necessarily imply the joint convergence of $(V^r(\cdot), \hat{Q}^r(\cdot))$. However, we claim that this component-wise convergence is sufficient for our purposes.

By Theorem 11.6.7 in [16], and by the convergence of $V^r(\cdot)$ and $\hat{Q}^r(\cdot)$ we have the tightness of the sequence $(V^r(\cdot), \hat{Q}^r(\cdot))$. Hence, by Prohorov's Theorem (Theorem 11.6.1 in [16]) we have that there exists a convergent subsequence $\{r_k\}$ for which

$$(V^{r_k}(\cdot), \hat{Q}^{r_k}(\cdot)) \Rightarrow (\hat{V}(\cdot), \hat{Q}(\cdot)), \quad (\text{A31})$$

for some process $(\hat{V}(\cdot), \hat{Q}(\cdot))$. Define $U^r(t) = \sqrt{N^{r_k}} \left(\frac{D_J^{r_k}(t)}{N^{r_k}} - \xi_J \mu t \right)$. Then, using the relation

$$Q_J^{r_k}(t) = Q_J^{r_k}(0) + A_J^{r_k}(t) - D_J^{r_k}(t), \quad (\text{A32})$$

or, alternatively,

$$U^{r_k}(t) = V^{r_k}(t) + Q_J^{r_k}(0) - Q_J^{r_k}(t), \quad (\text{A33})$$

and applying the continuous mapping theorem we have the convergence

$$(U^{r_k}(\cdot), V^{r_k}(\cdot)) \Rightarrow (\hat{U}(\cdot), \hat{V}(\cdot)), \quad (\text{A34})$$

where $\hat{U}(\cdot) = \hat{V}(\cdot) - \hat{Q}(\cdot)$. Since $U(\cdot)$ and $V(\cdot)$ are continuous with $U(0) = 0$ we can apply the corollary of [13] to obtain for the subsequence

$$\sqrt{N^{r_k}} W^{r_k}(\cdot) \Rightarrow W(\cdot), \quad (\text{A35})$$

where $W(t) = \frac{\hat{Q}(t)}{\xi_J \mu}$. Since the limit $\hat{Q}(\cdot)$ is independent of the subsequence chosen (and equal to $[X]^+(\cdot)$) we have the desired result. ■

10.2. Steady State Analysis

10.2.1. Stability Conditions To discuss steady state convergence, we first must address the question of stability, i.e. what are the conditions under which a steady state distribution exists as a proper random variable. For fixed parameters these conditions can be explicitly calculated using the formulae in [15]. However, these formulae are very complicated for calculation even for a simple two class system. Therefore we find the following theorem useful. In the theorem we use the notation λ_{jc}^r for the arrival rate of the “super class” consisting of classes $1, \dots, J-1$, i.e. $\lambda_{jc}^r = \sum_{i=1}^{J-1} \lambda_i$. Also, we denote by δ^r the probability of abandonment given wait in an $M/M/1+M$ system with arrival rate λ_{jc}^r , service rate $(N^r - K^r)\mu$ and abandonment rate μ . We denote by $\rho_{C,<J}^r$ the nominal load in this single server queue. i.e. $\rho_{C,<J}^r = \frac{\lambda_{jc}^r}{(N^r - K^r)\mu}$.

For the second part of the stability Proposition 10.2 we assume some regularity conditions on the threshold level K^r . In particular we assume that there exists a number $a \in [0, \infty)$, such that

$$\frac{\lambda^r}{R^r - K^r} \rightarrow a. \quad (\text{A36})$$

This condition is guaranteed to hold if $K^r = O(\sqrt{N^r})$. We say that a system is stable if there exists a unique stationary distribution.

Proposition 10.2 *Under assumption (15) we have that:*

1. *Fix r and assume $K^r > 0$. Then:*

(a) *The threshold system is stable if $\lambda^r < (N^r - K^r)\mu$.*

(b) *The system is unstable whenever $\lambda_j^r > (N^r - K^r)\mu - \lambda_{jc}^r \cdot \delta^r$.*

2. *Assume that $N^r = R^r + \Delta^r$ where $\Delta^r = o(R^r)$. Also, assume (A36). Then,*

(a) *If $K^r \neq o(N^r)$, there exists $r_1 > 0$ such that $\forall r > r_1$ the system is unstable .*

(b) *Otherwise, if $K^r = o(N^r)$, let $r_1 = \max\{r > 0 : \rho_{C,<J}^r \geq 1\}$. Then, for all $r > r_1$, $\delta^r \leq$*

$\frac{1}{(N^r - K^r)(1 - \rho_{C,<J}^r)}$, and in particular stability requires that $K^r \leq \Delta^r + O(1)$.

If $K^r \equiv 0$ (static priority), Condition 1.(a) is necessary and sufficient.

Remark: The advantage of writing stability conditions using δ^r is that δ^r has a known formula which can be also computed using existing software such as [20].

Proof: $Y^r(t)$ is not a Markovian process. However, proving that the state $N^r - K^r$ of Y^r is positive recurrent implies that the state $(Z^r + Q_1^r = N^r - K^r, Q_i^r = 0 : i = 2, \dots, J)$ of the underlying Markov process is positive recurrent. Also, the underlying Markov process is clearly irreducible and hence proving the positive recurrence of this state is sufficient for stability (see for example chapter 10 of [12]).

First, note that if $K^r \equiv 0$ then the result is straightforward. In this case the policy is work conserving policy and the sum process is the same Birth and Death process that describes the regular $M/M/N$ system.

Assume $K^r > 0$. For the sufficient conditions it is enough to use the coupling used for (10.1). It is clear that if the $M/M/N^r - K^r$ is stable then so is the threshold system which, by the construction in Proposition 10.1, is path wise dominated by the $M/M/N^r - K^r$ system.

For the necessary conditions we build a static priority system with abandonment and show that if it is unstable then the corresponding $M/M/N/\{K_i\}$ system is also unstable. Denote by S a static priority system with $N^r - K^r$ servers. All classes except for the lowest priority class J have a finite exponential patience with rate μ and class J has an infinite patience. Denote by $Y_S^r(t)$ the total number of customers in this system. Note that a system in which none of the customers of priorities $1, \dots, J-1$ wait before entering service (i.e. there is an infinite number of servers available to serve priorities $1, \dots, J$, and only $N - K$ are available to server class J) is equal in law to system S . Clearly the latter system outperforms the original system, and hence one can easily construct both systems from the same sample paths and have that for all $t \geq 0$, $Y^r(t) \geq Y_S^r(t)$. Hence, if $Y_S^r(t) \rightarrow \infty$ as $t \rightarrow \infty$ then $Y^r(t) \rightarrow \infty$ as $t \rightarrow \infty$. Hence, in the remaining of the proof we focus on the stability of system S .

System S can be modelled as a multi-dimensional Markov process with the coordinates $(Z^r + Q_1^r, Q_i^r = 0 : i = 2, \dots, J)$ where the notations have the same meaning as before. Let us look at this multidimensional when it is restricted to the states in which all $N^r - K^r$ servers are busy. The

restriction is formally obtained via a time-change argument, as customary in Markov Processes. See, for example, Chapter VII of [2]). Under this restriction the number of customers from the super class $(1, \dots, J-1)$ in this restricted process can be modelled by a Markov process, with the same law as an $M/M/1 + M$ queue. Hence, it has a unique stationary distribution.

Let δ^r to be the steady state probability of abandonment in this restricted process. This, in turn is equal to the probability of abandonment given wait in an $M/M/1 + M$ queue with arrival rate λ_J^c , service rate $(N^r - K^r)\mu$ and abandonment rate μ . The latter has a known formulae. As before, proving positive recurrence of Y_S^r is sufficient for the stability of the underlying multi-dimensional Markov process.

Thus, a trivial necessary condition for stability of system S is that

$$\lambda_J^r + \lambda_{J^c}^r(1 - \delta^r) \leq (N^r - K^r)\mu \quad (\text{A37})$$

Assume now that $K^r = o(N^r)$. Then, by (15) we have that there exists r_1 such that $\rho_{C, < J}^r < 1$ for all $r > r_1$. Then, using the identity $\lambda_{J^c}^r P\{Ab\} = \mu E[Q_{< J}^r(\infty)]$ (where $Q_{< J}^r(\infty)$ stands for the steady state queue length of the super class $1, \dots, J-1$), we have that

$$\delta^r = \frac{\mu}{\lambda} E[Q_{< J}^r(\infty) | Z^r(\infty) > N^r - K^r]. \quad (\text{A38})$$

But notice that

$$E[Q_{< J}^r(\infty) | Z^r(\infty) > N^r - K^r] \leq \frac{(\rho_{C, < J}^r)^2}{1 - \rho_{C, < J}^r}. \quad (\text{A39})$$

The latter is straightforward noting that the right side is average queue length of a non-abandonment $M/M/1$ with arrival rate $\lambda_{J^c}^r$ and service rate $(N^r - K^r)\mu$. After some simplification, we have that

$$\delta^r \leq \frac{\rho_{C, < J}^r}{(N^r - K^r)(1 - \rho_{C, < J}^r)} \quad (\text{A40})$$

This expression converges to zero as fast as $1/N^r$ by assumptions (15), (A36) and assuming that $K^r = o(N^r)$. Plugging this upper bound into (A37) results in the necessary condition: $K^r \leq \Delta^r + O(1)$.

It is now only left to consider the case in which $N^r = R^r + \Delta^r$, $\Delta^r = o(R^r)$ and $K^r \neq o(N^r)$. Assume there is a subsequence $\{r_k\}$ such that system S is stable for all $k \geq 1$. Then, we would necessarily have that

$$\lambda_J^{r_k} + \lambda_{J^c}^{r_k}(1 - \delta^r) \leq (N^{r_k} - K^{r_k})\mu$$

Consider two cases:

Case 1: $\lambda_{J^c}^r / (N^r - K^r)\mu \rightarrow \gamma > 1$. In this case, δ^r converges asymptotically to $1 - \frac{1}{\rho_{C, < J}}$ where $\rho_{C, < J} = \lim_{r \rightarrow \infty} \rho_{C, < J}^r$ (see for example [18]). By our assumption that $K^r \neq o(N^r)$, there exists a subsequence r_{k_j} and $0 < c < 1$ such that $\lim_{r_{k_j} \rightarrow \infty} \frac{(N^{r_{k_j}} - K^{r_{k_j}})}{N^r} = c$. For the subsequence r_{k_j} we have that

$$\lim_{j \rightarrow \infty} \frac{1}{N^{r_{k_j}}} (\lambda_J^{r_{k_j}} + \lambda_{J^c}^{r_{k_j}}(1 - \delta^{r_{k_j}})) \leq \lim_{j \rightarrow \infty} \frac{(N^{r_{k_j}} - K^{r_{k_j}})\mu}{N^{r_{k_j}}} \quad (\text{A41})$$

On this subsequence the limiting equation is

$$\hat{\lambda}_J + c\mu \leq c\mu \quad (\text{A42})$$

Which is a contradiction to the non-negligibility of class J assumption (15).

Case 2: $\lambda_{J^c}^r / (N^r - K^r)\mu \rightarrow \gamma \leq 1$. By [18] the probability of abandonment converges to 0 as $r_k \rightarrow \infty$.

Hence we would have that for the sequence r^k the stability equation (A37) can be written as

$$\lambda_J^r + \lambda_{J^c}^r - o(\lambda_{J^c}^r) \leq (N^r - K^r)\mu \quad (\text{A43})$$

or after dividing by μ this can be written as

$$K^r \leq \Delta^r + o(R^r), \quad (\text{A44})$$

which clearly contradicts the assumption on the size of K^r . ■

Define

$$S^r \triangleq X^r(\infty) = \frac{Y^r(\infty) - (N^r - K^r)}{\sqrt{N^r}} \quad (\text{A45})$$

where $Y^r(\infty)$ is the steady state distribution of the sum process in the r^{th} system.

One would expect that the steady state distribution of the diffusion process X of Theorem 10.1 would coincide with the limit of the sequence S^r . This is not immediate since an interchange of limits is involved. More formally, we want to show that

$$P\{X(\infty) \leq x\} \triangleq \lim_{t \rightarrow \infty} \lim_{r \rightarrow \infty} P\{X^r(t) \leq x\} = \lim_{r \rightarrow \infty} \lim_{t \rightarrow \infty} P\{X^r(t) \leq x\} \triangleq \lim_{r \rightarrow \infty} P\{S^r \leq x\} \quad (\text{A46})$$

We will show this in the following proposition where we again let $\rho_C^r = \frac{\lambda^r}{(N^r - K^r)\mu}$.

Proposition 10.3 (Steady State Convergence) *Under the notation above and assuming that*

$$\lim_{r \rightarrow \infty} \sqrt{N^r}(1 - \rho_C^r) = \beta, \quad 0 < \beta < \infty, \quad (\text{A47})$$

the following is true:

$$S^r \Rightarrow X(\infty), \quad (\text{A48})$$

where $X(\infty)$ is the steady state of the diffusion process spelled out in Proposition 10.1. Its distribution is given in Corollary 10.1.

Remark 10.2 *In analogy to Remark 10.1, whenever $K_J^r \ll \sqrt{r}$, Proposition 10.3 shows that the steady state number of customers in the $M/M/N/\{K_i\}$ system is approximately the same as in the associated $M/M/N$.*

Proof: Note that $Y^r(\infty)$ exists as a proper random variable according to Proposition 10.2 and under our choice of the parameters. Following the proof of Theorem 4 in [7] all we have to prove is the tightness of the sequence S^r . Recall systems (B) and (C) from the proof of Proposition 10.1. Then, since $M/M/N/\{K_i\}$ and (B) have the same law, it is enough to prove the tightness of the sequence $S_B^r \triangleq \frac{Y_B^r(\infty) - (N^r - K^r)}{\sqrt{N^r}}$. In addition, we create another coupling of X^r with an $M/M/N^r$ system (denoted by D) for which we define:

$$X_D^r(t) = \frac{Y_D^r(t) - (N^r - K^r)}{\sqrt{N^r}} \quad (\text{A49})$$

System (D) has the same total arrival rate as the $M/M/N/\{K_i\}$ system. We construct it in the same way as the threshold system by splitting the servers into two distinct pools and using the same preemption procedure as in the construction of System (B): For the three $N^r - K^r$ (of systems (B), (C) and (D)) create the departures from the same Poisson processes with thinning. Also for the K pools (in system (B) and (D)) create the departures from the same poisson process with thinning. Define

$$X_D^r(t) = \frac{Y_D^r(t) - (N^r - K^r)}{\sqrt{N^r}} \quad (\text{A50})$$

Clearly, by the same coupling arguments as in the proof of Proposition 10.1 we have path-wise domination $X_D^r(t) \leq X_C^r(t)$. And on the whole we have the path wise ordering

$$X_D^r(t) \leq X_B^r(t) \leq X_C^r(t) \quad \forall t \geq 0 \quad (\text{A51})$$

Define $S_C^r = X_C^r(\infty)$ and $S_D^r = X_D^r(\infty)$, where $X_C^r(\infty)$ and $X_D^r(\infty)$ are the steady state of X_C^r and X_D^r , respectively. We will compare the stationary threshold system with threshold K^r to both single class multi server stationary systems.

Note that since the constructed coupling preserves (A51) for every finite t it does so also for $t \rightarrow \infty$. Also, since under the conditions of the proposition both sequences S_C^r and S_D^r converge as $r \rightarrow \infty$, they are tight. The tightness of S_C^r implies that

$$\forall \epsilon > 0, \exists n_1 : P\{S_C^r \in [-n_1, n_1]\} > 1 - \frac{\epsilon}{2}. \quad (\text{A52})$$

The tightness of S_D^r implies that

$$\forall \epsilon > 0, \exists n_2 : P\{S_D^r \in [-n_2, n_2]\} > 1 - \frac{\epsilon}{2}. \quad (\text{A53})$$

Hence, by the ordering (A51) we have that

$$\forall \epsilon > 0 \exists n_1, n_2 : P\{S^r \in [-n_2, n_1]\} > 1 - \epsilon \quad (\text{A54})$$

With the tightness of $S^r = X^r(\infty)$ we have actually established the proposition. Here is why: Since $X^r(\infty)$ is tight, by Prohorov's Theorem it has a convergent subsequence $X^{r_k}(\infty)$. If we let

$(Z^{r_k}(0) + Q_1^{r_k}(0), Q_i^{r_k}(0) : i = 2, \dots, J)$ be distributed as $(Z^{r_k}(\infty) + Q_1^{r_k}(\infty), Q_i^{r_k}(\infty) : i = 2, \dots, J)$, then $(Z^{r_k}(t) + Q_1^{r_k}(t), Q_i^{r_k}(t) : i = 2, \dots, J)$ is a strictly stationary stochastic process. In particular $\{X^{r_k}(t), t \geq 0\}$ (which is a function of the multidimensional Markov process) is a strictly stationary stochastic process and by Proposition 10.1 we have $X^{r_k}(\cdot) \Rightarrow \hat{X}(\cdot)$, where $\hat{X}(\cdot)$ is the limiting diffusion process with $\hat{X}(0)$ having the stationary distribution of the limit of $X^{r_k}(0)$. However, since $X^{r_k}(\cdot)$ is stationary for each r_k so is the limit $\hat{X}(\cdot)$. Hence the limit of $X^{r_k}(\infty)$ must be the unique stationary distribution of $\hat{X}(\cdot)$. Since every subsequence of $X^{r_k}(\cdot)$ that converges must converge to this same limit, the sequence $X^r(\infty)$ itself must converge to this limit. ■

Corollary 10.4 *Under (15) if $\beta \leq 0$, there is no convergence of the sequence S^r .*

Proof: Let us assume that S^r does converge to a unique and finite limit S and that we start the r^{th} system with its stationary distribution S^r . $X^r(\cdot)$ is thus a stationary process with $X^r(t)$ having the stationary distribution for all $t \geq 0$. By the same arguments as above, and since we assume the convergence of S^r , we should have that $X^r(\cdot)$ converges to a limit $X(\cdot)$ as $r \rightarrow \infty$, and that $X^r(t)$ converges to the stationary distribution of X as $r \rightarrow \infty$.

First let us examine the case where $\beta < 0$: Then, for all M , there exists a subsequence $\{r_k\}$, $r_k > M$ such that $\rho_C^{r_k} > 1$, and by the coupling used in the proof of Proposition (10.1) there is no limit for $X^{r_k}(t)$ (since there is no limit for the corresponding sequence of single class C systems) and the process clearly diverges, contradicting the assumption on the convergence. Otherwise, if $\beta = 0$, we have a limit which is a diffusion process with infinitesimal drift function

$$m(x) = \begin{cases} 0 & x \geq 0 \\ -\mu x & x < 0 \end{cases} \quad (\text{A55})$$

See for example Theorem 4.2 of [9]. This is clearly a non-stationary process which leads to a contradiction to the assumption on the convergence of S^r . ■

For a sequence of $M/M/N$ single class queues, Halfin and Whitt showed the equivalence between

between a square root safety staffing rule and the convergence of the delay probabilities to a non-trivial limit. The following proposition is an analogous version of this equivalence for a sequence of $M/M/N^r/\{K^r\}$ systems.

Proposition 10.4 (Halfin-Whitt Analog) *Consider a sequence of $M/M/N^r/\{K_i^r\}$ systems indexed by $r = 1, 2, \dots$, with service rate μ for all classes and arrival rate λ_i^r for class i , $i = 1, \dots, J$, such that (15) holds. Then,*

$$P\{W_J^r(\infty) > 0\} \rightarrow \alpha_J, \quad 0 < \alpha_J < 1, \quad (\text{A56})$$

if and only if

$$\sqrt{N^r}(1 - \rho_C^r) \rightarrow \beta, \quad 0 < \beta < \infty, \quad (\text{A57})$$

where $\lambda^r = \sum_{i=1}^J \lambda_i^r$, $\rho_C^r = \frac{\lambda^r}{(N^r - K^r)\mu}$. In which case $\alpha_J = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}$, where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal distribution and density functions respectively.

Proof: The ‘if’ part is a direct result of the steady state convergence already proved. For the ‘only if’ part note the following: Since the threshold system is pathwise dominated from above by an $M/M/N^r - K^r$ system we have that, if $\beta = \infty$ then $P\{W_J^r(\infty) > 0\} \rightarrow 0$.

For the case in which $\beta = 0$, let us assume that steady state exists and $P\{W_J^r(\infty) > 0\} \rightarrow \alpha < 1$. Then by the continuity of the function $\alpha(\cdot)$ there exists $\beta' > 0$ such that

$$\alpha < \alpha(\beta') < 1. \quad (\text{A58})$$

We can then construct a threshold system with the same thresholds but with a total number of servers $M^r > N^r$, or more specifically take $M^r = N^r + \beta'\sqrt{N^r}$ to have $\sqrt{M^r}(1 - \rho_C^r) \rightarrow \beta'$. For the new system the ‘if’ direction applies and hence we will have the inequality (A58). Denote by $Y_{M^r}(t)$ the total number of customers in the system with M^r servers. Then, we can easily construct the sample paths such that $Y_{M^r}(t) - (M^r - K^r) \leq Y_{N^r}(t) - (N^r - K^r)$, $\forall t \geq 0$. Hence, we have a contradiction.

There is another case to consider in the ‘only if’ part. It is possible that the sequence $\sqrt{N^r}(1 - \rho_C^r)$ will fail to converge. In that case we would have at least two convergent subsequences converging to two different limits $\beta_1 \neq \beta_2$ (one of which might be ∞). But since the function $\alpha(\cdot)$ is strictly decreasing in its argument we would also have that $\alpha(\beta_1) \neq \alpha(\beta_2)$ and thus the sequence $P\{W_J^r(\infty) > 0\}$ would fail to converge. ■

Having the convergence of the probability of delay of class J , it remains to analyze the probabilities of delay for higher classes. In particular we would like to know what can be said about $P\{W_i^r(\infty) > 0\}$, $i = 1, \dots, J-1$. The answer is given in the following proposition.

Proposition 10.5 *For every $r > 0$ such that $\rho_C^r < 1$.*

$$1 \leq \frac{P\{W_i^r(\infty) > 0\}}{P\{W_J^r(\infty) > 0\} \cdot \prod_{j=i}^{J-1} (\rho_{\leq j}^r)^{K_{j+1}^r - K_j^r}} \leq \left(\frac{N^r}{N^r - K^r} \right)^{K^r}, \quad i = 1, \dots, J-1, \quad (\text{A59})$$

where $\rho_{\leq j}^r = \sum_{i=1}^j \frac{\lambda_i^r}{N^r \mu}$.

In particular, for $K^r = o(\sqrt{N^r})$ and assuming that $\alpha(\beta) > 0$ we have

$$P\{W_i^r(\infty) > 0\} \sim \alpha(\beta) \cdot \prod_{j=i}^{J-1} (\rho_{\leq j}^r)^{K_{j+1}^r - K_j^r}, \quad (\text{A60})$$

where $a_n \sim b_n$ if $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 1$.

Remark 10.3 In the case of $K^r = \Theta(\sqrt{N^r})$ the right bound converges by simple calculus to e^{d^2} where $d = \lim_{r \rightarrow \infty} \frac{K^r}{\sqrt{N^r}}$.

Remark 10.4 Note that the above implies that for any polynomially decreasing probability of delay for classes $i = 1, \dots, J-1$ and using the condition (15), it suffices to use thresholds logarithmically in r . In particular, since we will show that given wait, the waiting time of classes $i = 1, \dots, J-1$ is of order $1/r$, we have that for the formulation (16), it is sufficient to establish all our previous results for the case $K_J^r \ll \sqrt{r}$ and in general with $K_J^r \ll r^\gamma$.

Remark 10.5(Intuitive Explanation of Propositions 10.4 and 10.5) It is easy to understand the results by looking at the dynamics of the suggested policy in a simple two class case, where we let $K := K_2$. To explain (A56) we claim that the probability of delay for the low priority class 2 is approximately the same as the delay probability for a single class $M/M/N - K$ system. Then, it would remain to show that the probability of delay in the $M/M/N$ and $M/M/N - K$ systems are approximately equal. Indeed, one can show that under ITP the threshold K increases at most as a logarithmic function of R while the staffing level N obeys the relation $N = R + \beta R^\gamma$. By Halfin and Whitt [7] the probability of delay in the $M/M/N$ and $M/M/N^r - K_J^r$ will be approximately equal as long as K_J^r is orders of magnitude smaller than R^γ and hence (A56). To see why the delay probability of class J is approximately the same as for a single class $M/M/N - K$ system, we argue that the threshold priority policy is designed to leave almost only low-priority customers in the queue. These customers, in turn, have only $N - K$ servers available to serve them. More specifically, note that whenever less than $N - K$ of the servers are busy the total number of customers in system behaves like a single class $M/M/N - K$ queue. In contrast, whenever more than $N - K$ of the servers are busy, customers of the high priority class are served almost as if they are the only class in a single server queue with service capacity $(N - K)\mu$. By the comparability of the low priority class, this implies that the high priority class faces a light-traffic queue, for which the number of “customers in queue” in this single server queue will be of order $O(1)$ (to be precise, it is $\Theta(1/(1 - \rho_1))$). In particular, for the original system, the number of busy servers is $N - K + O(1)$. Hence, we expect the original system to operate approximately like a system with $N - K$ servers and no thresholds. In particular, we expect that the probability of finding more than $N - K$ busy servers would be approximately the same.

To understand (A60), note again that in the event that more than $N - K$ servers are busy, the high priority in the two class example will be served almost as if they are in a single server queue with capacity $(N - K)\mu$. In turn, their probability of delay (the probability that there are N busy servers) given at least $N - K$ busy servers is approximately equal to the probability of more than K customers waiting in the corresponding single server queue.

Proof: For the two-class case this can be proved by direct approximations of the formulae in [15].

However, we can exploit the structure of the model to prove the desired asymptotic equivalence.

The result is almost immediate using upper and lower bounds.

Let us look at priority class j . Given that class $j+1$ has to wait (i.e. the number of idle servers is smaller or equal to K_{j+1}) - the conditional probability of delay for class j equals to the probability that there would be additional $K_{j+1} - K_j$ busy servers or more.

Now, Let us look at the Markov process of the model restricted to the states in which more than $N^r - K_{j+1}^r$ servers are busy. Define a new process $\tilde{Y}^r(\cdot) = \{\tilde{Z}_j^r(\cdot), \tilde{Q}_1^r(\cdot), \dots, \tilde{Q}_j^r(\cdot)\}$, where $\tilde{Z}_j^r(\cdot)$ describes the number of busy servers above the level of $N^r - K_{j+1}^r$, and $\tilde{Q}_i^r(\cdot)$ is the number of class i customers in queue. Under our restriction $\tilde{Y}^r(\cdot)$ is also a Markov process. Denote its steady state by $\tilde{Y}^r(\infty) = \{\tilde{Z}_j^r(\infty), \tilde{Q}_1^r(\infty), \dots, \tilde{Q}_j^r(\infty)\}$. Also, because of the model structure, the probability in question for can be computed by

$$P\{W_j^r(\infty) > 0\} = P\{W_{j+1}^r(\infty) > 0\} \cdot P\{\tilde{Z}_j^r(\infty) + \sum_{i=1}^j \tilde{Q}_i^r(\infty) \geq K_{j+1}^r\}$$

To justify this, see, for example, Section 10.4 of [12] and the results therein.

Define

$$\pi_s = \sum_{z, q_1, \dots, q_j: z + \sum_{i=1}^j q_i = s} \pi_{z, q_1, \dots, q_j}, \quad s = N - K, \dots, N, \dots$$

to be the probability that the sum of the components of the restricted chain equals s , under its stationary distribution. Then, the cuts method implies for $s \in N - K, \dots$:

$$\begin{aligned} \pi_s \sum_{i=1}^j \lambda_i &\geq \pi_{s+1} (N - K_{j+1}) \mu \geq \pi_{s+1} (N^r - K^r) \mu \\ \pi_s \sum_{i=1}^j \lambda_i &\leq \pi_{s+1} N \mu \end{aligned} \tag{A61}$$

or alternatively

$$\begin{aligned} P\{\tilde{Z}_j^r(\infty) + \sum_{i=1}^j \tilde{Q}_i^r(\infty) \geq K_{j+1}\} &\leq \left(\frac{\sum_{i=1}^j \lambda_i}{(N-K)\mu} \right)^{K_j - K_{j+1}} \\ P\{\tilde{Z}_j^r(\infty) + \sum_{i=1}^j \tilde{Q}_i^r(\infty) \geq K_{j+1}\} &\geq \left(\frac{\sum_{i=1}^j \lambda_i}{N\mu} \right)^{K_j - K_{j+1}} \end{aligned} \tag{A62}$$

By induction, we have proved the desired result. By simple Taylor expansion the upper bound in (A59) converges to 1 if and only if K^r is $o(\sqrt{N^r})$. ■

11. Asymptotic Waiting Time Distribution

In this section we consider the prove of Proposition 5.2, which gives expression for steady state waiting time distributions. The result will follow the next two propositions. Proposition 11.1 below gives the asymptotic distribution for the waiting time of class J . Then, Proposition 11.2 deals with convergence of normalized version of the waiting times of classes $1, \dots, j-1$. Corollary 5.1, in turn, is a direct result of 11.2 applying Little's Law.

Proposition 11.1

$$\sqrt{N^r} W_J^r(\infty) \Rightarrow W_J, \text{ as } r \rightarrow \infty \quad (\text{B1})$$

where

$$W_J \sim \begin{cases} \exp(\xi_J \mu \beta) & \text{w.p. } \alpha(\beta) \\ 0 & \text{otherwise} \end{cases} \quad (\text{B2})$$

Proof: Having the convergence of $X^r(\infty)$ we can repeat the proof of (A29) with $Q^r(0) = Q^r(\infty)$ to obtain the desired result. ■

Proposition 11.2 Assume (15), then, for all $i = 1, \dots, J-1$,

$$N^r \cdot [W_i^r(\infty) | W_i^r(\infty) > 0] \Rightarrow [W_i | W_i > 0], \quad (\text{B3})$$

where $[W_i | W_i > 0]$ has the Laplace transform:

$$\begin{cases} \frac{\mu(1-\sigma_1)}{s+\mu(1-\sigma_1)}, & i = 1, \\ \frac{\mu(1-\sigma_i)(1-\tilde{\gamma}_i(s))}{s-\hat{\lambda}_i+\hat{\lambda}_i\tilde{\gamma}_i(s)}, & i = 2, \dots, J-1, \end{cases} \quad (\text{B4})$$

with $\sigma_i = \rho_{\leq i} = \lim_{r \rightarrow \infty} \sum_{j=1}^i \frac{\lambda_j^r}{N^r \mu}$, $\sigma_0 = 0$, $\hat{\lambda}_i = \lim_{r \rightarrow \infty} \frac{\lambda_i^r}{N^r}$, and

$$\tilde{\gamma}_i(s) = \frac{s+\mu}{2b_i\mu} + \frac{1}{2} - \sqrt{\left(\frac{s+\mu}{2b_i\mu} + \frac{1}{2}\right)^2 - \frac{1}{b_i}}, \quad (\text{B5})$$

for $b_i = \lim_{r \rightarrow \infty} \frac{\sum_{j=1}^{i-1} \lambda_j^r}{N^r}$. Also, the limits of the first and second moments of the conditional waiting time satisfy:

$$\begin{aligned} N^r E[W_i^r(\infty) | W_i^r(\infty) > 0] &\rightarrow [\mu(1-\sigma_i)(1-\sigma_{i-1})]^{-1}, \text{ and} \\ (N^r)^2 E[(W_i^r(\infty))^2 | W_i^r(\infty) > 0] &\rightarrow 2(1-\sigma_i\sigma_{i-1})[(\mu)^2(1-\sigma_i)^2(1-\sigma_{i-1})^3]^{-1}. \end{aligned} \quad (\text{B6})$$

Remark 11.1 Propositions 11.1 and 11.2 imply together that $E[W^r] = \sum_{i=1}^J \frac{\lambda_i^r}{\lambda^r} E[W_i^r] \approx \frac{\lambda_J^r}{\lambda^r} E[W_J^r] \approx \alpha(\beta) \frac{1}{\beta\sqrt{r}\mu}$ which is approximately the waiting time in a single class $M/M/N - K_J^r$ queue with arrival rate λ^r service rate μ and N^r agents. In particular, if $K_J^r \ll \sqrt{r}$ it is approximately the waiting time in an $M/M/N$ queue.

Remark 11.2 (Intuitive Explanation of Proposition 11.2) This result is based on an intuition similar to the one that explains Theorem 3.1; Consider the two class example. Then, the high priority customers experience light traffic and, given that they are delayed, they have a queue that is of order which is at most $O(1)$, and waiting time that is $\Theta(1/N)$. This is because, given that there are at least $N - K$ busy servers, the number of high priority customers behaves approximately like a single server queue with rate $(N - K)\mu$ and load that is strictly less than 1.

Proposition 5.2 states exactly that; in order to obtain a meaningful limit for the waiting time of the high priority this waiting time should be multiplied by at least N . The fact that, given that they are delayed, the high priority customers wait is analogous to a single server queue explains how the Laplace transforms can be easily derived from known Laplace transforms of the $M/G/1$ queue.

Proof: Let us focus on class $i, 1 \leq i < J$. We will prove the result through the $M/G/1$ reduction that was applied in both [15] and [8].

Step 1 (Limit for the $M/M/1$ Busy Period): Let us look at an $M/M/1$ queue with arrival rate $\lambda_i^- = \sum_{j=1}^{i-1} \lambda_j^r$ and service rate $N^r\mu$. Then, by known results (see for example [8]), $\tilde{\gamma}_i^r(s)$ - the Laplace transform of the busy period is given by:

$$\tilde{\gamma}_i^r(s) = \frac{N^r\mu + s + \lambda_i^- - \sqrt{N^r\mu + s + \lambda_i^- - 4\lambda_i^- N^r\mu}}{s\lambda_i^-}. \quad (\text{B7})$$

By simple algebra we can prove that

$$\tilde{\gamma}_i^r(s) \rightarrow \tilde{\gamma}_i(s), \text{ as } r \rightarrow \infty, \quad (\text{B8})$$

where $\tilde{\gamma}_i(s) = \lim_{r \rightarrow \infty} \tilde{\gamma}_i^r(s)$ and $\tilde{\gamma}_i(s)$ is given by (B5). Note that the convergence above is still valid if the service rate of the relevant $M/M/1$ is $(N^r - K^r)\mu$ where $K^r = o(N^r)$.

Step 2 (bounding): Following [15], note that given wait of class k their queue behaves like an $M/G/1$ queue with the G being the distribution of the busy period beginning with a class $j : j < i$ arriving to a system with $N - K_i$ busy servers and ends with a completion of service when there are $N - K_i - 1$ busy servers. The Laplace transform of this distribution G is denoted in [15] by $B_i^*(s)$, and its expectation is denoted by $E[B_i]$. Denote by $\phi_i^r(s)$ the Laplace transform of $W_i | W_i > 0$ in the r^{th} system. Then, by formula (17) in [15] we have that

$$\phi_i^r(s) = \frac{1 - B_i^*(s)}{(s - \lambda_i^r + \lambda_i^r B_i^*(s))} \frac{1 - \lambda_i E[B_i]}{E[B_i]} \quad (\text{B9})$$

G can be sample wise bounded from above by $G_{i, N^r - K^r}$ and from below by G_{i, N^r} . Hence we have by the previous step that

$$B_i^*(N^r s) \rightarrow \tilde{\gamma}_i(s), \text{ as } r \rightarrow \infty \quad (\text{B10})$$

and the convergence of the moments follows. Hence:

$$N^r E[B_i^*] \rightarrow \frac{1}{\mu(1 - \sigma_{i-1})}, \text{ as } r \rightarrow \infty \quad (\text{B11})$$

Now, by simple calculus, and since by (15) $\sigma_i < 1$ we have that

$$\phi_i^r(N^r s) \rightarrow \frac{\mu(1 - \sigma_i)(1 - \tilde{\gamma}_i(s))}{s - \hat{\lambda}_i + \hat{\lambda}_i \tilde{\gamma}_i(s)}. \quad (\text{B12})$$

The limiting transform is similar to the one obtained for the static priority case. Moments for the static priority case are given in [8] and their limits are easily calculated. ■

12. Performance Anslsysis For $\gamma < 1/2$

This part includes performance measures for the $M/M/N/\{K_i\}$ model under the Efficiency Driven Regime, i.e. for $0 < \gamma < 1/2$. The Efficiency Driven (ED) cab be characterized as follows: Consider a sequence of N -server queues, indexed by $r = 1, 2, \dots$. Define the *offered load* by $R^r = \frac{\lambda^r}{\mu}$, where λ^r is the arrival-rate and μ the service-rate. Without loss of generality, let $r = R^r$. The ED regime is achieved by letting $(N^r)^{1-\gamma}(1 - \rho^r) \rightarrow \beta$, as $r \uparrow \infty$, for some finite β .

We define the ED regime for a sequence of $M/M/N^r/\{K_i^r\}$ queues as follows: There exist $\gamma \in [0, 1/2)$ and $0 < \beta < \infty$, such that

$$\lim_{r \rightarrow \infty} (N^r)^{1-\gamma} (1 - \rho_C^r) = \beta. \quad (C1)$$

For purposes of optimization we will need to adapt some of the results of the previous sections to the case of the ED $M/M/N^r/\{K_i^r\}$ model. As before we assume (15), i.e. that class J is non-negligible.

12.1. Diffusion Limits

Since, by [7], the probability of delay in this regime converges to 1, we expect that the diffusion limit to be a reflected brownian motion as is the case with the conventional heavy traffic for multi-server queues. However, differently from conventional heavy traffic, this regime requires different scaling for different values of γ in order to obtain a non-degenerate limit.

Note that having ED limits for the relevant $M/M/N$ queue immediately translates into limits for our model using the same procedures as used in the proof of Proposition 10.1. The ED limits for a sequence of $M/M/N$ queues were not proved for a general $\gamma > 1/2$. In section 14 we adapt methods that were used in [5], to prove the desired results. In particular we prove the following:

Proposition 12.1 *Consider a sequence of $M/M/N$ system indexed by $N = 1, 2, \dots$, such that*

$$N^{1-\gamma} (1 - \rho^N) \rightarrow \beta, \text{ as } N \rightarrow \infty, \text{ } 0 < \beta < \infty. \quad (C2)$$

for some $\gamma \in [0, 1/2)$. Let $Y^N(t)$ be the normalized total number of customers in the N^{th} system at time t . Assume $\frac{Y^N(0) - N}{N^{1-\gamma}} \Rightarrow X(0)$, where $X(0) \geq 0$, a.s. Then,

$$\frac{Y^N(N^{2\gamma-1} \cdot) - N}{N^{1-\gamma}} \Rightarrow RBM(-\beta\mu, 2\mu), \text{ as } N \rightarrow \infty, \quad (C3)$$

where $RBM(-\beta\mu, 2\mu)$ is a Reflected Brownian Motion with infinitesimal drift $-\beta\mu$ and infinitesimal variance 2μ .

The following proposition summarizes the diffusion limit results for the ED $M/M/N/\{K_i\}$. Its proof is omitted since once the convergence of an ED sequence of $M/M/N$ queues is established, the proof for the $M/M/N/\{K_i\}$ model is analogous to that of the QED case.

Proposition 12.2 *Define*

$$X^r(t) = \frac{Y^r((N^r)^{1-2\gamma}t) - (N^r - K^r)}{(N^r)^{1-\gamma}}. \quad (\text{C4})$$

Assume that there exists $0 \leq \gamma < 1/2$ such that:

$$\lim_{r \rightarrow \infty} (N^r)^{1-\gamma} (1 - \rho_C^r) \rightarrow \beta, \quad 0 < \beta < \infty. \quad (\text{C5})$$

Also assume that $X^r(0) \Rightarrow X(0)$, as $r \rightarrow \infty$, where $X(0) \geq 0$. Then,

$$X^r(\cdot) \Rightarrow X(\cdot), \quad (\text{C6})$$

where $X(\cdot)$ is an $\text{RBM}(-\beta\mu, 2\mu)$. Finally,

$$\frac{1}{(N^r)^{1-\gamma}} Q_i^r((N^r)^{1-2\gamma} \cdot) \Rightarrow 0, \quad \text{as } r \rightarrow \infty, \quad i = 1, \dots, J-1. \quad (\text{C7})$$

Remark 12.1 The state space collapse in this case follows in the same manner as in the QED setting, using a bounding $M/M/1$ queue. The fact that this $M/M/1$ is not only scaled in space but also in time does not change the result.

12.2. Steady State

In the following proposition we adapt the steady state results of subsection 10.2 to the ED regime. Here we limit our discussion to thresholds $K^r = o((N^r)^\gamma)$. As will be shown in the next section (Asymptotic Optimality) we only need threshold that are logarithmic and this is clearly covered by $K^r = o((N^r)^\gamma)$ since $\gamma > 0$. Moreover, taking $K^r = o((N^r)^\gamma)$ simplifies the proof of the tightness that we need for convergence of the steady state distributions. The proof, being similar to the proofs of Propositions 10.3-10.5, is omitted.

Proposition 12.3 *Assume that there exists $0 \leq \gamma < 1/2$ such that*

$$(N^r)^{1-\gamma} (1 - \rho_C^r) \rightarrow \beta, \quad 0 < \beta < \infty, \quad \text{as } r \rightarrow \infty, \quad (\text{C8})$$

and $K^r = o((N^r)^\gamma)$. Then,

$$X^r(\infty) \Rightarrow X(\infty), \quad \text{as } r \rightarrow \infty, \quad (\text{C9})$$

where $X(\infty) \sim \exp(\beta)$. Also,

$$P\{W_J^r(\infty) > 0\} \rightarrow 1, \text{ as } r \rightarrow \infty, \quad (\text{C10})$$

and,

$$P\{W_i^r(\infty) > 0\} \sim \prod_{j=i}^{J-1} (\rho_j^r)^{K_{j+1}^r - K_j^r}. \quad (\text{C11})$$

Finally,

$$\begin{aligned} (N^r)^{-(1-\gamma)} Q_i^r(\infty) &\Rightarrow 0, i = 1, \dots, J-1, \text{ as } r \rightarrow \infty, \text{ and} \\ (N^r)^{-(1-\gamma)} Q_J^r(\infty) &\Rightarrow X^+(\infty), \text{ as } r \rightarrow \infty. \end{aligned} \quad (\text{C12})$$

Remark 12.2 Recall that for the proof of convergence of the steady state distribution in the QED case we had to prove first the tightness for the sequence $X^r(\infty)$. We achieved that by bounding our system from above and from below by two systems for which the tightness was known. By the same path-wise construction used before we can bound our system from above by an $M/M/m$ queue with $N^r - K^r$ servers and from below by an $M/M/m$ queue with N^r servers. Provided that $K^r = o((N^r)^\gamma)$ the tightness for both systems under our scaling is known, and the result follows by the same manner as before.

Remark 12.3 Having all the above, one can repeat the arguments given for $\gamma = 1/2$ to conclude that for $\gamma < 1/2$ we also have under ITP and SCS that

$$\frac{E[W^r]}{E[W_{\lambda^r, \mu}^{FCFS}(N^r)]} \rightarrow 1. \quad (\text{C13})$$

Regarding the waiting times of classes $i = 1, \dots, J-1$, Proposition 11.2 and its proof are not changed for the case $\gamma < 1/2$.

13. Adding Abandonment

In this section we prove the asymptotic optimality results for the model which includes abandonment as given in section 7. As in the non-abandonment case, we start with performance analysis of the ITP and SCS rules, first in the transient diffusion level and then in steady state. We end the

section with the proof of asymptotic optimality. Before proceeding we give the scaled versions of the optimization problem and the ITP and SCS rules. The formulation we consider here is given by

$$\begin{aligned} & \text{minimize } N \\ & \text{subject to } P\{Ab\} \leq \alpha^r, \\ & \quad P\{W_i > T_i^r\} \leq \alpha_i, \quad i = 1, \dots, J-1, \\ & \quad N \in \mathbb{Z}_+, \pi \in \Pi \end{aligned} \tag{F1}$$

To simplify the presentation of this case we restrict ourselves to the following assumption:

Assumption 13.1 $\alpha^r = \hat{\alpha}/\sqrt{r}$ and $T_i^r = \hat{T}_i/r^{\gamma_i}$, where $\gamma_i > 1/2$ for all $i = 1, \dots, J-1$.

To simplify the presentation of the results in this setting, we do not consider the general case with $\alpha^r = \hat{\alpha}/r^\gamma$ for arbitrary γ , but rather limit ourselves to $\gamma = 1/2$. The transition from $\gamma = 1/2$ to arbitrary γ however is as simple as in the non-abandonment case and the structure of the optimal policy does not change when changing γ .

As before, we assume w.l.o.g that classes $1, \dots, J-1$ are ordered in increasing order of T_i^r and that class J is the *Best Effort* class.

The ITP and SCS rule are given in the following definition:

Definition 13.2 *ITP and SCS for Abandonment Model*

- **Staffing:** Find the staffing level through the single class $M/M/N + M$ (or Erlang-A) model with arrival rate λ^r , service rate μ , abandonment rate θ_J and FCFS service. Specifically, let

$$N^{*r} = \text{Min}\{N \in \mathbb{Z}_+ : P\{Ab\}_{\lambda^r, \mu, \theta_J}^{FCFS}(N) \leq \alpha^r\}. \tag{F2}$$

- **Control:** Use the TP rule with the differences $\{K_{j+1}^r - K_j^r\}_{j \leq J-1}$ chosen recursively for $j = J-1, \dots, 1$ in the following manner::

— Compute

$$K_{j+1}^r - K_j^r = \left\lceil \frac{\ln(\alpha_j T_j^r / [P\{W_{j+1}^r > 0\} \hat{w}(N^{*r}, \sigma_j^r, \sigma_{j-1}^r)])}{\ln(\sigma_j^r)} \right\rceil \bigvee 0 \quad j = J-1, \dots, 1 \tag{F3}$$

where $\hat{w}(N^{*r}, \sigma_j^r, \sigma_{j-1}^r) = [N^{*r} \mu (1 - \sigma_j^r) (1 - \sigma_{j-1}^r)^{-1}]$.

— Set

$$P\{W_j^r > 0\} = P\{W_{j+1}^r > 0\}(\sigma_j^r)^{K_{j+1}^r - K_j^r}. \quad (\text{F4})$$

In the above we set $P\{W_j^r > 0\} = P\{W_{\lambda^r, \mu, \theta_j}^{FCFS}(N^{*r}) > 0\}$, and for two real numbers x and y , $x \vee y =: \max\{x, y\}$. The actual threshold values are then determined by setting $K_1^r = 0$.

Analogously to the non-abandonment case, we have the following lemma which is adapted from Zeltyn and Mandelbaum [10].

Lemma 13.3 *Consider the sequence λ^r and the sequence of staffing levels N^r determined through SCS. Then, $N^r - R \approx \beta\sqrt{R}$, for some $-\infty < \beta < \infty$. In particular, under SCS*

$$\sqrt{N^r}(1 - \rho^r) \rightarrow \beta. \quad (\text{F5})$$

13.1. Diffusion Limits

First we quote Theorem 2 from [6] for a sequence of $M/M/N + M$ queues. Denote by $\{Y^r(t), t \geq 0\}$ the total number in system in an $M/M/N^r + M$ system. Let

$$X^r(t) = \frac{Y^r(t) - N^r}{\sqrt{N^r}},$$

then we have the following:

Theorem 13.4 ([6], *Theorem 2*) *Consider a sequence of $M/M/N^r + M$ queues indexed by the superscript $r = 1, 2, \dots$. Let λ^r be the arrival rate in the r^{th} system. The service rate μ and the individual abandonment rate θ are independent of the index r . Let $\rho^r = \lambda^r / (N^r \mu)$, and assume that*

$$\lim_{r \rightarrow \infty} \sqrt{N^r}(1 - \rho^r) \rightarrow \beta, \quad -\infty < \beta < \infty. \quad (\text{F6})$$

Then, if $X^r(0) \Rightarrow X(0)$, then $X^r(\cdot) \Rightarrow X(\cdot)$ where $X(\cdot)$ is a diffusion process with drift

$$m(x) = \begin{cases} -(\beta + (\theta/\mu)x)\mu & x \geq 0 \\ -(\beta + x)\mu & x \leq 0 \end{cases}$$

and infinitesimal variance $\sigma^2 = 2\mu$.

Analogously to our previous notation let $M/M/N/\{K_i\} + M$ represent a system with N servers, thresholds $\{K_i\}$ with the addition of exponential patience. In the next two propositions we will show that the normalized and scaled overall number of customers in systems in the $M/M/N/\{K_i\} + M$ model converges to the same limit as in Theorem 13.4, with $\theta = \theta_J$ (which is the impatience rate of the lowest priority).

We consider a sequence of $M/M/N/\{K_i\} + M$ systems indexed by $r = 1, 2, \dots$. The policy is the same policy as in the non-abandonment case. A class i customer is served only if there are no customers of a higher priority j ($j < i$) waiting and the number of idle servers is greater than K_i^r . As before, we use the notation K^r to stand for the threshold of the lowest priority (i.e. $K^r = K_J^r$), and define a “nominal” load: $\rho_C^r = \frac{\lambda^r}{N^r - K^r}$.

As before, let $Q_i^r(t)$ stand for the queue length of class i at time t in the r^{th} system, $Z^r(t)$ stands for the number of busy servers at time t in the r^{th} system, and $Y^r(t)$ is the overall number of customers in system, i.e. $Y^r(t) = Z^r(t) + \sum_{i=1}^J Q_i^r(t)$.

Proposition 13.1 (State Space Collapse) *Assume (15) and that*

$$\lim_{r \rightarrow \infty} \sqrt{N^r} (1 - \rho_C^r) \rightarrow \beta, \quad -\infty < \beta < \infty. \quad (\text{F7})$$

Then, as $r \rightarrow \infty$,

$$\begin{aligned} \frac{1}{\sqrt{N^r}} Q_i^r(\cdot) &\Rightarrow 0, \quad i = 1, \dots, J-1, \\ \frac{1}{\sqrt{N^r}} [(N^r - K^r) - Z^r(\cdot)]^- &\Rightarrow 0, \quad \text{and} \\ \frac{1}{N^r} [(N^r - K^r) - Z^r(\cdot)]^+ &\Rightarrow 0. \end{aligned} \quad (\text{F8})$$

Proof: for the first two limits the proof is omitted since it is similar to the proof in the no-abandonment case. To Show that the third limit applies we will use bounding as before. Assume we start $[(N^r - K^r) - Z^r]^+$ from zero. Then, this process can be bounded from above by a birth and death process with birth rates $\lambda_i = (N - K - i)\mu, i = 0, \dots, N - K$ and death rates $\mu_i = \lambda$. By [9] the fluid limit of the bounding process is zero and hence the result. ■

Proposition 13.2 *Assume (15) and*

$$\lim_{r \rightarrow \infty} \sqrt{N^r} (1 - \rho_C^r) \rightarrow \beta, \quad -\infty < \beta < \infty. \quad (\text{F9})$$

If $X^r(0) \Rightarrow X(0)$, then,

$$X^r(\cdot) = \frac{Y^r(\cdot) - (N^r - K^r)}{\sqrt{N^r}} \Rightarrow X(\cdot), \quad \text{as } r \rightarrow \infty, \quad (\text{F10})$$

where X is a diffusion process with infinitesimal drift given by

$$m(x) = \begin{cases} -(\beta + (\theta_J/\mu)x)\mu & x \geq 0 \\ -(\beta + x)\mu & x \leq 0 \end{cases}$$

and infinitesimal variance $\sigma^2 = 2\mu$.

Proof: In this proof we employ the same approach that was used in [1] for the proof of the diffusion limit. We write the proof for the two-class case. The proof is similar for arbitrary number of classes as will be explained at the end of the proof.

First, like in the proof of Proposition 10.1, we define a system with two server pools: *The $N - K$ pool* and *The K pool*. For simplicity of notation we will call them from now on pools 1 and 2, respectively. Whenever a server in pool 1 completes service and there are any customers in service in pool 2 we preempt a customer from pool 2 and pass it to pool 1. This system has the same law as the original system. Denote by $I_k^r(t)$ and $Z_k^r(t)$ the number of idle servers and the number of busy servers respectively in pool k ($k = 1, 2$) at time t . Also, let $Q^r(t)$ be the total number of customers in queue (i.e. $Q^r(t) = Q_1^r(t) + Q_2^r(t)$).

Consider a Poisson process with rate $(N - K)\mu$, and create the service completions using this Poisson process in the following manner: A jump in this Poisson process creates a departure from pool 1 with probability $\frac{Z_1^r(t)}{N^r - K^r}$, and does not result in a departure, otherwise.

Then, the total number of customers in the system $Y^r(t)$ admits the following dynamics:

$$\begin{aligned} Y^r(t) &:= Q^r(t) + Z_1^r(t) + Z_2^r(t) \\ &= Y^r(0) + A^r(t) - \mathcal{N}_1(\mu(N - K)) + \mathcal{N}_1\left(\mu \int_0^t I_1^r(s) ds\right) - \mathcal{N}_2\left(\mu \int_0^t Z_2^r(s) ds\right) \\ &\quad - \sum_{l=1}^2 \mathcal{N}_l^a\left(\theta_l \int_0^t Q_l(s) ds\right), \end{aligned} \quad (\text{F11})$$

where $\mathcal{N}_k$, $k = 1, 2$ and $\mathcal{N}_l^a$, $l = 1, 2$ are independent Poisson processes with rate 1, and $A^r(t)$ is a poisson process with rate λ^r independent of all the other processes.

Define $\mathcal{F}^r(t)$ to be the following σ -algebra:

$$\mathcal{F}^r(t) = \sigma \{Q_k^r(0); Z_k^r(0), A_k^r(t), \mathcal{N}_l^a(t), \mathcal{N}_j(t); k = 1, 2, l = 1, 2, j = 1, 2\} \vee \mathcal{N},$$

where $\mathcal{N}$ denotes the family of P -null sets, and introduce the filtration $\mathbb{F}^r = (\mathcal{F}^r(t), t \geq 0)$. Clearly, the processes $Q^r(\cdot)$, $Z_k^r(\cdot)$ and $I_k^r(\cdot)$, $k = 1, 2$, are $\mathbb{F}^r$ adapted. Then, $Y^r(t)$ admits the following decomposition:

$$Y^r(t) = Y^r(0) + \lambda^r t - \mu(N - K)t + \mu \int_0^t I_1^r(s)ds - \mu \int_0^t Z_2^r(s)ds - \sum_{l=1}^2 \theta_l \int_0^t Q_l^r(s)ds + M^r(t), \quad (\text{F12})$$

where $M^r = (M^r(t), t \geq 0)$ is an $\mathbb{F}^r$ -locally square-integrable martingale, that satisfies $M^r = M_A^r - M_1^r + M_{I_1}^r - M_{Z_2}^r - \sum_{l=1}^2 M_{Q_l}^r$, where all the above are $\mathbb{F}^r$ -locally square-integrable martingales with respective predictable quadratic variations:

$$\langle M_A^r \rangle(t) = \lambda^r t, \quad (\text{F13})$$

$$\langle M_1^r \rangle(t) = (N^r - K^r)\mu t, \quad (\text{F14})$$

$$\langle M_{I_1}^r \rangle(t) = \mu \int_0^t I_1^r(s)ds, \quad (\text{F15})$$

$$\langle M_{Z_2}^r \rangle(t) = \mu \int_0^t Z_2^r(s)ds, \text{ and} \quad (\text{F16})$$

$$\langle M_{Q_l}^r \rangle(t) = \theta_l \int_0^t Q_l^r(s)ds, \quad l = 1, 2. \quad (\text{F17})$$

Note that (F12) can be rewritten as

$$\begin{aligned} Y^r(t) = & Y^r(0) + \lambda^r t - \mu(N - K)t + \mu \int_0^t I_1^r(s)ds - \mu \int_0^t Z_2^r(s)ds - \\ & + \theta_2 \int_0^t Q_1^r(s) + Q_2^r(s) + Z_2^r(s)ds + \int_0^t (\theta_2 - \theta_1)Q_1^r(s) + \theta_2 Z_2^r(s)ds + M^r(t). \end{aligned} \quad (\text{F18})$$

Also, by definition,

$$\begin{aligned} Q_1^r(t) + Q_2^r(t) + Z_2^r(t) &= [Y^r(t) - (N^r - K^r)]^+ \\ I_1^r(t) &= [Y^r(t) - (N^r - K^r)]^- \end{aligned} \quad (\text{F19})$$

Finally, note that $Z_2^r(t) = [N^r - K^r - Z^r(t)]^+$. Hence, by Proposition (13.1),

$$\begin{aligned} \frac{1}{\sqrt{N^r}} Q_1^r(\cdot) &\Rightarrow 0, \\ \frac{1}{\sqrt{N^r}} Z_2^r(\cdot) &\Rightarrow 0, \end{aligned} \quad (\text{F20})$$

as $r \rightarrow \infty$. After normalizing and scaling we have that

$$\begin{aligned} X^r(t) &= X^r(0) - \beta\mu t + \mu \int_0^t [X^r(s)]^- ds + \theta_2 \int_0^t [X^r(s)]^+ ds \\ &\quad + \epsilon^r(t) + \frac{M^r(t)}{\sqrt{N^r}} + o(1), \end{aligned} \quad (\text{F21})$$

where $\sup_{t \leq T} |\epsilon^r(t)| \xrightarrow{P} 0$. We claim that

$$\begin{aligned} \left\{ M_A^r/\sqrt{N^r}, M_1^r/\sqrt{N^r}, M_{I_1}^r/\sqrt{N^r}, M_{Z_2}^r/\sqrt{N^r}, M_{Q_1}^r/\sqrt{N^r}, M_{Q_2}^r/\sqrt{N^r} \right\} \\ \Rightarrow \{ \sqrt{\mu}b_a, \sqrt{\mu}b_1, 0, 0, 0, 0 \}, \end{aligned} \quad (\text{F22})$$

where b_a and b_1 are independent standard Brownian motions. By the continuous mapping theorem, the latter would imply that $M^r/\sqrt{N^r}$ converges to $\sqrt{\mu}b_a - \sqrt{\mu}b_1$, which is a Brownian motion with zero drift and variance 2μ . Since $[\cdot]^+$ and $[\cdot]^-$ are Lipschitz continuous functions we have by Gronwall's inequality that $X^r(t)$ is a continuous function of $X^r(0) - \beta\mu t + \epsilon^r(t) + \frac{M^r(t)}{\sqrt{N^r}} + o(1)$. The result now follows from the continuous mapping theorem.

It is still left to establish (F22). First note that by the Functional Law of Large Numbers (FLLN), as $r \rightarrow \infty$,

$$\left\langle \frac{M_A^r}{\sqrt{N^r}} \right\rangle(t) \Rightarrow \mu t, \text{ as} \quad (\text{F23})$$

$$\left\langle \frac{M_1^r}{\sqrt{N^r}} \right\rangle(t) \Rightarrow \mu t. \quad (\text{F24})$$

By Proposition, 13.1 we have that, as $r \rightarrow \infty$

$$\left\langle \frac{1}{\sqrt{N^r}} M_{Z_2}^r \right\rangle(t) \Rightarrow 0, \quad (\text{F25})$$

$$\left\langle \frac{1}{\sqrt{N^r}} M_{Q_l}^r \right\rangle(t) \Rightarrow 0, \quad l = 1, 2. \quad (\text{F26})$$

Also,

$$\left\langle \frac{1}{\sqrt{N^r}} M_{I_1}^r \right\rangle(t) \Rightarrow 0. \quad (\text{F27})$$

The latter follow from the argument that $I_1^r(t)$ can be pathwise bounded from below by the number of idle servers in an $M/M/N - K/N - K$ loss system, for which the result can be easily proved using [9].

Note that the independence of M_A^r and M_1^r together with the inequality $\langle M, N \rangle \leq \sqrt{\langle M \rangle \langle N \rangle}$ imply that all covariations converge to zero. Also, note that since the jumps of all the above martingales are bounded by 1 we have also that for each $T > 0$,

$$\lim_{r \rightarrow \infty} E \left[\sup_{t \leq T} \left| \frac{1}{N^r} M^r(t) - \frac{1}{N^r} M^r(t-) \right| \right] = 0 \quad (\text{F28})$$

Hence, we can apply Theorem 7.1.4 from [4] to obtain the result. To prove the result for an arbitrary number of classes it is enough to construct the decomposition of Y^r (F12). The rest readily follows. ■

13.1.1. Steady State By [6], the process X defined in Proposition 13.2 has a unique stationary distribution whose density is given by:

$$f(x) = \begin{cases} \sqrt{\theta_J/\mu} \cdot h(\beta\sqrt{\mu/\theta_J}) \cdot w(-\beta, \sqrt{\mu/\theta_J}) \frac{\phi(x+\beta)}{\phi(\beta)} & x \leq 0 \\ \sqrt{\theta_J/\mu} \cdot h(\beta\sqrt{\mu/\theta_J}) \cdot w(-\beta, \sqrt{\mu/\theta_J}) \frac{\phi(x\sqrt{\theta_J/\mu} + \beta\sqrt{\mu/\theta_J})}{\phi(\beta\sqrt{\mu/\theta_J})} & x > 0 \end{cases}$$

where the hazard function h is defined by

$$h(x) = \frac{\phi(x)}{1 - \Phi(x)}$$

and

$$w(x, y) = \left[1 + \frac{h(-xy)}{yh(x)} \right]^{-1}. \quad (\text{F29})$$

Proposition 13.3 *Assume (15) and*

$$\lim_{r \rightarrow \infty} \sqrt{N^r} (1 - \rho_C^r) \rightarrow \beta, \quad -\infty < \beta < \infty. \quad (\text{F30})$$

Then

$$X^r(\infty) \Rightarrow X(\infty), \quad \text{as } r \rightarrow \infty. \quad (\text{F31})$$

where $X^r(\infty)$ and $X(\infty)$ are the steady state of X^r and X as defined in Proposition 13.2.

Proof: In this case there is no problem of stability since the abandonments stabilize the system. Hence, $X^r(\infty)$, exists for all $r = 1, 2, \dots$. Having the tightness of the sequence Y^r , the proof follows in the same manner as the proof of Theorem 10.3. To prove the tightness we will again construct two systems that will constitute stochastic lower and upper bounds on our system. Define U^r to be an $M/M/(N^r - K^r) + M$ system with arrival rate $\lambda^r = \sum_{i=1}^J \lambda_i^r$, service rate μ and abandonment rate $\underline{\theta} = \min_{i \in 1, \dots, J} \theta_i$. Define L^r to be an $M/M/N^r - K^r/N^r - K^r$ loss system. We denote by $Y_U^r(t)$ and $Y_L^r(t)$ the total number of customers in systems U^r and L^r respectively. Let O^r stand for an $M/M/N^r/\{K_i^r\} + M$ system with the server pool decomposed into two pools of sizes $N - K$ and K and with the same preemption scheme used in the construction of system B in the proof of Proposition 10.1. By the same argument used in the non-abandonment case, O^r has the same probability law as the original $M/M/N^r/\{K_i^r\} + M$ system. Let $Y^r(t)$ stand for the total number of customers in system O^r at time t .

In the following, we fix r and hence omit the superscript for simplicity of notation. We will show that:

$$Y_L(t) \leq_{st} Y(t) \leq_{st} Y_U(t), t \geq 0. \quad (\text{F32})$$

To show (F32), we use sample path coupling. For systems U and L and for the $N - K$ pool of system O , we create the departures from the same Poisson process with thinning, as we did in the proof of Proposition 10.1. The abandonments for systems O and U are also created from a joint same Poisson process with thinning: i.e. whenever there are i customers in system U and $j_k, k = 1, \dots, J$ customers from class k in queue in system O , we create the next abandonment from a Poisson process with rate $\max\{i \cdot \underline{\theta}, \sum_{k=1}^J j_k \theta_k\}$. Then, we create an abandonment in system U with probability $\frac{i \underline{\theta}}{\max\{i \cdot \underline{\theta}, \sum_{k=1}^J j_k \theta_k\}}$ and an abandonment in system O with probability $\frac{\sum_{k=1}^J j_k \theta_k}{\max\{i \cdot \underline{\theta}, \sum_{k=1}^J j_k \theta_k\}}$. Note that whenever $\sum_{k=1}^J j_k \geq i$, the next abandoning event will be an abandonment from system O with probability 1.

For simplicity, suppose that all 3 systems are initialized with $N - K$ customers in service and none in queue. An arrival will not alter the state of system L while it will increase the total number

of customers in both systems O and U . So, the ordering is still preserved. Now, if there are no customers in the K pool of system O the creation of the service completions from the same Poisson process will preserve the order. Otherwise, if there are any customers in service at the K pool, the next service completion is more likely to happen in system O , but this will not violate inequality F32.

Assume that there are i customers in queue in system O and $j = i$ in system U . Then, by our construction, any abandonment in the U system will cause an abandonment in O and the ordering is preserved.

By [6] we have the tightness of the normalized and scaled sequence $Y_U^r(\infty)$. By [11] we have the tightness of the normalized and scaled sequence $Y_L^r(\infty)$. The rest follows as in the proof of Theorem 10.3. ■

Corollary 13.1 *Assume (15) and*

$$\lim_{r \rightarrow \infty} \sqrt{N^r} (1 - \rho_C^r) \rightarrow \beta, \quad -\infty < \beta < \infty. \quad (\text{F33})$$

Then,

$$P\{W_J^r(\infty) > 0\} = P\{Z^r(\infty) \geq N^r - K^r\} \rightarrow w(-\beta, \sqrt{\mu/\theta_J}), \quad \text{as } r \rightarrow \infty, \quad (\text{F34})$$

where $w(x, y)$ is defined according to (F29). In particular,

$$P\{W_J^r > 0\} \approx P\{W_{\lambda, \mu, \theta_J}^{FCFS}(N^r - K_J^r) > 0\}. \quad (\text{F35})$$

The next proposition is analogous to Proposition 10.5 for the non-abandonment case. However, in the context of abandonments we have a result that is somewhat weaker in the sense that we do not find an exact asymptotic expression for the probability of delay of the high priority, but rather an asymptotic upper bound.

Proposition 13.4 (Probability of Delay) For every $r > 0$

$$\frac{P\{W_i^r(\infty) > 0\}}{P\{W_J^r(\infty) > 0\} \cdot \prod_{k=i}^{J-1} (\rho_k^r)^{K_{k+1}^r - K_k^r}} \leq \left(\frac{N^r}{N^r - K^r} \right)^{K^r}. \quad (\text{F36})$$

In particular for $K^r = o(\sqrt{N^r})$ and assuming $w(-\beta, \sqrt{\mu/\theta}) > 0$ we have

$$P\{W_i^r(\infty) > 0\} = O \left(w(-\beta, \sqrt{\mu/\theta}) \cdot \prod_{k=i}^{J-1} (\rho_k^r)^{K_{k+1}^r - K_k^r} \right), \quad (\text{F37})$$

where $\rho_{\leq k}^r = \sum_{i=1}^k \frac{\lambda_i^r}{N^r \mu}$.

Proof: By the same considerations as in the non-abandonment case we have that

$$P\{W_i^r(\infty) \geq 0 | W_{i+1}^r(\infty) \geq 0\} \leq \left(\frac{\sum_{j=1}^i \lambda_j^r}{(N^r - K^r) \mu} \right)^{K_{i+1} - K_i} \quad (\text{F38})$$

The proof is completed as in the case without abandonment. ■

Proposition 13.5 (Waiting Time for Classes $i=1, \dots, J-1$) Under the conditions of Proposition 13.1 we have that

$$\limsup_{r \rightarrow \infty} N^r E[W_i^r | W_i^r > 0] \leq [\mu(1 - \sigma_i)(1 - \sigma_{i-1})]^{-1}, \forall i = 1, \dots, J-1, \quad (\text{F39})$$

where, as before, $\sigma_i = \lim_{r \rightarrow \infty} \frac{\sum_{k=1}^i \lambda_k^r}{N^r \mu}$.

The proof uses the same argument used in the proof of Proposition 11.2 and is hence omitted. Essentially, the idea is to bound this system by a related system where we have the same abandonment rate for class J but with $\theta_i = 0$ for $i = 1, \dots, J-1$. Then, one can repeat the arguments used in the proof of Proposition 11.2.

Corollary 13.2 (Probability of Abandonment) Denote by $P_k^r\{Ab\}$ the probability of abandonment for class k . Then,

$$\lim_{r \rightarrow \infty} \sqrt{N^r} P_k^r\{Ab\} = \Delta_k, 0 \leq \Delta_k < \infty, \quad (\text{F40})$$

where Δ_k is given by

$$\Delta_k = \begin{cases} \xi_k^{-1} [\sqrt{\theta_k/\mu} \cdot h(\beta \sqrt{\mu/\theta_k}) - \beta] \cdot w(-\beta, \sqrt{\mu/\theta_k}) & k = J \\ 0 & \text{Otherwise.} \end{cases} \quad (\text{F41})$$

Here a_k is equal to $\lim_{r \rightarrow \infty} \frac{\lambda^r}{\lambda_k^r}$. In particular,

$$\lim_{r \rightarrow \infty} \frac{P^r\{Ab\}}{P^r\{Ab\}_{\lambda, \mu, \theta_J}^{FCFS}(N^r - K_J^r)} = 1, \quad (\text{F42})$$

where $P^r\{Ab\}$ is the overall probability of abandonment.

Proof: The proof follows from the identity $\lambda_J P_J^r\{Ab\} = \theta_J E[Q_J^r(\infty)]$. We claim that there exists M and r_0 such that for all $r > r_0$ the sequence $E[\frac{1}{\sqrt{N^r}} Q_J^r(\infty)]$ can be uniformly bounded by M . This follows from the construction of the bounding system U^r in the proof of Proposition 13.3 and [6]. By the dominant convergence theorem we have the convergence

$$E[Q_J^r(\infty)] \rightarrow E[X(\infty)^+]. \quad (\text{F43})$$

The proof is completed by taking $E[X(\infty)]$ from [6]. Equation (F42) follows also from [6] where the expression for the limiting probability of abandonment for the

$M/M/N + M$ model are given. ■

We are now in position to prove the asymptotic optimality of ITP and SCS for the model with abandonment. First, the scaled version of ITP and SCS is given as follows:

Proposition 13.6 *Asymptotic Optimality of ITP and SCS for the Abandonment Case*

Consider the problem (F1). Then, ITP and SCS as given in definition 13.2 are asymptotically optimal.

Proof:

First, we establish a lower bound for the overall number of abandonments. We can restrict our attention to preemptive policies. Since all random variables involved here are exponential, allowing preemption cannot damage the performance when looking at the overall abandonment rate. Denote by A , a system with the arrival, service and abandonment parameters as defined in section 5 (In this stage A is not equipped with any routing policy). Denote by B a system with the same arrival

and service parameters but such that the patience parameters are the same for all classes and are equal to

$$\underline{\theta} = \min_{i=1,\dots,J} \theta_i.$$

Under any non-idling policy, system B behaves (in the sense of the overall abandonment) as a single class $M/M/N + M$. We wish to show that system B with a non-idling policy is a lower bound for any preemptive policy in system A .

Now, note that for any non-idling policy, the average length of the excursions, for the total number of customers in system, below the level of N is equal for systems A and B . Now, let us focus on the excursions above N (the positive excursions): it is clear (and can be proved by simple coupling arguments), that the positive excursions in system B are stochastically larger than the positive excursions in system A . Furthermore, when visiting state N , the probability of starting a positive excursion is the same for both systems.

Denote by Y_i the steady state overall number of customers in system i , $i \in \{A, B\}$, by Z_i the steady state number of busy servers, and $P_i\{Ab\}$ the steady state probability of abandonment in system i . Then, for any non-idling policy

$$P\{Y_A \geq N\} \leq P\{Y_B \geq N\} \quad (\text{F44})$$

Moreover, since the negative excursions have the same law, we have that

$$E[Z_A|Y_A < N] = E[Z_B|Y_B < N] \quad (\text{F45})$$

Hence, we have that

$$\begin{aligned} E[Z_A] &= E[Z_A|Y_A < N]P\{Y_A < N\} + NP\{Y_A \geq N\} \\ &\leq E[Z_B|Y_B < N]P\{Y_B < N\} + NP\{Y_B \geq N\} = E[Z_B]. \end{aligned} \quad (\text{F46})$$

But, by Little's Law

$$E[Z_i] = \frac{\lambda}{\mu}(1 - P_i\{Ab\}),$$

and hence we have that

$$P_A\{Ab\} \geq P_B\{Ab\}. \quad (\text{F47})$$

So, system B with non-idling policy constitutes a lower bound for system A under any policy. In particular it constitutes a lower bound for our system with respect to the overall probability of abandonment.

Hence, a lower bound staffing level is given exactly by let

$$N^* = \text{Min}\{N \in \mathbb{Z}_+ : P\{Ab\}_{\lambda, \mu, \theta_J}^{FCFS}(N) \leq \alpha^r\}. \quad (\text{F48})$$

By corollary 13.2 the global abandonment rate is asymptotically achieved using ITP. Note that by Markov's inequality

$$P\{W_i^r > T_i^r\} \leq P\{W_i^r > 0\} \frac{E[W_i^r | W_i^r > 0]}{T_i^r}. \quad (\text{F49})$$

The threshold defined by the ITP rule grow at most as a logarithm of r so that the ITP rule, Proposition 13.5 and corollary 13.1 imply that the bounds for the individual waiting time constraints are asymptotically achieved. ■

14. Efficiency Driven $M/M/N$

In Section 12, we introduced the diffusion limit for the Efficiency Driven $M/M/N/\{K_i\}$ model. The result there is heavily based on having an Efficiency Driven limit for the single class $M/M/N$ queue.

In the next proposition we consider a sequence of $M/M/N$ queues where, for simplicity of notation, we use the number of servers as the index. We wish to examine the limits obtained in the Efficiency Driven regime. In particular, we explore the limit when i.e. we fix δ , $1/2 < \delta \leq 1$ is fixed and when λ^N grow with N in the following manner:

$$N^\delta(1 - \rho^N) \rightarrow \beta, \quad 0 < \beta < \infty, \quad \text{as } N \rightarrow \infty. \quad (\text{G1})$$

Our aim is to prove convergence of the process $Q^N(t)$ (which stands for the total number of customers in system N at time t) to a Reflected Brownian Motion. This result was proved in [17] for the particular case in which $\delta = 1$. Essentially, the limit we obtain here is the same as would

be obtained in the conventional heavy traffic regime where the number of servers, N , is held fixed and the load is increases to one.

Essentially, in order to obtain convergence, it suffices to prove that the time that the process Q^N spends below N becomes negligible as N grows indefinitely. Since the positive part is clearly the same as in the case of an $M/M/1$ queue with fast arrivals and fast services, the result will follow by a time change argument.

The proof of the next proposition is an adaptation of a proof used in [5] (see the proof of part 3 of Theorem 6.2 there. A brief version of the proof can be also found in Garnett et al. [6], where most of the details are omitted).

Let $X^N(t)$ be the scaled process, i.e.

$$X^N(t) = \frac{Q^N((N^{2\delta-1}t) - N)}{N^\delta} \quad (\text{G2})$$

Remark 14.1 The condition $X(0) \geq 0$ is necessary for the limit process to be continuous on $[0, \infty)$. Otherwise, we would have a limit process that is continuous only on the open interval $(0, \infty)$. See [5] and the references therein for more details on this kind of limits.

Proof of Proposition 12.1: The time changed process, when restricting the process to be positive, is the same as an $M/M/1$ queue with fast arrivals and fast service and converges by known results (see for example [9]) to the desired limit. Formally, denote by $\tau_+^N(t)$ and $\tau_-^N(t)$ the time the process spends above zero and below zero respectively, i.e.

$$\tau_+^N(t) = \int_0^t 1_{\{X^N(s) \geq 0\}} ds, \quad (\text{G3})$$

$$\tau_-^N(t) = \int_0^t 1_{\{X^N(s) < 0\}} ds, \quad (\text{G4})$$

Then,

$$X^N \circ \tau_+^N \Rightarrow RBM(-\beta\mu, 2\mu), \quad (\text{G5})$$

where $f \circ g$ is the composition map (i.e. $f \circ g(t) = f(g(t))$). By the random time change theorem (see for example section 13.2 in [16]) all that is left to prove is that

$$\tau_-^N(t) \Rightarrow 0. \quad (\text{G6})$$

Let us examine the process $Q^N(N^{2\delta-1}t)$. Let A_i^N be the length of the i^{th} period in which there is no queue (i.e. $Q^N \leq 0$). Also let B_i^N be the length of the i^{th} busy period (i.e. $Q^N > 0$ during this times). Let $C_i^N = A_i^N + B_i^N$, $i = 1, 2, \dots$ be the length of the i^{th} cycle, where a cycle consists of a busy period and a non-busy period. By the Markovian structure of the process $\{C_i^N\}_{i=1}^\infty$ is a sequence of I.I.D random variables.

Let $\sigma^N(T)$ be the number of cycles that begin until time T , or formally

$$\sigma^N(T) = \min\{n : \sum_{i=1}^n C_i^N > T\} \quad (\text{G7})$$

Then, $\sigma^N(T)$ is a stopping time with respect to the sequence $\{C_i^N\}$. What we are seeking to prove is that

$$\lim_{N \rightarrow \infty} P\left\{\sum_{i=1}^{\sigma^N(T)} A_i^N > \epsilon\right\} = 0. \quad (\text{G8})$$

We will prove the convergence of $\sum_{i=1}^{\sigma^N(T)} A_i^N$ to zero in $\mathcal{L}^1$, which in turn implies convergence in probability. We will assume for now that $Q^N(0) = 0$, so that C_1^N will have the same distribution as any other C_i^N . We will relax this assumption later. Note that $N^\delta(1 - \rho^N) \rightarrow \beta$ implies that $N\mu - \lambda \sim N^{1-\delta}$. Now, B_i^N is just a busy period in an $M/M/1$ queue with accelerated time scale. Hence,

$$E[B_i^N] = \frac{1}{N^{2\delta-1}(N\mu - \lambda)} \sim \frac{1}{\beta N^\delta}. \quad (\text{G9})$$

$N^\delta(1 - \rho^N) \rightarrow \beta$ also implies that $\sqrt{N}(1 - \rho^N) \rightarrow 0$ and hence, following [5] and due to the time acceleration, we also have that

$$E[A_i^N] = O\left(\frac{1}{N^{2\delta-1/2}h(0)}\right) = o\left(\frac{1}{N^\delta}\right),$$

where h is the hazard rate function of a standard normal r.v (i.e. $h(x) = \phi(x)/(1 - \Phi(x))$). Hence, we have that $E[C_i^N] \sim \frac{1}{\beta N^\delta}$. From here, following exactly pages (64-67) of [5], with $\sqrt{N}$ replaced by N^δ , $h(-\beta)$ replaced by β and B_i^N replaced by A_i^N , we can conclude that

$$\lim_{N \rightarrow \infty} E \left[\sum_{i=1}^{\sigma^N(T)} A_i^N \right] = 0.$$

It is only left to remove the assumption that $Q^N(0) = 0$: If $X(0) > 0$ a.s. the result clearly holds with a limit that is continuous on $[0, \infty)$. So, let us assume that $X(0) = 0$. Whenever $Q^N(0) > 0$ the result clearly holds since the time spent below zero would be stochastically smaller than in the case with $Q^N(0) = 0$. The only problem is when $Q^N(0) < 0$ (remember that we are still dealing with the case in which $X(0) = 0$ which means that $Q^N(0) = o(N^{-\delta})$).

We will prove that if $Q^N(0) < 0$ and $X(0) = 0$

$$\lim_{N \rightarrow \infty} E[A_1^N] = 0, \tag{G10}$$

and hence the negative part still disappears in the limit. In particular, denote by V_N^{N-k} the expected time it takes for the process to arrive from $N - k$ to N . Then

$$V_N^{N-k} \leq E[A_i^N] \frac{1 - \left(\frac{\lambda^N}{\lambda^N + (N-k+1)\mu} \right)^k}{1 - \left(\frac{\lambda^N}{\lambda^N + (N-k+1)\mu} \right)}. \tag{G11}$$

The above is obtained by a simple adaptation of pages (67-68) in [5]. Now, $E[A_i^N] = o(\frac{1}{N^\delta})$ and the result follows. ■

References

- [1]Armony M. 2005. Dynamic routing in large-scale Service systems with heterogenous servers. *Queueing Systems* **51** 287-329.
- [2]Blumenthal R.M. 1992. *Excursions of Markov Processes*. Birkhäuser, Boston.
- [3]Borst S., A. Mandelbaum, M. Reiman. 2004. Dimensioning large call centers. *Operations Research* **52**(1) 17-34.
- [4]Ethier, S.N., T.G. Kurtz. 1985. *Markov Processes, Characterization and Convergence*. John Wiley & Sons, New York, NY.

- [5] Garnett, O. 1998. Designing a telephone call center with impatient customers. Masters Thesis. Technion - The Israeli Institute of Technology, Haifa, Israel.
- [6] Garnett O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing and Service Operations Management* **4**(3) 208-227.
- [7] Halfin S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29**(3) 567-587.
- [8] Kella O., U. Yechiali. 1985. Waiting times in the non-preemptive priority M/M/c queue. *Communications in Statistics - Stochastic Models* **1**(2) 357-262.
- [9] Mandelbaum A., G. Pats. 1995. State-dependent queues: approximations and applications. *IMA Volumes in Mathematics and Its Applications F. Kelly and R. J. Williams, eds.* **71** 239-282.
- [10] Mandelbaum A., S. Zeltyn. 2006. Staffing many server queues with impatient customers: Constraint satisfaction in call centers. Working Paper. Technion - The Israeli Institute of Technology, Haifa, Israel.
- [11] Massey A.W., B.R. Wallace. 2006. An optimal design of the M/M/C/K queue for call centers. *Queueing Systems*. to appear.
- [12] Meyn S.P., R.L. Tweedie. 1993. *Markov Chains and Stochastic Stability*. Springer-Verlag, London.
- [13] Puhalskii A. 1994. On the Invariance principle for the first passage time, *Math. Oper. Res.* **19**(4) 946 - 954.
- [14] Puhalskii A., M. Reiman. 2000. The multiclass GI/PH/N queue in the halfin-whitt regime. *Advances in Applied Probability* **32** 564-595.
- [15] Schaack C., R. Larson. 1986. An N-server cutoff priority queue. *Operations Research* **34**(2) 257-266.
- [16] Whitt W. 2002. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer-Verlag, New York.
- [17] Whitt W. 2003. How multiserver queues scale with growing congestion-dependent Demand. *Operations Research* **51**(4) 531-542.
- [18] Whitt W. 2004. Efficiency driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, 50(10) 1449-1461.
- [19] Whitt W. 2005. Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Mathematics of Operations Research* 30(1) 1-27.

[20]4 Call Centers Software. Downloadable from ie.technion.ac.il/serveng