Staffing and Control of Large-Scale Service Systems with Multiple Customer Classes and Fully Flexible Servers: Technical Appendix

Itay Gurvich[2] Mor Armony[3] Avishai Mandelbaum[4]

In this technical appendix we provide proofs for the various results stated in the manuscript titled: "Staffing and Control of Large-Scale Service Systems with Multiple Customer Classes and Fully Flexible Servers".

The notational convention is as follows: for any stochastic process B, $B(t)$ corresponds to the value of the process at time t. $B(\infty)$ denotes the process in steady-state, and both B amd $B(\cdot)$ denote the entire process. Note that this is somewhat different from the notational convention used in the body of the paper, but is convenient here due to some statements which connect transient processes to their steady-state counterparts.

A Asymptotic Feasibility in the QED Regime

In this section, we prove Theorem 3.1 which states the asymptotic feasibility of SRS and TP in the QED regime. The proof consists of several steps. The limits of the steady state performance measures for the TP 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 proof is presented in two subsections: Subsection A.1 below establishes a functional central limit theorem (FCLT) for the overall number of customers in system under the TP scheduling policy. As corollaries we obtain convergence for the queue lengths and waiting times of the different classes.

Section A.2 focuses on steady state analysis. In subsection A.2.1 we give a simple set of necessary conditions and a set of sufficient conditions for existence of steady state under the TP policy. These conditions are less tight than the conditions of [28] but they are much simpler to check and provide

[2]Graduate School of Business, Columbia University, ig2126@columbia.edu

[3]Stern School of Business, New York University, marmony@stern.nyu.edu

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

insight on the system behavior under the TP 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 A.5 then gives asymptotic expressions for the probabilities of delay of the higher priority classes $1, ..., J-1$.

In what follows we only assume that $K^r = O(\sqrt{N^r})$ and the results apply immediately for the case $K^r = o(\sqrt{N^r})$. To deal with this more general case the heavy traffic condition (10) is replaced by $\lim_{r\to\infty}\sqrt{N^r}(1-\rho_C^r) = \beta\,,\, 0<\beta<\infty$, where $\rho_C^r \stackrel{\triangle}{=} \frac{\lambda^r}{(N^r-K^r)\mu}$. The results can be easily generalized to include $K^r > O(\sqrt{N^r})$ with the heavy traffic condition (10) replaced by $\lim_{r\to\infty}\sqrt{N^r-K^r}(1-\rho_C^r) = \beta\,,\, 0<\beta<\infty$, and with the normalizing factor being $\sqrt{N^r-K^r}$ instead of $\sqrt{N^r}$.

A.1 Diffusion Limits

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

$$\frac{1}{N^r}A_j^r(t) \Rightarrow \hat{\lambda}_j t, \tag{A1}$$

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

$$\frac{1}{\sqrt{N^r}}(A_j^r(t) - \lambda_j^r t) \Rightarrow BM(0, \hat{\lambda}_j), \tag{A2}$$

where $BM(0,\hat{\lambda}_j)$ is Brownian motion with drift 0 and infinitesimal variance $\hat{\lambda}_j$. Also, recall that

$$Y^r(t) = Z^r(t) + \sum_{i=1}^{J} Q_j^r(t) \tag{A3}$$

is the total number of customers in the r^{th} system at time t.

Finally, for $r = 1, 2, ...$ define the centered and scaled process

$$X^r(t) = \frac{Y^r(t) - (N^r - K^r)}{\sqrt{N^r}}. \tag{A4}$$

Proposition A.1. *Assume (9), (10) and that $X^r(0) \Rightarrow X(0)$, where $\Rightarrow$ stands for weak convergence. Then*

$$X^r \Rightarrow X, \tag{A5}$$

where X is a diffusion process with infinitesimal drift given by

$$m(x) = \begin{cases} -\beta\mu & x \geq 0 \\ -(\beta + x)\mu & x \leq 0 \end{cases} \tag{A6}$$

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

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 of the proof. 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 [33]) to conclude the proof.

Definition of system B:

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

$$X_B^r(t) = \frac{Y_B^r(t) - (N^r - K^r)}{\sqrt{N^r}} \tag{A7}$$

and

$$X_C^r(t) = \frac{Y_C^r(t) - (N^r - K^r)}{\sqrt{N^r}} \tag{A8}$$

By our assumption that $\lim_{r\to\infty} \sqrt{N^r}(1-\rho_C^r) = \beta\,,\ 0<\beta<\infty$ we have from [15] 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 $r \to \infty$. 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 [34]). 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 $j \le l$). 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 [20])

$$D_k(T) = \mathcal{N}\left(\int_0^T Z^r_{K^r,B}(\tau)\mu d\tau\right) \tag{A9}$$

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

By the construction of the sample paths we have that for all $T \ge 0$ 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

$$\sup_{0\le t\le T} \|Y_B^r(t) - Y_C^r(t)\| \le \mathcal{N}\left(\int_0^T Z^r_{K^r,B}(\tau)\mu d\tau\right) \tag{A10}$$

or,

$$\sup_{0\le t\le T} \|X_B^r(t) - X_C^r(t)\| \le \frac{1}{\sqrt{N^r}}\mathcal{N}\left(\int_0^T Z^r_{K^r,B}(\tau)\mu d\tau\right) \tag{A11}$$

Provided that

$$\frac{1}{\sqrt{N^r}}\mathcal{N}\left(\int_0^T Z^r_{K^r,B}(\tau)\mu d\tau\right) \Rightarrow 0\,, \tag{A12}$$

and applying the convergence together theorem the proposition follows.

To establish (A12) 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 $Z^r_{K^r}(t) + Q_1^r(t)$ 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 $Z^r_{K^r,B}(t)+Q_1^r(t)$. Assume that at time $t \ge 0$ 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 $Z^r_{K^r,B}(t)+Q_1^r(t)$ 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 $t \ge 0$, $Z^r_{K^r}(t) + Q_1^r(t)$ can be path wise bounded by the associated $M/M/1$.

By (9) this M/M/1 is under-loaded and by Theorems 4.1 and 4.2 of [20]

$$\frac{1}{\sqrt{N^r}}\int_0^T Z^r_{K^r,B}(\tau)\mu d\tau \Rightarrow 0. \tag{A13}$$

Since the Poisson process $\mathcal{N}\left(\int_0^T Z^r_{K^r,B}(\tau)\mu d\tau\right)$ admits the decomposition (see for example [25])

$$\mathcal{N}\left(\int_0^T Z^r_{K^r,B}(\tau)\mu d\tau\right) = \int_0^T Z^r_{K^r,B}(\tau)\mu d\tau + M^r(T) \tag{A14}$$

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 (A5). 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 server higher priority customers (this time with thresholds). Finally, one would need to show that $Z^r_{K^r,B}(t) + \sum_{i=1}^{J-1} Q^r_i(t)$ can be bounded by an under-loaded $M/M/1$ queue and hence the proof follows.

■

Corollary A.1. *Let $X(\cdot)$ be the diffusion process described in Proposition A.1. Then the steady-state distribution of $X(\infty)$ has a density $f(\cdot)$ which satisfies:*

$$f(x) = \begin{cases} \exp\{-\beta x\}\alpha(\beta), & x \geq 0 \\ \\ \frac{\phi(\beta+x)}{\Phi(\beta)}(1-\alpha(\beta)), & x < 0 \end{cases} \tag{A15}$$

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

Proof: This result follows directly from [15]. ■

A consequence of the last proof is that $X^r(t)$ (the scaled and normalized process of the overall number of customers in system) becomes sufficient in describing the asymptotic behavior of the J dimensional process $(Z_1(\cdot) + Q_1(\cdot), Q_2(\cdot), ..., Q_J(\cdot))$. This state space collapse property of the $M/M/N/\{K_i\}$ model is summarized by the following corollary:

Corollary A.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*

$$\begin{aligned} &\tfrac{1}{\sqrt{N^r}}\mathcal{E}^r \Rightarrow 0 \\ &\tfrac{1}{\sqrt{N^r}}Q^r_i \Rightarrow 0,\ \forall i \leq J-1 \\ &\tfrac{1}{\sqrt{N^r}}Q^r_J \Rightarrow X^+ \end{aligned} \tag{A16}$$

Proof: Note that $\mathcal{E}^r(t) + Q^r_1(t)$ is just $Z^r_K(t) + \sum_{i=1}^{J-1} Q^r_i(t)$, hence the result follows from the proof of Proposition A.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.

Corollary A.3. *Let $W^r_i(t)$ be the virtual waiting time process for class i. Then, if there exists $-\infty < c < \infty$, such that*

$$\sqrt{N}\left(\frac{\lambda^r_J}{N^r} - a_J\mu\right) \rightarrow c\,, \tag{A17}$$

then

$$\sqrt{N^r} W^r_J \Rightarrow \frac{1}{a_J \mu}[X]^+. \tag{A18}$$

Proof: By the FCLT for the arrivals and by (A17) we have the convergence

$$V^r(t) = \sqrt{N^r}(\frac{A^r_J(t)}{N^r} - a_J \mu t) \Rightarrow V(t), \tag{A19}$$

where $V(t) = \hat{A}(t) + ct$ and $\hat{A}$ is $BM(0, \hat{\lambda}_J)$. Define $\hat{Q}^r = \frac{1}{\sqrt{N^r}} Q^r_J$. Then, by corollary A.2 we have that $\hat{Q}^r \Rightarrow [X]^+$.

The convergence of V^r and $\hat{Q}^r$ does not necessarily imply the joint convergence of $(V^r, \hat{Q}^r)$. However, following [34], we claim that the component-wise convergence is sufficient for our purposes.

By Theorem 11.6.7 in [33], and by the convergence of V^r and $\hat{Q}^r$ we have the tightness of the sequence $(V^r, \hat{Q}^r)$. Hence, by Prohorov's Theorem (Theorem 11.6.1 in [33]) we have that there exists a convergent subsequence $\{r_k\}$ for which

$$(V^{r_k}, \hat{Q}^{r_k}) \Rightarrow (\hat{V}, \hat{Q}), \tag{A20}$$

for some process $(\hat{V}, \hat{Q})$. Define $U^r(t) = \sqrt{N^{r_k}}(\frac{D^{r_k}_J(t)}{N^{r_k}} - a_J \mu t)$. Then, using the relation

$$Q^{r_k}_J(t) = Q^{r_k}_J(0) + A^{r_k}_J(t) - D^{r_k}_J(t), \tag{A21}$$

or, alternatively,

$$U^{r_k}(t) = V^{r_k}(t) + Q^{r_k}_J(0) - Q^{r_k}_J(t), \tag{A22}$$

and applying the continuous mapping theorem we have the convergence

$$(U^{r_k}, V^{r_k}) \Rightarrow (\hat{U}, \hat{V}), \tag{A23}$$

where $\hat{U} = \hat{V} - \hat{Q}$. Since U and V are continuous with $U(0) = 0$ we can apply the corollary of [24] to obtain for the subsequence

$$\sqrt{N^{r_k}} W^{r_k} \Rightarrow W, \tag{A24}$$

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

A.2 Steady State Analysis

A.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 [28]. 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^r_{Jc} for the arrival rate of the "super class" consisting of classes $1, ..., J-1$, i.e $\lambda^r_{Jc} = \sum_{i=1}^{J-1} \lambda_J$. Also, we denote by δ^r the probability of abandonment given wait in an $M/M/1+M$ system with arrival rate λ^r_{Jc}, service rate $(N^r - K^r)\mu$ and abandonment rate μ. We denote by $\rho^r_{C,<J}$ the nominal load in this single server queue. i.e. $\rho^r_{C,<J} = \frac{\lambda^r_{Jc}}{(N^r-K^r)\mu}$.

For the second part of the stability Proposition A.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

$$\frac{\lambda^r}{R^r - K^r} \to a. \tag{A25}$$

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 A.2. *Under assumption (9) 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 $\lambda^r_J > (N^r - K^r)\mu - \lambda^r_{Jc} \cdot \delta^r$.*
2. *Assume that $N^r = R^r + \Delta^r$ where $\Delta^r = o(R^r)$. Also, assume (A25). 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 $r_1 = \max\{r > 0 : \rho^r_{C,<J} \geq 1\}$. Then, for all $r > r_1$, $\delta^r \leq \frac{1}{(N^r-K^r)(1-\rho^r_{C,<J})}$, 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 [39].

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 $(Z^r + Q_1^r = N^r - K^r, Q_i^r = 0 : i = 2, ..., J)$ 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 Theorem 5.5.3 in [27]).

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 (A.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 A.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 $1, ..., J-1$ wait before entering service (i.e. there is an infinite number of servers available to serve priorities $1, ..., J$, 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 $Y_S^r(t) \to \infty$ as $t \to \infty$ then $Y^r(t) \to \infty$ as $t \to \infty$. 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 $(Z^r + Q_1^r, Q_i^r = 0 : i = 2, ..., J)$ 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 [7]). Under this restriction the number of customers from the super class $(1, ..., J-1)$ 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

$$\lambda_J^r + \lambda_{Jc}^r(1-\delta^r) \leq (N^r - K^r)\mu \tag{A26}$$

Assume now that $K^r = o(N^r)$. Then, by (9) we have that there exists r_1 such that $\rho_{C,<J}^r < 1$ for all $r > r_1$. Then, using the identity $\lambda_{Jc}^r P\{Ab\} = \mu E[Q_{<J}^r(\infty)]$ (where $Q_{<J}^r(\infty)$ stands for the steady state queue length of the super class $1, ..., J-1$), we have that

$$\delta^r = \frac{\mu}{\lambda} E[Q_{<J}^r(\infty) | Z^r(\infty) > N^r - K^r]. \tag{A27}$$

But notice that

$$E[Q_{<J}^r(\infty)|Z^r(\infty) > N^r - K^r] \leq \frac{(\rho_{C,<J}^r)^2}{1-\rho_{C,<J}^r}. \tag{A28}$$

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

$$\delta^r \leq \frac{\rho_{C,<J}^r}{(N^r - K^r)(1-\rho_{C,<J}^r)} \tag{A29}$$

This expression converges to zero as fast as $1/N^r$ by assumptions (9), (A25) and assuming that $K^r = o(N^r)$. Plugging this upper bound into (A26) 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

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

Consider two cases:

Case 1: $\lambda^r_{Jc}/(N^r - K^r)\mu \to \gamma > 1$. In this case, δ^r converges asymptotically to $1 - \frac{1}{\rho_{C,<J}}$ where $\rho_{C,<J} = \lim_{r\to\infty} \rho^r_{C,<J}$ (see for example [35]). By our assumption that $K^r \neq o(N^r)$, there exists a subsequence r_{k_j} and $0 < c < 1$ such that $\lim_{r_{k_j}\to\infty} \frac{(N^{r_{k_j}} - K^{r_{k_j}})}{N^r} = c$. For the subsequence r_{k_j} we have that

$$\lim_{j\to\infty} \frac{1}{N^{r_{k_j}}}(\lambda_J^{r_{k_j}} + \lambda_{J^c}^{r_{k_j}}(1 - \delta^{r_{k_j}}) \leq \lim_{j\to\infty} \frac{(N^{r_{k_j}} - K^{r_{k_j}})\mu}{N^{r_{k_j}}} \tag{A30}$$

On this subsequence the limiting equation is

$$\hat{\lambda}_J + c\mu \leq c\mu \tag{A31}$$

Which is a contradiction to the non-negligibility of class J assumption (9).

Case 2: $\lambda^r_{Jc}/(N^r - K^r)\mu \to \gamma \leq 1$. By [35] the probability of abandonment converges to 0 as $r_k \to \infty$. Hence we would have that for the sequence r^k the stability equation (A26) can be written as

$$\lambda^r_J + \lambda^r_{Jc} - o(\lambda^r_{Jc}) \leq (N^r - K^r)\mu \tag{A32}$$

or after dividing by μ this can be written as

$$K^r \leq \Delta^r + o(R^r), \tag{A33}$$

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

Define

$$S^r = \frac{Y^r(\infty) - (N^r - K^r)}{\sqrt{N^r}} \tag{A34}$$

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 A.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

$$P\{X(\infty) \leq x\} \stackrel{\triangle}{=} \lim_{t\to\infty}\lim_{r\to\infty} P\{X^r(t) \leq x\} = \lim_{r\to\infty}\lim_{t\to\infty} P\{X^r(t) \leq x\} \stackrel{\triangle}{=} \lim_{r\to\infty} P\{S^r \leq x\} \tag{A35}$$

We will show this in the following proposition.

Proposition A.3. *(**Steady State Convergence**) Under the notation above and assuming that*

$$\lim_{r\to\infty} \sqrt{N^r}(1-\rho_C^r) = \beta\,, \quad 0<\beta<\infty, \tag{A36}$$

the following is true:

$$S^r \Rightarrow X(\infty), \tag{A37}$$

where $X(\infty)$ is the steady state of the diffusion process spelled out in Proposition A.1. Its distribution is given in Corollary A.1.

Proof: Note that $Y^r(\infty)$ exists as a proper random variable according to Proposition A.2 and under our choice of the parameters. Following the proof of Theorem 4 in [15] all we have to prove is the tightness of the sequence S^r. Recall systems (B) and (C) from the proof of Proposition A.1. Then, since $M/M/N/\{K_i\}$ and (B) have the same law, it is enough to prove the tightness of the sequence $S_B^r \triangleq \dfrac{Y_B^r(\infty)-(N^r-K^r)}{\sqrt{N^r}}$. In addition, we create another coupling of X^r with an $M/M/N^r$ system (denoted by D) for which we define:

$$X_D^r(t) = \frac{Y_D^r(t)-(N^r-K^r)}{\sqrt{N^r}} \tag{A38}$$

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

$$X_D^r(t) = \frac{Y_D^r(t)-(N^r-K^r)}{\sqrt{N^r}} \tag{A39}$$

Clearly, by the same coupling arguments as in the proof of Proposition A.1 we have path-wise domination $X_D^r(t) \le X_C^r(t)$. And on the whole we have the path wise ordering

$$X_D^r(t) \le X_B^r(t) \le X_C^r(t) \ \forall t \ge 0 \tag{A40}$$

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 (A40) for every finite t it does so also for $t \to \infty$. Also, since under the conditions of the proposition both sequences S_C^r and S_D^r converge as $r \to \infty$, they are tight. The tightness of S_C^r implies that

$$\forall \epsilon > 0,\ \exists n_1 : P\{S_C^r \in [-n_1, n_1]\} > 1 - \frac{\epsilon}{2}. \tag{A41}$$

The tightness of S_D^r implies that

$$\forall \epsilon > 0,\ \exists n_2 : P\{S_D^r \in [-n_2, n_2]\} > 1 - \frac{\epsilon}{2}. \tag{A42}$$

Hence, by the ordering (A40) we have that

$$\forall \epsilon > 0\, \exists\, n_1, n_2 : P\{S^r \in [-n_2, n_1]\} > 1 - \epsilon \tag{A43}$$

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 $(Z^{r_k}(0) + Q_1^{r_k}(0), Q_i^{r_k}(0) : i = 2, ..., J)$ be distributed as $(Z^{r_k}(\infty) + Q_1^{r_k}(\infty), Q_i^{r_k}(\infty) : i = 2, ..., J)$, then $(Z^{r_k}(t) + Q_1^{r_k}(t), Q_i^{r_k}(t) : i = 2, ..., J)$ 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 A.1 we have $X^{r_k} \Rightarrow \hat{X}$, where $\hat{X}$ 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} is stationary for each r_k so is the limit $\hat{X}$. Hence the limit of $X^{r_k}(\infty)$ must be the unique stationary distribution of $\hat{X}$. Since every subsequence of X^{r_k} that converges must converge to this same limit, the sequence $X^r(\infty)$ itself must converge to this limit. ■

Corollary A.4. *Under (9) 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 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 converges to a limit X as $r \to \infty$, and that $X^r(t)$ converges to the stationary distribution of X as $r \to \infty$.

First let us examine the case where that $\beta < 0$: 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 (A.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

$$m(x) = \begin{cases} 0 & x \geq 0 \\ -\mu x & x < 0 \end{cases} \tag{A44}$$

See for example Theorem 4.2 of [20]. This is clearly a non-stationary process which leads to a contradiction to the assumption on the convergence of S^r.

■

The next two propositions (A.4 and A.5) prove Theorem 3.1.

Proposition A.4. *(**Halfin-Whitt Analog**) Consider a sequence of $M/M/N^r/\{K_i^r\}$ systems indexed by $r = 1, 2, ...,$ with service rate μ for all classes and arrival rate λ_i^r for class i, $i = 1, ..., J$, such that (9) holds. Then,*

$$P\{W_J^r(\infty) > 0\} \to \alpha_J, \quad 0 < \alpha_J < 1, \tag{A45}$$

if and only if

$$\sqrt{N^r}(1 - \rho_C^r) \to \beta \,, 0 < \beta < \infty, \tag{A46}$$

where $\lambda^r = \sum_{i=1}^{J} \lambda_i^r$, $\rho_C^r = \dfrac{\lambda^r}{(N^r - K^r)\mu}$. *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 $P\{W_J^r(\infty) > 0\} \to 0$.

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

$$\alpha < \alpha(\beta') < 1. \tag{A47}$$

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 $\sqrt{M^r}(1 - \rho_C^r) \to \beta'$. For the new system the 'if' direction applies and hence we will have the inequality (A47). 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 $P\{W_i^r(\infty) > 0\}$, $i = 1, ..., J-1$. The answer is given in the following proposition.

Proposition A.5. *For every $r > 0$ such that $\rho_C^r < 1$.*

$$1 \leq \frac{P\{W_i^r(\infty) > 0\}}{P\{W_J^r(\infty) > 0\} \cdot \prod_{j=i}^{J-1} (\rho_{\leq j}^r)^{K_{j+1}^r - K_j^r}} \leq \left(\frac{N^r}{N^r - K^r} \right)^{K^r}, \; i = 1, ..., J-1, \tag{A48}$$

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

$$P\{W_i^r(\infty) > 0\} \sim \alpha(\beta) \cdot \prod_{j=i}^{J-1} (\rho_{\leq j}^r)^{K_{j+1}^r - K_j^r}, \tag{A49}$$

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

Remark A.1. In the case of $K^r = \Theta(\sqrt{N^r})$ the right bound converges by simple calculus to e^{d^2} where $d = \lim_{r \to \infty} \frac{K^r}{\sqrt{N^r}}$.

Proof. For the two-class case this can be proved by direct approximations of the formulae in [28]. 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.

Let us look at the Markov process of the model restricted to the states in which more than $N^r - K^r_{j+1}$ servers are busy. Define a new process $\tilde{Y}^r = \{\tilde{Z}^r_j, \tilde{Q}^r_1, ..., \tilde{Q}^r_j\}$, where $\tilde{Z}^r_j$ describes the number of busy servers above the level of $N^r - K^r_{j+1}$, and $\tilde{Q}^r_i$ is the number of class i customers in queue. Under our restriction $\tilde{Y}^r$ is also a Markov process. Denote its steady state by $\tilde{Y}^r(\infty) = \{\tilde{Z}^r_j(\infty), \tilde{Q}^r_1(\infty), ..., \tilde{Q}^r_j(\infty)\}$. Also, because of the model's structure, the probability in question for can be computed by

$$P\{W^r_j(\infty) > 0\} = P\{W^r_{j+1}(\infty) > 0\} \cdot P\{\tilde{Z}^r_j(\infty) + \sum_{i=1}^{j} \tilde{Q}^r_i(\infty) \geq K^r_{j+1}\}$$

To justify this, see, for example, Section 10.4 of [23] and the results therein.

Define

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

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 $s \in N - K, ...$:

$$\begin{aligned} &\pi_s \textstyle\sum_{i=1}^{j} \lambda_i \geq \pi_{s+1}(N - K_{j+1})\mu \geq \pi_{s+1}(N^r - K^r)\mu \\ &\pi_s \textstyle\sum_{i=1}^{j} \lambda_i \leq \pi_{s+1} N\mu \end{aligned} \tag{A50}$$

or alternatively

$$\begin{aligned} &P\{\tilde{Z}^r_j(\infty) + \textstyle\sum_{i=1}^{j} Q^r_i(\infty) \geq K_{j+1}\} \leq \left(\frac{\sum_{i=1}^{j} \lambda_i}{(N-K)\mu}\right)^{K_j - K_{j+1}} \\ &P\{\tilde{Z}^r_j(\infty) + \textstyle\sum_{i=1}^{j} \tilde{Q}^r_i(\infty) \geq K_{j+1}\} \geq \left(\frac{\sum_{i=1}^{j} \lambda_i}{N\mu}\right)^{K_j - K_{j+1}} \end{aligned} \tag{A51}$$

By induction we have proved the desired result. By simple Taylor expansion the upper bound in (A48) converges to 1 if and only if K^r is $o(\sqrt{N^r})$. ■

B Asymptotic Waiting Time Distribution (QED)

In this section we consider the prove of Proposition 3.1, which gives expression for steady state waiting time distributions. The result will follow the next two propositions. Proposition B.1 below gives the asymptotic distribution for the waiting time of class J. Then, Proposition B.2 deals with convergence of normalized version of the waiting times of classes $1, .., j-1$. Corollary 3.1, in turn, is a direct result of B.2 applying Little's Law.

Proposition B.1.

$$\sqrt{N^r} W_J^r(\infty) \Rightarrow W_J, \ \ as\ r \to \infty \tag{B1}$$

where

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

Proof. Having the convergence of $X^r(\infty)$ we can repeat the proof of (A18) with $Q^r(0) = Q^r(\infty)$ to obtain the desired result. ■

Proposition B.2. *Assume (9), then, for all $i = 1, ..., J-1$,*

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

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

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

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

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

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

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

Proof: Let us focus on class $i, 1 \le i < J$. We will prove the result through the $M/G/1$ reduction that was applied in both [28] and [19].

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 [18]), $\tilde{\gamma}_i^r(s)$ - the Laplace transform of the busy period is given by:

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

By simple algebra we can prove that

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

where $\tilde{\gamma}_i(s) = \lim_{r\to\infty} \tilde{\gamma}_i^r(s)$ 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 [28], 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 [28] 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 [28] we have that

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

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

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

and the convergence of the moments follows. Hence:

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

Now, by simple calculus, and since by (9) $\sigma_i < 1$ we have that

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

The limiting transform is similar to the one obtained for the static priority case. Moments for the static priority case are given in [19] and their limits are easily calculated. ■

C Asymptotic Feasibility in the Efficiency Driven Regime

This part includes performance measures for the $M/M/N/\{K_i\}$ model under the Efficiency Driven Regime. The results are used in Remark 3.5 and are required for the asymptotic optimality results.

The Efficiency Driven (ED) cab be characterized as follows: Consider a sequence of N-server queues, indexed by $r = 1, 2, \ldots$. 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 $(N^r)^\delta(1-\rho^r) \to \beta$, as $r \uparrow \infty$, for some finite β and $1 \geq \delta > 1/2$.

Analogously to (10), we define the ED regime for a sequence of $M/M/N^r/\{K_i^r\}$ queues as follows: There exist $\delta \in (1/2, 1]$ and $0 < \beta < \infty$, such that

$$\lim_{r\to\infty} (N^r)^\delta (1-\rho_C^r) = \beta. \tag{C1}$$

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 (9), i.e. that class J is non-negligible.

C.1 Diffusion Limits

Since, by [15], 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 δ 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 A.1. The ED limits for a sequence of $M/M/N$ queues were not proved for a general $\delta > 1/2$. In section G we adapt methods that were used in [12], to prove the desired results. In particular we prove the following:

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

$$N^\delta(1-\rho^N) \to \beta, \text{ as } N \to \infty, \ 0 < \beta < \infty. \tag{C2}$$

Let $Q^N(t)$ be total number of customers in the N^{th} system at time t. Assume $\frac{Q^N(0)}{N^\delta} \Rightarrow X(0)$, where $X(0) \geq 0$, a.s. Then,

$$X^N \Rightarrow RBM(-\beta\mu, 2\mu), \text{ as } N \to \infty, \tag{C3}$$

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. 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 C.2. *Define*

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

Assume that there exists $\delta > 1/2$ such that:

$$\lim_{r\to\infty} (N^r)^\delta (1 - \rho_C^r) \to \beta, \quad 0 < \beta < \infty. \tag{C5}$$

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

$$X^r \Rightarrow X, \tag{C6}$$

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

$$\frac{1}{(N^r)^\delta} Q_i^r((N^r)^{2\delta-1}t) \Rightarrow 0, \ \text{as } r \to \infty, \quad i = 1, ..., J-1. \tag{C7}$$

Remark C.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.

C.2 Steady State

In the following proposition we adapt the steady state results of subsection A.2 to the ED regime. Here we limit our discussion to thresholds $K^r = o((N^r)^{1-\delta})$. As will be shown in the next section (Asymptotic Optimality) we only need threshold that are logarithmic and this is clearly covered by $K^r = o((N^r)^{1-\delta})$ since $\delta < 1$. Moreover, taking $K^r = o((N^r)^{1-\delta})$ 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 A.3-A.5, is omitted.

Proposition C.3. *Assume that there exists* $1 \geq \delta > 1/2$ *such that*

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

and $K^r = o((N^r)^{1-\delta})$. *Then,*

$$X^r(\infty) \Rightarrow X(\infty), \text{ as } r \to \infty, \tag{C9}$$

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

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

and,

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

Finally,

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

Remark C.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 $K^r = o((N^r)^{1-\delta}$ the tightness for both systems under our scaling is known, and the result follows by the same manner as before.

D Constraint Satisfaction in the QED Regime

In this section we provide the proof of the asymptotic optimality of SRS and TP with respect to the constraint satisfaction problem (1), as is stated in Theorem 4.1. First, below is Proposition D.1 which is an adaption of a result from [8]. This, together with Theorem 3.1, leads to a simple proof of Theorem 4.1.

Proposition D.1. *($M/M/N$ **Staffing.**) Consider an $M/M/N$ system with arrival rate λ^r and fixed service rate μ. We are interested in finding*

$$N^* := \min\{N : P\{W^r(\infty) > 0\} \leq \alpha\} \tag{D1}$$

where $0 < \alpha < 1$. Assume that we have a sequence of arrival rates λ^r. Then, the staffing sequence $N^r = \lambda^r/\mu + \beta\sqrt{\lambda^r/\mu}$ is asymptotically optimal with respect to (D1) in the sense of the previous definition, where $\beta = \beta(\alpha)$.

Proof: From [15] and the monotonicity of the function $P(\cdot)$ it follows that $\{N^r\}$ is asymptotically feasible if and only if

$$\sqrt{N^r}(1-\rho^r) \to \bar{\beta}, \text{ as } r \to \infty, \; 0 < \beta \leq \bar{\beta} < \infty. \tag{D2}$$

So for each $\beta > \epsilon > 0$ and staffing sequence $N^r = \lambda^r/\mu + (\beta - \epsilon)\sqrt{\lambda^r/\mu}$ there exists r_0 such that for all $r \geq r_0$, $P^r\{W^r > 0\} > \alpha$. Hence, we would necessarily have that for all $r > r_0$, $(N^r)^* \geq \lambda^r/\mu + (\beta - \epsilon)\sqrt{\lambda^r/\mu}$. Therefore, for each $\beta > \epsilon > 0$ we have that

$$\liminf_{r\to\infty} \frac{\beta\sqrt{\lambda^r/\mu}}{(\beta-\epsilon)\sqrt{\lambda^r/\mu}} \geq \liminf_{r\to\infty} \frac{\beta\sqrt{\lambda^r/\mu}}{(N^r)^* - \underline{N}^r} \geq 1 \tag{D3}$$

taking ϵ to zero, and by the definition of asymptotic optimality the proof is concluded. ■

Having Proposition D.1 we can proceed to defining the asymptotically optimal solution for (1).

Proof of Theorem 4.1: Define

$$M^r = \min\{N : P\{W_i^r(\infty) > 0\} \leq \alpha_J, \quad \forall i = 1, ..., J\}. \tag{D4}$$

Denote by $(N^*)^r$ the optimal solution to (1) then clearly $M^r \leq (N^*)^r$. Now, (D4) is equivalent to a single class $M/M/N$ constrained staffing problem. By Proposition D.1, the asymptotically optimal staffing level is $\lambda^r/\mu + \beta(\alpha_J)\sqrt{\lambda^r/\mu}$.

For α_i^r, $i = 1, ..., J-1$, that decrease polynomially with r we have by Proposition A.5 that α_i^r, $i = 1, ..., J-1$, are achieved by logarithmic thresholds. Proposition A.4 guarantees that staffing the system with M servers and using logarithmic thresholds asymptotically achieves α_J. Hence, the lower bound is asymptotically achieved. ■

E Cost Minimization (QED and ED)

In this section we provide a proof for the asymptotic optimality of SRS and TP to the cost minimization problem (2), as is stated in Theorem 4.2 and Remark 4.4. Before presenting the solution to (2) it is necessary to adapt an important theorem from [8] to our setting. In [8], the authors show how different costs lead to the three different regimes: *Efficiency Driven* (or *ED*), *QED* and *Quality Driven*. We omit from our discussion the *Quality Driven* regime and hence we will not use the general results of [8], but rather their conclusions with respect to the *ED* and *QED* regimes.

Theorem E.1. *(Theorems 6.1 and 7.1: Borst, Mandelbaum & Reiman 2002) Consider a sequence of $M/M/N^r$ systems, indexed by $r = 1, 2, ...,$ with arrival rate λ^r and fixed service rate μ. Assume that the staffing cost per server per time unit is given by s^r. A customer waiting one unit of time incurs a cost of c^r. We are interested in finding*

$$(N^r)^* := \arg \min_{N > \lambda^r/\mu} \{s^r N^r + c^r E[W^r(\infty)]\} \tag{E1}$$

For a sequence a^r, we say that $a^r \sim a$ if $\lim_{r\to\infty} \frac{a^r}{a} = 1$.

Then:

- *Assume $s^r \sim 1$ and $c^r \sim c$. Then, the staffing sequence $N^{*r} = R^r + (y^r)^*(c)\sqrt{R}$ is asymptotically optimal, where*

$$(y^r)^*(c) \equiv y^*(c) = arg \min_{y>0} \left\{ y + \frac{cP(y)}{y} \right\}. \tag{E2}$$

- *Assume $s^r \sim 1$ and $c^r \sim cJ^r$ where $J^r = o(1)$. Then the staffing sequence $N^r = R^r + (y^r)^*(c)\sqrt{R^r}$ is asymptotically optimal, where*

$$(y^r)^*(c) = arg \min_{y>0} \left\{ y + \frac{c}{y} J^r \right\} = \sqrt{cJ^r}. \tag{E3}$$

Proof of Theorem 4.2 and Remark 4.4:

Step 1 (Lower Bound): Since we have a common μ, the long run average number of customers in queue is minimized, for afixed N, by any work conserving policy. For all work conserving policies the average number of customers in queue is equal. This gives us a lower bound on the target function since

$$\sum_{i=1}^{J} c_i^r \lambda_i^r E[W_i^r(\infty)] \geq c_J^r \cdot \sum_{i=1}^{J} \lambda_i^r E[W_i^r(\infty)] = c_J^r \cdot \sum_{i=1}^{J} E[Q_i^r(\infty)] \geq c_J^r \cdot E[Q^r(\infty)], \tag{E4}$$

where $Q^r(\infty)$ is the steady state queue length in an $M/M/N^r$ system with $\lambda^r = \sum_{i=1}^J \lambda_i^r$. Then, as a lower bound for the staffing problem we can take the solution of

$$\begin{aligned} \text{minimize} \quad & c_J^r E[Q^r(\infty)] + N^r \\ \text{subject to} \quad & P\{W^r(\infty) > 0\} \leq \alpha_J \\ & N^r \in \mathbb{Z} \end{aligned} \tag{E5}$$

Let M_1^r be the solution of the unconstrained problem

$$\begin{aligned} \text{minimize} \quad & c_J^r E[Q^r(\infty)] + N \\ & N^r \in \mathbb{Z} \end{aligned} \tag{E6}$$

Let M_2^r be the solution of the constrained staffing problem:

$$\begin{aligned} \text{minimize} \quad & N \\ \text{subject to} \quad & P\{W^r(\infty) > 0\} \leq \alpha_J \\ & N^r \in \mathbb{Z} \end{aligned} \tag{E7}$$

By [8], the cost function is strictly convex and unimodal and the feasible region for (E5) is the interval $[M_2^r, \infty)$ the solution (M^r) to the above problem (E5) will equal $\max\{M_1^r, M_2^r\}$.

Consider the following three cases:

Case 1: $\gamma_J = 0 \Rightarrow M^r = \lambda^r/\mu + \beta\sqrt{\lambda^r/\mu}$, where $\beta = \max\{(y^r)^*, \beta(\alpha_J)\}$ (where $(y^r)^*(d_J) = arg\min_{y>0}\left\{y + \frac{d_J P(y)}{y}\right\}$. For the lower bound we have that:

$$\frac{1}{\sqrt{M^r}}\left[C^r(M^r) - \underline{N^r}\right] = \frac{1}{\sqrt{M^r}}\left[c_J E[Q^r(\infty)] + \beta\sqrt{\lambda^r/\mu}\right] \sim \left[c_J \alpha(\beta)\frac{1}{\beta} + \beta\right]. \tag{E8}$$

Under the proposed choice of the thresholds we have by propositions A.5 and 3.1 that

$$\frac{1}{\sqrt{M^r}} c_i^r E[Q_i^r(\infty)] \to 0, \text{ as } r \to \infty, \tag{E9}$$

and

$$\frac{1}{\sqrt{M^r}} E[Q_J^r(\infty)] \to \alpha(\beta)\frac{1}{\beta}, \text{ as } r \to \infty. \tag{E10}$$

Hence, the lower bound is achieved, that is,

$$\lim_{r\to\infty} \frac{\sum_{i=1}^J c_i^r E[Q_i^r(\infty)] + \beta\sqrt{R^r}}{c_J E[Q^r(\infty)] + \beta\sqrt{R^r}} = 1. \tag{E11}$$

Case 2: $\gamma_J < 0, \alpha_J < 1$. In this case $N^r = R^r + \beta(\alpha_J)\sqrt{R^r}$. The lower bound cost is equivalent to $\beta\sqrt{R^r}$ and under the given thresholds we again have that

$$\frac{1}{\sqrt{N^r}} c_i^r E[Q_i^r(\infty)] \to 0, \tag{E12}$$

and hence the lower bound is achieved.

Case 3: $\gamma_J < 0,\ \alpha_J = 1 \Rightarrow M^r = \lambda^r/\mu + (y^r)^* \sqrt{\lambda^r/\mu}$ where $(y^r)^* \to 0$, as $r \to \infty$. Due to the restriction $\gamma_j > -1$, we have by Theorem (E.1) that there exists an $1 > \delta > 1/2$ such that $y_\lambda^* \sqrt{\lambda^r/\mu} = \Theta(N^{1-\delta})$. In particular we have that $\delta = 1/2(1-\gamma_J)$. Staffing with M^r and choosing the appropriate logarithmic threshold would still lead to

$$\lim_{r\to\infty} (M^r)^\delta (1-\rho^r) = \lim_{r\to\infty} (M^r)^\delta (1-\rho_C^r) \to \beta, \text{ as } r \to \infty, \tag{E13}$$

and hence the overall lower bound for the normalized cost is $\theta\left((M^r)^{1/2(1+\gamma_J)}\right)$.

By the choice of α_i^* we have that, as $r \to \infty$,

$$\frac{1}{(M^r)^{1/2(1+\gamma_J)}} c_i^r E[Q_i^r(\infty)] \Rightarrow 0, \tag{E14}$$

and

$$\frac{1}{(M^r)^{1/2(1+\gamma_J)}} c_J^r E[Q_J^r(\infty)] \Rightarrow d_J \frac{1}{\beta}. \tag{E15}$$

Again the lower bound is asymptotically achieved.

■

Corollary 4.1 follows immediately by noting that in Theorem 4.2 $\alpha_i * \equiv 1$, and hence the staffing problem reduces to the single class dimensioning problem, and the routing is a simple static priority scheme.

F Adding Abandonment

In this section we prove the asymptotic optimality results for the model which includes abandonment. These results are stated in Theorem 5.1, Lemma F.1 and Theorem F.2. The proof of optimality of the TP control follows the same steps as the non-abandonment case. First, in subsection F.1 we prove convergence of the scaled and normalized process, describing the overall number of customers in system, to a one-dimensional diffusion process. In subsection F.1.1 we deduce convergence of steady state distributions. In particular we have asymptotic expression for the probabilities of delay and abandonment for all customers classes. Subsection F.2 concludes this part by incorporating the two previous subsections to establish a proof for Theorem 5.1. Lemma F.1 and Theorem F.2 are stated and proved in subsection F.2.1.

F.1 Diffusion Limits

First we quote Theorem 2 from [13] 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 F.1. *([13], **Theorem 2**) Consider a sequence of $M/M/N^r + M$ queues indexed by the superscript $r = 1, 2, \ldots.$ 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*

$$\lim_{r\to\infty} \sqrt{N^r}(1-\rho^r) \to \beta\,,\ -\infty < \beta < \infty. \tag{F1}$$

Then, if $X^r(0) \Rightarrow X(0)$, then $X^r \Rightarrow X$ where X 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 $\sigma^2 = 2\mu$.

In analogy 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 F.1, 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 $r = 1, 2, \ldots.$ 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. $K^r = K^r_J$), and define a"nominal" load: $\rho^r_C = \frac{\lambda^r}{N^r - K^r}$.

As before, let $Q^r_i(t)$ 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. $Y^r(t) = Z^r(t) + \sum_{i=1}^{J} Q^r_i(t)$.

Proposition F.1. *(**State Space Collapse.**) Assume (9) and that*

$$\lim_{r\to\infty} \sqrt{N^r}(1 - \rho^r_C) \to \beta, \quad -\infty < \beta < \infty. \tag{F2}$$

Then, as $r \to \infty$,

$$\frac{1}{\sqrt{N^r}} Q^r_i \Rightarrow 0, \ i = 1, ..., J-1,$$

$$\frac{1}{\sqrt{N^r}} [(N^r - K^r) - Z^r]^- \Rightarrow 0, \ \textit{and} \tag{F3}$$

$$\frac{1}{N^r} [(N^r - K^r) - Z^r]^+ \Rightarrow 0.$$

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 $\lambda_i = (N - K - i)\mu, i = 0, ..., N - K$ and death rates $\mu_i = \lambda$. By [20] the fluid limit of the bounding process is zero and hence the result. ■

Proposition F.2. *Assume (9) and*

$$\lim_{r\to\infty} \sqrt{N^r}(1 - \rho^r_C) \to \beta, -\infty < \beta < \infty. \tag{F4}$$

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

$$X^r = \frac{Y^r - (N^r - K^r)}{\sqrt{N^r}} \Rightarrow X, \ \textit{as } r \to \infty, \tag{F5}$$

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 [4] 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 A.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:

$$\begin{aligned} Y^r(t) &:= Q^r(t) + Z_1^r(t) + Z_2^r(t) \\ &= Y^r(0) + A^r(t) - \mathcal{N}_1(\mu(N-K)) + \mathcal{N}_1\left(\mu\int_0^t I_1^r(s)ds\right) - \mathcal{N}_2\left(\mu\int_0^t Z_2^r(s)ds\right) \\ &- \sum_{l=1}^{2}\mathcal{N}_l^a\left(\theta_l\int_0^t Q_l(s)ds\right), \end{aligned} \tag{F6}$$

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 $\sigma-$algebra:

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

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, Z_k^r and I_k^r, $k = 1,2$, are $\mathbb{F}^r$ adapted. Then, $Y^r(t)$ admits the following decomposition:

$$Y^r(t) = Y^r(0) + \lambda^r t - \mu(N-K)t + \mu\int_0^t I_1^r(s)ds - \mu\int_0^t Z_2^r(s)ds - \sum_{l=1}^{2}\theta_l\int_0^t Q_l^r(s)ds + M^r(t), \tag{F7}$$

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:

$$\langle M_A^r \rangle (t) = \lambda^r t, \tag{F8}$$

$$\langle M_1^r \rangle (t) = (N^r - K^r)\mu t, \tag{F9}$$

$$\left\langle M_{I_1}^r \right\rangle (t) = \mu \int_0^t I_1^r(s)ds, \tag{F10}$$

$$\left\langle M_{Z_2}^r \right\rangle (t) = \mu \int_0^t Z_2^r(s)ds, \text{ and} \tag{F11}$$

$$\left\langle M_{Q_l}^r \right\rangle (t) = \theta_l \int_0^t Q_l^r(s)ds, \quad l = 1, 2. \tag{F12}$$

Note that (F7) can be rewritten as

$$\begin{aligned} Y^r(t) =& Y^r(0) + \lambda^r t - \mu(N-K)t + \mu \int_0^t I_1^r(s)ds - \mu \int_0^t Z_2^r(s)ds - \\ &+ \theta_2 \int_0^t Q_1^r(s) + Q_2^r(s) + Z_2^r(s)ds + \int_0^t (\theta_2 - \theta_1)Q_1^r(s) + \theta_2 Z_2^r(s)ds + M^r(t). \end{aligned} \tag{F13}$$

Also, by definition,

$$\begin{aligned} &Q_1^r(t) + Q_2^r(t) + Z_2^r(t) = [Y^r(t) - (N^r - K^r)]^+ \\ &I_1^r(t) = [Y^r(t) - (N^r - K^r)]^- \end{aligned} \tag{F14}$$

Finally, note that $Z_2^r(t) = [N^r - K^r - Z^r]^+$. Hence, by Proposition (F.1),

$$\begin{aligned} &\frac{1}{\sqrt{N^r}} Q_1^r \Rightarrow 0, \\ &\frac{1}{\sqrt{N^r}} Z_2^r \Rightarrow 0, \end{aligned} \tag{F15}$$

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

$$\begin{aligned} X^r(t) = X^r(0) - \beta\mu t + \mu \int_0^t \left[X^r(s)\right]^- ds + \theta_2 \int_0^t \left[X^r(s)\right]^+ ds \\ + \epsilon^r(t) + \frac{M^r(t)}{\sqrt{N^r}} + o(1), \end{aligned} \tag{F16}$$

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

$$\begin{aligned} \left\{ M_A^r/\sqrt{N^r}, M_1^r/\sqrt{N^r}, M_{I_1}^r/\sqrt{N^r}, M_{Z_2}^r/\sqrt{N^r}, M_{Q_1}^r/\sqrt{N^r}, M_{Q_2}^r/\sqrt{N^r} \right\} \\ \Rightarrow \{\sqrt{\mu} b_a, \sqrt{\mu} b_1, 0, 0, 0, 0\}, \end{aligned} \tag{F17}$$

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 (F17). First note that by the Functional Law of Large Numbers (FLLN), as $r\to\infty$,

$$\left\langle \frac{M_A^r}{\sqrt{N^r}} \right\rangle(t) \Rightarrow \mu t, \text{ as} \tag{F18}$$

$$\left\langle \frac{M_1^r}{\sqrt{N^r}} \right\rangle(t) \Rightarrow \mu t. \tag{F19}$$

By Proposition, F.1 we have that, as $r\to\infty$

$$\left\langle \frac{1}{\sqrt{N^r}} M_{Z_2}^r \right\rangle(t) \Rightarrow 0, \tag{F20}$$

$$\left\langle \frac{1}{\sqrt{N^r}} M_{Q_l}^r \right\rangle(t) \Rightarrow 0, \quad l=1,2. \tag{F21}$$

Also,

$$\left\langle \frac{1}{\sqrt{N^r}} M_{I_1}^r \right\rangle(t) \Rightarrow 0. \tag{F22}$$

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 [20].

Note that the independence of M_A^r and M_1^r together with the inequality $\langle M,N\rangle \le \sqrt{\langle M\rangle\langle N\rangle}$ 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 $T>0$,

$$\lim_{r\to\infty} E\left[\sup_{t\le T}\left|\frac{1}{N^r}M^r(t)-\frac{1}{N^r}M^r(t-)\right|\right]=0 \tag{F23}$$

Hence, we can apply Theorem 7.1.4 from [10] to obtain the result. To prove the result for an arbitrary number of classes it is enough to construct the decomposition of Y^r (F7). The rest readily follows.

■

F.1.1 Steady State

By [13], the process X defined in Proposition F.2 has a unique stationary distribution whose density is given by:

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

where the hazard function h is defined by

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

and

$$w(x, y) = \left[1 + \frac{h(-xy)}{yh(x)}\right]^{-1}. \tag{F24}$$

Proposition F.3. *Assume (9) and*

$$\lim_{r\to\infty} \sqrt{N^r}(1 - \rho_C^r) \to \beta, \ -\infty < \beta < \infty. \tag{F25}$$

Then

$$X^r(\infty) \Rightarrow X(\infty), \ \text{as } r \to \infty. \tag{F26}$$

where $X^r(\infty)$ and $X(\infty)$ are the steady state of X^r and X as defined in Proposition F.2.

Proof: In this case there is no problem of stability since the abandonments stabilize the system. Hence, $X^r(\infty)$, exists for all $r = 1, 2, ...$ Having the tightness of the sequence Y^r, the proof follows in the same manner as the proof of Theorem A.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 $\underline{\theta} = \min_{i\in 1,...,J} \theta_i$. 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 A.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:

$$Y_L(t) \leq_{st} Y(t) \leq_{st} Y_U(t),\ t \geq 0. \tag{F27}$$

To show (F27), 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 A.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 $j_k,\ k=1,...,J$ 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 $\frac{i\theta}{max\{i\cdot\underline{\theta}, \sum_{k=1}^{J} j_k\theta_k\}}$ 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 F27.

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 [13] we have the tightness of the normalized and scaled sequence $Y_U^r(\infty)$. By [22] we have the tightness of the normalized and scaled sequence $Y_L^r(\infty)$. The rest follows as in the proof of Theorem A.3. ■

Corollary F.1. *Assume (9) and*

$$\lim_{r\to\infty} \sqrt{N^r}(1-\rho_C^r) \to \beta,\ -\infty < \beta < \infty. \tag{F28}$$

Then,

$$P\{W_J^r(\infty) > 0\} = P\{Z^r(\infty) \geq N^r - K^r\} \to w(-\beta, \sqrt{\mu/\theta_J}),\ \ as\ r\to\infty, \tag{F29}$$

where $w(x,y)$ is defined according to (F24).

The next proposition is analogous to Proposition A.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 F.4. ***(Probability of Delay)*** *For every $r > 0$*

$$\frac{P\{W_i^r(\infty) > 0\}}{P\{W_J^r(\infty) > 0\} \cdot \prod_{k=i}^{J-1} (\rho_k^r)^{K_{k+1}^r - K_k^r}} \le \left(\frac{N^r}{N^r - K^r} \right)^{K^r} . \tag{F30}$$

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

$$P\{W_i^r(\infty) > 0\} = O\left(w(-\beta, \sqrt{\mu/\theta}) \cdot \prod_{k=i}^{J-1} (\rho_k^r)^{K_{k+1}^r - K_k^r} \right), \tag{F31}$$

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

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

$$P\{W_i^r(\infty) \ge 0 | W_{i+1}^r(\infty) \ge 0\} \le \left(\frac{\sum_{j=1}^i \lambda_j^r}{(N^r - K^r)\mu} \right)^{K_{i+1} - K_i} \tag{F32}$$

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

Corollary F.2. ***(Probability of Abandonment)*** *Denote by $P_k^r\{Ab\}$ the probability of abandonment for class k. Then,*

$$\lim_{r \to \infty} \sqrt{N^r} P_k^r\{Ab\} = \Delta_k \, , 0 \le \Delta_k < \infty, \tag{F33}$$

where Δ_k is given by

$$\Delta_k = \begin{cases} a_k^{-1} [\sqrt{\theta_k/\mu} \cdot h(\beta\sqrt{\mu/\theta_k}) - \beta] \cdot w(-\beta, \sqrt{\mu/\theta_k}) & k = J \\ 0 & Otherwise. \end{cases} \tag{F34}$$

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

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 F.3 and [13]. By the dominant convergence theorem we have the convergence

$$E[Q_J^r(\infty)] \to E[X(\infty)^+]. \tag{F35}$$

The proof is completed by taking $E[X(\infty)]$ from [13]. ■

F.2 Asymptotic Optimality - Cost Minimization

Proof of Theorem 5.1: 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 stead 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

$$P\{Y_A \geq N\} \leq P\{Y_B \geq N\} \tag{F36}$$

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

$$E[Z_A|Y_A < N] = E[Z_B|Y_B < N] \tag{F37}$$

Hence, we have that

$$\begin{aligned} E[Z_A] &= E[Z_A|Y_A < N]P\{Y_A < N\} + NP\{Y_A \geq N\} \\ &\leq E[Z_B|Y_B < N]P\{Y_B < N\} + NP\{Y_B \geq N\} \quad = \quad E[Z_B]. \end{aligned} \tag{F38}$$

But, by Little's Law

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

and hence we have that

$$P_A\{Ab\} \geq P_B\{Ab\}. \tag{F39}$$

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.

Having the lower bound, we can now proceed with the asymptotic optimality. By our condition on the c_i's and θ_i's and by Proposition F.2, we have that the lower bound which is $c_J \lambda P_B\{Ab\}$ is asymptotically achieved, for any logarithmic threshold level. Finally, if the abandonment costs of high priorities grow polynomially with r we can still obtain the lower bound by using thresholds such that the probability of delay for class i is an $o(1/N^{\gamma_i})$. ■

F.2.1 Asymptotic Optimality - Constraint Satisfaction

Another problem of interest, in analogy to (1), is the problem of constraint satisfaction. Particularly, this problem deals with determining the minimal staffing required to ensure that the probability of *abandonment* for class i customers does not exceed a certain level η_i. More precisely, we consider the following problem: Given $\eta_1, ..., \eta_J$ in the open interval $(0, 1)$, find a staffing level N and a policy π such that:

$$\begin{aligned} &\text{minimize} \quad N \\ &\text{subject to} \quad P_\pi(Ab_i) \leq \eta_i, \quad i = 1, ..., J, \\ &\qquad\qquad\quad N \in \mathbb{Z}_+, \quad \pi \in \Pi. \end{aligned} \tag{F40}$$

For this problem, attention is restricted to the case in which the bounds on abandonment probabilities η_i, $i = 1, ..., J$ are fixed (independently of system size). Under this configuration, the different classes are allowed to have different service rates (in particular we assume that the service time of class i customers is exponentially distributed with rate μ_i). The optimal policy in this case gives rise to a *Pool Decomposition* solution. That is, it is asymptotically optimal to decompose the N servers into J groups of servers, such that class i customers are served by group i servers only. The case of $\eta_i's$ which scale with the system load is left for future research.

In the following, we assume, without loss of generality, that the classes are ordered such that $\eta_1 \geq \eta_2 \geq ... \geq \eta_J$. Also, we assume that class J is non-negligible (condition (9)). Before stating the asymptotic optimality of the pool decomposition policy we provide a lower bound on the staffing cost, as stated in the following Lemma.

Lemma F.1. *Consider a multi-class system where class i customers have exponential service times and patience with rates μ_i and θ_i, respectively. Additionally, class i arrival rate is λ_i^r. Denote by $(N^r)^*$ the optimal solution to (F40). Then,*

$$(N^r)^* \geq \sum_{i=1}^{J} \frac{\lambda_i^r}{\mu_i}(1-\eta_i). \qquad \text{(F41)}$$

Proof: For this proof, and since r is fixed, we omit the superscript for simplicity of notation. Define $T_i(t)$ to be the cumulative time dedicated to service of class i customers up to time t. Define $R_i(t)$ to be the number of abandonments from class i until time t, $A_i(t)$ to be the number of arrivals to class i until time t, $Q_i(t)$ to be the class i queue length at time t, and $D_i(t)$ the number of service completions from class i until time t. Also, define $T(t) = \sum_{i=1}^{J} T_i(t), D(t) = \sum_{i=1}^{J} D_i(t), Q(t) = \sum_{i=1}^{J} Q_i(t), R(t) = \sum_{i=1}^{J} R_i(t)$. Finally, $P\{Ab\}$ is the overall probability of abandonment: $P\{Ab\} = \sum_{i=1}^{J} \frac{\lambda_i}{\lambda} P\{Ab_i\}$.

Then, we have:

$$\alpha_i \geq P_i\{Ab\} = \lim_{t\to\infty} \frac{R_i(t)}{A_i(t)} = \lim_{t\to\infty} \frac{A_i(t) - D_i(t) + Q_i(t)}{A_i(t)} = \lim_{t\to\infty} 1 - \frac{\mu_i T_i(t)}{\lambda_i t}, \quad \forall i = 1, ..., J. \qquad \text{(F42)}$$

To justify the last equality, note the following:

- $A_i(t) \to \infty$, as $t \to \infty$. Also $A_i(t) = \lambda_i(t) + M_A(t)$ where $M_A(t)$ is a locally square-integrable martingale for which $\lim_{t\to\infty} \frac{M_A(t)}{A_i(t)} \to 0$.
- Steady state exists and hence $\lim_{t\to\infty} \frac{Q_i(t)}{A_i(t)} \to 0$.
- $D_i(t) = \mu_i T_i(t) + M_T(t)$ where $M_T(t)$ is a locally square-integrable martingale for which $\lim_{t\to\infty} \frac{M_T(t)}{A_i(t)} \to 0$.
- $T_i(t) \to \infty$, as $t \to \infty$ (otherwise $\lim_{t\to\infty} \frac{R_i(t)}{A_i(t)} \to 1$). Hence $\lim_{t\to\infty} \frac{D_i(t)}{A_i(t)} = \lim_{t\to\infty} \frac{\mu_i T_i(t)}{\lambda_i(t)}$.

Note that (F42) can be rewritten as

$$\lim_{t\to\infty} 1 - \mu_i T_i(t)/\lambda_i t \leq \alpha_i$$

or, alternatively, as

$$N^* \geq \lim_{t\to\infty} \sum_{i=1}^{J} T_i(t)/t \geq \sum_{i=1}^{J} \frac{\lambda_i}{\mu_i}(1-\alpha_i).$$

■

Having the lower bound, we proceed to the definition of asymptotic optimality associated with this problem:

Definition: Consider the problem (F40). Then, the sequence $\{N^r, \pi^r\}$ is **asymptotically optimal** with respect to $\bar{\lambda}^r$ if:

- it is asymptotically feasible: $\limsup_{r\to\infty} P_{\pi^r}\{Ab_i\} \leq \eta_i \,, \forall i = 1, ..., J$, and,
- any other asymptotically feasible sequence of policies $\{\bar{N}^r, \hat{\pi}^r\}$ satisfies $\liminf_{r\to\infty} \dfrac{\hat{N}^r - \underline{N}^r}{N^r - \underline{N}^r} \geq 1$,

where $\underline{N}^r = \sum_{i=1}^{J} \frac{\lambda_i^r}{\mu_i}(1-\eta_i)$.

The pool decomposition policy, stated in the following theorem is an immediate result of the lower bound.

Theorem F.2. *Consider a sequence of multi-class $M/M/N^r + M$ systems indexed by r. Suppose that class i arrival rate is λ_i^r, their service rate is μ_i and their abandonment rate is θ_i. Also, suppose that the upper bound on class i abandonment probability is η_i. Then, the following policy is asymptotically optimal with respect to (F40): Let the total staffing level be $N^r = \sum_{i=1}^{J} \lfloor \frac{\lambda_i^r}{\mu_i}(1-\eta_i) \rfloor$, and partition this server pool into distinct pools of sizes $N_i^r = \lfloor \frac{\lambda_i^r}{\mu_i}(1-\eta_i) \rfloor$, $i = 1, \ldots J$. Serve class i customers only by servers of pool i (i.e. convert the system into J single class systems).*

Proof: The proposed policy and staffing are clearly asymptotically feasible. Each class is now served in a single class $M/M/N + M$ with *ED* staffing, and, by the choice of the staffing, the probability of abandonment for class i is α_i (see for example [35]). The optimality is immediate since the lower bound is approached from below. ■

G Efficiency Driven $M/M/N$

In Section C, 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:

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

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 [36] 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 [12] (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. [13], where most of the details are omitted).

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

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

Remark G.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 [12] and the references therein for more details on this kind of limits.

Proof of Proposition C.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 [20]) 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.

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

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

Then,

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

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 [33]) all that is left to prove is that

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

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 $C_i^N = A_i^N + B_i^N$, $i = 1, 2, ...$ 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

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

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

$$\lim_{N\to\infty} P\{\sum_{i=1}^{\sigma^N(T)} A_i^N > \epsilon\} = 0. \tag{G8}$$

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 $N^\delta(1-\rho^N) \to \beta$ 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,

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

$N^\delta(1-\rho^N) \to \beta$ also implies that $\sqrt{N}(1-\rho^N) \to 0$ and hence, following [12] 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 [12], 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

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

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 $X(0) = 0$

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

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

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

The above is obtained by a simple adaptation of pages (67-68) in [12]. Now, $E[A_i^N] = o(\frac{1}{N^\delta})$ and the result follows. ■