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

Itay Gurvich[1] Mor Armony[2] Avishai Mandelbaum[3]

Abstract

We study large-scale service systems with multiple customer classes and many statistically identical servers. The following question is addressed: How many servers are required (staffing) and how does one match them with customers (control) in order to minimize cost or maximize profit, subject to quality of service (QoS) constraints? We tackle this question by characterizing scheduling and staffing schemes that are asymptotically optimal in the limit, as system load grows to infinity. The main asymptotic regime considered is the many-server heavy-traffic Quality and Efficiency Driven (QED) regime. The Efficiency Driven (ED) regime is also studied. In the QED regime, which was formally introduced by Halfin and Whitt, a delicate balance is obtained between server efficiencies and quality of service. This balance is enabled by the economies of scale associated with the system size.

Our main findings are: a) *Decoupling* of staffing and control, namely (i) Staffing disregards the multi-class nature of the system and is analogous to the staffing of a single class system with the same aggregate demand and the *lowest priority* class cost and QoS parameters, and (ii) Class level service differentiation is obtained by using a simple *threshold-priority* (TP) control (with state-independent thresholds), b) *Robustness* of the staffing and control rules: In the QED regime, our proposed Square-Root Safety (SRS) staffing rule and TP control are asymptotically optimal with respect to various problem formulations and model assumptions. c) The QED and ED regimes are obtained as solutions of the joint staffing and control problem rather than as assumptions.

Acknowledgement: The authors thank Rami Atar and Haya Kaspi for many helpful comments.

[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

1 Introduction

Modern service systems thrive to provide customers with personalized service, which is customized to the customers needs. Recent trends include self selecting market segmentation, multi-lingual customer support, and customized cross-sales offerings. With this growing level of service customization, the variety of services provided by any given organization is increasingly high. This variety requires service personnel to possess a large skill set. It has been long recognized that in order to avoid over-staffing it is important to cross-train customer service representatives and maintain server flexibility. However, to take full advantage of this high flexibility level, one needs to make efficient customer-server assignments and sensible staffing and cross-training decisions. This latter challenge is now receiving increasing attention, and is where this work's contribution lies.

One way to think of a cross-trained or flexible server environment is as an opportunity to resolve a common conflict between an organization marketing and operational goals; generally, from a marketing point of view, resource allocation should be based on customer relationship management (CRM) criteria such as revenue generating potential and quality of service commitments. A natural environment to support CRM is the dedicated service resources one. However, from an operational viewpoint, this typically implies an inefficient use of resources. Instead, we suggest that in large systems, fully cross-trained servers can provide an efficient service to multiple customer classes; at the same time, carefully allocating these servers dynamically among the different classes facilitates the appropriate requirements of CRM.

Our work is largely motivated by modern call centers which often consist of dozens, hundreds or even thousands of agents, and who thrive to meet a large variety of customers needs. Examples include direct banking, multi-lingual services, and help desks. In such centers, a customer class may be characterized by its members special service needs, their relative importance to the organization, or their quality of service expectations or guarantees. We model such systems by a multi-class multi-server queue with many servers, which we call the V-design model. This model is depicted in Figure 1.

With respect to the V-design model we ask the following question: How many servers are required (staffing) and how does one match them with customers (control) in order to maximize profit (or minimize cost), subject to class-level quality of service constraints?

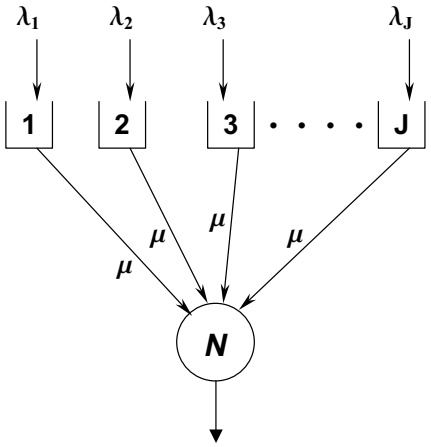

Figure 1: The V Model - multiple customers classes and a single server type.

The staffing and control decisions are generally made at different time scales. While the control decisions are made on-line in real-time, the staffing decisions are often made on a weekly basis, or even less frequently. Generally speaking, this implies that when staffing decisions are made, the information about the arrival rate (call volume) is incomplete, and is based on forecasts. Therefore, to avoid under- or over-staffing, it is desirable to come up with staffing rules that require only limited demand information. We show that to make the staffing decision it is sufficient to know only *aggregate* demand information instead of class-level forecasts.

The demand uncertainty becomes even more of an issue when the service is being performed by a third party who does not have access to demand information. This problem appears to be of increasing importance due to the proliferation of call-center outsourcing. As it is often the case with sub-contracting, the uncertainty associated with future demand together with information asymmetry can cause incentive misalignments between the two parties, which may result in system inefficiencies (e.g. [9]). To resolve these inefficiencies a mechanism needs to be designed that would enforce the multidimensional demand to be shared truthfully. But such multidimensional signalling problems are notoriously hard. Our insight reduces the problem into a one-dimensional one, that may be more tractable.

Even though the staffing and control decisions involve different time scales, it is important to consider these problems together in a common framework; if the on-line assignment of servers to

customer classes is not optimal, many more servers may be required. Additionally, if the performance evaluation of the system under a given on-line control is incomplete, staffing levels may be either too high or too low. Finally, some on-line control rules are, by nature, staffing level dependent. Nevertheless, due to the relative complexity of the joint staffing-control problem, they have generally been considered separately in the literature. Recent exceptions include [2, 3, 1, 4, 6, 17, 31] as well as this current paper.

Our approach in addressing the staffing and control question is an asymptotic one; specifically, we characterize scheduling and staffing schemes that are asymptotically optimal as the aggregate arrival rate increases to infinity. The main asymptotic framework considered in this work is the many-server heavy-traffic regime, first introduced by Halfin and Whitt [15]. We refer to this regime as the QED (Quality and Efficiency Driven) regime. Systems that operate in the QED regime enjoy an unusual combination of high efficiencies together with high quality of service. In addition to the QED regime, we also consider the so called "Efficiency Driven (ED) Regime" in which server efficiency is emphasized over service quality.

1.1 Main Results

This paper's main results are:

1. The *joint* problem of staffing and control may be *decoupled* into two separate problems where:

   (a) The *staffing* level is the same as in a single class system with a common *total* arrival rate, and cost and quality of service characteristics which are equal to the those of the *lowest priority* customers.

   (b) The on-line *control* provides quality of service differentiation between the various customer classes via a *threshold priority* (TP) scheduling rule. The thresholds associated with this rule are state-independent and their values are easily determined as a function of the system parameters.

   Thus, the staffing rule has the desirable property that it only requires partial demand information. Particularly, no class-level arrival rate information is needed. When these arrival rates become known in real-time the control decisions make full use of this new information.

2. Robustness of staffing and control: The Square-Root Safety (SRS) staffing rule (of the form: $R + \beta\sqrt{R}$, where R is the system load and β is a constant) together with the TP control are shown to be asymptotically optimal (in the QED regime) for a variety of problem formulations and model assumptions, including constraint satisfaction, cost minimization and profit maximization, with or without customer abandonment.

3. The ED and the QED regimes are obtained as solutions to the staffing and control problems, rather than as assumptions. That is, we show that in order to optimize with respect to costs or profits, subject to quality of service constraints, one should choose staffing levels that are consistent with these asymptotic regimes. The specific regime is determined by the problem parameters.

The *simplicity* of the suggested staffing rule is of great importance. A-priory, staffing decisions that need to take into consideration the service requirement of multiple customer classes can potentially be very complex. Our result, that only total arrival rate and the lowest priority service characteristics are needed, simplifies the staffing decision tremendously. Moreover, the *form* of this SRS staffing rule as a function of these two parameters is also very simple. (Note that in order for a square-root staffing rule to be useful, the system load R must be predicted at an accuracy level under which square-root deviations are significant. In contrast, [17, 6] consider a framework in which deviations of this magnitude are insignificant).

The dynamic control we propose of matching servers to customers is based on priorities and thresholds. In a nutshell, customer classes are ordered with respect to service priorities. A customer of a certain priority can enter service only if there are no higher priority customers waiting, and the number of idle servers exceeds a class-dependent threshold. The role of the thresholds is to ensure that enough servers are available to serve *future* arrivals of *higher* priorities. The thresholds can be easily adjusted to provide the right level of service. The dependence of the quality of service on the threshold is illustrated in Table 1.

1.2 A Two-class example

The example depicted in Table 1 has two customer classes with class 1 having higher priority over class 2. Here, the number of servers N is assumed to be large and to satisfy $N = R + \beta\sqrt{R}$, for

$\beta > 0$ (recall that R is the system load). Additionally, ρ_1 corresponds to the traffic intensity of class 1 customers, and W_i is the steady-state virtual waiting time associated with class i ($i = 1, 2$). Finally, $\alpha(\beta)$ is a decreasing function of β given in (3), and $a_N = \Theta(b_N)$ denotes that a_N and b_N are of the same order of magnitude.[4]

♯	Threshold K	$\sim P\{W_1 > 0\}$	$\sim P\{W_2 > 0\}$	$E[W_1 \mid W_1 > 0]$	$E[W_2 \mid W_2 > 0]$
A	0	$\alpha(\beta)$	$\alpha(\beta)$	$\Theta(\frac{1}{N})$	$\Theta(\frac{1}{\sqrt{N}})$
B	b	$\alpha(\beta) \cdot \rho_1^b$	$\alpha(\beta)$	$\Theta(\frac{1}{N})$	$\Theta(\frac{1}{\sqrt{N}})$
C	$c \cdot \ln N$	$\alpha(\beta) \cdot \rho_1^{c \ln N}$	$\alpha(\beta)$	$\Theta(\frac{1}{N})$	$\Theta(\frac{1}{\sqrt{N}})$
D	$d \cdot \sqrt{N}, \quad d < \beta$	$\Theta(\alpha(\beta - d)\rho_1^{d\sqrt{N}})$	$\alpha(\beta - d)$	$\Theta(\frac{1}{N})$	$\Theta(\frac{1}{\sqrt{N}})$

Table 1: Service Levels as a function of the threshold for a two-class V model.

Note how the threshold level affects the quality of service of the two classes. Specifically, with respect to the probability of delay, when using a threshold of size 0 (static priorities) both classes enjoy a QED service, which is typified by delay probabilities that are strictly between 0 and 1. As the threshold level increases to a positive constant, class 2 delay probability stays the same, while class 1 receives a higher quality of service. As the threshold increases even more, class 1 delay probability approaches 0 (if the number of servers is large), which is a characteristic of the Quality Driven (QD) regime. In contrast, the expected wait given that a customer is delayed does not change with the level of the threshold. For this measure of quality of service class 2 always experiences QED service, while class 1 enjoys waiting times which are consistent with the QD regime.

1.3 Paper structure

The rest of the paper is organized as follows: The introduction is concluded with a brief literature review. We formally introduce the joint staffing and control problem in Section 2. The threshold-priority (TP) rule and the corresponding queueing model (denoted by $M/M/N/\{K_i\}$) as well as the square-root safety (SRS) staffing are also introduced in this section. Section 3 establishes the asymptotic feasibility of our proposed joint staffing and control policies. Section 4 then shows the asymptotic optimality of TP and SRS. In Section 5 staffing and control are discussed for the

[4] More formally, we say that $a_N = \Theta(b_N)$ if $0 < \lim_{N\to\infty} a_N/b_N < \infty$.

extension of the original model in which customers may abandon before their service starts. To conclude, Section 6 summarizes the results and suggests directions for further research.

Due to the technical nature of our results, our approach in their presentation is to state them formally and precisely, followed by an intuitive explanation. The formal proofs appear in a technical appendix to this paper [14].

1.4 Literature Review

There is extensive literature dealing with the V-Model both in terms of performance analysis and in terms of performance optimization and control. However, little work has been done on the staffing problem and especially on the combined solution of staffing and control. Next we mention only the papers most closely related to our work.

Exact steady-state performance analysis of the V-Model under non-preemptive priorities (without thresholds) is given in Federgruen and Groenvelt [11]. The threshold-priority scheme $M/M/N/\{K_i\}$ was analyzed by Schaack and Larson [28]. The latter provides steady-state expressions for the different performance measures associated with this model.

In terms of control, Yahalom and Mandlebaum [37] consider the multi-server V-Model with Poisson arrival streams and identically distributed exponential service time for all customer classes. They conjecture that a threshold-priority policy is optimal in terms of minimizing discounted holding costs in the long run. However, in contrast to [28], the thresholds depend on the system state. It is important to note that [37] discussed the *structure* of the scheduling policy but it does not address the problem of *choosing* the appropriate threshold *levels*. The dependence of the thresholds on the state of the system makes the choice of the thresholds a very complicated task.

Note that the threshold-priority policy is not work conserving, in the sense that lower priority customers may be not allowed to enter service even though some of the servers are idle. When restricted to work conserving policies and a single server, it can be proved by simple interchange arguments (see [32]) that the $c\mu$ rule is optimal even when classes have different service requirements. As for multi-server, the authors of [11] proved that the $c\mu$ rule is optimal among all *work-conserving* policies for the multi-class $M/M/N$ queue with linear holding costs. However, this policy is sub-optimal when allowing for intensional idleness.

Under conventional heavy traffic, the V Model, as well as more complicated scenarios, are

amenable to analysis. Van Mieghem [30] analyzed the single server V Model under heavy traffic and proved the asymptotic optimality of the so-called Generalized $c\mu$ (or $Gc\mu$) rule . Later, in [21], a generalization of this policy was proved to be optimal under conventional heavy traffic for convex holding cost functions and for a very rich family of network topologies, including the V model.

Limits in the *QED* regime for the V Model were first introduced in Puhalskii and Reiman [25]. Here, the authors considered the more general setting of $GI/PH/N$, under FIFO and static priorities service disciplines. Later, Armony and Maglaras [2] and [3] were the first to consider a control problem in the *QED* regime. The authors consider a call center with two classes of service: real-time and postponed service with guaranteed delay. The resulting system is a V-design call center, with two customer classes, and a single server pool. For this system, the authors in [2, 3] devise a routing algorithm which is asymptotically optimal in the sense that it (asymptotically) minimizes the waiting time of the real-time service while complying with the delay bound of the postponed one. This non-probabilistic delay bound is the main feature that differentiates their setting from ours. They finally determine that the square-root safety-staffing rule is also asymptotically optimal under the criteria of minimizing staffing costs, while maintaining pre-specified performance measures.

Another control optimization problem of the V Model in the *QED* regime was considered in Atar et al. [5]. A Brownian control problem is constructed for the V model under exponentially distributed service times and where all customer classes have exponential patience. For linear discounted queueing costs it is shown that under particular assumptions (such as equal service rates) the asymptotically optimal policy leads in the QED regime to a limit that is a one-dimensional diffusion. This gives a structural insight about the asymptotic performance of the optimal policy but it does not provide a specific policy to obtain this performance. Similar problem is studied in [16]. There, the authors characterize structural properties of the asymptotically optimal routing policies, in the QED regime. As a special case, they show that if all customers share the same service and abandonment rates, then the cμ rule is asymptotically optimal.

2 Model Formulation

Consider a large service system which is modelled as a multi-class queueing system with J customer classes and N statistically identical servers. Customers of class i arrive according to a Poisson process with rate λ_i, independently of other classes. Service times are assumed to be exponential with rate μ for all customer classes. Class i delayed customers wait in an infinite buffer queue i.

The description below assumes that customers do not abandon (the model which includes abandonment is described in Section 5). Two general forms joint staffing and control problems are considered: One is the minimization of the number of servers subject to a target service level constraint. The other is a cost minimization problem which seeks to minimize waiting plus staffing costs, subject to quality of service constraints. For both problem formulations, the quality of service constraints are in terms of an upper bound on the steady state probability that a class i customer waits before starting service (delay probability). Let W_i be the steady-state waiting time of class i customers, and denote class i delay probability by $P\{W_i > 0\}$. Let α_i be class i target delay probability. It is assumed, without loss of generality, that the classes are ordered in an increasing order of α_i: $\alpha_1 \leq \alpha_2 \leq ... \leq \alpha_J$.

For a given staffing level N, a control policy is a set of rules that determine how to match calls with servers at any given time. Let Π be the set of all non-preemptive non-anticipative control policies. Given that a control policy $\pi \in \Pi$ is used, let $P_\pi\{W_i > 0\}$ be the delay probability of class i customers. The joint staffing and control problem is then stated as follows: Given $0 < \alpha_1 \leq \alpha_2 \leq ... \leq \alpha_J < 1$,

$$\begin{array}{ll} \text{minimize} & N \\ \text{subject to} & P_\pi(W_i > 0) \leq \alpha_i, \quad i = 1, ..., J, \\ & N \in \mathbb{Z}_+, \quad \pi \in \Pi. \end{array} \tag{1}$$

A solution to (1) should involve both the staffing level N, as well as the control rule that obtains the performance level constraints.

The alternative problem formulation that includes the waiting time cost as part of the objective

function is: Given $0 < \alpha_1 \leq \alpha_2 \leq ... \leq \alpha_J \leq 1$ and $c_1 \geq c_2 \geq ... \geq c_J \geq 0$,

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{J} c_i \lambda_i E[W_i] + N \\ \text{subject to} & P_\pi(W_i > 0) \leq \alpha_i\,, \quad i = 1, ..., J, \\ & N \in \mathbb{Z}_+, \quad \pi \in \Pi. \end{array} \tag{2}$$

Here one assumes linear waiting costs for all classes, i.e. a unit waiting time of a class i customer incurs a cost of c_i. We show that if classes are ordered in decreasing order of their cost (i.e. $c_1 \geq c_2 \geq ... \geq c_J$), then the same type of policy that asymptotically minimizes (1) is also the solution to (2). Note that in (2) the values of the α_i's are allowed to be equal to 1. If all α_i's are equal to 1 (2) becomes a pure cost minimization problem.

As an alternative to (2), consider a profit maximization problem in which a customer of class i generates a revenue which is decreasing in her waiting time. Specifically, suppose that the revenue generated by a class i customer is $r_i - c_i E[W_i]$. Consequently, the revenue rate generated by class customers is $r_i\lambda_i - c_i\lambda_i E[W_i]$. Since the factor $r_i\lambda_i$ is independent of the choice of staffing and control, the resulting problem of profit maximization is equivalent to (2).

For a general control policy $\pi \in \Pi$, let $Z(\pi; t)$, $Q_i(\pi; t)$, and $Y(\pi; t) = Z(\pi; t) + \sum_{i=1}^{J} Q_j(\pi; t)$ be the number of busy servers, the number of class i customers in queue, and the total number in the system at time t, respectively. We omit π from the notation whenever it is clear from the context which control policy is used. Also, t is replaced by a $\cdot$ when the entire process is considered. Finally, the time argument is omitted when referring to steady-state quantities.

Our proposed solution to the joint staffing and control problem is the Square-Root Safety (SRS) staffing rule and Threshold-Priority (TP) control. These rules are described as follows:

- **Staffing - the SRS Rule:** Let R be the system load (that is, $R = \lambda/\mu$), and let

$$\alpha(\beta) \stackrel{\triangle}{=} \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}, \tag{3}$$

  be the Halfin-Whitt delay probability function, where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal distribution and density functions, respectively. Also, let $\beta(\cdot) = \alpha^{-1}(\cdot)$ (that is $\beta(\cdot)$ is the inverse of $\alpha(\cdot)$). Then, according to the SRS rule, the staffing level should be of the form:

$$N = R + \beta\sqrt{R}, \tag{4}$$

  where β satisfies $\beta \geq \beta(\alpha_J)$. More specific values of β are proposed in Section 4.

- **Control - the TP rule:** Upon a customer arrival or a service completion, assign the head-of-the-line class i customer to an idle server if and only if (1) queue j is empty for all higher priority classes j (i.e. $j < i$), and (2) the number of idle servers exceeds K_i. Here, $0 = K_1 \leq K_2 \leq ... \leq K_J$ are the threshold levels. We denote the queueing model associated with this policy as $M/M/N/\{K_i\}$. The particular values of the thresholds are determined according to the recursive formula:

$$K_i - K_{i-1} = \left\lceil \frac{\ln \alpha_{i-1}^* - \ln \alpha_i^*}{\ln \rho_{\leq i-1}} \right\rceil, \quad i = 2, \ldots, J$$
$$K_1 \equiv 0, \tag{5}$$

where

$$\rho_{\leq i-1} \triangleq \frac{\sum_{j=1}^{i-1} \lambda_j}{N\mu}, \tag{6}$$

and $\alpha_i^* \leq \alpha_i$. More concrete values of α_i^* are spelled out in Sections 3 and 4.

For the $M/M/N/\{K_i\}$ model, the J dimensional state descriptor $\{Z(t) + Q_1(t), Q_i(t) : i = 2, ..., J\}$ is a continuous time Markov chain (CTMC). For this system, class i delay probability can be stated in terms of the system state (due to the PASTA property) as follows:

$$P_{TP}\{W_i > 0\} = P_{TP}\{Z \geq N - K_i\} \tag{7}$$

2.1 The QED asymptotic framework

Our approach is solving the joint staffing and control problem is asymptotic. Specifically, we consider a sequence of systems with increasing arrival rates, and characterize staffing and control schemes which are *asymptotically* optimal, as the arrival rates increase to ∞. The original system of interest is assumed to be a member in this sequence. If the total arrival rate for this system is sufficiently large, then an asymptotically optimal policy is likely to be nearly optimal for this original system.

There are several reasons why it makes sense to consider an asymptotic approach to this problem instead of an exact one. First, it is clear from [37] that an optimal control policy that minimizes waiting costs must be highly dependent on system parameters and system state. Particularly, implementing such a control is difficult due to the large state-space and the large number of system

parameters. Even if attention is restricted to the threshold-priority (TP) rule, the actual threshold values need to be determined. In addition, for staffing purposes, one would need to evaluate the system performance given different values of N. An exact approach would lead to very complicated expressions, and is not likely to provide useful and general insights.

Following the asymptotic approach, we consider a sequence of systems indexed $r = 1, 2, ...$ (to appear as superscript) with an increasing total arrival rate $\lambda^r = \sum_{i=1}^{J} \lambda_i^r$ and a fixed service rate $\mu^r \equiv \mu$. Let $R^r = \lambda^r/\mu$ be the total system load, then, without loss of generality, we assume that the index r is selected such that

$$r \equiv R^r. \tag{8}$$

The arrival rates to the different classes may be quite general. We only assume that the arrival rate of the lowest priority is comparable to λ^r for each r. More formally, we assume that there are J numbers $a_k \geq 0$, $k = 1, ..., J$, with $\sum_{k=1}^{J} a_k = 1$, such that the arrival rate of each class behaves according to the following rule:

$$\lim_{r\to\infty} \frac{\lambda_k^r}{\lambda^r} = a_k,\ k = 1, ..., J; \quad a_J > 0,\ a_i \geq 0,\ i = 1, ..., J-1. \tag{9}$$

The specifics of the TP policy in the asymptotic framework are as follows: Suppose that the r^{th} system is staffed with N^r servers and the customers are routed according to J thresholds given by $0 = K_1^r \leq K_2^r \leq ... \leq K_J^r \stackrel{\triangle}{=} K^r$. The appropriate staffing and threshold levels are determined (in section 4) to asymptotically optimize (1) and (2), where both $\alpha_1, ..., \alpha_{J-1}$ and $c_1, ..., c_J$ are allowed to scale with r. For a sequence of $M/M/N^r/\{K_i^r\}$ systems, we say that it operates in the QED regime if the number of servers grows with r in the following manner[5]:

$$\lim_{r\to\infty} \sqrt{N^r}(1-\rho^r) = \beta\,,\ 0 < \beta < \infty \tag{10}$$

and

$$K^r = o(\sqrt{N^r}).$$

[5]Note that if $K^r = O(\sqrt{N^r})$ (where, $a^r = O(b^r)$ means that $\limsup_{r\to\infty} a^r/b^r < \infty$), then, to characterize the QED regime, (10) needs to be replaced by the condition: $\lim_{r\to\infty} \sqrt{N^r}(1-\rho_C^r) = \beta', 0 < \beta' < \infty$, where $\rho_C^r \stackrel{\triangle}{=} \frac{\lambda^r}{(N^r-K^r)\mu}$. Clearly, if, $K^r = o(\sqrt{N^r})$ (where $a^r = o(b^r)$ means that $\lim_{r\to\infty} a^r/b^r = 0$), then $\beta' = \beta$. For simplicity of presentation, we restrict our exposition to the case $K^r = o(\sqrt{N^r})$. Most results follow through to the case of $K^r = O(\sqrt{N^r})$.

Here,

$$\rho^r \triangleq \frac{\lambda^r}{N^r \mu}. \tag{11}$$

3 Asymptotic Feasibility of TP and SRS

In this section we establish that our proposed Square-Root Safety (SRS) staffing rule and Threshold-Priority (TP) control are asymptotically feasible in the QED regime for the problems (1) and (2).

We start by defining asymptotic feasibility. For $r = 1, 2, ...$, consider a sequence of systems with fixed number of customer classes J and a fixed service rate. Let $\bar{\lambda}^r = \{\lambda_1^r, ..., \lambda_J^r\}$ be a sequence of arrival rates with a total arrival rate $\lambda^r = \sum_{i=1}^J \lambda_i^r$ which is increasing to ∞ as $r \to \infty$. Let (N^r, π^r) be a joint staffing and control pair associated with the r^{th} system.

Definition: The sequence $\{(N^r, \pi^r)\}$ is **asymptotically feasible** with respect to $\bar{\lambda}^r$ and $\bar{\alpha}^r = (\alpha_1^r, ..., \alpha_J^r)$ if, the following condition applies:

$$\limsup_{r \to \infty} \frac{P_{\pi^r}\{W_i^r > 0\}}{\alpha_i^r} \leq 1 \,, \forall i = 1, ..., J.$$

The following theorem formally states the asymptotic feasibility of SRS and TP.

Theorem 3.1. ***(Asymptotic Feasibility of TP and SRS)*** *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 class i arrival rate λ_i^r, $i = 1, ..., J$, which satisfy (9) (that is, class J is comparable). For class i, $i = 1, ..., J-1$, suppose that the delay bound α_i^r decreases polynomially in r (that is, $\alpha_i^r = \alpha_i r^{\xi_i}$ for some $\alpha_i > 0$ and $\xi_i \leq 0$), while for class J assume that α_J^r is independent of r ($\alpha_J^r \equiv \alpha_J$). In addition, suppose that for all r large enough, we have $\alpha_1^r < \alpha_2^r < ... < \alpha_J$. Let $\rho^r = \frac{\lambda^r}{N^r \mu}$, and suppose that $K^r = o(\sqrt{N^r})$. Then, the quality of service guarantee of class J is asymptotically satisfied, that is,*

$$P\{W_J^r > 0\} \to \alpha_J, \quad \text{as } r \to \infty, \quad 0 < \alpha_J < 1, \tag{12}$$

if and only if the staffing levels satisfy

$$\sqrt{N^r}(1 - \rho^r) \to \beta, \quad \text{as } r \to \infty, \quad 0 < \beta < \infty, \tag{13}$$

where, $\beta = \beta(\alpha_J)$. In addition,

$$P\{W_i^r > 0\} \sim \alpha(\beta) \cdot \prod_{j=i}^{J-1} (\rho_{\leq j}^r)^{K_{j+1}^r - K_j^r}, \quad i = 1, ..., J-1, \tag{14}$$

where $\rho^r_{\leq j} = \frac{\sum_{i=1}^{j} \lambda_i^r}{N^r \mu}$, *and* $a^r \sim b^r$ *means that* $\lim_{r\to\infty} a^r/b^r = 1$. *Finally, if, in addition, all classes are non-negligible, i.e.* $\lambda_i^r/\lambda^r \to a_i > 0$, $i = 1, ..., J$, *and* α_i^r *is independent of* r *for all* $i = 1, ..., J$, *then the quality of service constraints of all classes are asymptotically satisfied, that is,*[6]

$$P\{W_i^r > 0\} \to \alpha_i^*, \quad as \quad r\to\infty, \quad 0 < \alpha_i^* \leq \alpha_i,\ i = 1, ..., J-1, \tag{15}$$

if and only if the threshold values satisfy the recursive relationship

$$K_{i+1}^r - K_i^r \to \frac{\ln \alpha_i^* - \ln \alpha_{i+1}^*}{\ln \rho_{\leq i}}, \quad as\ r\to\infty, \quad \forall i = 2, \ldots, J, \tag{16}$$

where $\rho_{\leq i} = \lim_{r\to\infty} \frac{\sum_{j=1}^{i} \lambda_j^r}{N^r \mu}$.

Remark 3.1. (Intuitive Explanation of Theorem 3.1) It is easy to understand the result by looking at the dynamics of the suggested policy in a simple two class case. To explain the equivalence between (12) and (13) we claim that the probability of delay for the low priority class is approximately the same as the delay probability for a single class $M/M/N-K$ system. Note that the delay probability in both these systems is the probability of having at least $N-K$ busy servers. For the $M/M/N-K$ system the equivalence between (12) and (13) was established in [15].

To see why the two systems have similar delay probabilities, 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 $N-K$ 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 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

[6]Note that, when using TP, it may be impossible to obtain delay probabilities which are *exactly* equal to the upper bounds α_i, because the thresholds must be integers. This is the reason why α_i^* replaces α_i in (15) and (16).

expect that the probability of finding more than $N-K$ busy servers would be approximately the same.

To understand (14), 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. The equivalence between (15) and (16) immediately follows from (14).

Remark 3.2. Note that the equivalence between (12) and (13) is independent of the thresholds values, as long as $K^r = o(\sqrt{N^r})$. In other words, if the control policy is TP with arbitrary threshold values that satisfy $K_J^r = K^r = o(\sqrt{N^r})$, then SRS guarantees that the delay probability constraint of class J is satisfied.

Remark 3.3. Notice that since $\alpha_J < 1$, the QED regime is obtained due to the feasibility requirements of (1) and (2) rather than an assumption.

Remark 3.4. The asymptotic feasibility of SRS and TP is true provided that the thresholds $K_1^r, ..., K_J^r$ are computed according to (5), with α_i^* taken to be α_i; that is, even if some of the classes are of negligible proportion, or if (16) does not hold, then the limiting delay probability is less than or equal to the target of α_i. If, in addition, all classes are of comparable size and the thresholds satisfy (16), then the delay probability is obtained as an equality in the limit. Finally, note that if K_i^r, $i = 1, .., J$ are calculated according to the recursive relationship (5) then, indeed, $K^r = o(\sqrt{N^r})$.

Remark 3.5. (Asymptotic Feasibility in the Efficiency Driven Regime) Suppose that some of the customer classes have no constraints on their delay probability, while others still do. In particular, this implies that $\alpha_J = 1$ and $\alpha_1 < 1$. In this case the following staffing and routing rules are asymptotically feasible: Staff with $N^r = R^r + \beta R^{1-\delta}$, for some $1/2 < \delta \leq 1$, and route according to TP, where the thresholds are still determined by (5). The value of δ is determined according to Remark 4.4. As suggested by its name, the efficiency driven regime is a regime in which the servers have high efficiencies, but quality of servers suffers, as is suggested by the fact that virtually *all* customers of some of the classes will wait ($\alpha_J = 1$).

Theorem 3.1 evaluates the delay probability for the different customer classes under the SRS and TP policies. But what about the actual waiting time, given that a customer is indeed delayed? Table 1 in the introduction suggests that for the high priority classes this wait is of order of $\Theta(1/N)$, while for the lowest priority class it is of the order of $\Theta(1/\sqrt{N})$. According to the following proposition, this is indeed the case. The proposition also provides expressions for the limiting distribution of the normalized waiting times (conditional on a positive wait). Notice that these limiting distributions do not depend on the particular threshold values, and for $i = 1, ..., J-1$ they do not depend on β.

Proposition 3.1. *Under the assumptions of Theorem 3.1 and assuming that SRS are TP are used, the steady state waiting time of the lowest priority class J satisfies:*

$$\sqrt{N^r} W_J^r \Rightarrow W_J, \quad as\ r \to \infty, \tag{17}$$

where W_J has the simple distribution:

$$W_J \sim \begin{cases} \exp(a_J \mu \beta) & w.p.\ \alpha(\beta), \\ 0 & otherwise. \end{cases} \tag{18}$$

In addition, the steady state waiting times of the higher priorities $i = 1, ..., J-1$ satisfy:

$$N^r \cdot [W_i^r | W_i^r > 0] \Rightarrow [W_i | W_i > 0], \quad as\ r \to \infty, \tag{19}$$

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

$$\begin{cases} \frac{\mu(1-\sigma_1)}{s+\mu(1-\sigma_1)}, & i = 1, \\ \\ \frac{\mu(1-\sigma_i)(1-\tilde{\gamma}_i(s))}{s-\hat{\lambda}_i+\hat{\lambda}_i\tilde{\gamma}_i(s)}, & i = 2, ..., J-1, \end{cases} \tag{20}$$

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

$$\tilde{\gamma}_i(s) = \frac{s+\mu}{2b_i\mu} + \frac{1}{2} - \sqrt{\left(\frac{s+\mu}{2b_i\mu} + \frac{1}{2}\right)^2 - \frac{1}{b_i}}, \tag{21}$$

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

$$\begin{gathered} N^r E[W_i^r | W_i^r > 0] \to [\mu(1-\sigma_i)(1-\sigma_{i-1})]^{-1}, \quad as\ r \to \infty, \quad and \\ (N^r)^2 E[(W_i^r)^2 | W_i^r > 0] \to 2(1-\sigma_i\sigma_{i-1}) \left[(\mu)^2(1-\sigma_i)^2(1-\sigma_{i-1})^3\right]^{-1}, \quad as\ r \to \infty. \end{gathered} \tag{22}$$

Remark 3.6. (Intuitive Explanation of Proposition 3.1) 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 3.1 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.

To understand (17) and (18) note that the overall number of customers in queue is asymptotically similar to number of customers in a single class multi-server queue in the QED regime. For the latter, it is known that the queue length is $\Theta(\sqrt{N})$. Since, as noted above, all high priorities have queue that is $O(1)$ we can deduce that the queue of the low priority would be $\Theta(\sqrt{N})$. This, in turn, implies (due to a distributional version of Little's law) that their waiting time is $\Theta(1/\sqrt{N})$.

As a direct consequence of Proposition 3.1 one can conclude that the order of magnitude of the queue lengths associated with the higher priority classes is $\Theta(1)$. The details are stated in the following corollary.

Corollary 3.1. *Under the assumptions of Theorem 3.1, and assuming that SRS and TP are used, the class level queue lengths for the high priority classes satisfy* $E[Q_i^r|Q_i^r>0]=\Theta(\lambda_i^r/N^r)$, $i=1,2,...,J-1$. *In particular, for* $i=1,...,J-1$, *and using the notation of Proposition 3.1,*

$$E[Q_i^r|Q_i^r>0]=\lambda_i^r E[W_i^r|W_i^r>0]\to\hat{\lambda}_i\left[\mu(1-\sigma_i)(1-\sigma_{i-1})\right]^{-1},\quad \text{as } r\to\infty,\quad \text{and}$$
$$E[Q_i^r]\sim\tfrac{\lambda_i^r}{N^r}P\{W_i^r>0\}\left[\mu(1-\sigma_i)(1-\sigma_{i-1})\right]^{-1}. \tag{23}$$

An important implication of Proposition 3.1 and Corollary 3.1 is that if queue lengths are scaled by $1/\sqrt{N^r}$, only the queue length of the lowest priority J does not disappear in the limit as $r\to\infty$. This essentially implies that, when r is very large, it is sufficient to know the total queue

length in order to deduce the class level queue lengths. This result is summarized in the following proposition.

Proposition 3.2. ***(State Space Collapse)*** *Under the assumptions of Theorem 3.1, and assuming that SRS and TP are used,*

$$\begin{aligned} &\tfrac{1}{\sqrt{N^r}} Q_i^r \Rightarrow 0,\ i = 1, ..., J-1 \\ &\tfrac{1}{\sqrt{N^r}} Q_J^r \Rightarrow X, \end{aligned} \tag{24}$$

where X is a non-negative random variable with a density $f(\cdot)$ which satisfies: $f(x) = \exp\{-\beta x\}\alpha(\beta)$, for $x > 0$, and $P\{X = 0\} = 1 - \alpha(\beta)$.

4 Asymptotic Optimality

In this section we establish the asymptotic optimality of SRS and TP as a joint staffing and control solution to the problems (1) and (2). In contrast with the feasibility discussion of Section 3, here each of the problems (1) and (2) require a somewhat different treatment, and are therefore discussed separately. Some common features, though, are discussed first.

First, note that our analysis is done under the assumption that the lowest priority class is comparable (9). Without this condition, many of our results are incorrect. Second, the definition of asymptotic optimality is common between the two problems. One would expect that, in order to insure stability, a reasonable staffing level would be of at least the order of λ^r. Hence, different staffing level propositions are expected to all be of the same order of magnitude. Therefore, in order to obtain a meaningful form of asymptotic optimality one must compare *normalized* staffing costs that measure the difference between the actual staffing costs and a base cost of the order of λ^r, which is a lower bound of the staffing cost.

To define asymptotic optimality, let $\bar{K}^r = \{K_1^r, ..., K_J^r\}$ and $\bar{\lambda}^r = \{\lambda_1^r, ..., \lambda_J^r\}$ the thresholds and arrival rates in the r^{th} system. Note that $R^r = \lambda^r/\mu$ is a lower bound on the value of the objective function in both (1) and (2), because at least R^r servers are required for stability. For (1) and (2), let the cost function be

$$C^r(N^r, \pi^r) = N^r, \quad \text{and} \quad C^r(N^r, \pi^r) = c_1^r \lambda_1^r EW_1^r + ... + c_J^r \lambda_J^r EW_J^r + N^r,$$

respectively, when the system is staffed with N^r servers and is controlled by π^r.

Definition: The sequence $\{N^r, \pi^r\}$ is **asymptotically optimal** with respect to $\bar{\lambda}^r$, $\bar{\alpha}^r = (\alpha_1^r, ..., \alpha_J^r)$, and $C^r(\cdot, \cdot)$ if the following two conditions apply:

- *Asymptotic feasibility*: $\limsup_{r\to\infty} \frac{P_{\pi^r}\{W_i^r > 0\}}{\alpha_i^r} \leq 1, \quad \forall i = 1, ..., J$; and
- *Asymptotic optimality*: For any other sequence of policies $\{\hat{N}^r, \hat{\pi}^r\}$ that is asymptotically feasible we have

$$\liminf_{r\to\infty} \frac{C^r(\hat{N}^r, \hat{\pi}^r) - R^r}{C^r(N^r, \pi^r) - R^r} \geq 1$$

4.1 Constraint Satisfaction

We will now turn to the solution of (1). Here the cost function reduces to the staffing costs, i.e $C^r(N^r, \pi^r) = N^r$. The following theorem states the asymptotic optimality of SRS and TP as a solution for (1).

Theorem 4.1. *Under the assumptions of Theorem 3.1, the following combined staffing and routing policy is asymptotically optimal for the problem (1):*
Staff according to SRS: $N^r = R^r + \beta\sqrt{R^r}$, where $R^r = \sum_{i=1}^J \lambda_i^r/\mu$, and $\beta = \beta(\alpha_J)$ ($\beta(\cdot)$ is the inverse of (3)). Route according to TP with the threshold determined by the following recursive relation:

$$\begin{aligned} K_i^r - K_{i-1}^r &= \left\lceil \frac{\ln \alpha_{i-1}^r - \ln \alpha_i^r}{\ln \rho_{\leq i-1}^r} \right\rceil, \quad i = 2, \ldots, J, \\ K_1^r &\equiv 0 \end{aligned} \tag{25}$$

The proof of Theorem 4.1 follows as a direct consequence of [8], which deals with staffing a single class system modelled as an $M/M/N$. The reduction of the problem (1) to a single class problem is to be expected due to the state-space collapse property of the TP policy (stated in Proposition 3.2), that guarantees that almost only low priority customers wait in queue. This is the essence of the proof of Theorem 4.1. More details are given in Remark 4.1.

Remark 4.1. (Intuitive Explanation of Theorem 4.1) This theorem is an immediate consequence of Theorem 3.1. To see this, let us consider the following reasoning: An intuitive lower

bound for the required staffing would be to solve a different constraint satisfaction problem, in which all classes are treated as a single class with the the same probability of delay constraint α_J. By [8] we know that in the single class case the optimal staffing level is given by the SRS with the corresponding β. Theorem 3.1 ensures, then, that using the same lower bound staffing level for the original multi-class system in combination with logarithmic thresholds will be asymptotically feasible. Hence, the lower bound is achieved and the policy is asymptotically optimal.

4.2 Cost Minimization

The following section deals with the cost minimization problem (2). This problem is more involved than the constraint satisfaction one (1) because the waiting time steady-state distribution is part of both the objective function and the constraints. Solving the problem (2) in its most generality is difficult. However, under the natural assumption that customers who incur a higher holding cost also have a tighter delay bound, the problem becomes more tractable. The latter is the form of the problem considered here.

In what follows we claim that, when the costs and delay bounds are in the right order, SRS and TP are asymptotically optimal for (2) with the appropriate choice of the multiplier β and the thresholds $K_1^r, ..., K_J^r$. The reason these rules work in this case is simple. In a nutshell, note that the total waiting costs in the objective function of (2) can be rewritten as $c_1^r EQ_1 + ... + c_J^r EQ_J$ (due to Little's law). Also note that, given a fixed staffing level, the total queue length across all customer classes does not vary much, as long as servers are not idle for too long. Consequently, in order to minimize costs, one should attempt to concentrate as much of the queue length as possible in the class with the lowest waiting cost. This is the lower priority class J. As noted in Proposition 3.2, this is exactly what the TP policy obtains. With cost parameters which are polynomial in r, the thresholds need to be more carefully selected such that, indeed, the waiting cost associated with the lowest priority class would be the only dominant cost factor in the limit as $r \to \infty$. Now, given that the thresholds are selected such that this latter goal is obtained, it is immediate that the staffing level should depend on the lowest priority waiting cost and delay bound only.

When considering the problem (2) one needs to balance between the waiting cost and the delay probability constraints. This requires refinement of the staffing and control proposals made in section 2. Particularly, it may be required to choose a value of β in (4) which is greater than $\beta(\alpha)$.

Similarly, the values of α_i^r, $i = 1, ..., J$ in (25) need to be changed into α_i^{r*} which are less than or equal to the original α values. The exact way in which these values are modified is described in Theorem 4.2 which states the asymptotic optimality of SRS and TP with respect to the cost minimization problem (2).

Theorem 4.2. *Consider the joint staffing and control problem (2), under the assumptions of Theorem 3.1. In addition, assume that the waiting cost coefficients scale with r in a polynomial manner:* $c_i^r = d_i \cdot r^{\gamma_i}, \quad i = 1, ..., J$, *where* $d_i > 0$ *and* $\gamma_i \geq 0$, $i = 1, ..., J-1$, *while* $\gamma_J = 0$. *Assume further that* $c_1^r \geq c_2^r \geq ... \geq c_J^r$. *Then, the following staffing and control proposition is asymptotically optimal:*

- *Staff according to SRS:* $N^r = R^r + \beta\sqrt{R^r}$, *where*

$$\beta = \max\{y^*(d_J), \beta(\alpha_J)\}, \tag{26}$$

  and $y^*(d_J) = arg\min_{y>0}\left\{y + \frac{d_J\alpha(y)}{y}\right\}$.[7]

- *Route according to TP with threshold levels that satisfy*

$$K_i^r - K_{i-1}^r = \left\lceil \frac{\ln \alpha_{i-1}^{r*} - \ln \alpha_i^{r*}}{\ln \rho_{\leq i-1}^r} \right\rceil \quad i = 2, \ldots, J,$$
$$K_1^r \equiv 0, \tag{27}$$

  where

$$\alpha_J^{r*} = \alpha(\beta),$$

  and, for $i = 1, ..., J-1$,

$$\alpha_i^{r*} = \min\{\alpha_i^r, \frac{1}{(N^r)^{\gamma_i}}, \alpha_J^{r*}\}, \tag{28}$$

  Finally, ties are resolved according to the cμ rule.

Remark 4.2. (Intuitive Explanation of Theorem 4.2) To further understand the result, let us examine the special case with two classes for which $\alpha_1 = \alpha_2 = 1$ (a pure cost minimization problem). Note that a lower bound for the cost of this system is the cost associated with a single class multi-server system with rate $\lambda = \lambda_1 + \lambda_2$, with the same staffing cost and a waiting cost of c_J

[7] Approximations of the function $y^*(\cdot)$ are given in [8].

per waiting unit time. For this single class system one can find an asymptotically optimal staffing level using [8]. Denote this staffing level by N^*.

For the original two-class system, given that all servers are busy, the queue of the high priority is of order $\Theta(1)$. Also, by using a logarithmic threshold, one can ensure that the probability of high-priority delay would be approximately $r^{-\gamma_1}$ (see (14)). Hence, since the waiting cost is at most polynomial in r, the waiting cost for the high priority is $\Theta(1)$.

After applying the thresholds the queue will comprise almost entirely from low priority customers. Moreover, by Proposition 3.1 and Corollary 3.1, it will be of approximately the same length as the queue in the lower bound system. These two results imply that the overall waiting cost would be the same as in the lower bound system. Since we have used the same staffing level, we have achieved the lower bound.

Remark 4.3. Notice that if the constraints on the probability of delay are removed, the remaining problem is a pure cost minimization problem. For this latter problem the optimal policy, as conjectured by [37], is one with state dependent thresholds. In Theorem 4.2, however, we show that asymptotically, the state *independent* threshold policy TP is optimal. The intuitive explanation for why the state-dependence disappears in the limit is the economies of scale associated with the many-server system. For a fixed N, The state dependency of the thresholds is aimed at protecting against a situation where a lower priority queue gets too long and expensive because of the reservation for higher priorities. However, in large systems, one can combine high quality of service for the high priorities with very little harm to the low priorities.

Remark 4.4. (Asymptotic Regimes) Note that if $\alpha_J < 1$, then the proposed staffing level follows the square root safety (SRS) staffing rule, which is consistent with the QED regime. Similarly, if $\alpha_J = 1$ but $\gamma_J = 0$ (which implies that the holding cost associated with class J does not scale with r) then the SRS rule still applies. In contrast, if $\alpha_J = 1$ and $-1 < \gamma_J < 0$, then TP is still asymptotically optimal (with threshold levels that are determined according to (27) and (28)) but with staffing rule that satisfies $N^r = R^r + \sqrt{d_J}\sqrt{(R^r)^{1+\gamma_J}}$, which is consistent with the Efficiency Driven (ED), and not with the QED, regime (see Remark 3.5).

Corollary 4.1. *($c\mu$ **Optimality:**) Consider the problem (2) without any constraints (or, equivalently, assume that $\alpha_i = 1, \forall i = 1, ..., J$). Also, assume that $c_i^r = c_i, \forall i = 1, ..., J, \forall r \geq 1$ (that*

is, assume that c_i^r does not change with r). Then, the $c\mu$ rule is asymptotically optimal among all non-preemptive policies (work-conserving and non-work conserving).

4.3 Discussion: Practical Considerations and the Cost of Cross-Training

Theorems 4.1 and 4.2 describe asymptotically optimal staffing and control rules for a sequence of systems with respect to the problems (1) and (2). In reality, one considers, not a sequence of systems, but a single system with arrival rate forecasts, a list of delay bounds and staffing and waiting costs estimates. How should one determine the staffing level and routing rule for this system? Our theorems suggest, that if the total arrival rate is large, class J is comparable in size, and its waiting cost is of magnitude not greater than the staffing cost, then SRS and TP are nearly-optimal.

When implementing SRS and TP with respect to a particular system the choice of both the square-root safety staffing multiplier β and the thresholds associated with TP requires careful consideration; the flexibility in choosing these parameters emanates from the degrees of freedom associated with the choice of the parameters $\gamma_1, ..., \gamma_J$, and $d_1, ..., d_J$. Particularly, even if one is able to accurately estimate the cost parameters: $c_1, ..., c_J$, and the total system load R, there are still infinitely many valid values of d_i and γ_i, $i = 1, ..., J$, that satisfy $c_i = d_i R^{\gamma_i}$ (recall the assumptions that $c_i = d_i r^{\gamma_i}$ and that $R \equiv r$ (8)). The values of β and $K_2, ..., K_J$, in turn, are quite sensitive with respect to the choice of γ_i and d_i, so these, indeed, should be carefully selected.

To understand the main tradeoffs involved with the choice of this model's parameters consider the following two class example: Suppose that a two class system has a total load of $R = 100$, service rate $\mu = 1$, and class level arrival rates: $\lambda_1 = \lambda_2 = 50$. Assume that no constraints on delay probability are given, that is, $\alpha_1 = \alpha_2 = 1$. Consider two extreme scenarios both with the same staffing cost of 1, the same class 2 waiting cost $c_2 = 1$, but one in which class 1 waiting cost satisfies $c_1 = 2$, while in the other $c_1 = 100$. Since, the staffing and class 2 waiting cost are fixed, both systems will have the same staffing level (see Theorem 4.2). However, in terms of control, we expect that the second scenario will require more service effort dedicated towards class 1, which translates into a higher threshold. Exactly how much higher depends on the actual values of the cost parameters.

To be more specific, consider the second scenario (with $c_1 = 100$), and suppose that $c_1 =$

$d_1 R^{\gamma_1} = 100$. If $\gamma_1 = 0$, then Theorem 4.2 recommends that no threshold should be used. On the other hand, if $\gamma_1 = 1$, then the Theorem recommends a threshold which is equal to 5. Higher levels of γ_1 will translate into even higher thresholds. So which one is the right value of γ_1? To consider this question, recall that the choice of threshold in Theorem 4.2 is such that the waiting cost associated with the higher priority classes is negligible with respect to the low priority delay cost. Let us examine class i waiting cost in this example: $c_i \lambda_i E W_i = c_i \lambda_i E[W_i | W_i > 0] P(W_i > 0)$. For class 2, this translates into an expression of the order of $\sqrt{R} = 10$. For class 1, $\lambda_1 E[W_1 | W_1 > 0] = O(1)$ (due to Proposition 3.1), and $c_1 P(W_1 > 0) \approx d_1 R^{\gamma_1} R^{-\gamma_1} = d_1$. Using the criterion that class 1 waiting cost should be negligible with respect to the cost associated with the delay cost of class 2 we conclude that the if $c_1 = 100$, then $(\gamma_1, d_1) = (1, 1)$ is a better choice than $(\gamma_1, d_1) = (0, 100)$.

So how does one indeed choose the values of $\gamma_1, ..., \gamma_J$? As illustrated in the above example, it is important to first determine the value of γ_J. Suppose that all costs are normalized such that the staffing cost is 1. As long as the waiting cost is comparable with the staffing cost, it makes sense to choose $\gamma_J = 0$. In this case, class J waiting cost is of order $c_J \cdot \sqrt{R}$. Once γ_J is selected, a reasonable choice for γ_i, $i = 1, ..., J-1$ is the lowest non-negative value that makes the cost associated with class i negligible with respect to the cost of class J. This is equivalent to choosing the smallest value of γ_i such that $d_i = c_i / R^{\gamma_i}$ is negligible with respect to $c_J \cdot \sqrt{R}$. To get a sense of the order of magnitudes of the quantities in question, note that, for reasonably large call centers, $\sqrt{R}$ is in the range of 10 - 30. Requiring that $d_i = c_i / R^{\gamma_i}$ be negligible with respect to $c_J \cdot \sqrt{R}$ then means that d_i / c_J should be of order 1 or less. Finally, note that to obtain the optimal threshold and staffing values for a particular system it is best to search over a range of values for these parameters. Our guidelines above provide a good starting point for such search procedures.

In addition to determining staffing levels and routing rules, the results in this paper may be readily used to examine questions related to the cross training of servers; More specifically, one may consider two system configurations for a multi-class service system. The first configuration is the multi-I design, in which each class is served by a pool of dedicated servers. The second is the V-design, in which all servers are cross-trained to provide service to all customers, regardless of their class. Clearly, to provide the same level of service, fewer servers are needed in the V-design versus the multi-I design. But, by how much is the system size reduced? Also, what is the highest cost of cross-training that would justify switching from a multi-I design to a V-design?

To illustrate how to address questions related to cross training, consider a two class example. Suppose that both classes have the same service rate μ, and arrival rates: λ_1, λ_2, with loads R_1 and R_2, respectively. Suppose that one wishes to satisfy the delay bounds $P(W_i(\infty) > 0) \leq \alpha_i$, $i = 1, 2$, where $0 < \alpha_1 < \alpha_2 < 1$. According to the I-design, one would generate two dedicated server pools of sizes N_1 and N_2, where $N_i = R_i + \beta_i\sqrt{R_i}$, and $\beta_i = \beta(\alpha_i)$, $i = 1, 2$. Note that $\beta_1 > \beta_2$. On the other hand, if the V-design is used, one would instead choose to have a single server pool of size $N = R + \beta_2\sqrt{R}$, where $R = R_1 + R_2$. It is easily seen that the total number of servers under the latter configuration is less than that number under the I-design. More precisely, the difference satisfies

$$N_1 + N_2 - N = \beta_1\sqrt{R_1} + \beta_2\sqrt{R_2} - \beta_2\sqrt{R} > \beta_2(\sqrt{R_1} + \sqrt{R_2} - \sqrt{R}). \tag{29}$$

In particular, the size advantage of the V-design becomes more significant as system load increases.

An additional, very significant advantage of the V-design versus the I-design is in terms of the actual waiting times for those customers who end up waiting. Specifically, while, for the I-design, the positive waiting times, for both classes, are of order $1/\sqrt{R}$ (i.e. both classes experience QED service), those waiting times associated with class 1 customers are of order $1/R$ (i.e. class 1 experiences QD service) under the V-design. For large systems, this difference in quality of service is very significant and should not be ignored.

In addition to size and order of magnitude comparisons, our results can also be implemented to examine the question of what are the values of the cost of cross training that would justify using the V-design instead of the I-design. For $i = 1, 2$, let s_i, $i = 1, 2$ be the cost per server of pool i in the I-design. Also, let s be the cost of a server in the V-design, where $s > \max\{s_1, s_2\}$. Note that s_1, s_2 and s include training of the servers, their salaries, and other variable costs associated with staffing the system. If one disregards waiting costs, it is easy to see, by comparing the staffing costs associated with both designs, that if

$$s \leq \frac{s_1(R_1 + \beta_1\sqrt{R_1}) + s_2(R_2 + \beta_2\sqrt{R_2})}{R + \beta_2\sqrt{R}}, \tag{30}$$

then the V-design is more cost effective. If, in addition, one takes waiting costs into consideration, then the benefits of cross training are even greater, and, hence, a higher cost of cross training can be tolerated.

5 Adding Abandonment

One common characteristic of some service systems, including call centers, is customers tendency to abandon if their service does not start soon enough. It has been shown in [13] that including the abandonment feature in models of systems whose customers may indeed renege is crucial. In this section, we show that for the V-design our proposed TP control remains asymptotically optimal with respect to the profit maximization / cost minimization problem even in the presence of customers abandonment. This result reinforces the *robustness* of this threshold priority rule with respect to optimization criteria and model assumptions.

The model: Consider a model identical in features to the one introduced in Section 2, with the additional assumption that customers of class i will abandon (independently of all other processes) after an exponential time with rate θ_i, if her service does not start by then. Denote by $P_\pi\{Ab_i\}$ the steady state probability of abandonment for class i, under a control policy $\pi \in \Pi$, where Π is the set of all non-preemptive, non-anticipative control policies. Finally, assume that an abandonment of a class i customer incurs a cost of c_i.

The problem of minimizing weighted abandonment costs plus staffing costs is given as follows:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{J} c_i \lambda_i P_\pi\{Ab_i\} + N \\ & N \in \mathbb{Z}_+, \quad \pi \in \Pi. \end{array} \tag{31}$$

As an alternative to the cost minimization problem (31) one could also consider a profit maximization problem in which a customer of class i contributes a revenue of r_i. In this case, one is interested in maximizing $\sum_{i=1}^{J} r_i \lambda_i (1 - P_\pi\{Ab_i\}) - N$, which is clearly equivalent to (31) with $r_i = c_i$.

The pure control problem of minimizing the total number of abandonments (with no cost differentiation) was addressed in [29]. In [29] the authors proved that the non-preemptive policy that stochastically minimizes the number of customers lost during a finite interval of time belongs to the class of *stochastic earliest deadline* policies. Specifically, in the exponential patience setting their result implies that the optimal policy is such that it admits customers into service in order of their average patience. i.e. it always serves first the waiting customers with the shortest patience (or highest patience parameter θ). In addiion, when restricted to non-idling policies the optimal policy is a static priority policy where customers are served in decreasing order of θ_i. This gives the structure of the optimal policy but does not give an explicit optimal policy. Moreover, [29] does

not address the possibility of deliberately idling some servers.

5.1 Asymptotic Optimality - Cost Minimization

In this section, we consider finding an asymptotically optimal *control* for the problem (31) given the SRS staffing rule $N = R + \beta\sqrt{R}$, for $-\infty < \beta < \infty$. That is, we take the SRS staffing as given, and consider the control problem alone. Following this section results, the joint staffing and control problem should be immediately resolved once the asymptotic optimality of SRS staffing is established for the single class Erlang-A (M/M/N+M) model. This latter setting is currently being investigated by the authors of [38].

Our attention in this section is restricted to the case where customers abandonment cost is positively correlated with their impatience. In particular, we assume that $c_i \geq c_j$ if and only if $\theta_i \geq \theta_j$. Specifically, without further loss of generality, we assume that customer classes are in decreasing order of their abandonment cost ($c_1 \geq c_2 \geq ... \geq c_J$) and their abandonment rate ($\theta_1 \geq \theta_2 \geq ... \geq \theta_J$). Additionally, our attention is restricted to systems in which the abandonment cost of the low priority is comparable with the staffing costs.

Definition: Consider a sequence of systems indexed by r, $r = 1, 2, ...$, defined in an analogous manner to the asymptotic framework described in section 2.1, with the additional assumption that class i abandonment rate is θ_i. In particular, the abandonment parameters do not scale with r. For $r = 1, 2, ...$, suppose that $R^r \equiv r$, and let $N^r = R^r + \beta\sqrt{R^r}$, $\infty < \beta < \infty$, and let $C^r(\pi^r)$ be the total abandonment plus staffing cost associated with the sequence of scheduling policies $\{\pi^r\}$; i.e.

$$C^r(\pi^r) = \sum_{i=1}^{J} c_i^r \lambda_i^r P_{\pi^r}\{Ab_i^r\} + N^r.$$

The sequence $\{\pi^r\}$ is said to be **asymptotically optimal** with respect to $\bar{\lambda}^r$, $\bar{\theta}$, N^r and $C^r(\cdot)$ if for any other sequence of policies $\{\hat{\pi}^r\}$ we have that

$$\liminf_{r\to\infty} \frac{C^r(\hat{\pi}^r) - N^r}{C^r(\pi^r) - N^r} \geq 1$$

With respect to the cost minimization problem (31) we show that the TP control is asymptotically optimal, where the thresholds associated with this rule are dependent on the cost and patience parameters. This is stated in the following Theorem.

Theorem 5.1. *Consider a sequence of multi-class multi-server systems, with class-level abandonment rates: $\theta_1, \theta_2, ..., \theta_J$. Consider the problem (31), and assume that all of the following three conditions holds:*

(a) $c_1^r \geq c_2^r \geq ... \geq c_J^r$, *and* $\theta_1 \geq \theta_2 \geq ... \geq \theta_J$.

(b) The cost parameters c_i, $i = 1, ..., J-1$ *are allowed to grow polynomially with* r, *but* c_J *is fixed (i.e* $c_i^r = c_i \cdot r^{\gamma_i}$, $\gamma_i \geq 0$, $i = 1, ..., J-1$, *and* $\gamma_J = 0$*).*

(c) Class J has non-negligible demand. In particular, $\lim_{r\to\infty} \frac{\lambda_J^r}{\lambda^r} \to a_J > 0$.

In addition, assume that the system is staffed according to the SRS rule with $N^r = R^r + \beta\sqrt{R^r}$, *for some* $-\infty < \beta < \infty$. *Then the TP control is asymptotically optimal, with priorities given in decreasing order of* c_i, *and the threshold satisfy:*

$$K_i^r - K_{i-1}^r = \left\lceil \frac{\ln \alpha_{i-1}^r - \ln \alpha_i^r}{\ln \rho_{\leq i-1}^r} \right\rceil, \quad i = 2, ..., J, \tag{32}$$

where $K_1^r = 0$, $\alpha_i^r = (N^r)^{-\gamma_i}$, *and* $\alpha_J^r = w(-\beta, \sqrt{\mu/\theta_J})$, *with* $w(x, y) = \left[1 + \frac{h(-xy)}{yh(x)}\right]^{-1}$, *and* $h(x) = \frac{\phi(x)}{1-\Phi(x)}$.

The delay probability and the probability of abandonment, provided that SRS and TP are used can also be computed. They are given in the Technical appendix.

6 Conclusions and further research

We study large scale service systems with multiple customer classes and fully flexible servers. For such systems we investigate the question of how many servers are needed and how to match them with customers so as to minimize operating and delay costs (or maximize profit) subject to constraints on class level delay probabilities. We find that a square-root safety (SRS) staffing rule and a threshold-priority (TP) control are asymptotically optimal for a variety of problem formulations and model assumptions. In particular, these rules are very robust with respect to changes in either costs or constraints parameters. Moreover, we suggest that the staffing level determined by the SRS rule depends on the overall system demand, and the performance bounds and costs associated with the lower priority class of customers only. This implies that even if the

costs and performance bounds associated with the higher priority customers are non-constant or are unknown, the staffing level does not change.

Several directions for future research may be considered:

1. With respect to service systems with abandonment, the restrictive assumption that customer impatience is positively correlated with the waiting cost may be weakened.

2. Also with abandonment, one may consider a constraint satisfaction problem (analogous to (1)) with bounds on the abandonment probability. For the case that these bounds do not scale with system size we find that the ED regime is optimal, and that a server pool decomposition is asymptotically optimal, in which each customer class has a pool of servers dedicated to it. This result is reported in the Technical appendix. As future research, it is interesting to consider the case in which the bounds on the abandonment probability do scale with r. The latter assumption appears to complicate analysis considerably.

3. Other cost criteria or constraints may be considered such as those associated with tail probabilities of the waiting times.

4. Finally, The V-design model studied in this paper is important in its own right. However, the insights gained from studying this model are also useful in studying more complicated network structures. Specifically, [1, 4] study the "Inverted-V" model, in which a single class of customers may be served by multiple pools of servers, who are differentiated by their processing speed. We intend to use the V and and Inverted-V models as building blocks in studying network topologies such as the N-design; in the N-design some server pools are fully flexible, whereas the others are dedicated to a single customer class.

References

[1] Armony, M. "Dynamic routing in large-scale service systems with heterogeneous servers". Preprint. 2004.

[2] Armony M., Maglaras C., "On customer contact centers with a call-back option: customer decisions, routing rules and system design", *Operations Research*, **52**(2), pp. 271-292. 2004.

[3] Armony M., Maglaras C., "Contact centers with a call-back option and real-time delay information", *Operations Research*, **52**(4), pp. 527-545. 2004.

[4] Armony, M. and Mandelbaum, A. "Staffing of large service systems: The case of a single customer class and multiple server types". Preprint. 2004.

[5] Atar R., Mandelbaum A., Reiman M., "Scheduling a Multi-Class Queue with Many Exponential Servers: Asymptotic Optimality in Heavy Traffic", *Ann. Appl. Prob.*, **14**(3), pp. 1084-1134. 2004.

[6] A. Bassamboo, J.M. Harrison and A. Zeevi. "Design and control of a large call center: Asymptotic analysis of an LP-based method". Preprint. 2004.

[7] Blumenthal R.M., *"Excursions of Markov Processes"*, Birkhäuser, 1992.

[8] Borst S., Mandelbaum A. and Reiman M., "Dimensioning Large Call Centers", *Operations Research*, 52(1), pp. 17-34, 2004.

[9] Cachon, G.P. and Lariviere, M.A., "Contracting to assure supply: how to share demand forecasts in a supply chain", *Management Science*, **47**(5), pp. 629-646, 2001.

[10] Ethier, S.N.and Kurtz, T.G., *"Markov Processes,Characterization and Convergence"*,John Wiley & Sons, 1985.

[11] Federgruen A., Groenvelt H. "M/G/c Queueing Systems With Multiple Customer Classes: Characterization and Control of Achievable Performance Under Nonpreemptive Priority Rules", *Management Science*, **34**, pp. 1121-1138. 1988.

[12] Garnett, O. "Designing a telephone call center with impatient customers". Masters Thesis, Technion - Israel Institute of Technology, 1998.

[13] Garnett O., Mandelbaum A. and Reiman M., "Designing a Call Center with Impatient Customers", *Manufacturing and Service Operations Management* , **4**(3), pp. 208-227. 2002.

[14] Gurvich I., Armony M. and Mandelbaum A., "Staffing and control of large-scale service systems with multiple customer classes and fully flexible servers: Technical appendix", preprint, 2004.

[15] Halfin S., Whitt W., "Heavy-Traffic Limits for Queues with Many Exponential Servers", *Operations Research*, **29**, pp. 567-587. 1981.

[16] J.M. Harrison and A. Zeevi, "Dynamic scheduling of a multiclass queue in the Halfin and Whitt heavy traffic regime", *Operations Research*, to appear. 2004.

[17] Harrison J.M., Zeevi A., "A Method for Staffing Large Call Centers Based on Stochastic Fluid Models", *Manufacturing and Service Operations Management*, to appear. 2005.

[18] Kleinrock L., *"Queueing Systems"*, Volume II, CH. 1, pp. 1-22, John Wiley & Sons, 1976.

[19] Kella O., Yechiali U., "Waiting Times in the Non-Preemptive Priority M/M/c Queue", *Communications in Statistics - Stochastic Models*, **1**(2), pp. 357-262, 1985.

[20] Mandelbaum A., Pats G., "State-Dependent Queues: Approximations and Applications", In F. Kelly and R. Williams, editors, *Stochastic Networks*, **71**, pp 239–282. Proceedings of the IMA, 1995.

[21] Mandelbaum A., Stolyar A., "Scheduling Flexible Servers with Convex Delay Costs: Heavy-Traffic Optimality of the Generalized $c\mu$-Rule", *Operations Research*, to appear. 2004.

[22] Massey A.W., Wallace B.R., "An Optimal Design of the M/M/C/K Queue for Call Centers", *Queueing Systems*, to appear. 2004.

[23] Meyn S.P. and Tweedie R.L., *"Markov Chains and Stochastic Stabiliy"*, Springer, 1993.

[24] Puhalskii A. "On the Invariance Principle For the First Passage Time", *Mathematics of Operations Research*, **19**, Nov. 1994.

[25] Puhalskii A., Reiman M., "The Multiclass GI/PH/N Queue in the Halfin-Whitt Regime", *Advances in Applied Probablity*, **32**, pp. 564-595, 2000.

[26] Quinzii, M. and Rochet, J-C., "Multidimensional signalling", *Journal of Mathematical Economics*, **14**, pp. 261-284, 1985.

[27] Resnick S., *"Adventures in Stochastic Process"*, Birkhäuser, 1992.

[28] Schaack C., Larson R., "An N-Server Cutoff Priority Queue", *Opererations Research*, **34**(2), pp. 257-266, 1986.

[29] Towsley D., Panwar S.S., "Optimality of the Stochastic Earliest Deadline Policy for the G/M/c Queue Serving Customers with Deadlines", COINS Technical Report 91-61, Univ. Massachusetts, Aug. 1991

[30] Van Mieghem J.A., "Dynamic scheduling with convex delay costs: the generalized $c\mu$ rule", *Annals of Applied Probability*, **5**, pp. 809-833, 1995.

[31] Wallace R.B., Whitt W., "A Staffing Algorithm for Call Centers with Skill-Based Routing". working paper, 2004.

[32] Walrand J., *"An Introduction to Queueing Networks"*, CH. 8, pp. 254-260, New Jersey, Prentice-Hall, 1988.

[33] Whitt W., *"Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues"*, Springer, 2002.

[34] Whitt W., "Heavy-Traffic Limits For the $G/H_2^*/n/m$ Queue", *Mathematics of Operations Research*, to appear. 2004.

[35] Whitt W., "Efficiency Driven Heavy-Traffic Approximations for Many-Server Queues with Abandonments". *Management Science*, to appear. 2004.

[36] Whitt W., "How Multiserver Queues Scale with Growing Congestion-Dependent Demand". *Operations Research*, **51**(4), pp. 531-542, 2003.

[37] Yahalom T., Mandelbaum A., "Optimal Scheduling of a Multi-Server Multi-Class Non-Preemptive Queueing System", Preprint, 2004.

[38] Mandelbaum A. and Zeltyn S., "Dimensioning call centers with abandonment", preprint, 2004.

[39] 4 Call Centers Software. Downloadable from ie.technion.ac.il/serveng