Design and Control of the M/M/N Queue with Multi-Type Customers and Many Servers

Itay Gurvich

Design and Control of the M/M/N Queue with Multi-Type Customers and Many Servers

Research Thesis

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Operations Research and Systems Analysis

Itay Gurvich

Submitted to the Senate of the Technion –
Israel Institute of Technology

Tamuz, 5764 – Haifa – July, 2004

This research thesis was conducted under the supervision of Professor Avishai Mandelbaum in the Industrial Engineering Faculty in the Technion. I would like to express my deep gratitude for his invaluable devotion; for many long hours of discussions during which he led me through the research problems while maintaining the right balance between his role as an advisor and his commitment to the promotion of my personal and academic development; and for sharing with me his knowledge and ideas. Our joint work was of great significance for me, and for that I am grateful.

I would like to thank, as well, Professor Mor Armoni from the Stern School of Business in New York University, who during her one-year stay at the Technion was practically an additional advisor for me. Her considerable help was essential for the success of this research.

I thank Professor Haya Kaspi from the Industrial Engineering Faculty in the Technion, who despite her tight schedule found the time to answer my questions and give me useful advices.

Finally, the generous financial help of the Technion's Graduate School is gratefully acknowledged.

Contents

List of Tables

List of Figures

Abstract

We analyze The V-Model of Skills Based Routing. This is a model in which servers are homogeneous, and there are J customer classes having the same service requirements. With respect to the V-Model we ask the following questions:

1. Given a fixed number of servers, how to schedule servers to the different customer classes so as to optimize system performance, and
2. How many servers are required in order to minimize staffing and waiting costs while maintaining pre-specified performance goals.

We address these questions by first characterizing a scheduling scheme and staffing scheme that are asymptotically optimal as the arrival rate increases to infinity. The asymptotic optimality is in the sense that the policy (asymptotically and stochastically) minimizes the steady-state waiting and staffing costs while satisfying a pre-specified waiting probability in steady-state, asymptotically as the arrival rate grows large.

The main asymptotic framework considered in this paper is the many-server heavy-traffic regime formally introduced by Halfin and Whitt. We refer to this regime as the QED (Quality and Efficiency Driven) regime. In the concluding sections, we extend the V-Model by adding abandonment and considering optimization of staffing and control under certain cost structures. To conclude, we briefly introduce some ongoing research about the N Model of Skills Based Routing.

List of Symbols

$\Rightarrow$	Weak Convergence
E	Expectation
P	Probability Measure
$\mathbb{Z}_+$	Positive Integers
$\sim$	$a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to \infty$
o	$a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$
Θ	$a_n = \Theta(b_n)$ if $a_n/b_n \to c$, as $n \to \infty$, $-\infty < c < \infty$
O	$a_n = O(b_n)$ if $\exists c : \limsup a_n/b_n \leq c$, as $n \to \infty$
$\langle M \rangle$	Quadratic Variation Process of M
$\langle M, N \rangle$	Quadratic Covariation Process of M and N
X^+	The Maximum Between X and Zero
X^-	The Minus of the Minimum Between X and Zero
$Q_i(t)$	Queue Length of Class i Customers at Time t
$W_i(t)$	Virtual Waiting Time of Class i Customers at Time t
$Q_i(\infty)$	Queue Length of Class i Customers in Steady State
$W_i(\infty)$	Waiting Time of Class i Customers in Steady State
μ	The Service Rate Common For all Customer Classes
λ_i^r	Arrival Rate of Class i in the r^{th} System

N^r	Number of Servers in the r^{th} System
K_i^r	Threshold of Class i in the r^{th} System
K^r	The Threshold of the Lowest Priority Class (J), K_J^r
ρ_C^r	$\lambda^r/((N^r - K^r)\mu)$
$\rho_{\leq i}^r$	$\sum_{j=1}^{i} \lambda_i^r/(N^r\mu)$
θ_i	Patience Parameter of Class i Customers

List of Acronyms

QED	Quality and Efficiency Driven
ED	Efficiency Driven
$M/M/\{K_i\}$	V-Model with Thresholds K_i, $i = 1, ..., J$

1 Introduction[1]

In modern service systems it is common to have multiple classes of customers and multiple server types (skills). The customer classes are differentiated according to their service needs. The server types are characterized by the subset of customer classes that they can adequately serve and the quality of service that they can devote to each such class. An important example of such large scale service systems are multi-skill call/contact-centers. Such centers are often characterized by multiple classes of calls (classified according to type or level of service requested, langauge spoken, perceived value of customers, etc.). To match the various service needs of those customers, call centers often consist of hundreds or even thousands of customer service representatives (CSRs). These CSRs have different skills, depending on the call classes that they can handle, and the speed in which they do it.

There are three main issues to address when dealing with the operations management of large-scale service systems. Given a forecast of the customers' arrival rates and their service requirements, these issues are:

- **Design:** The long-term problem of determining the class partitioning of customers, and the types of servers; this typically includes overlapping skills (i.e. servers that can handle more than one class of customers, and classes that can be served by several server types).
- **Staffing:** The short-term problem of determining how many servers are needed of each type, in order to deal with the given demand. These server types may be of overlapping skills. (In addition, there is a scheduling problem which determines the shift structure for the system, as well as the determining of who are the actual servers that would work in these shifts. The last two issues will not be discussed in this work.)
- **Control:** The on-line problem of customer routing and server scheduling that involves the assignment of customers to the appropriate server upon service completion or a customer's arrival.

These three problems are all interrelated and should, therefore, be discussed in conjunction with one another. Yet, because of the complexity involved in addressing all these three combined, they are typically addressed hierarchically and unilaterally in the literature.

Even when one addresses the three issues separately, a general solution for all possible system configurations is currently out of reach. Instead, we approach the problem by studying a

[1]The introduction is adapted from the paper by Armony and Mandelbaum [2], with the authors' approval.

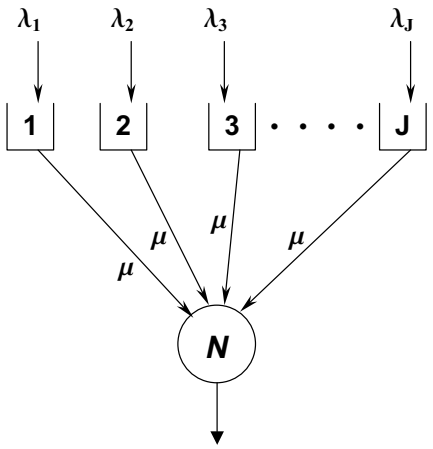


Figure 1: The V Model - multiple customers classes and a single server type.

relatively simple model in order to gain insight to the more general model. The model we focus on in this work is the V-design. This is a system design in which servers are homogeneous, and there are J customer classes having the same service time requirements. The V-design is depicted in Figure 1.

With respect to the V-design we ask the following two questions:

1. Given a fixed number of servers, how to schedule servers to the different customer classes so as to optimize system performance, and
2. How many servers are required in order to minimize staffing and waiting costs while maintaining pre-specified performance goals.

We address these questions by first characterizing a scheduling scheme and staffing scheme that are asymptotically optimal as the arrival rate increases to infinity. The asymptotic optimality is in the sense that the policy (asymptotically and stochastically) minimizes the steady-state waiting and staffing costs while satisfying a pre-specified waiting probability in steady-state, asymptotically as the arrival rate grows large.

The main asymptotic framework considered in this paper is the many-server heavy-traffic regime acknowledged by Erlang [18], Jagerman [35] and ultimately introduced and formalized by Halfin and Whitt [30]. We refer to this regime as the QED (Quality and Efficiency Driven) regime. Systems that operate in the QED regime enjoy a rare combination of high efficiency together with high quality of service. More formally, consider a sequence of systems of a fixed design and an increasing arrival rate λ. Suppose that the total service capacity

of each system in the sequence exceeds λ by a safety capacity of order $\sqrt{\lambda}$. In particular, the traffic intensity (or server efficiency) approaches 1 as $\lambda \to \infty$ (i.e. the system goes to heavy traffic). On the other hand, the high quality aspect of the QED regime may be seen through it's following alternative characterization: Suppose that as $\lambda \to \infty$, the limiting waiting probability is non-trivial (ie. it is strictly between 0 and 1). This high performance, which is typically impossible to achieve for systems in heavy traffic, is obtained here due to the economies of scale associated with the large number of servers. The two characterizations of the QED regime are shown to be equivalent in various settings, as first established in [30]. (See the literature review, Section 1.1.1 for more details). In the present work we establish this equivalence for the V-Model.

1.1 Literature Review

1.1.1 The QED regime: asymptotic theory of many-server queues.

The QED regime has been given much attention in the last few years, especially the "I^k"-model, which corresponds to multiple independent queues, each with its own devoted server pool (no overlap in skills). For a formal description, 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. The QED regime is achieved by choosing λ^r and N^r so that $\sqrt{N^r}(1-\rho^r) \to \beta$, as $r \uparrow \infty$, for some finite β. Here $\rho^r = R^r/N^r$. When customers have infinite patience ρ^r may be interpreted as the long-run servers' utilization and $0 < \beta < \infty$. Otherwise, ρ^r is the offered load per server and $-\infty < \beta < \infty$. Equivalently, the staffing level is approximately given by

$$N^r \approx R^r + \beta\sqrt{R^r}, \quad -\infty < \beta < \infty\,. \tag{1}$$

Yet another equivalent characterization is a non-trivial limit (within $(0,1)$) of the fraction of *delayed* customers. The latter equivalence was established for GI/M/N [30], GI/D/N [36] and M/M/N with exponential patience [27].

Due to the desirable features of the QED regime, it has enjoyed recently considerable attention in the literature. Yet the regime was explicitly recognized already in Erlang's 1923 paper (that appeared in [18]) which addresses both Erlang-B (M/M/N/N) and Erlang-C (M/M/N) models. Later on, extensive related work took place in various telecom companies but little has been openly documented, as in Sze [59] (who was actually motivated by AT&T call centers operating in the QED regime). A precise characterization of the asymptotic expansion of the blocking probability, for Erlang-B in the QED regime, was given in Jagerman [35]; see also Whitt [66], and then Massey and Wallace [46] for the analysis of finite buffers. But the operational significance of the QED regime, in particular its balancing of "service and econ-

omy" via a non-trivial delay probability, was first discovered and formalized by Halfin and Whitt [30]: within the GI/M/N framework, they analyzed the scaled number of customers, both in steady state and as a stochastic process. Recent generalizations for non-exponential service times were made by Whitt [67, 68]. Convergence of the scaled queueing process, in the more general GI/PH/N setting, was established By Puhalskii and Reiman in [50]. Application of QED queues to modelling and staffing of telephone call centers and communication networks, taking into account customers' impatience, can be found in Garnett et al. [27] and Fleming et al. [21], respectively. The optimality of the QED regime, under revenue maximization or constraint satisfaction, is discussed in Borst et al. [11] and in [41, 4, 5, 2]. Readers are referred to Sections 4 and 5.1.4 of Gans et al. [22] for a survey of the QED regime, both practically and academically.

It is important to note that the QED regime differs in significant ways from the conventional (or "classical") heavy traffic regime. Indeed, QED combines light and heavy traffic characteristics. For example, in conventional heavy traffic, the theory of which has been well established (see for example Chen & Yao [16]), essentially all customers are delayed prior to service. In the QED regime, on the other hand, a non-trivial fraction is served immediately upon arrival. Also, conventional heavy traffic can be achieved by setting $N^r \approx R^r + \beta$, for some constant β, rather than the square-root form in (1).

To conclude this part we would like to use the characterization through the probability of delay to partition the spectrum of operational regimes. We identify three operational regimes:

- **Efficiency Driven (ED)**: Almost all the customers wait - $P\{Wait > 0\} \approx 1$.
- **Quality and Efficiency Driven (QED)**: a non trivial fraction of the customers wait -$P\{Wait > 0\} \approx \alpha$, $0 < \alpha < 1$.
- **Quality Driven:** Only a negligible portion of the customers wait - $P\{Wait > 0\} \approx 0$.

In the rest of the paper, we will focus mainly on the *QED* and *ED* regimes. For more details on the different operational regimes readers are referred to [22].

1.1.2 Skill-based Routing

Of the three issues related to the management of large-scale service system, the control problem has received the most attention in the literature. Specifically, for a given design, and staffing levels, researchers have proposed routing and / or scheduling schemes that are either

optimal or near-optimal. Alternatively, researchers have considered commonly used routing schemes (such as fixed priority rules, or dedicated servers per customer class) and computed the relevant performance measures. Examples for both criteria include: **Exact analysis** (Kella and Yechiali [39], Federgruen and Groenvelt [20], Schaack and Larson [55], Brandt and Brandt [14], Gans and Zhou [25], Armony and Bambos [3], Rykov [53], Luh and Viniotis [40], and de Véricourt and Zhou [17], **Asymptotic analysis - "conventional" heavy traffic** (Harrison [31], Bell and Williams [9], Glazebrook and Niño-Mora [28], Teh and Ward [71] and Mandelbaum and Stolyar [44]) and **Asymptotic analysis - QED regime** (Armony and Maglaras [4, 5], Harrison and Zeevi [32], and Atar et al. [6, 7]).

1.1.3 Staffing Rules

The staffing problem in the single-class, single-type case has also gained a lot of attention in the literature. However, things are quite different in the case of multiple types of servers, as is the case dealt with in [2]. The problem of determining how many servers of each type are required is very difficult. This is especially true if skills overlap. In the latter case, one wishes to take advantage of the flexibility of the servers who have multiple skills, but these servers are typically more costly. The most common approaches taken by researchers to tackle the staffing problem are:

- a) **Heuristical bounds:** Using heuristics to achieve performance bounds by analyzing simpler (but related) systems (Examples include Borst and Seri [12], Whitt [65], and Jennings et al. [37]),
- b) **Stability Staffing:** Staffing levels that guarantee system stability (Examples include Bambos and Walrand [8], Gans and van Ryzin [23], Armony and Bambos [3]), and
- c) **Cost minimizing staffing:** For a given routing scheme, find the staffing level that minimizes personnel costs while guaranteeing certain performance bounds, or alternatively, such staffing levels that minimize personnel costs plus operating costs. (Examples include Borst et al. [11], Perry and Nilsson [48], Stanford and Grassmann [54] and Shumsky [56]).

The common thread among these approaches is that they all focus on the QED regime, which corresponds to the so-called square-root staffing rule (due to the form of the staffing rule $N \approx R + \beta\sqrt{R}$ of equation (1)). Although these approaches arrive at the QED regime from different points of view, they seem to produce similar results (at least for the single-class, single-skill case). We next expand on some of the different approaches.

Heuristic bounds: The approach of achieving staffing through heuristic bounds is based on reducing the original system to a simpler system for which performance measures can be easily obtained. The simpler system can then offer staffing levels for the original system, with guaranteed upper or lower bound on performance. Two directions have been taken by different researchers.

Borst and Seri [12] determine bounds for the number of agents required to offer a given level of service by considering two systems: one in which servers are dedicated to a single customer class ("I^k" design), and the other is a "V" design, in which all servers can serve all customers. For the former they apply the well known formulae of the performance of an M/M/N system to identify an upper bound on the number of agents needed. For the "V" design they use results concerning the achievable performance of multi-server systems. This produces a lower bound, due to the maximum flexibility that applies in this system. Today, more is known, both about the staffing in the "I" design (the square-root staffing rule, for example), and about the achievable region (Glazebrook and Niño-Mora [28] provide performance bounds for the system with "V" design and different service requirements for different customer classes). Therefore, one may be able to a) obtain tighter staffing bounds, and b) apply this approach to more general designs.

The second approach taken by researchers (see, for example, Whitt [65] and Jennings et al. [37]) is achieving performance bounds by considering an infinite server system. In a single class infinite server system ($M/G/\infty$), the number of busy servers found by an arriving customer has a Poisson distribution with mean $R = \frac{\lambda}{\mu}$, and the heuristic assumes that in large finite-server systems, this number is *nearly* Poisson if delays are not prevalent. In turn, a Poisson random variable with mean R is approximately a normally distributed random variable with mean R and standard deviation $\sqrt{R}$. Then, given a target waiting probability of α, one chooses the number of servers, N, to be $N = R + \beta\sqrt{R}$ such that

$$\alpha = 1 - \Phi(\beta).$$

This is justified by

$$\begin{aligned} P\{\text{Wait} > 0\} &= P\{\text{Number of busy servers} \geq N\} \\ &\approx P\{R + Z\sqrt{R} \geq R + \beta\sqrt{R}\} = 1 - \Phi(\beta). \end{aligned}$$

Here Z denotes a standard normal random variable, and the PASTA property ensures the first equality. This heuristically justifies the applicability of the square-root safety-staffing rule, and for small values of $P\{\text{Wait} > 0\}$, the heuristic's recommendation essentially matches that of QED regime. In order to apply this approach in the multi-type case, one could perhaps use performance measures of an infinite server, multiclass queueing system with a given design. One example in which such performance measures have been computed is Alanyali and Hajek [11].

Stability staffing: If one is to apply the square-root staffing principle to complex multi-skill settings, a natural question arises: how to define the offered load, beyond which one needs to add a safety staffing. If the service rate of a class i customer is independent of the type of server (that is, if $\mu_{ij} = \mu_i$ for all j), then the offered load is easily calculated according to $R = \sum_i \frac{\lambda_i}{\mu_i}$. However, if μ_{ij} is different for different j's, then the offered load will generally depend on the fraction of class i customers who are served by each server type j. In fact, in this case, the offered load should be thought of as a vector of offered loads, whose entries are the offered loads on each server pool. Armony and Bambos [3] propose a definition for the vector of offered loads that is independent of the routing rule, and is based on a solution of an optimization problem. Specifically, [3] sets up a mixed integer program (MIP) whose solution specifies the number of servers required of each skill, in order to minimize personnel costs, while ensuring that system in *stable* (hence the name stability staffing). The solution of this MIP (which for large systems may be approximated by a solution to a linear program) can be treated as the vector of offered loads for staffing purposes.

Cost minimizing staffing: This approach takes the point of view of minimizing staffing costs with respect to various constraints such as class-dependant bounds on the mean waiting times, and on the probability of waiting more than pre-specified time. Alternatively, it seeks to minimize total cost which is a sum of the staffing costs and the costs associated with waiting. Both these points of view were taken in Borst et al. [11] for the staffing of a single class call center. The authors in [11] formalized the optimality of the square-root staffing principle. They verified the robustness and accuracy of this form of N as $R + \beta\sqrt{R}$, and showed how the actual value of β depends on the particular model and performance criteria used.

For the M/M/N model, [11] shows that the square-root principle is essentially asymptotically optimal, for large heavily-loaded call centers ($\lambda \uparrow \infty$, $N \uparrow \infty$). There is an ample evidence, however, that the principle is applicable much more broadly [27, 37, 50, 12, 4, 5, 2]. Given the applicability of the square-root safety-staffing rule in many different settings, it is natural to examine its applicability in the presence of multiple customer class and / or server types. Note, however, that even in the single class case, there are situations in which the QED regime is not optimal. For example, Jelenkovic et al. [36] show that for the G/D/N queue, if the inter-arrival time distribution is heavy tailed, then the appropriate safety-staffing is of larger order than square-root of the offered load. Having said that, we note that as long as the arrival process is Poisson, or other renewal processes with light-tail inter-arrival time distribution, the QED regime appears to be very robust.

Given the complexity of general large-scale service systems, it is difficult to assess the applicability of the square-root safety-staffing rule to these systems. An approach that may lead to simple staffing rules is that of looking at *simple* routing schemes (that may optimize other performance criteria - other than cost minimization). This approach has already lead to a sim-

ple staffing rule in a particular multi-class setting (see [4] and [5]); indeed, conjecture that it will be widely applicable in more general settings. In fact, we find it to be very useful for the model studied in this paper.

The approaches described above are all quite promising; at the same time they each have their own subtleties and challenges. The staffing rules obtained using these different approaches may turn out to be quite different; nevertheless, they all enjoy the potential of producing interesting and useful results. In this work, we take the third approach of minimizing staffing costs while maintaining performance levels within pre-specified bounds. Moreover, we manage to solve both routing and staffing problems simultaneously, a task which is commonly beyond reach. An example in Harrison and Zeevi [33], in which the authors suggest an algorithm for determining optimal routing and staffing levels when the target is to minimize overall abandonment costs in a general multi-class multi-type setting.

1.1.4 Design

On the design front, even less has been done. Ganz and Zhou [24] develop a dynamic programming (DP) model of long term server hiring that admits a general class of controls. There, the lower level routing problem is explicitly modelled as the core of the DP's one-period cost function, and the optimal hiring policies are characterized as analogues to "order-up-to" policies in the inventory literature. Other studies we are aware of focus on design for flexibility that results from the cross-training of service reps. Such is the paper by Wallace and Whitt [62] which shows how performance is improved with the increase of flexibility. In particular [62] shows that the biggest improvement in performance is obtained when replacing a completely specialized system (with no overlapping skills) by a system with little flexibility. For more on design for flexibility in service systems see Aksin and Karaesmen [1] and references therein. There is also much work on design for flexibility in the context of manufacturing systems. For an outline of the existing approaches and a survey of the literature on the subject see Hopp and Van-Oyen [34] and the references therein.

1.1.5 The V Model

A particular case of the skills-based routing model is the so-called General V Model: several customer types are served by one pool of servers. The importance of the General-V case is that it isolates the scheduling problem of different types of customers between a group of statistically identical servers. In [72] the authors considered the multi-server V-Model with Poisson arrival streams and identically distributed exponential service time for all customer

classes, $\mu_i = \mu$, $\forall i = 1, 2, ..., J$, where J is the number of classes. They considered the optimization problem of minimizing discounted holding costs in the long run, where a class i customer has a holding cost of c_i per unit of time. Assume, without loss of generality, that the c_i's are ordered in descending order ($c_1 \geq c_2 ... \geq c_J$). Then the optimal policy is a threshold policy where the control rule is:
Upon finishing a service, a server chooses to serve a class i customer if there are no customers in all higher-priority queues and there are more than $K_i(x)$ idle servers where $0 \leq K_1 \leq K_2 \leq .. \leq K_J \leq N$, N is the number of servers and $x = (x_1, ..., x_J)$ is the state descriptor in which $x_i, i > 1$ describes the number of class i customers in queue and x_1 describes the number of class 1 customers in queue plus the total number of customers in service (regardless of their class identity). Within each queue, service follows a First-Come First-Served (FCFS) order. It is important to note that [72] establishes structure for the optimal scheduling policy but it does not address the problem of choosing the appropriate threshold levels $K_i(x), i = 1, 2, ..., J$. The dependence of the thresholds on the state of the system makes the choice of the thresholds a very complicated task.

Note that the proposed policy is not work conserving, namely 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 [63]) that the $c\mu$ rule is optimal also when classes have different service requirements. The $c\mu$ rule is a static priority rule that assigns priorities according to $c_i\mu_i$ values: the higher the value of $c_i\mu_i$ the higher the priority of class i. Here c_i is the cost incurred by a priority i customer waiting one unit of time, and μ_i is the service rate of priority i customers. As for multi-server, the authors of [20] 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. This policy is clearly suboptimal when allowing imposed idleness.

Under conventional heavy traffic the V Model, as well as much more complicated scenarios, are amenable to analysis. Van Mieghem [61] analyzed the single server V Model under heavy traffic and proved the asymptotic optimality of the so-called Generalized $c\mu$ (or *Gc*μ) rule: upon completion of service, a server chooses to serve next a customer of class i^* for which

$$i^* = \arg\max_i C_i'(W_i(t))\mu_i \ ,$$

where $C_i(t)$ is the (convex) cost incurred by a priority i customer waiting t units of time and $W_i(t)$ is the waiting time at time t of the oldest customer in queue i. Later, in [44], 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 [50].

Here, the authors considered the more general setting of $GI/PH/N$. They allow each class to a have a different Phase-Type service time distribution. convergence of the scaled queueing process to multi-dimensional diffusion limits. In particular, they consider the limits of the V Model under FIFO and non-preemptive priority schemes.

Armony et al. [4] and [5] 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. Callers self-select their type of service, given information on their expected 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 [4, 5] devise a routing algorithm which is asymptotically optimal (in the sense that it minimizes real-time delays, while guaranteeing the delay bound for the postponed service), and determine that the square-root safety-staffing rule is optimal under the criteria of minimizing staffing costs, while maintaining pre-specified performance measures (such as average waiting time, and the fraction of callers who wait more than a certain length of time). Hence, [4, 5] are examples in which both the routing and the staffing problems are solved jointly, and where the square-root staffing principle applies.

Another control problem of the V Model in the *QED* regime was considered in Rami et al. [6]. A Brownian Control Problem is constructed for the two class V model under exponentially distributed service times and where both customer classes have exponential patience. For linear discounted queueing costs it is shown that under particular assumptions (such as $\mu_1 = \mu_2$) 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 suggest a specific policy to obtain this performance. In our work we show that under certain cost functions, or alternatively under certain constraints, a threshold policy in which the thresholds are state *in*dependent is optimal in some asymptotic sense. To this end, we would like to introduce the $M/M/\{K_i\}$ model.

1.1.6 The $M/M/\{K_i\}$ Model

$M/M/\{K_i\}$ is a multi-class multi-server system in which different customer classes are served according to a threshold policy, where thresholds (on the number of idle servers) are constants and state independent. This service discipline is also motivated by applications in police and ambulance dispatching, hospital bed management, communication channel allocation, and many other priority queueing systems in which it is desirable to retain a "strategic reserve" of servers for higher priority customers. See Schaack and Larson [55] for further discussion of the relevance of the model to the dispatch of police cars. Also, see Rege and Sengupta [51] for some interesting examples of application of the model to communications.

More formally, the model assumes that customers from class i ($i = 1, ..., J$) arrive into an N-Server queueing system according to a homogeneous Poisson process, with arrival rate λ_i customers per unit time. All Poisson streams are independent. Service time is assumed to be exponential with mean $1/\mu$, independent of the priority of the customer or the identity of the server. The service discipline is assumed to be non-preemptive. The threshold based scheduling rule is then as follows:
Assign an idle server to a priority i customer only if there are more than K_i idle servers and all higher priority queues are empty ($0 = K_1 \leq K_2 \leq ... \leq K_J$). Customers that upon arrival can not be served immediately are backlogged in a queue dedicated to their priority class. Each queue is depleted in a FCFS manner.

Our notation of the model is based on the notation $M/M/\{N_i\}$, used in [55], where it denotes a Marokovian (Poisson) input, exponential service times, and a set of server *cutoffs* $\{N_i\}$. In our notation K_i denotes the *least* number of *idle* servers needed before accepting a priority i customer to service whereas, in the notation used in [55], N_i denotes the maximum number of busy servers beyond which priority i customers are not accepted to service. Thus, $K_i + N_i = N$, $i = 1, ..., J$.

An exact analysis of the model, including the probability of delay for each priority type as well as the Laplace transforms for their waiting times, was performed in [55], using an $M/G/1$ reduction, i.e. utilizing the fact that, given wait, the queue of class i customers behaves like an $M/G/1$ queue where the Laplace transform of G can be obtained by a sequence of recursive equations. The recursive equations obtained in [55] translate into quite complicated expression even for the two class case. Let P_n be the steady state probability that n servers are busy ($0 \leq n \leq N$). Also let K be the threshold level for the low priority customers (that is, low priority customers will enter service only if there are more than K idle servers). Finally, let $M = N - K$. The stability conditions of the system, as given in [55], are:

$$\frac{\lambda_1}{N\mu} < 1 \tag{2}$$

and

$$\lambda_2 h_1(M) < 1 \,, \tag{3}$$

where

$$h_1(M) = \frac{1}{M\mu} + \sum_{k=2}^{N-M} \frac{{\lambda_1}^{k-1}}{\mu^k \prod_{j=0}^{k-1}(M+j)} + \frac{{\lambda_1}^{N-M}}{(N\mu - \lambda)\,\mu^{N-M} \prod_{l=1}^{N-M}(N-l)} \,. \tag{4}$$

Under these conditions and the PASTA property we have by [55] that the probability of delay for high priority customers is:

$$P_N = P_0 \cdot \left(\frac{\lambda_1 + \lambda_2}{2}\right)^M \cdot \frac{{\lambda_1}^N}{\mu} \cdot \frac{1}{N!} \cdot \frac{1}{1 - \frac{\lambda_1}{N\mu}} \cdot \frac{1}{1 - \lambda_2 h_1(M)} \,, \tag{5}$$

where

$$P_0 = \left[\sum_{n=0}^{M-1}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^n\frac{1}{n!}+\sum_{n=M}^{N-1}\left(\frac{\lambda_1+\lambda_2}{2}\right)^M\left(\frac{\lambda_1}{\mu}\right)^n\frac{1}{n!}\frac{1}{1-\lambda_2 h_1(M)}\right.$$
$$\left.+\left(\frac{\lambda_1+\lambda_2}{2}\right)^M\left(\frac{\lambda_1}{\mu}\right)^N\frac{1}{N!}\frac{1}{1-\frac{\lambda_1}{N\mu}}\frac{1}{1-\lambda_2 h_1(M)}\right]^{-1}.$$

Define P_{M+} to be delay probability for low priority customers. Then, by PASTA, P_{M+} is equal to the probability that M or more servers are busy, i.e.

$$P_{M+} = P_0\left(\sum_{n=M}^{N-1}\left(\frac{\lambda_1+\lambda_2}{2}\right)^M\left(\frac{\lambda_1}{\mu}\right)^n\frac{1}{n!}\frac{1}{1-\lambda_2 h_1(M)}+\left(\frac{\lambda_1+\lambda_2}{2}\right)^M\left(\frac{\lambda_1}{\mu}\right)^N\frac{1}{N!}\frac{1}{1-\frac{\lambda_1}{N\mu}}\frac{1}{1-\lambda_2 h_1(M)}\right). \quad (6)$$

We denote by W_i, $i=1,2$, the waiting time of class i customers. Then,

$$E[W_2] = \frac{F}{2h_1(M)}\frac{1}{1-\lambda_2 h_1(M)} \quad (7)$$

where

$$F = \frac{\lambda_1/\mu^{N-M}}{\prod_{j=0}^{N-M-1}(M+j)}\frac{2N\mu}{(N\mu-\lambda_1)^3}+2\left(\sum_{k=1}^{N-M-1}\frac{\lambda_1/\mu^k}{\prod_{j=0}^{k-1}(M+j)}h_1(M+k)^2+h_1(M)^2\right) \quad (8)$$

In this work we will show how these complicated expressions translate asymptotically into simpler forms. In the following section we summarize our results and illustrate some of them through a simple example of a two-class V Model.

2 Summary of our Results

The asymptotically optimal routing policy we propose is a threshold type priority policy. According to the proposed policy a class i customer is admitted to service only if there are more than K_i servers idle. The asymptotic optimality is in terms of the steady state holding costs and delay constraints. The proposed policy allows one to differentiate between the probabilities of delay for the different customers in a quite delicate manner. This is achieved by choosing the appropriate sizes of the thresholds, which asymptotically turns out to be a rather simple function of the problem parameters.

Moreover, we propose an asymptotically optimal staffing level that goes with the proposed policy and gives rise to the QED regime. We show that for reasonable constraints on the probability of delay, this staffing level is determined easily as a function of the overall load on the system and the parameters of the lowest priority class. We deduce that for an unconstrained,

linear holding costs problem, the $c\mu$ rule is asymptotically optimal among all non-anticipating policies: work conserving and non-work conserving.

We extend part of the results in Garnett et. al. [27] to the case of non-preemptive priorities with thresholds where we allow the different classes to have different patience parameters. Moreover, we show that under certain setting the threshold policy is optimal when one wishes to minimize abandonment costs.

2.1 An Illustrative Example

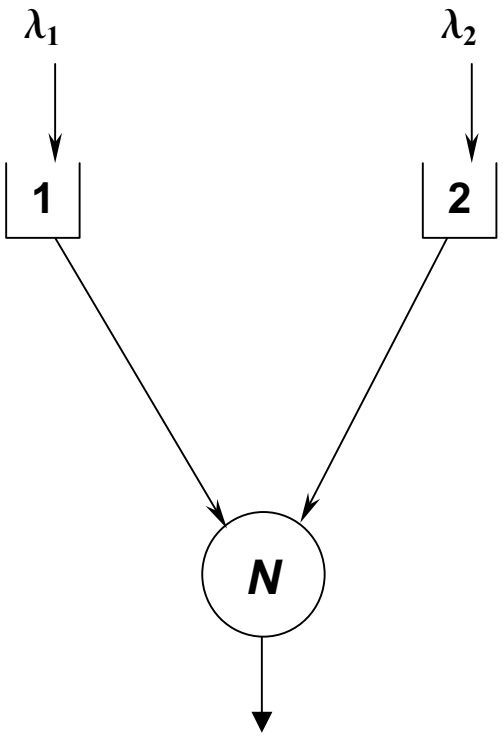


Figure 2: An Example of a V Model

To clarify our results, let us consider a particular case of the V Model where there are two customer classes: high priority and low priority (see Figure 2.1). The two classes are denoted by 1 and 2 respectively. This model, despite its relative simplicity, already provides some interesting insights.

The model specifics are as follows: Both classes have a common exponentially distributed service time with rate μ. The arrival process for priorities 1 and 2 are Poisson with arrival

rates λ_1 and λ_2 respectively and we denote by λ the total arrival rate (i.e. $\lambda = \lambda_1 + \lambda_2$). The only restriction we impose on the arrival rates is that λ_2 is comparable to λ, i.e. as λ increases to infinity we assume that $\lambda_2/\lambda \to a_2$ where $a_2 > 0$. The reason will become clearer in the following sections. For the meanwhile we just point out that, as long as λ_2 is comparable to λ, the high priority customers will experience light traffic regardless of whether λ_1 is comparable to λ or not. Assume that N, the total number of servers, is determined by the square root safety staffing rule mentioned previously ($N = R + \beta\sqrt{R}$ with $R = \frac{\lambda_1+\lambda_2}{\mu}$), and therefore the system operates in the QED regime. We will show later that the QED regime is a direct outcome of the optimization problem, similarly to [11]. That is, the QED regime is an outcome rather than an assumption.

Once the staffing level has been determined there is still one degree of freedom in the model parameters - the choice of K_2, the threshold level on the number of idle servers before serving class 2 customers.

We will show how one can carefully adjust the threshold level to obtain different delay probabilities for both classes.

Despite the fact that this model assumes non-preemption, one expects an asymptotic equivalence of preemptive and non-preemptive strategies, as suggested in [6]. This enables the following insight: while the low priority customers could experience different service levels under the three different regimes, depending on the magnitude of the threshold, the high priority customers will, regardless of the threshold chosen, experience a Quality Driven service level. This becomes clear by noting that, under the preemptive policy, the high priority customers do not "see" the low priority customers and therefore experience a system with substantially more staffing than needed to accommodate their workload.

2.1.1 Steady State Analysis

The formulae obtained by the recursive equations in [55] are very complicated even for the two-class case. We will show in the following sections that asymptotic analysis does not require the explicit use of these formulae.

For the asymptotic analysis we consider a sequence of $M/M/\{K_i\}$ system, indexed by a superscript r that denotes the r^{th} system. For example: $E[W_1^r]$ stands for the average waiting time of high priority customers in the r^{th} system. For the meanwhile let us assume that our system is staffed according to the *Square-Root Staffing* rule, i.e. λ^r and N^r scale in a manner that $\sqrt{N^r}(1-\rho^r) \to \beta$, $0 < \beta < \infty$, as $r \to \infty$.

As a preliminary observation, note that if all customers are served FIFO (with no differentiation between both classes) the asymptotic probability of delay as given in [30] is

$$\alpha(\beta) = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}, \tag{9}$$

in which $\Phi(\cdot)$ and $\phi(\cdot)$ are the normal distribution and density function respectively.

Having this observation we can now describe what happens as we create service level differentiation between the two classes, using the $M/M/\{K_i\}$ model.

Table 1 is based on Section 4.2 and it summarizes the relations between the threshold level and the service level, as experienced asymptotically by both customer classes. Here, ρ_1^r stands for the load generated by the high priority class, i.e. $\rho_1^r = \frac{\lambda_1^r}{N^r\mu}$

For two sequences $\{a_n\}_{n=1}^{\infty}$, $\{b_n\}_{n=1}^{\infty}$, we say that a_n is $\Theta(b_n)$ if $\lim_{n\to\infty} \frac{a_n}{b_n} \to a$, where a is a finite constant; When $a = 1$ we say that $a_n \sim b_n$. Also, b, c and d, are constant that do not scale with r.

♯	Threshold K	$\sim P\{W_1^r > 0\}$	$\sim P\{W_2^r > 0\}$	$E[W_1^r \mid W_1^r > 0]$	$E[W_2^r \mid W_2^r > 0]$
A	0	$0 < \alpha(\beta) < 1$	$0 < \alpha(\beta) < 1$	$\Theta(\frac{1}{N})$	$\Theta(\frac{1}{\sqrt{N}})$
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	$d \cdot \sqrt{N}$	$\Theta(\alpha(\beta - d)\rho_1^{d\sqrt{N}})$	$\alpha(\beta - d)$	$\Theta(\frac{1}{N})$	$\Theta(\frac{1}{\sqrt{N}})$

Table 1: A Two-Class V Model: Service Levels for Both Classes

The β in Table 1 is obtained via the Halfin Whitt limit for N servers and arrival rates equal to $N - \beta\sqrt{N}$. For stability reasons (see Section 4.2.1 we assume that $d < \beta$. In case D we have that the probability of delay of the high priority is such that

$$\frac{P\{W_1^r > 0\}}{\alpha(\beta - d)(\rho_1^r)^{d\sqrt{N^r}}} \to \eta,\ 1 \le \eta \le e^{d^2}$$

Remarks on the probability of delay and the waiting time distribution:

1. Case A in Table 1 is simply a two class static priority system. In this system the probability of delay for both customer classes should be the same (equal to P_N) and asymptotically given by $\alpha(\beta)$ from (9).

2. In all cases, $P\{W_2 > 0\}$ is equal to the probability of the event "more than or exactly $N-K$ servers busy". This probability is clearly smaller than the probability of the same

event in a single class system with the same overall arrival rate and $N-K$ servers. Also, it is higher than the probability of finding N servers busy in a single class system with the same overall arrival rate and N servers. This procedure results in tight bounds for $P\{W_2 > 0\}$ and in turn it implies that $P\{W_2^r > 0\} \to \alpha(\beta)$. Actually, the same reasoning works for any threshold that is $o(\sqrt{N})$, but it does not work for Case D. In Proposition 4.2.1 we give the result for the general threshold case.

3. Note that in all cases, A-C, the probability of delay for the high priority is of the form $\alpha(\delta) \cdot (\rho_1^r)^{K^r}$, where δ equals β for cases A-C and β_1 for case D. Hence, the probability of delay for high priority is always proportional to ρ_1 to the power of the threshold. The result for the general case is given in Proposition 4.2.2.

4. In all cases, A-D, we can prove that $\sqrt{N^r}W_2^r$ converges to a mixture of an exponential random variable and a point mass at the origin (see corollary 4.2.4. Proposition 4.2.3 gives asymptotic Laplace transforms and average waiting times for the high priorities.

Note that Table 1 does not cover the case where the threshold is proportional to N. The reason is the instability of the system when the threshold is proportional to N. We give a simple set of necessary and sufficient conditions for stability in Section 4.2.1.

2.1.2 Static Priorities

In this section we briefly consider the case of non-preemptive static priorities, which is a particular case of the $M/M/\{K_i\}$ model where all K_i's are taken to be zero (see Case A in Table 1). Due to its relative simplicity, the static priority setting is very useful for developing intuition that would apply later for the more general $M/M/\{K_i\}$ model.

The performance of the non-preemptive static priorities case was analyzed in Kella and Yechiali [39]. The average waiting times for this model are given by:

$$E[W_1] = \pi[N\mu(1-\rho_1)]^{-1}, \quad E[W_2] = \pi[N\mu(1-\rho_2)(1-\rho_1)]^{-1}$$

where

$$\pi = \frac{(\lambda/\mu)^N}{N!(1-\rho)}\left[\sum_{i=0}^{N-1}\frac{(\lambda/\mu)^i}{i!} + \frac{(\lambda/\mu)^N}{N!(1-\rho)}\right]^{-1},$$

and $\lambda = \lambda_1 + \lambda_2$, $\rho_1 = \lambda_1/N\mu$, $\rho_2 = \lambda_2/N\mu$ and $\rho = \rho_1 + \rho_2$.

Based on an $M/G/1$ reduction, [39] gives moments and Laplace transforms for $W|W > 0$ of the different classes in the static priority V-Model. These expressions will be used later on to

determine the limiting expressions of $W|W > 0$ in the threshold system. As explained above, the probability of delay for both customer classes is the same (equal to P_N) and asymptotically given by formula (9) for $\alpha(\beta)$.

Preemptive Vs. Non-Preemptive

In Atar et al. [6], the authors suggested the asymptotic equivalence of preemptive and non-preemptive policies for the V-Model. Again, through this simple setting the equivalence (mainly for the low priorities) can be illustrated in quite a convincing manner.

The following conclusions can be drawn:

1. Probability of Delay:
   The probability of delay for low priority will remain the same under preemptive and non-preemptive regimes disciplines (since the Birth and Death process representing the total number in system is the same for both disciplines). Consider again a sequence of systems indexed by r such that $\lambda^r \to \infty$ and both classes are comparable. Choose the staffing level in a manner that $\sqrt{N^r}(1-\rho^r) \to \beta$, $0 < \beta < \infty$, where $\rho^r = \lambda^r/(N^r\mu)$. Under the preemptive discipline, the delay probability for the high priority is given by the Erlang-C Formula for an $M/M/N$ system with arrival rate equal to λ_1 and it converges to zero at rate
   $$O\left(\frac{e^{-N^r\rho_1}(N^r\rho_1)^{N^r}}{N^r!(1-\rho_1)}\right).$$
   where $\rho_1 = \lim_{r\to\infty}\frac{\lambda_1^r}{N^r\mu}$. For two positive sequences $\{a_n\}$, and $\{b_n\}$, we say that a_n is $O(b_n)$ if for some constant $c \geq 0$, $\limsup a_n/b_n \leq c$. This convergence can be established by simple manipulations using the approximations for Poisson tails given in [29].

2. Queue Lengths:
   The sum processes (i.e. the total number of customers in system) for both preemptive and non-preemptive disciplines are equal in law to the respective single class $M/M/N$ system with arrival rate $\lambda_1 + \lambda_2$. Therefore, by [30] we have that
   $$\frac{1}{\sqrt{N^r}}(Q_2^r + Q_1^r) \Rightarrow Q$$
   where Q_i^r denotes the steady state number of type i customers in the r^{th} queue and Q is a proper random variable. For both the preemptive and the non-preemptive cases we claim that,
   $$\frac{1}{\sqrt{N^r}}Q_1^r \Rightarrow 0$$

and therefore

$$\frac{1}{\sqrt{N^r}}Q_2^r \Rightarrow Q$$

which means that in the limit the low priority queue is equal in law for both preemptive and non-preemptive regimes.

To support the claim that $\frac{1}{\sqrt{N^r}}(Q_1^r) \Rightarrow 0$, note that given wait the high priority queue behaves like an under-loaded $M/M/1$ queue. So, even in the worst case scenario (when the delay probability equals 1) we have that $E[Q_1] = E[Q_1|Q_1 > 0] = O(1)$. This is true for both preemptive and non-preemptive regimes and therefore high priority will disappear from queue under the normalization above.

3. Average Waiting Time:
Low Priority: The order of the average waiting time of the low priority in the non-preemptive case is given in Table 1. For the preemptive case we can calculate explicitly the waiting time of the law priority which is given by

$$E[W_2] = \frac{1}{\lambda_2}\left[(\lambda_1+\lambda_2)E[\overline{W}] - \lambda_1 E[W_1]\right].$$

Here, W_1 and W_2 remain as before and $\overline{W}$ is the waiting time in a single class $M/M/N$ system with $\lambda = \lambda_1 + \lambda_2$. But there is no need for explicit calculations. We have claimed above that the asymptotic queue length of low priority is the same for both the preemptive and non-preemptive regimes. Based on this formula it can be shown that the average waiting time of the low priority class is the same under the preemptive and the non-preemptive regimes..

High Priority: As shown in Table 1, the waiting time of the high priority customers under the non-preemptive regime is $\Theta(1/N^r)$. Under the preemptive regime it will no longer be $\Theta(1/N^r)$ but instead it will behave like

$$O\left(\frac{e^{-N^r\rho_1}(N^r\rho_1)^N}{N^r!(1-\rho_1)}\frac{1}{N^r}\right).$$

The latter is a direct consequence of the rate of convergence of the delay probability.

To summarize, we have illustrated by a simple two-class example the asymptotic equivalence of preemptive and non-preemptive regimes. While the waiting time of high priority will improve significantly under the preemptive regime, the waiting time of the low priority will hardly suffer any deterioration, in particular their waiting time is asymptotically the same under the preemptive and non-preemptive regimes.

The discussion in this example was limited to questions of performance analysis of the $M/M/\{K_i\}$ model. In this work, we show that the $M/M/\{K_i\}$ model is of great relevance. In particular we solve concurrently the questions of staffing and control of the V Model and show that using $M/M/\{K_i\}$ is asymptotically optimal. We now turn to a formal presentation of the $M/M/\{K_i\}$ model.

2.2 Thesis Outline

We formally introduce the $M/M/\{K_i\}$ Model in Section 3. Section 4 contains complete transient and steady state analysis of the $M/M/\{K_i\}$ model in the QED regime. In particular we show that the diffusion limit of the overall number of customers in system, when properly scaled and normalized, converges to a diffusion limit with a piecewise linear drift. The result is obtained through a collapse of the multi-dimensional state-space into a single dimension - the overall number of customers in system.

In the context of steady state analysis we give, in Subsection 4.2, a set of necessary and sufficient conditions for stability and present asymptotic steady state performance measures for the probability of delay and waiting time distributions for all customer classes.

Section 5 contains adaptation of the result of Section 4 to the case of Efficiency Driven $M/M/\{K_i\}$.

We conclude the first part of the thesis in Section 6 with the solution to both optimization problems (10) and (11). We show that the $M/M/\{K_i\}$ is the asymptotically optimal policy for both the constraint satisfaction and the cost minimization problems.

In Section 7 we present some extensions to the V-Model. The main extension presented is the introduction of abandonments into the V-Model. In particular, we assume that customers of class i have exponential patience with rate θ_i. We give complete analysis, transient and steady state, for the $M/M/\{K_i\}$ model in the new setting. Also we prove, that under certain cost structures the $M/M/\{K_i\}$ policy minimizes the overall abandonment cost. We conclude this section with a brief presentation of ongoing research about the N Model.

We now proceed to the formulation of the V-Model under the $M/M/\{K_i\}$ policy.

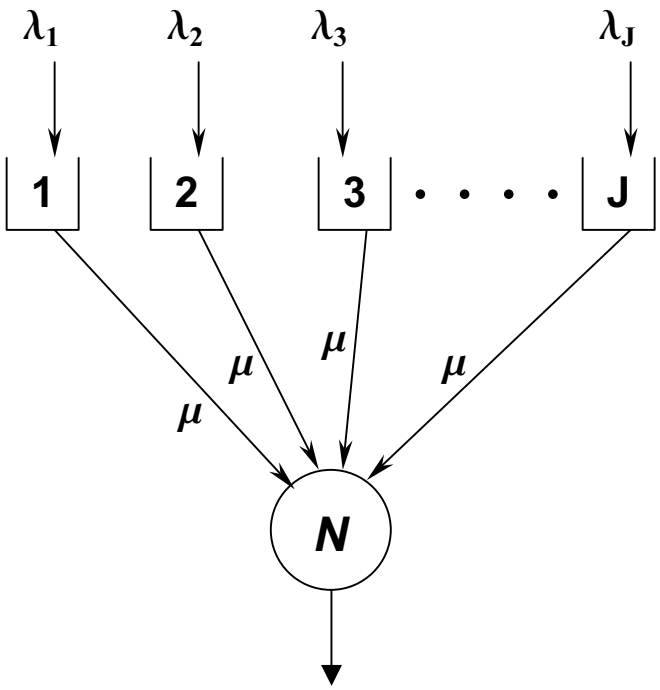


Figure 3: The V Model

3 Model Formulation

Consider the system described in Figure 3 with J customer classes and a single server type. 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.

We wish to minimize the staffing cost while maintaining a target service level constraint. The service level performance measure that we study is the steady state probability that a class i customer waits before starting service. Denote this steady state probability by $P\{W_i > 0\}$ for class i, and let α_i be its target delay probability. Assume that the classes are ordered in an increasing order of α_i: $\alpha_1 \leq \alpha_2 \leq ... \leq \alpha_J$, namely class 1 are the highest priority and class J are the lowest priority.

Let Π be the set of all non-preemptive non-anticipative scheduling policies. Given a policy $\pi \in \Pi$, let $P_\pi\{W_i > 0\}$ be the steady state probability that a customer from class i is delayed

before his service starts. The staffing problem is then stated as follows:

$$\begin{array}{ll} \text{minimize} & N \\ \text{subject to} & P_\pi(W_i > 0) \leq \alpha_i,\ 0 < \alpha_i < 1, \quad i = 1, ..., J, \text{ for some } \pi \in \Pi; \\ & N \in \mathbb{Z}_+ \end{array} \tag{10}$$

As the arrival rate to the system increases, we will also allow $\alpha_i,\ i = 1, ..., J-1$, to converge to zero in a certain manner.

Another problem formulation is to assume linear waiting costs for all classes, i.e. a unit waiting time of a class i customer incurs a cost of c_i. Now, assume that classes are ordered in decreasing order of their cost, i.e. $c_1 \geq c_2 \geq ... \geq c_J$; again, 1 is the highest priority and J is the lowest. Then, we will show that the same type of policy that asymptotically minimizes (10) is also the solution to the following problem:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{J} c_i \lambda_i E[W_i] + N \\ \text{subject to} & P_\pi(W_i > 0) \leq \alpha_i\,, \quad i = 1, ..., J, \text{ for some } \pi \in \Pi; \\ & N \in \mathbb{Z}_+ \end{array} \tag{11}$$

Under the second formulation we allow the α_i's to be equal to 1. If all α_i's are equal to 1 (11) becomes a pure cost minimization problem. We will also consider the case where the coefficients $c_i,\ i = 1, ..., J$, are allowed to scale with the overall arrival rate, λ, in certain manners.

Notation:

For the $M/M/\{K_i\}$ model, we denote by $Z(t)$ the number of busy servers at time t, and by $Q_i(t)$ the number of class i customers in queue at time t. Then, the J dimensional vector $\{Z(t) + Q_1(t), Q_i(t) : i = 2, ..., J\}$ gives a Markovian description of the system. The probability of delay for class i can be stated in terms of the system state. In particular, due to the PASTA property, if the model's parameters are such that steady state exists then

$$P_\pi\{W_i > 0\} = P\{Z(t) \geq N - K_i\} \tag{12}$$

In addition, though not essential for the description of the system, we would like to define $Z_i(t)$ to be the number of busy servers above the level of $N - K_{J-i+1}$ servers, or equivalently $Z_i(t) = [Z(t) - N - K_{J-i+1}]^+$, where

$$[Z(t) - N - K_{J-i+1}]^+ \triangleq ((Z(t) - (N - K_{J-i+1})) \vee 0).$$

The rest of the thesis is organized as follows: In the following section we will analyze limits (Diffusion and Steady State) for the $M/M/\{K_i\}$ model in the QED regime. In Section 5, we

will make a short excursion through Efficiency Driven analysis of the $M/M/\{K_i\}$ model. We will conclude in Section 6 by showing (relying mainly on [11]), how different cost structures or constraints give rise asymptotically to either the Efficiency Driven or QED regimes.

4 QED Asymptotic Framework

As mentioned before, we have (by [72]) that the optimal policy, when trying to minimize waiting costs, is to use a threshold policy. The optimal thresholds are state dependent which makes the use of the $M/M/\{K_i\}$ policy suboptimal for fixed number of servers N. Moreover, even if the $M/M/\{K_i\}$ policy was optimal we would still have to determine the staffing level and the optimal thresholds. We could have made use of the work done in [55] to solve the optimization problem (10) by direct enumeration for systems of reasonably small size. However, as shown for the two class case this is very complicated, time consuming and is not likely to provide any further useful insights. Instead, we take an asymptotic approach which finds asymptotically optimal staffing rules for systems with high demand. To this end, we consider a sequence of systems indexed by $r = 1, 2, ...$ (to appear as a superscript) with increasing demand values $\lambda^r \to \infty$ as $r \to \infty$ and a fixed service rate μ. All other quantities that are associated with the r^{th} system will be denoted with a superscript r. We 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

$$\lim_{r\to\infty} \frac{\lambda_k^r}{\lambda^r} = a_k,\ k = 1, ..., J; \quad a_J > 0,\ a_i \geq 0,\ i = 1, ..., J-1 \tag{13}$$

We consider a sequence of $M/M/\{K_i\}$ systems indexed by r. The r^{th} system is staffed with N^r servers and the customers are routed according to J thresholds given by $K_1^r \leq K_2^r \leq ... \leq K_J^r \stackrel{\triangle}{=} K^r$, where $K_1^r \equiv 0$; i.e. an arriving class i customer will enter service immediately upon arrival only if there are more than K_i^r idle servers. Upon a service completion, if there are k idle servers, admit into service the first customer from class i^*, where $i^* = \min_i\{K_i^r < k,\ i^{th}$ queue is not empty$\}$

The appropriate staffing level will be determined according to the solution of the optimization problems (10) and (11) given in Section 6. For the time being we assume that the number of servers grows with r in the following manner:

$$\lim_{r\to\infty} \sqrt{N^r}(1 - \rho_C^r) = \beta\,,\ 0 < \beta < \infty \tag{14}$$

where

$$\rho_C^r = \frac{\lambda^r}{(N^r - K^r)\mu}. \tag{15}$$

For simplicity of presentation of the results **We restrict our performance analysis to** $K^r = o(N^r)$. All the results that follow hold also in the case where $K^r \neq o(N^r)$ (unless stated otherwise) with the heavy traffic condition (14) 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 $\sqrt{N^r}$. For example, $X^r(t)$ in Section will be defined as $\dfrac{Y^r(t) - (N^r - K^r)}{\sqrt{N^r - K^r}}$.

Remarks:

- **Since we restrict ourselves to** $K^r = o(N^r)$, (14) implies that $\rho^r \stackrel{\triangle}{=} \lambda^r/(N^r\mu)$ converges to 1 as $r \to \infty$, and the system is in *heavy traffic*.
- Looking at the "super-class" composed from classes 1 through $J-1$, (13) implies that

$$\rho^r_{1\to J-1} \stackrel{\triangle}{=} \frac{\sum_{i=1}^{J-1} \lambda_i^r}{(N^r\mu)} \to \delta < 1 \tag{16}$$

  By (14), this also means that $\frac{\sum_{i=1}^{J-1} \lambda_i^r}{(N^r - K^r\mu)} \to \delta$. Equivalently we can say that under the $M/M/\{K_i\}$ policy with the staffing implied by (14), all classes, except for class J, can experience light traffic, or *Quality Driven* service.

- Note that (14) is typically different from the heavy traffic condition of [30], which is given by

$$\lim_{r\to\infty} \sqrt{N^r}(1 - \rho^r) = \beta', 0 < \beta' < \infty. \tag{17}$$

  Still, $\beta = \beta'$ whenever $K^r = o(\sqrt{N^r})$.

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 FLLN and FCLT we have

$$\frac{1}{N^r} A_j^r(t) \Rightarrow \hat{\lambda}_j t \tag{18}$$

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

$$\frac{1}{\sqrt{N^r}}(A_j^r(t) - \lambda_j^r t) \Rightarrow BM(0, \hat{\lambda}_j). \tag{19}$$

Also, define

$$Y^r(t) = Z^r(t) + \sum_{i=1}^{J} Q_j^r(t) \tag{20}$$

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

4.1 Diffusion Limits for the $M/M/\{K_i\}$ model

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

$$X^r(t) = \frac{Y^r(t) - (N^r - K^r)}{\sqrt{N^r}} \tag{21}$$

Theorem 4.1.1 *Assume (13), (14) and that $X^r(0) \Rightarrow X(0)$. Then*

$$X^r \Rightarrow X \tag{22}$$

where X is a diffusion process with infinitesimal drift given by

$$m(x) = \begin{cases} -\beta\mu & x \geq 0 \\ -(\beta + x)\mu & x \leq 0 \end{cases} \tag{23}$$

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

Remarks:

- ρ_C^r is the load on the system assuming only $N^r - K^r$ servers. Note that $N^r - K^r$ is the capacity available for all customers disregarding the priority they have. Of course, when K^r is $o(\sqrt{N^r})$ this centering is asymptotically equivalent to centering around N^r.
- Whenever $K^r = o(\sqrt{N^r})$ we will have that $\beta = \beta'$ (where β' was defined in equation (17). Hence a sequence of threshold systems such that $\beta = \beta'$, converges weakly to the same limit as a sequence of $M/M/N$ queues with β'. We will address this question in detail when dealing with queues and waiting of the different priorities, but we can already conclude that the overall number in system is asymptotically indifferent to reservation of servers for high priorities, as long as the reservation is $o(\sqrt{N})$.

Proof: For simplicity we will 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 our $M/M/\{K_i\}$ system (denoted by (A)). In the second step we will use a coupling argument and the convergence together theorem (Theorem 11.4.7 in [64]) to conclude the proof.

Definition of system B:
Split the server pool into two distinct pools: one with $N^r - K^r$ servers and the other with K^r

servers. Throughout the proof we will denote these two pools by *"The $N-K$ Pool"* and the *"The K pool"* respectively.

Use the following routing policy: as long as the total number in system is below $N-K$ route all customers to the $N-K$ pool. When the system is above $N-K$ (i.e. there are more than $N-K$ servers busy) route any arriving high priority customer to the K pool. If there are any customers in service in the K pool upon a service completion in the $N-K$ pool preempt one of these customers and assign to him/her 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 constructed so that the total number in system process will have the same sample paths and the same probability law. Thus, proving the convergence of (B) will result in the desired convergence for (A).

Let us further introduce 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 $Z_K^r(t)$ the number of busy servers from the K pool in system B. As before define

$$X_B^r(t) = \frac{Y_B^r(t) - (N^r - K^r)}{\sqrt{N^r}} \tag{24}$$

and

$$X_C^r(t) = \frac{Y_C^r(t) - (N^r - K^r)}{\sqrt{N^r}} \tag{25}$$

By our assumption that $\lim_{r\to\infty} \sqrt{N^r}(1-\rho_C^r) = \beta\,,\, 0 < \beta < \infty$ we have from [30] that $X_C^r \Rightarrow X$.

Coupling:
We will now couple system (B) with (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 result will follow by the convergence together theorem. In the following paragraphs we fix $r > 0$ and eliminate the superscript from the notation.

Then, we create the coupled sample paths in the following manner:

We use the same sample path of arrivals for both systems. For simplicity let us assume that both systems are initiated with $N-K$ servers busy and an arrival of a customer. As long as $Y_B(t) > N-K$ and $Y_C(t) > N-K$ we can create the departures for system C as well as for

the $N-K$ pool of system B from a common Poisson process with rate $(N-K)\mu$. System B will have also departures from the K pool generated by a different 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 [67]). i.e. If system B is at level j at a departure epoch and system C is in level l, then the candidate departure event generated from the Poisson process with rate $l\mu$ is an actual departure for system B with probability j/l (recall that $j \leq l$).

During the epoch in which system B is below $N-K$ the distance between the two systems in consideration can only decrease. If the two systems meet they will proceed together until they hit $N-K$ for the first time.

Denote by $D_K(t)$ the departures from the K pool up to time t. Then, we can write (see for example [43])

$$D_k(t) = \mathcal{N}\left(\int_0^T Z^r_{K^r}(t)\mu dt\right) \tag{26}$$

Where, $\mathcal{N}$ is a unit Poisson process.

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

$$\sup_{0\leq t\leq T} \|Y^r_B(t) - Y^r_C(t)\| \leq \mathcal{N}\left(\int_0^T Z^r_{K^r}(t)\mu dt\right) \tag{27}$$

or,

$$\sup_{0\leq t\leq T} \|X^r_B(t) - X^r_C(t)\| \leq \frac{1}{\sqrt{N^r}}\mathcal{N}\left(\int_0^T Z^r_{K^r}(t)\mu dt\right) \tag{28}$$

Proving

$$\frac{1}{\sqrt{N^r}}\mathcal{N}\left(\int_0^T Z^r_{K^r}(t)\mu dt\right) \Rightarrow 0\,, \tag{29}$$

and applying the convergence together theorem leads to the desired result.

To establish (29) it is enough to show that for each r, $Z^r_K(t) + Q_1(t)$ can be path wise bounded by an $M/M/1$ queue with arrival rate $\lambda^r = \lambda^r_1$ 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 $Z^r_{K^r}(t) + Q^r_1(t)$ 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 $Z^r_{K^r}(t)+Q^r_1(t)$. Assume that at time $t \geq 0$ both systems are not empty (in particular assume that $Z^r_k(t) + Q^r_1(t) = j > 0$, i.e. 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 $Z^r_{K^r}(t) + Q^r_1(t)$ will result in a service completion in the $M/M/1$ with probability $\frac{N^r-K^r}{N^r-K^r+j}$. Thus we have proved that for all $t \geq 0$, $Z^r_{K^r}(t) + Q^r_1(t)$ can be path wise bounded by the associated $M/M/1$.

By (13) this M/M/1 is under-loaded and by Theorems 4.1 and 4.2 of [43] its scaled version converges to zero. Since the poisson process $\mathcal{N}\left(\int_0^T Z^r_{K^r}(t)\mu dt\right)$ admits the decomposition (see for example [50])

$$\mathcal{N}\left(\int_0^T Z^r_{K^r}(t)\mu dt\right) = \int_0^T Z^r_{K^r}(t)\mu dt + M^r(t) \tag{30}$$

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 (22). To prove the result for an arbitrary number of classes it is enough to note that in the general case $Z^r_{K^r}(t) + \sum_{i=1}^J Q^r_i(t)$ can also be bounded by an under-loaded $M/M/1$ queue and hence the proof follows.

■

Corollary 4.1.2 *Let $X(\cdot)$ be the diffusion process described in (4.1.1). Then the steady-state distribution of X has a density $f(\cdot)$ which satisfies:*

$$f(x) = \begin{cases} \exp\{-\beta x\}\alpha(\beta) & x \geq 0 \\ \\ \frac{\phi(\beta+x)}{\Phi(\beta)}(1-\alpha(\beta)) & x < 0 \end{cases} \tag{31}$$

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

This result follows from [30].

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 $(Z_1 + Q_1, Q_2, ..., Q_J)$. We call the collapse of the dimensionality of the problem *State-Space Collapse*. The state space collapse of the $M/M/\{K_i\}$ model is summarized by the following corollary:

Corollary 4.1.3 ***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*

$$\begin{aligned} &\tfrac{1}{\sqrt{N^r}}\mathcal{E}^r(t) \Rightarrow 0 \\ &\tfrac{1}{\sqrt{N^r}}Q_i^r(t) \Rightarrow 0,\ \forall i \leq J-1 \\ &\tfrac{1}{\sqrt{N^r}}Q_J^r(t) \Rightarrow X^+ \end{aligned} \tag{32}$$

Proof: Note that $\mathcal{E}^r(t) + Q_1^r(t)$ is just $Z_K^r(t) + \sum_{i=1}^{J-1} Q_i^r(t)$, hence the result follows from the proof of Theorem 4.1.1.

■

The next corollary show how to obtain the limit of the virtual waiting time for class J as a function of the limit queue length process X.

Corollary 4.1.4 *Let $W_i^r(t)$ be the virtual waiting time process for class i. If*

$$\exists -\infty < c < \infty\ : \sqrt{N}(\frac{\lambda_J^r}{N^r} - a_J\mu) \to c\,, \tag{33}$$

then

$$\sqrt{N^r}W_J^r \Rightarrow \frac{1}{a_J\mu}[X]^+. \tag{34}$$

Proof:

By the FCLT for the arrivals and by (33) we have the convergence

$$V^r(t) = \sqrt{N^r}(\frac{A_J^r(t)}{N^r} - a_J\mu t) \Rightarrow V(t) \tag{35}$$

Where $V(t) = \hat{A}(t) + ct$ and $\hat{A}$ is $BM(0, \hat{\lambda}_J)$.

Define $\hat{Q}^r = \frac{1}{\sqrt{N^r}}Q_J^r$. Then, by corollary (4.1.3) we have that $\hat{Q}^r \Rightarrow [X]^+(t)$.

The convergence of V^r and $\hat{Q}^r$ does not imply the joint convergence of $(V^r, \hat{Q}^r)$. However, following [67], we claim that the component wise convergence is enough for our purposes.

By Theorem 11.6.7 in [64], 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 we have that there exists a convergent subsequence $\{r_k\}$ for which

$$(V^{r_k}, \hat{Q}^{r_k}) \Longrightarrow (\hat{V}, \hat{Q}), \tag{36}$$

for some process $(\hat{V}, \hat{Q})$.

Define $U^r(t) = \sqrt{N^{r_k}}(\frac{D_J^{r_k}(t)}{N^{r_k}} - a_J\mu t)$. Then, using the relation

$$Q_J^{r_k}(t) = Q_J^{r_k}(0) + A_J^{r_k}(t) - D_J^{r_k}(t), \tag{37}$$

or alternatively

$$U^{r_k}(t) = V^{r_k}(t) + Q_J^{r_k}(0) - Q_J^{r_k}(t), \tag{38}$$

and applying the continuous mapping theorem we have the convergence

$$(U^{r_k}, V^{r_k}) \Rightarrow (\hat{U}, \hat{V}), \tag{39}$$

where $\hat{U} = \hat{V} - \hat{Q}$.

Since U and V are continuous with $U(0) = 0$ we can apply the corollary of [49] to obtain for the subsequence

$$\sqrt{N^{r_k}} W^{r_k}(t) \Rightarrow W(t) \tag{40}$$

where $W(t) = \frac{\hat{Q}(t)}{a_J\mu}$.

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

4.2 Steady State Analysis

4.2.1 Stability Conditions

First we should 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 [55]. However, as we have shown in Section 2.1, these formulae are very complicated for calculation even for two classes. Therefore we find the following theorem useful. In the following theorem we use the notation $\lambda^r_{J^c}$ for the arrival rate of the "super class" S consisting of classes $1, ..., J-1$, i.e $\lambda^r_{J^c} = \sum_{i=1}^{J-1} \lambda_J$. Also, we denote by δ^r the probability of abandonment given wait ($P^r\{ab|W^r > 0\}$) in an $M/M/1+M$ system with arrival rate $\lambda^r_{J^c}$, 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. $\rho^r_{C,<J} = \frac{\lambda^r_{J^c}}{(N^r-K^r)\mu}$.

For the second part of the Theorem 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

$$\frac{\lambda^r}{R^r - K^r} \to a. \tag{41}$$

We say that a system is stable is there exists a unique stationary distribution.

Under these notations we have the following:

Theorem 4.2.1 *Under assumption (13) 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 $\lambda^r_J > (N^r - K^r)\mu - \lambda^r_{J^c} \cdot \delta^r$.*
2. *Assume that $N^r = R^r + \Delta^r$ where $\Delta^r = o(R^r)$. Also, assume (41). 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 calculated using existing software such as [73].

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 [52]).

First, the case where $K^r \equiv 0$ is clear since this is a 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 (4.1.1). It is clear that if the $M/M/N^r - K^r$ is stable than so is the threshold system which, by the construction in Theorem 4.1.1, is path wise dominated by the $M/M/N^r - K^r$ system.

For the necessary conditions we build a static priority system with abandonments and show that if it is not stable then the corresponding $M/M/\{K_i\}$ system is not stable. 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, the J^{th} class has infinite patience. Note the following: If we assume that none of the customers of priorities $1, ..., J-1$ waits before entering service (i.e. there is infinite capacity for all priorities except for J) then the system we would have is equal in law to system S. We 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 [10]).

Then, 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.

Define by 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.

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

$$\lambda_J^r + \lambda_{J^c}^r(1-\delta^r) \leq (N^r - K^r)\mu \tag{42}$$

Assume now that $K^r = o(N^r)$. Then, by (13) we have that there exists r_1 such that $\rho_{C,<J}^r < 1$ for all $r > r_1$. Then, using the identity $\lambda_{J^c}^r P\{ab\} = \mu E[Q_{<J}^r]$ (where $Q_{<J}^r$ stands for the steady state queue length of the super class $1, ..., J-1$), we have that

$$\delta^r = \frac{\mu}{\lambda} E[Q_{<J}^r | Z^r > N^r - K^r] \tag{43}$$

But

$$E[Q_{<J}^r | Z^r > N^r - K^r] \leq \frac{(\rho_{C,<J}^r)^2}{1-\rho_{C,<J}^r}. \tag{44}$$

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

$$\delta^r \leq \frac{\rho_{C,<J}^r}{(N^r - K^r)(1-\rho_{C,<J}^r)} \tag{45}$$

This expression converges to zero as fast as $1/N^r$ by assumptions (13), (41 and assuming that $K^r = o(N^r)$. Plugging this upper bound into (42) results in the necessary condition: $K^r \leq \Delta^r + O(1)$.

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

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

Now, consider two cases:

Case 1: $\lambda_{J^c}^r/(N^r - K^r)\mu \to \gamma > 1$. In this case, δ^r converges asymptotically to $1 - \frac{1}{\rho_{C,<J}}$ where $\rho_{C,<J} = \lim_{r\to\infty} \rho_{C,<J}^r$ (see for example [69]).

By our assumption that $K^r \neq o(N^r)$, there exists a subsequence r_{k_j} and $0 < c < 1$ such that $\lim_{r_{k_j}\to\infty} \frac{(N^{r_{k_j}} - K^{r_{k_j}})}{N^r} = c$.

On the subsequence r_{k_j} we have that

$$\lim_{j\to\infty}\frac{1}{N^{r_{k_j}}}(\lambda_J^{r_{k_j}}+\lambda_{J^c}^{r_{k_j}}(1-\delta^{r_{k_j}})\leq\lim_{r_{k_j}\to\infty}\frac{(N^{r_{k_j}}-K^{r_{k_j}})\mu}{N^{r_{k_j}}} \tag{46}$$

On this subsequence the limiting equation is

$$\hat{\lambda}_J+c\mu\leq c\mu \tag{47}$$

Which is a contradiction to our non-negligibility of class J assumption (13).

Case 2: $\lambda_{J^c}^r/(N^r-K^r)\mu\to\gamma\leq 1$ By [69] 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 (42) can be written as

$$\lambda_J^r+\lambda_{J^c}^r-o(\lambda_{J^c}^r)\leq(N^r-K^r)\mu \tag{48}$$

or after dividing by μ this can be written as

$$K^r\leq\Delta^r+o(R^r) \tag{49}$$

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

4.2.2 Convergence of Steady State Distributions

Define

$$S^r=\frac{Y^r(\infty)-(N^r-K^r)}{\sqrt{N^r}}=X^r(\infty) \tag{50}$$

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

We should expect that the limiting distribution of the diffusion process X in Theorem 4.1.1 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

$$P\{X(\infty)\leq x\}\stackrel{\triangle}{=}\lim_{t\to\infty}\lim_{r\to\infty}P\{X^r(t)\leq x\}=\lim_{r\to\infty}\lim_{t\to\infty}P\{X^r(t)\leq x\} \tag{51}$$

We will show this in the following theorem.

Theorem 4.2.2 *Under the notation above and assuming*

$$\lim_{r\to\infty}\sqrt{N^r}(1-\rho_C^r)=\beta\,,\quad 0<\beta<\infty, \tag{52}$$

the following is true:

$$S^r\Rightarrow X(\infty). \tag{53}$$

where X is the limit process from Theorem 4.1.1 with steady state as given in 4.1.2.

Proof: Note that $Y^r(\infty)$ exists as a proper random variable according to Theorem 4.2.1 and under our choice of the parameters. Following the proof of Theorem 4 in [30] all we have to prove is the tightness of the sequence S^r. Recall systems (B) and (C) from the proof of Theorem 4.1.1. Then, since S^r and B^r have the same law, it is enough to prove the tightness of the sequence S^r_B. In addition we create another coupling of X^r with a $M/M/N^r$ system (denoted by D) and for which we define:

$$X^r_D(t) = \frac{Y^r_D(t) - (N^r - K^r)}{\sqrt{N^r}} \tag{54}$$

Construct system (D) 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)) systems) 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

$$X^r_D(t) = \frac{Y^r_D(t) - (N^r - K^r)}{\sqrt{N^r}} \tag{55}$$

Clearly, by the same coupling arguments as in the proof of Theorem 4.1.1 we have path-wise domination $X^r_D(t) \leq X^r_C(t)$. And on the whole we have the path wise ordering

$$X^r_D(t) \leq X^r_B(t) \leq X^r_C(t) \ \forall t \geq 0 \tag{56}$$

Define $S^r_C = X^r_C(\infty)$ where X^r_C is as defined in (25) and S^r_D in the same way for the $M/M/N$ system constructed above. We will compare the stationary threshold system with threshold K^r to both single class multi server stationary systems.

Since the constructed coupling preserves (56) for every finite t it does so also for $t \to \infty$. Since under the conditions of the theorem both sequences S^r_C and S^r_D converge, they are tight. The tightness of S^r_C implies that

$$\forall \epsilon > 0 \, \exists n_1 : P\{S^r_C \in [-n_1, n_1]\} > 1 - \frac{\epsilon}{2} \tag{57}$$

The tightness of S^r_D implies that

$$\forall \epsilon > 0 \, \exists n_2 : P\{S^r_D \in [-n_2, n_2]\} > 1 - \frac{\epsilon}{2} \tag{58}$$

and by the ordering (56) we have that

$$\forall \epsilon > 0 \, \exists \, n_1, n_2 : P\{S^r \in [-n_2, n_1]\} > 1 - \epsilon \tag{59}$$

With the tightness of $S^r = X^r(\infty)$ we have actually established the theorem.

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 Theorem 4.1.1 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 4.2.3 *Under (13) 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. 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 and that $X^r(t)$ converges to the stationary distribution of X.

First let us assume 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 in (4.1.1) there is no limit for $X^{r_k}(t)$ 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

$$m(x) = \begin{cases} 0 & x \geq 0 \\ -\mu x & x < 0 \end{cases} \tag{60}$$

See for example Theorem 4.2 of [43]. This is clearly a non-stationary process and this is again a contradiction to the assumption on the convergence of S^r.

■

Corollary 4.2.4

$$\sqrt{N^r} W_J^r(\infty) \Rightarrow W_J, \tag{61}$$

where

$$W_J \sim \begin{cases} \exp(a_J \mu \beta) & w.p.\, \alpha(\beta) \\ 0 & otherwise \end{cases} \tag{62}$$

Proof. Having the convergence of $X^r(\infty)$ we can repeat the proof of (34) with $Q^r(0) = Q^r(\infty)$ to obtain the desired result. ■

Proposition 4.2.1 ***Halfin-Whitt Analog***

Consider a sequence of $M/M/\{K_i\}$ 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 (13 holds. Then,

$$P\{W_J^r(\infty) > 0\} \to \alpha_J, \quad 0 < \alpha_J < 1, \tag{63}$$

iff

$$\sqrt{N^r}(1 - \rho_C^r) \to \beta\,,\ 0 < \beta < \infty, \tag{64}$$

where $\lambda^r = \sum_{i=1}^{J} \lambda_i^r$ and $\rho_C^r = \dfrac{\lambda^r}{(N^r - K^r)\mu}$,

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 path wise dominated from above by an $M/M/N^r - K^r$ system we have that, if $\beta = \infty$ then $P\{W_J^r > 0\} \to 0$.

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

$$\alpha < \alpha(\beta') < 1. \tag{65}$$

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 $M^r = N^r + \beta'\sqrt{N^r}$ and we will have that $\sqrt{M^r}(1 - \rho_C^r) \to \beta'$. For the new system the 'if' direction applies and hence we will have the inequality (65). 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 $P\{W_J^r > 0\}$ 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 4.2.2 *For every $r > 0$ such that $\rho_C^r < 1$.*

$$1 \leq \frac{P\{W_i^r(\infty) > 0\}}{P\{W_J^r(\infty) > 0\} \cdot \prod_{j=i}^{J-1} (\rho_{\leq j}^r)^{K_{j+1}^r - K_j^r}} \leq \left(\frac{N^r}{N^r - K^r}\right)^{K^r}, \tag{66}$$

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

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

$$P\{W_i^r(\infty) > 0\} \sim \alpha(\beta) \cdot \prod_{j=i}^{J-1} (\rho_{\leq j}^r)^{K_{j+1}^r - K_j^r} \tag{67}$$

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

Remarks:

- In the case of $K^r = \Theta(\sqrt{N^r})$ the right bound converges by simple calculus to e^{d^2} where $d = \lim_{r\to\infty} \frac{K^r}{\sqrt{N^r}}$
- Note that the above implies that for thresholds that are of the form $c \log N$ the probability of delay is asymptotically polynomial, i.e. it is of the form $d\frac{1}{N^\gamma}$ where $d,\ \gamma > 0$.

Proof. For the two-class case this can be proved by direct approximations of the formulae in [55]. 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 $N^r - K_{j+1}^r$ servers are busy. Define a new process $\tilde{Y}^r = \{\tilde{Z}_j^r, \tilde{Q}_1^r, ..., \tilde{Q}_j^r\}$, where $\tilde{Z}_j^r$ describes the number of busy servers above the level of $N^r - K_{j+1}^r$, and $\tilde{Q}_i^r$ is the number of class i customers in queue. Under our restriction $\tilde{Y}^r$ is also a Markov process. Denote it's steady

state by $\tilde{Y}^r(\infty) = \{\tilde{Z}_j^r(\infty), \tilde{Q}_1^r(\infty), ..., \tilde{Q}_j^r(\infty)\}$. Also, because of the model's structure, the probability we are looking for can be calculated by

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

To justify this, see, for example, Section 10.4 of [47] 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 $s \in N-K, ...$:

$$\begin{aligned} &\pi_s \textstyle\sum_{i=1}^{j} \lambda_i \geq \pi_{s+1}(N - K_{j+1})\mu \geq \pi_{s+1}(N^r - K^r)\mu \\ &\pi_s \textstyle\sum_{i=1}^{j} \lambda_i \leq \pi_{s+1} N\mu \end{aligned} \tag{68}$$

or alternatively

$$\begin{aligned} &P\{\tilde{Z}_j^r(\infty) + \textstyle\sum_{i=1}^{j} Q_i^r(\infty) \geq K_{j+1}\} \leq \left(\frac{\sum_{i=1}^{j}\lambda_i}{(N-K)\mu}\right)^{K_j - K_{j+1}} \\ &P\{\tilde{Z}_j^r(\infty) + \textstyle\sum_{i=1}^{j} \tilde{Q}_i^r(\infty) \geq K_{j+1}\} \geq \left(\frac{\sum_{i=1}^{j}\lambda_i}{N\mu}\right)^{K_j - K_{j+1}} \end{aligned} \tag{69}$$

By induction we have proved the desired result. By simple Taylor expansion the upper bound in (66) converges to 1 if and only if K^r is $o(\sqrt{N^r})$. ■

In proposition 4.2.4 we have shown the convergence of $\sqrt{N^r}W_J^r(\infty)$ to a mixture of an exponential r.v. and a point mass in the origin. Equivalently we could say that the waiting time of class J is $\Theta(1/\sqrt{N^r})$. In the next proposition we show that given wait, the waiting time of all the other classes ($\{W_i(\infty)|W_i(\infty) > 0\}$, $i = 1, ..., J-1$) are $O(1/N^r)$. Furthermore, we give expressions for the Laplace transforms and moments for limits of $W_i(\infty)|W_i(\infty) > 0$ for all $i = 1, ..., J-1$.

Proposition 4.2.3 *Assume (13), then, for all* $i = 1, ..., J-1$

$$N^r \cdot [W_i^r | W_i^r > 0] \Rightarrow [W_i | W_i > 0] \tag{70}$$

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

$$\frac{\mu(1-\sigma_i)(1-\tilde{\gamma}(s))}{s - \hat{\lambda}_i + \hat{\lambda}_i\tilde{\gamma}(s)} \tag{71}$$

where $\sigma_i = \lim_{r\to\infty} \sum_{j=1}^{i} \frac{\lambda_i^r}{N^r\mu}$, *and*

$$\tilde{\gamma}_i(s) = \frac{s+\mu}{2b_i\mu} + \frac{1}{2} - \sqrt{\left(\frac{s+\mu}{2b_i\mu} + \frac{1}{2}\right)^2 - \frac{1}{b_i}} \tag{72}$$

where $b_i = \lim_{r\to\infty} \dfrac{\sum_{j=1}^{i-1} \lambda_j^r}{N^r}$
Also,

$$\begin{aligned} &N^r E[W_i^r | W_i^r > 0] \to [\mu(1-\sigma_i)(1-\sigma_{i-1})]^{-1} \\ &(N^r)^2 E[(W_i^r)^2 | W_i^r > 0] \to 2(1-\sigma_i\sigma_{i-1}) \left[(\mu)^2(1-\sigma_i)^2(1-\sigma_{i-1})^3\right]^{-1} \end{aligned} \tag{73}$$

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 [55] and [39].

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

$$\tilde{\gamma}^r(s) = \frac{N^r\mu + s + \lambda_i^- - \sqrt{N^r\mu + s + \lambda_i^- - 4\lambda_i^- N^r\mu}}{s\lambda_i^-} \tag{74}$$

By simple algebra we can prove that

$$\tilde{\gamma}_i^r(s) \to \tilde{\gamma}(s) \tag{75}$$

Where $\tilde{\gamma}_i(s) = \lim_{r\to\infty} \tilde{\gamma}_i^r(s)$ and $\tilde{\gamma}_i(s)$ is given by (72).

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 [55], 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 G is denoted in [55] by $B_i^*(s)$, and it's 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 [55] we have that

$$\phi_i^r(s) = \frac{1 - B_i^*(s)}{(s - \lambda_i^r + \lambda_i^r B_i^*(s))} \frac{1 - \lambda_i E[B_i]}{E[B_i]} \tag{76}$$

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

$$B_i^*(N^r s) \to \tilde{\gamma}_i(s), \tag{77}$$

and the convergence of the moments follows. Hence:

$$N^r E[B_i^*] \to \frac{1}{\mu(1-\sigma_{i-1})} \tag{78}$$

Now, by simple calculus, and since by (13) $\sigma_i < 1$ we have that

$$\phi_i^r(N^r s) \to \frac{\mu(1-\sigma_i)(1-\tilde{\gamma}_i(s))}{s - \hat{\lambda}_i + \hat{\lambda}_i \tilde{\gamma}_i(s)} \tag{79}$$

The limiting transform is similar to the one obtained for the static priority case. The pre-limit moments for the static priority case are given in [39] and their limits are easily calculated.

Corollary 4.2.5 *for $i = 1, ..., J-1$ we have that $E[Q_i^r | Q_i^r > 0] = O(\lambda_i^r / N^r)$. In particular,*

$$E[Q_i^r | Q_i^r > 0] = \lambda_i^r E[W_i^r | W_i^r > 0] \to \hat{\lambda}_i \left[\mu(1-\sigma_i)(1-\sigma_{i-1})\right]^{-1},$$

and (80)

$$E[Q_i^r] \sim \tfrac{\lambda_i^r}{N^r} P\{W_i^r > 0\} \left[\mu(1-\sigma_i)(1-\sigma_{i-1})\right]^{-1}$$

where, as before, $\hat{\lambda}_i = \lim_{r\to\infty} \lambda_i^r / N^r$

Proof: This is a direct application of Theorem (4.2.3) using Little's Law.

■

We would like to conclude this section with a theorem that summarizes important result proved throughout the section.

Theorem 4.2.6 *Consider a sequence of $M/M/\{K_i\}$ 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 (13) holds. Let $\rho_C^r = \frac{\lambda^r}{(N^r - K^r)\mu}$. Then,*

$$P\{W_J^r(\infty) > 0\} \to \alpha_J, \quad 0 < \alpha_J < 1, \tag{81}$$

iff

$$\sqrt{N^r}(1-\rho_C^r) \to \beta\,,\ 0<\beta<\infty, \tag{82}$$

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.*

Corollary 4.2.7 *Let* $\rho^r = \frac{\lambda^r}{N^r\mu}$. *If in addition to the conditions of Theorem 4.2.6,* $K^r = o(\sqrt{N^r})$, *then*

$$P\{W_J^r(\infty)>0\} \to \alpha_J, \quad 0<\alpha_J<1, \tag{83}$$

iff

$$\sqrt{N^r}(1-\rho^r) \to \beta\,,\ 0<\beta<\infty, \tag{84}$$

Moreover,

$$P\{W_i^r(\infty)>0\} \sim \alpha(\beta)\cdot\prod_{j=i}^{J-1}(\rho_{\le j}^r)^{K_{j+1}^r-K_j^r} \tag{85}$$

and if, in addition, all classes are non-negligible, i.e. $\lambda_j^r/\lambda^r \to a_j > 0,\ j=1,...,J$, *then*

$$P\{W_i^r(\infty)>0\} \to \alpha_i,\ 0<\alpha_i<1,\ i=1,...,J-1, \tag{86}$$

iff

$$K_{j+1}^r - K_j^r \to \frac{\ln \alpha_j/\alpha_{j+1}}{\ln \rho_{\le j}}, \quad \forall j=2,\dots,J, \tag{87}$$

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

5 Efficiency Driven $M/M/\{K_i\}$

Analogously to the characterization of the QED regime given in the introduction, we can characterize the Efficiency Driven (ED) regime as follows:

Consider a sequence of N-server queues, indexed by $r = 1, 2, \ldots$. Define the *offered load* by $R = \frac{\lambda^r}{\mu}$, where λ^r is the arrival-rate and μ the service-rate. 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 (14), we define the ED regime for a sequence of $M/M/\{K_i\}$ queues as follows:

$$\exists 0 < \beta < \infty\,, \quad 1 \geq \delta > 1/2 \,:\, \lim_{r\to\infty} (N^r)^\delta(1-\rho_C^r) \to \beta, \quad 0 < \beta < \infty \tag{88}$$

For purposes of optimization we will need to adapt some of the results of the previous sections to the case of the ED $M/M/\{K_i\}$ model.

As before we assume (13), i.e. that class J is non-negligible.

5.1 Diffusion Limits

Since by [30] the probability of delay in this regime converges to 1, we expect that the diffusion limits will be reflected brownian motions as is the case with the conventional heavy traffic for multi-server queues.

However, to differ 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 Theorem 4.1.1.

The ED limits for a sequence of $M/M/N$ queues where not proved for a general $\delta > 1/2$, in the appendix we adapt methods that were used in [26], to prove the desired results. In particular we prove the following:

Proposition 5.1.1 *Consider a sequence of $M/M/N$ system indexed by $N = 1, 2, ...$, such that*

$$N^\delta(1-\rho^N) \to 0 < \beta < \infty, \tag{89}$$

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,

$$X^N(t) \Rightarrow RBM(-\beta\mu, 2\mu) \tag{90}$$

We omit the proofs of the following theorems. Having the convergence of an ED sequence of $M/M/N$ queues, the proofs for the $M/M/\{K_i\}$ model are the same as for the QED case.

The following theorem summarizes the diffusion limit results for the ED $M/M/\{K_i\}$.

Theorem 5.1.1 *Define*

$$X^r(t) = \frac{Y^r((N^r)^{2\delta-1}t) - (N^r - K^r)}{(N^r)^\delta}. \tag{91}$$

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

$$\lim_{r\to\infty} (N^r)^\delta (1 - \rho_C^r) \to \beta, \quad 0 < \beta < \infty. \tag{92}$$

and $X^r(0) \Rightarrow X(0)$, $X(0) \geq 0$. Then,

$$X^r \Rightarrow X, \tag{93}$$

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

Also:

$$\frac{1}{N^\delta} Q_i^r((N^r)^{2\delta-1}t) \Rightarrow 0,\ i = 1, ..., J-1. \tag{94}$$

Remark: 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 influence the result.

5.2 Steady State

In the following theorem we adapt the steady state results of the previous section to the ED case. 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.

Theorem 5.2.1 *Assume that there exists* $1 \geq \delta > 1/2$ *such that*

$$(N^r)^{\delta}(1-\rho_C^r) \to \beta, \quad 0 < \beta < \infty. \tag{95}$$

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

$$N^{-\delta}Y^r(\infty) \Rightarrow X(\infty), \tag{96}$$

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

$$P\{W_J^r(\infty) > 0\} \to 1, \tag{97}$$

and, for every $r > 0$

$$P\{W_i^r(\infty) > 0\} \sim \prod_{j=i}^{J-1} (\rho_j^r)^{K_{j+1}^r - K_j^r}, \tag{98}$$

$$\begin{array}{l} (N^r)^{-\delta}Q_i^r(\infty) \Rightarrow 0, i = 1, ..., J-1; \\ (N^r)^{-\delta}Q_J^r(\infty) \Rightarrow X^+(\infty). \end{array} \tag{99}$$

Remark: 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 $Y^r(\infty)$. 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.

■

6 Asymptotic Optimality

6.1 Definition

In this section we consider the solution for (10) and (11) under condition (13). As in [11], for a meaningful form of asymptotic optimality one needs to compare normalized staffing costs which measure the difference between the actual staffing costs and a base cost of order λ^r, which is a lower bound of the staffing cost.

First, following [2] we will define asymptotic optimality. Let $\bar{K}^r = \{K_1^r, ..., K_J^r\}$ and $\bar{\lambda}^r = \{\lambda_1^r, ..., \lambda_J^r\}$ the vectors of the thresholds and arrival rates in the r^{th} system. Also, let $\underline{N^r}$ be the minimal number of servers required in the r^{th} system to ensure stability (i.e. $\underline{N^r} = \lceil \lambda^r/\mu \rceil$). Let $\underline{C^r}$ be the staffing cost when using $\underline{N^r}$ servers.

Let $C^r(N^r, \pi^r)$ be the cost function in the r^{th} system when the system is equipped with N^r servers and controlled by, π^r.

Definition: The sequence $\{N^r, \pi^r\}$ is **asymptotically optimal** with respect to $\bar{\lambda}^r$ if, when used for the system, the following two conditions apply:

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

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

6.2 Constraint Satisfaction

We will now turn to the solution of (10). Here the cost function reduces to the staffing costs, i.e $C^r(N^r, \pi^r) = N^r$. The results that follow are direct consequences of [11] for the single class case of $M/M/N$. The original work done in [11] in the context of *Constraint Satisfaction* covers general constraints on the waiting costs. Since we have a very simple constraint on the probability of delay (i.e. $P\{Wait > 0\} \leq \alpha\}$) we can establish a simpler property of optimality than the one stated in [11].

Proposition 6.2.1 $M/M/N$ ***Staffing.*** *Consider an $M/M/N$ system with arrival rate λ^r and fixed service rate μ. We are interested in finding*

$$(N^*)^r := \min\{N : P\{W^r > 0\} \leq \alpha\} \tag{100}$$

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 in the sense of the previous definition, where β is such that $P(\beta) = \alpha$. $P(\cdot)$ is the Halfin-Whitt function

$$\alpha(\beta) = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}.$$

Proof: Note that [30] and the monotonicity of the function $P(\cdot)$ imply that the asymptotically feasible region is the following:

$$P^r\{W^r > 0\} \to \alpha', 0 \le \alpha' \le \alpha, \tag{101}$$

iff

$$\sqrt{N^r}(1 - \rho^r) \to \bar{\beta},\ 0 < \beta \le \bar{\beta} < \infty \tag{102}$$

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 $r \ge r_0$, $P^r\{W^r > 0\} > \alpha$. Hence, we would necessarily have that for all $r > r_0$, $(N^r)^* \ge \lambda^r/\mu + (\beta - \epsilon)\sqrt{\lambda^r/\mu}$. Therefore, for each $\beta > \epsilon > 0$ we have that

$$\liminf_{r\to\infty} \frac{\beta\sqrt{\lambda^r/\mu}}{(\beta - \epsilon)\sqrt{\lambda^r/\mu}} \ge \liminf_{r\to\infty} \frac{\beta\sqrt{\lambda^r/\mu}}{(N^r)^* - \underline{N^r}} \ge 1 \tag{103}$$

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

Having Proposition 6.2.1 we can proceed to defining the asymptotically optimal solution for (10).

Theorem 6.2.1 *For (10) and under condition (13), the following combined staffing and routing policy is asymptotically optimal:*
Assume (without loss of generality) that for all $i \ne j$, $\alpha_i \ne \alpha_j$ (otherwise we can merge classes i and j for which $\alpha_i = \alpha_j$) and that the classes are ordered in increasing order of α_i. Choose $N^r = R^r + P^{-1}(\alpha_J)\sqrt{R^r}$ where $R^r = \sum_{i=1}^{J} \lambda_i^r/\mu$, and let $P(\beta)$ be the Halfin-Whitt delay function given by $P(\beta) = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}$. Route according to threshold priorities with the threshold determined by the following recursive relation:

$$\begin{aligned} K_i^r - K_{i-1}^r &= \left\lceil \frac{\ln\alpha_{i-1} - \ln\alpha_i}{\ln\rho^r_{\le i-1}} \right\rceil \quad i = 2, \ldots, J \\ K_1^r &\equiv 0 \end{aligned} \tag{104}$$

Remark: Note that our assumption that $\alpha_J < 1$ rules out the case of efficiency driven staffing. It appears that better non-asymptotic results are required to handle the case of $\alpha_J \approx 1$.

Proof. Define

$$M^r = \{\min N : P\{W_i^r > 0\} \leq \alpha_J, \quad \forall i = 1, ..., J\}. \tag{105}$$

If we denote by $(N^*)^r$ the optimal solution to (10) then clearly $M \leq N^{r^*}$. Now, (105) is equivalent to a single class $M/M/N$ constrained staffing problem. For this problem the asymptotically optimal staffing is given by Proposition 6.2.1 and it equals $\lambda^r/\mu + P^{-1}(\alpha_J)\sqrt{\lambda^r/\mu}$.

For α_i^r, $i = 1, ..., J-1$, that decrease polynomially with r we have by Proposition 4.2.2 that α_i^r, $i = 1, ..., J-1$, are achieved by logarithmic thresholds. Proposition 4.2.1 guarantees that staffing the system with M servers and using logarithmic thresholds asymptotically achieves α_J. Hence, the lower bound is asymptotically achieved. ■

6.3 Cost Minimization

Before presenting the solution to (11) it is necessary to adapt an important theorem from [11] to our setting. In [11], 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 [11], but rather their conclusions with respect to the *ED* and *QED* regimes.

Theorem 6.3.1 *(Theorems 6.1 and 7.1: Borst, Mandelbaum & Reiman 2002) Consider a sequence of $M/M/N$ systems, indexed by $r = 1, 2, ...$, with arrival rate λ^r and fixed service rate μ. Assume that an agent's salary is a function of r given by s^r. A customer waiting one unit of time incurs a cost of c^r. We are interested in finding*

$$(N^r)^* := \arg \min_{N > \lambda^r/\mu} \{s^r N^r + c^r E[W^r]\} \tag{106}$$

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*

$$(y^r)^*(c) \equiv y^*(c) = arg \min_{y>0} \left\{ y + \frac{cP(y)}{y} \right\} \tag{107}$$

approximations for the staffing function $y^*(\cdot)$ are given in [11].

- *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*

$$(y^r)^*(c) = arg\min_{y>0}\left\{y + \frac{c}{y}J^r\right\} = \sqrt{cJ^r}. \tag{108}$$

The following Theorem deals with the cost minimization problem (11. Recall that when we remove the constraints on the probability of delay the remaining problem is a pure cost minimization problem. For this problem the optimal policy, as established in [72], is one with state dependent thresholds. In the following theorem, however, we show that asymptotically the state independent threshold policy is optimal. The intuitive explanation for this phenomenon is what we call Economies of Scale. The state dependence of the optimal policy is aimed at protecting against a situation where lower class queue gets too long and expensive because of the reservation for higher priorities. However, in large systems, we can combine high quality service for the high priorities with very little harm to low priorities.

Theorem 6.3.2 *Consider (11) and assume that the waiting cost coefficients scale with r in a polynomial manner: $c_i^r = d_i \cdot r^{\gamma_i}, \gamma_i \geq 0,\ i = 1, ..., J-1, -1 < \gamma_J \leq 0$. Recall that our assumptions are such that $c_1^r \geq c_2^r \geq ... \geq c_J^r$. Also, $\alpha_1^r \leq \alpha_2^r \leq ... \leq \alpha_J$, where $\alpha_i^r,\ i = 1, ..., J-1$ are allowed to decrease polynomially with r while α_J is fixed.*

Then, the following is asymptotically optimal:

Staff with $N^r = R^r + \beta\sqrt{R^r}$, where

$$\beta = \max\{(y^r)^*(d_J), P^{-1}(\alpha_J)\} \tag{109}$$

Where $(y^r)^*(d_J) = arg\min_{y>0}\left\{y + \frac{d_J P(y)}{y}\right\}$, whenever $\gamma_J = 0$ and $(y^r)^*(d_J) = \sqrt{d_J J^r}$, otherwise.

Route using $M/M/\{K_i\}$ with

$$\begin{aligned} K_i^r - K_{i-1}^r &= \left\lceil \frac{\ln\alpha_{i-1}^* - \ln\alpha_i^*}{\ln\rho_{i-1}^r} \right\rceil \quad i = 2, \ldots, J, \\ K_1^r &\equiv 0, \end{aligned} \tag{110}$$

where for $i = 1, ..., J$

$$\alpha_i^* = \alpha_i \wedge \left(\frac{1}{N^{\gamma_i - 1/2(1+\gamma_J)+\epsilon}}\right), \tag{111}$$

and $\epsilon > 0$ can be arbitrarily small.

Finally, ties are resolved according to the $c\mu$ rule.

Proof:

Step 1 (Lower Bound): Since we have a common μ, the long run average number of customers in queue is minimized, for fixed 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

$$\sum_{i=1}^{J} c_i^r \lambda_i^r E[W_i^r] \geq c_J^r \cdot \sum_{i=1}^{J} \lambda_i^r E[W_i^r] = c_J^r \cdot \sum_{i=1}^{J} E[Q_i^r] \geq c_J^r \cdot E[Q^r]. \tag{112}$$

Where Q^r is the steady state queue length in a $M/M/N^r$ 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

$$\begin{array}{ll} \text{minimize} & c_J^r E[Q^r] + N \\ \text{subject to} & P\{W^r > 0\} \leq \alpha_J \\ & N \in \mathbb{Z} \end{array} \tag{113}$$

Let M_1 be the solution of the unconstrained problem

$$\begin{array}{ll} \text{minimize} & c_J^r E[Q^r] + N \\ & N \in \mathbb{Z} \end{array} \tag{114}$$

Let M_2 be the solution of the constrained staffing problem:

$$\begin{array}{ll} \text{minimize} & N \\ \text{subject to} & P\{W^r > 0\} \leq \alpha_J \\ & N \in \mathbb{Z} \end{array} \tag{115}$$

By [11], the cost function is strictly convex and unimodal and the feasible region for (113) is the interval $[M_2, \infty)$ the solution (M^r) to the above problem (113) will equal $\max\{M_1, M_2\}$.

Now, we have three cases:
Case 1: $\gamma_J = 0 \Rightarrow M^r = \lambda^r/\mu + \beta\sqrt{\lambda^r/\mu}$, where $\beta = \max\{(y^r)^*, P^{-1}(\alpha_J)\}$ (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:

$$\frac{1}{\sqrt{M^r}}\left[C^r(M^r) - \underline{N^r}\right] = \frac{1}{\sqrt{M^r}}\left[c_J E[Q^r] + \beta\sqrt{\lambda^r/\mu}\right] \sim \left[c_J \alpha(\beta)\frac{1}{\beta} + \beta\right] \tag{116}$$

Under the proposed choice of the thresholds we have by propositions 4.2.2 and 4.2.5 that

$$\frac{1}{\sqrt{M^r}} c_i^r E[Q_i^r] \to 0, \tag{117}$$

and

$$\frac{1}{\sqrt{M^r}} E[Q_J^r] \to \alpha(\beta)\frac{1}{\beta}, \tag{118}$$

Hence, we have that the lower bound is achieved. i.e.

$$\lim_{r\to\infty} \frac{\sum_{i=1}^{J} c_i^r E[Q_i^r] + \beta\sqrt{R^r}}{c_J E[Q^r] + \beta\sqrt{R^r}} = 1. \tag{119}$$

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

$$\frac{1}{\sqrt{N^r}} c_i^r E[Q_i^r] \to 0 \tag{120}$$

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 $\lambda \to \infty$. Due the restriction $\gamma_j > -1$, we have by (6.3.1) 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

$$\lim_{r\to\infty} (M^r)^\delta (1-\rho^r) = \lim_{r\to\infty} (M^r)^\delta (1-\rho_C^r) \to, \beta \tag{121}$$

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

By the choice of α_i^* we have that

$$\frac{1}{(M^r)^{1/2(1+\gamma_J)}} c_i^r E[Q_i^r] \Rightarrow 0, \tag{122}$$

and

$$\frac{1}{(M^r)^{1/2(1+\gamma_J)}} c_J^r E[Q_J^r] \Rightarrow d_J \frac{1}{\beta}. \tag{123}$$

Again the lower bound is asymptotically achieved.

■

Note that in this theorem we restricted our attention to $\gamma_J > -1$. This rules out cases for which the solution of the single class dimensioning problem would result in $N^r \approx R^r + b(R^r)$, where $b(\cdot)$ is a sub-polynomial function of R (i.e. $b(x) = o(x^\delta), \forall \delta > 0$.

Corollary 6.3.3 $c\mu$ ***Optimality:*** *Assume for (11) that* $\alpha_i = 1, \forall i = 1, ..., J$*, and that* $c_i^r = c_i, \forall i = 1, ..., J, \forall r \geq 1$*. Then, the* $c\mu$ *rule is asymptotically optimal among all non-preemptive policies (work-conserving and non-work conserving), and the corresponding optimal staffing is given by* $R^r + \beta\sqrt{R^r}$*, where*

$$\beta = arg \min_{y>0} \left\{ y + \frac{cP(y)}{y} \right\}.$$

Proof: In the previous theorem we would have that $\alpha_i* \equiv 1$, in (111), and hence the staffing problem reduces to the single class dimensioning problem, and the routing is static priority.

■

7 Some Extensions

7.1 Adding Abandonment

7.1.1 Model Formulation

The previous sections were aimed at generalizing the dimensioning results of [11] to the multi-class case, by characterizing the asymptotically optimal staffing and control. Moreover, the optimal staffing in the multi-class case was derived directly from the optimal staffing in [11]. Such dimensioning results are still not available for the single class $M/M/N+M$ queue, i.e. a single class queue where all customers have finite exponential patience with rate θ. However, much can be said about the control problem in the multi-class $M/M/N+M$ (now, class i has finite exponential patience with rate θ_i). Furthermore, it turns out that under certain settings the minimization of weighted abandonment costs gives rise to the $M/M/\{K_i\}+M$ model (designating now a multi-class multi-server system with thresholds and impatient customers) as the asymptotically optimal policy in the *QED* regime - this makes this section a natural continuation of the previous sections.

The setting: As before, we consider a multi-class queue with a single type of servers attending all customer-classes. The service times of different customers are i.i.d exponentially distributed random variables with rate μ for all customer classes. As before, class i customers arrive according to a Poisson process with rate λ_i, and we still assume non-negligibility of the low priority class 13). The patience of each customer is defined as the maximal time this customer will wait before abandoning the system. A customer does not leave after her service starts. In our setting, the patience of each customer is an exponentially distributed random variable with rate θ_i, $0<\theta_i<\infty$ for class i, and is independent of all other processes.

We denote by $P_i\{Ab\}$ the steady state probability of abandonment for class i. Note, that in this case the system is always stable due to the impatience of the customers. Also, we assume that an abandonment of a class i customer incurs a cost of c_i, where $c_i, i=1,...,J=1$ are allowed to grow with system size (to be made precise later).

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

$$\begin{array}{ll}\text{minimize} & \sum_{i=1}^{J} c_i\lambda_i P_i\{Ab\}+N \\ & N\in\mathbb{Z}_+\end{array} \tag{124}$$

The pure control problem of minimizing the total number of abandonments (with no cost differentiation) was addressed in [60]. In [60] 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). Moreover, 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, [60] does not say if the optimal policy is idling or non-idling.

Another problem of interest, in analogy to (11), is the problem of constraint satisfaction. i.e. we wish to determine the minimal staffing required to ensure that the probability of abandonment for class i customers does not exceed a certain level α_i. More formally, we consider the following problem:

$$\begin{array}{ll} \text{minimize} & N \\ \text{subject to} & \\ & P_i^{\pi}(Ab) \leq \alpha_i,\ 0 < \alpha_i < 1, \quad i = 1, ..., J, \text{ for some } \pi \in \Pi \\ & N \in \mathbb{Z}_+ \end{array} \tag{125}$$

Where, as before, Π is the set of all non-preemptive non-anticipative scheduling policies. We have shown in Section 6 that when the constraints on the probability of delay are fixed (i.e. when α_i's in (11) are not allowed to scale with system size) the asymptotically optimal staffing leads to the *QED* regime. However, for (125), fixed α_i's lead to the *ED* regime, which suggests a very simple pool decomposition policy, i.e. decompose the V-Model into J I models, each serving a different class of customers. The real challenge, then, is to solve (125) when the α_i's are allowed to scale with system size. We leave this question open for future research.

As mentioned before, we will prove that under certain settings a threshold policy with fixed thresholds is asymptotically optimal in the *QED* regime. To this end, we would like first to analyze diffusion and steady state limits for the $M/M/\{K_i\} + M$ model.

7.1.2 Diffusion Limits

First we quote Theorem 2 from [27] for a sequence of $M/M/N + M$ queues.
Denote by $\{Y^r(t), t \geq 0\}$ the total number in system in an $M/M/N + M$ system. Let

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

then we have the following:

Theorem 7.1.1 ***[27], Theorem 2*** *Consider a sequence of $M/M/N+M$ queues indexed by the superscript $r = 1, 2, ...$ Let λ^r and N^r be, respectively, the arrival rate and the number of servers in the r^{th} system. The service rate μ and the individual abandonment rate θ are independent of the index r. Let $\rho^r = \lambda^r/(N^r\mu)$.*

Assume that

$$\lim_{r\to\infty} \sqrt{N^r}(1-\rho^r) \to \beta\,,\ -\infty < \beta < \infty \tag{126}$$

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 $\sigma^2 = 2\mu$.

In the next two propositions we will show that the normalized and scaled overall number of customers in systems in the $M/M/\{K_i\}+M$ model converges to the same limit as in 7.1.1, with $\theta = \theta_J$ (which is the impatience rate of the lowest priority).

We consider a sequence of $M/M/\{K_i\}$ systems indexed by $r = 1, 2, ...$ 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 bigger than K_i^r. As before, we use the notation K^r to stand for the threshold of the lowest priority (i.e. $K^r = K_J^r$), and define a"nominal" load: $\rho_C^r = \frac{\lambda^r}{N^r - K^r}$.

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

Proposition 7.1.1 ***State Space Collapse.*** *Assume (13) and*

$$\lim_{r\to\infty} \sqrt{N^r}(1-\rho_C^r) \to \beta,\ -\infty < \beta < \infty. \tag{127}$$

Then, as $r \to \infty$,

$$\begin{aligned} &\tfrac{1}{\sqrt{N^r}} Q_i^r \Rightarrow 0,\ i = 1, ..., J-1, \\ &\tfrac{1}{\sqrt{N^r}}[(N^r - K^r) - Z^r]^- \Rightarrow 0, \\ &\tfrac{1}{N^r}[(N^r - K^r) - Z^r]^+ \Rightarrow 0. \end{aligned} \tag{128}$$

Proof: for the first two limits the proof is omitted since it is similar to the proof in the no-abandonment case. Still, we would like to prove

$$\frac{1}{N^r}[(N^r - K^r) - Z^r]^+ \Rightarrow 0 \tag{129}$$

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 [43] the fluid limit of the bounding process is zero and hence the result. ■

Theorem 7.1.2 *Assume (13) and*

$$\lim_{r\to\infty} \sqrt{N^r}(1 - \rho_C^r) \to \beta, \ -\infty < \beta < \infty. \tag{130}$$

If $X^r(0) \Rightarrow X(0)$*, Then,*

$$X^r(t) = \frac{Y^r(t) - (N^r - K^r)}{\sqrt{N^r}} \Rightarrow X \tag{131}$$

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 $\sigma^2 = 2\mu$*.*

Proof.

In this proof we employ the same approach that was used in [2] 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 4.1.1, 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 at time t. Also, let $Q^r(t)$ be the total number of customers in queue (i.e. $Q^r(t) = Q_1^r(t) + Q_2^r(t)$).

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

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

$$
\begin{aligned}
Y^r(t) &:= Q^r(t) + Z_1^r(t) + Z_2^r(t) \\
&= Y^r(0) + A^r(t) - \mathcal{N}_1(\mu(N-K)) + \mathcal{N}_1\left(\mu\int_0^t I_1^r(s)ds\right) - \mathcal{N}_2\left(\mu\int_0^t Z_2^r(s)ds\right) \\
&- \sum_{l=1}^{2}\mathcal{N}_l^a\left(\theta_l\int_0^t Q_l(s)ds\right)
\end{aligned}
\tag{132}
$$

Where $\mathcal{N}_k,\ k=1,2$ and $\mathcal{N}_l^a,\ l=1,2$ are independent unit Poisson processes, 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:

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

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:

$$
Y^r(t) = Y^r(0) + \lambda^r t - \mu(N-K)t + \mu\int_0^t I_1^r(s)ds - \mu\int_0^t Z_2^r(s)ds - \sum_{l=1}^{2}\theta_l\int_0^t Q_l^r(s)ds + M^r(t),
\tag{133}
$$

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 martingales are $\mathbb{F}^r-$locally square-integrable martingales with respective predictable quadratic variations:

$$
\left\langle M_A^r \right\rangle (t) = \lambda^r t, \tag{134}
$$

$$
\left\langle M_1^r \right\rangle (t) = (N^r - K^r)\mu t, \tag{135}
$$

$$
\left\langle M_{I_1}^r \right\rangle (t) = \mu\int_0^t I_1^r(s)ds \tag{136}
$$

$$
\left\langle M_{Z_2}^r \right\rangle (t) = \mu\int_0^t Z_2^r(s)ds \tag{137}
$$

$$
\left\langle M_{Q_l}^r \right\rangle (t) = \theta_l\int_0^t Q_l^r(s)ds, \quad l=1,2. \tag{138}
$$

we can rewrite (133) also as

$$
\begin{aligned}
Y^r(t) =& Y^r(0) + \lambda^r t - \mu(N-K)t + \mu\int_0^t I_1^r(s)ds - \mu\int_0^t Z_2^r(s)ds - \\
&+ \theta_2\int_0^t Q_1^r(s) + Q_2^r(s) + Z_2^r(s)ds + \int_0^t (\theta_2-\theta_1)Q_1^r(s) + \theta_2 Z_2^r(s)ds + M^r(t),
\end{aligned}
\tag{139}
$$

By definition,

$$
\begin{aligned}
Q_1^r(t) + Q_2^r(t) + Z_2^r(t) &= [Y^r(t) - (N^r - K^r)]^+ \\
I_1^r(t) &= [Y^r(t) - (N^r - K^r)]^-
\end{aligned}
\tag{140}
$$

Also, note that $Z_2^r(t) = [N^r - K^r - Z^r]^+$. Hence, by (7.1.1),

$$
\begin{aligned}
\frac{1}{\sqrt{N^r}} Q_1^r &\Rightarrow 0 \\
\frac{1}{\sqrt{N^r}} Z_2^r &\Rightarrow 0
\end{aligned}
\tag{141}
$$

After normalization and scaling we have that

$$
\begin{aligned}
X^r(t) = X^r(0) - \beta\mu t + \mu\int_0^t [X^r(s)]^- ds + \theta_2\int_0^t [X^r(s)]^+ ds \\
+ \epsilon^r(t) + \frac{M^r(t)}{\sqrt{N^r}} + o(1),
\end{aligned}
\tag{142}
$$

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

$$
\begin{aligned}
\left\{M_A^r/\sqrt{N^r}, M_1^r/\sqrt{N^r}, M_{I_1}^r/\sqrt{N^r}, M_{Z_2}^r/\sqrt{N^r}, M_{Q_1}^r/\sqrt{N^r}, M_{Q_2}^r/\sqrt{N^r}\right\} \\
\Rightarrow \{\sqrt{\mu}b_a, \sqrt{\mu}b_1, 0, 0, 0, 0\},
\end{aligned}
\tag{143}
$$

where b_a and b_1 are independent standard Brownian motions. Hence, by the continuous mapping theorem we would have that $M^r/\sqrt{N^r}$ converges to $\sqrt{\mu}b_a - \sqrt{\mu}b_1$. This 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 prove (143). First note that by the Functional Law of Large Numbers (FLLN), as $r \to \infty$,

$$\left\langle \frac{M_A^r}{\sqrt{N^r}} \right\rangle (t) \Rightarrow \mu t, \tag{144}$$

$$\left\langle \frac{M_1^r}{\sqrt{N^r}} \right\rangle (t) \Rightarrow \mu t, \tag{145}$$

By Proposition, 7.1.1 we have that

$$\left\langle \frac{1}{\sqrt{N^r}} M_{I_1}^r \right\rangle (t) \Rightarrow \mu t, \tag{146}$$

$$\left\langle \frac{1}{\sqrt{N^r}} M_{Z_2}^r \right\rangle (t) \Rightarrow 0 \tag{147}$$

$$\left\langle \frac{1}{\sqrt{N^r}} M_{Q_l}^r \right\rangle (t) \Rightarrow 0, \quad l = 1, 2. \tag{148}$$

The independence of M_A^r and M_1^r together with the inequality $\langle M, N \rangle \leq \sqrt{\langle M \rangle \langle N \rangle}$ imply that all covariations converge to zero.

Also, note that since the jumps of all the above martingales are bounded by 1 we have also that for each $T > 0$,

$$\lim_{r \to \infty} E \left[\sup_{t \leq T} \left| \frac{1}{N^r} M^r(t) - \frac{1}{N^r} M^r(t-) \right| \right] = 0 \tag{149}$$

Hence, we can apply Theorem 7.1.4 from [19] to obtain the result. To prove the result for an arbitrary number of classes it is enough to re-build the decomposition of Y^r. The rest readily follows.

■

7.1.3 Steady State

By [27], the process X defined in Theorem 7.1.2 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

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

Theorem 7.1.3 *Assume (13) and*

$$\lim_{r\to\infty} \sqrt{N^r}(1 - \rho_C^r) \to \beta, \; -\infty < \beta < \infty. \tag{150}$$

Then

$$X^r(\infty) \Rightarrow X(\infty) \tag{151}$$

where $X^r(\infty)$ and X are as defined in Theorem 7.1.2,

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 4.2.2. 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 a $M/M/(N^r - K^r) + M$ 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 $U^r(t)$ and $L^r(t)$ the total number of customers in systems U^r and L^r respectively. Let Y^r stand for our $M/M/\{K_i\}$ with abandonment system and recall that $Y^r(t)$ stands for the total number of customers in system at time t.

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

$$L(t) \leq_{st} Y(t) \leq_{st} U(t), \; t \geq 0. \tag{152}$$

To show (152), we use sample path coupling. For system U and L and for the $N-K$ pool of system Y, we create the departures from the same Poisson process with thinning, as we did in the proof of 4.1.1. The abandonments for systems Y and U will be also created from the 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 Y, we create the next abandonment from a Poisson process with rate $max\{i\cdot\underline{\theta},\sum_{k=1}^{J}j_k\theta_k\}$. Then, we create an abandonment in system U with probability $\frac{i\underline{\theta}}{max\{i\cdot\underline{\theta},\sum_{k=1}^{J}j_k\theta_k\}}$ and an abandonment in system Y 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 Y with probability 1.

For simplicity, lets start all 3 system 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 Y and U. So, the ordering is still preserved. Now, If there are no customers in the K pool of system Y 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 Y, but this can only preserve the order.

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

By [27] we have the tightness of the sequence $U^r(\infty)$. By [46] we have the tightness of the sequence $L^r(\infty)$.

The rest follows as in the proof of Theorem 4.2.2. ■

Corollary 7.1.4 *Assume (13) and*

$$\lim_{r\to\infty}\sqrt{N^r}(1-\rho_C^r)\to\beta,\ -\infty<\beta<\infty. \tag{153}$$

Then,

$$P\{W_J^r>0\}=P\{Z^r\geq N^r-K^r\}\to w(-\beta,\sqrt{\mu/\theta}) \tag{154}$$

where $w(x,y)$ was defined in Theorem 7.1.2.

The next proposition is analog to Proposition 4.2.2 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 7.1.2 ***(Probability of Delay)*** *For every* $r > 0$

$$\frac{P\{W_i^r(\infty) > 0\}}{P\{W_J^r(\infty) > 0\} \cdot \prod_{k=i}^{J-1} (\rho_k^r)^{K_{k+1}^r - K_k^r}} \leq \left(\frac{N^r}{N^r - K^r} \right)^{K^r} . \tag{155}$$

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

$$P\{W_i^r(\infty) > 0\} = O\left(w(-\beta, \sqrt{\mu/\theta}) \cdot \prod_{k=i}^{J-1} (\rho_k^r)^{K_{k+1}^r - K_k^r} \right) \tag{156}$$

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

Proof:

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

$$P\{W_i \geq 0 | W_{i+1} \geq 0\} \leq \left(\frac{\sum_{j=1}^{i} \lambda_j^r}{(N^r - K^r)\mu} \right)^{K_{i+1} - K_i} \tag{157}$$

The proof is completed as in the case without abandonment.

Corollary 7.1.5 ***Probability of Abandonment***

Denote by $P_k^r\{Ab\}$ *the probability of abandonment for class* k*. Then*

$$\lim_{r \to \infty} \sqrt{N^r} P_k^r\{Ab\} = \Delta_k \, , 0 \leq \Delta_k < \infty \tag{158}$$

where Δ_k *is given by*

$$\Delta_k = \begin{cases} a_k^{-1} [\sqrt{\theta_k/\mu} \cdot h(\beta \sqrt{\mu/\theta_k}) - \beta] \cdot w(-\beta, \sqrt{\mu/\theta_k}) & k = J \\ 0 & Otherwise. \end{cases} \tag{159}$$

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

7.1.4 Asymptotic Optimality - Cost Criterion

In this section, we consider the solution to a specific setting of (124). In the no-abandonment case our optimality results were based heavily on the existing results for the single class case (as given in [11]). However, at this point, there are no such results for the abandonment case. Hence, we limit our discussion to asymptotic optimality of the threshold policy under given staffing levels and do not include staffing recommendations.

Also, we restrict our attention to a certain setting of the problem in which the $c_i \geq c_j$ whenever $\theta_i \geq \theta_j$. i.e. the case where the value the system gives to the different customer classes is proportional to their relative patience.

We restrict ourselves to systems which are in the *QED* regime. Formally we consider systems in which the staffing level of the r^th system, N^r, is such that

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

Actually, the results are applicable also to the Efficiency Driven regime, but we do not discuss this here, since we would need for that purpose analysis of $M/M/\{K_i\}$ model with abandonments in the ED regime. We leave this part for future research. However, we would like to mention that even in the particular case of the Efficiency Driven regime, in which $\rho^r \to c > 1$, we can show that the thresholds have a very positive impact on the abandonment cost.

We now turn the definition of asymptotic optimality in the context of abandonment cost.

Definition: Assume

$$\sqrt{N^r}(1-\rho^r) \to \beta,\ 0 < \beta < \infty. \tag{160}$$

The sequence $\{\pi^r\}$ is asymptotically optimal with respect to $\bar{\lambda}^r$, if for any sequence of policies $\{\hat{\pi}^r\}$ we have that

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

where $C^r(\pi^r, N^r) = \sum_{i=1}^{J} c_i^r \lambda_i^r P_i^{\pi^r}\{Ab\}$

Under this definition, we have the following:

Theorem 7.1.6 *Consider a sequence of multi-class multi-server systems, with the i^{th} class customers have finite patience with rate θ_i.*

Assume that all of the following three conditions holds:

(a) The θ_i's are such that $c_i^r \geq c_j^r$ whenever $\theta_i \geq \theta_j$, where we allow c_i, $i = 1, ..., J-1$ to grow polynomially with r (i.e $c_i^r = c_i \cdot r^{\gamma_i}$, $\gamma_J = 0$).

(b) Condition (13) holds.

(c) The staffing is such that

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

Then, serving the classes in decreasing order of $c_i\theta_i$, and according to threshold priorities asymptotically minimizes C^r. In particular, choose the threshold such that the probability of delay for class i, $i = 1, ..., J-1$ is $\frac{1}{N^{\gamma_i}}$. Or, equivalently, choose

$$K_i^r - K_{i-1}^r = \left\lceil \frac{\ln \alpha_{i-1} - \ln \alpha_i}{\ln \rho_{i-1}^r} \right\rceil \tag{161}$$

where $\alpha_i = N^{-\gamma_i}$.

Proof:

First, we would like to create a lower bound for the overall number of abandonments. We can restrict our attention to preemptive policies. Since all r.v's 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 before.

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

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

Under any non-idling policy, system B behaves (in the sense of the overall abandonment) as a single class $M/M/N + M$.

Now, note that for any non-idling policy, the average length of the excursions 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 N, the probability of starting a positive excursion is the same for both systems.

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

$$P\{Y_A \geq N\} \leq P\{Y_B \geq N\} \tag{162}$$

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

$$E[Z_A|Y_A < N] = E[Z_B|Y_B < N] \tag{163}$$

Hence, we have that

$$\begin{array}{rcl} E[Z_A] & = & E[Z_A|Y_A < N]P\{Y_A < N\} + NP\{Y_A \geq N\} \\ & \leq & E[Z_B|Y_B < N]P\{Y_B < N\} + NP\{Y_B \geq N\} \quad = \quad E[Z_B] \end{array} \tag{164}$$

But, by Little's Law

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

and hence we have that

$$P_A\{Ab\} \geq P_B\{Ab\} \tag{165}$$

So, system B with non-idling policy constitutes a lower bound for our system with respect to the overall probability of abandonment.

Having the lower bound, we can not proceed with the asymptotic optimality.

By our condition on the c_i's and θ_i's and by Proposition 7.1.5, we have that the lower bound which is $c_J \lambda P_B\{Ab\}$ is asymptotically achieved, for any logarithmic threshold level.

Now, if we allow the abandonment costs of high priorities to grow polynomially with r we can still achieve the lower bound by using thresholds such that the probability of delay for class i is an $o(1/N^{\gamma_i - 1/2})$. ■

7.1.5 Asymptotic Optimality - Constraint Satisfaction

In this section, we wish to tackle the question of staffing a large call center when the optimization problem is given by (125). We give here the result for the case in which $0 < \alpha_i < 1$, $i = 1, ..., J$ are fixed (independently of system size). Under this configuration we allow the different classes to have different service rate (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. i.e. decompose the N servers into J groups of servers, such that class i customers will be served only by group i servers. It is, however, interesting to examine the question what is the optimal policy in the general case, in which the α_i's are allowed to vary with the system size. It seems that this more general question requires different tools than those used in this work and hence we leave it for future research.

In the following, we assume, without loss of generality, that the classes are ordered such that $\alpha_i \geq \alpha_k$ whenever $i \geq k$. Also, we still demand that condition (13) holds.

Before proving the optimality of the pool decomposition policy we would like to state the definition of asymptotic optimality in this context. In the non-abandonment case we centered the cost around the lower bound $\lceil \lambda/\mu \rceil$ in order to get a meaningful result. This is clearly not a lower bound in the case with abandonment.

Lemma 7.1.7 *Consider a multi-class system where class i customers have exponential service times and patience with rates μ_i and θ_i, respectively.*

Fix $r > 0$ and denote by $(N^r)^*$ the optimal solution to (125). Then,

$$(N^r)^* \geq \sum_{i=1}^{J} \frac{\lambda_i^r}{\mu_i}(1 - \alpha_i). \tag{166}$$

where, as before, $\lambda^r = \sum_{i=1}^{J} \lambda_i^r$,

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: $P\{Ab\} = \sum_{i=1}^{J} \frac{\lambda_i}{\lambda} P_i\{Ab\}$.

Then, we have:

$$\alpha_i \geq P_i\{Ab\} = \lim_{t\to\infty} \frac{R_i(t)}{A_i(t)} = \lim_{t\to\infty} \frac{A_i(t) - D_i(t) + Q_i(t)}{A_i(t)} = \lim_{t\to\infty} 1 - \frac{\mu_i T_i(t)}{\lambda_i t}, \quad \forall i = 1, ..., J. \tag{167}$$

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 $M(t)$ 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 $\lim_{t\to\infty} \frac{D_i(t)}{A_i(t)} = \lim_{t\to\infty} \frac{\mu_i T_i(t)}{\lambda_i(t)}$.

We can re-write the last equation 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:

Definition: Suppose that the sequence $\{N^r, \pi^r\}^*$ is an optimal solution of (125) with respect to the sequence of arrival rates vector $\bar{\lambda}^r$. Then the sequence $\{N^r, \pi^r\}$ is asymptotically optimal if when used for the system,

- $\limsup_{r\to\infty} P_i^r\{Ab_i\} \leq \alpha_i\,, \forall i = 1, ..., J$, and,
- if we take any other sequence of policies $\{N_2^r, \pi_2^r\}$ then $\liminf_{r\to\infty} \frac{N_2^r - \underline{N^r}}{N^r - \underline{N^r}} \geq 1$,

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

The following is an immediate result of the lower bound.

Proposition 7.1.3 *For (125) the following policy is asymptotically optimal (in the sense of the last definition):*

Partition the server pool into distinct pools of sizes $N_i, i = 1, \ldots J$, *where* $N_i = \lfloor \frac{\lambda_i}{\mu_i}(1-\alpha_i) \rfloor$.

Let class i *customers be served only by the servers of the* i^{th} *pool (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 [69]).

The optimality is immediate since the lower bound is approached from below.

■

Remark: Similar procedures are likely to work for more general network topologies in which the service times are also dependent of the servers (i.e. when class i customers are served by server type j with rate μ_{ij}.

Remark: Performance measures for the proposed policy can be found in Whitt [69].

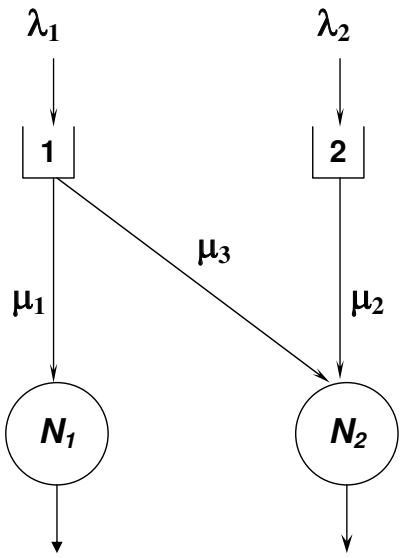


Figure 4: The N Model

7.2 Ongoing Research - An N Model

In the previous sections of this paper we analyzed the V Model which constitutes a single server type and multiple customers classes. Armony and Mandelbaum [2] analyzed the case of multiple server types and a single customer class (denoted the $\bigwedge$ Model) and established the asymptotic optimality of a certain staffing and routing scheme. These two models can be thought of as building blocks for the more general multi-class multi-type systems. In this section we introduce some ongoing research that considers a more general, but still relatively simple system. In particular, we consider the N Model depicted in Figure 4, that can be thought of as a combination of the V Model and the $\bigwedge$ Model. While the V Model isolates the scheduling problem (which customer to admit into service upon a service completion) and the $\bigwedge$ Model isolates the routing problem (which server to choose upon a customer arrival), the N Model combines scheduling and routing.

The N model constitutes two customer classes (1 and 2) arriving according to independent poisson processes with arrival rate λ_i for class i, $i = 1, 2$. A class 2 customer can be served only by a type 2 server and his service will take on average $1/\mu_1$ units of time. Class 1 customers can be served by both server types. A service of a class 1 customer by a type 1 server will take, on average, $1/\mu_1$ units of time. service of a class 1 customer by a type 2 server will take, on average, $1/\mu_3$ units of time

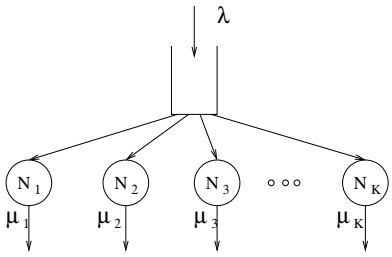


Figure 5: The $\bigwedge$ Model

An analytical method for calculation of steady state performance measures was given by Stanford and Grassmann [54] [54] which consider a special case in which $\mu_1 = \mu_2 = \mu_3 = \mu$. The policy analyzed is the following:

Type 2 servers give non-preemptive priority to class 2 customers. In the event that no class 2 customers are waiting and all type 1 servers are busy, a type 2 server that becomes available would select a waiting class 1 customer for service, if any is present. A class 1 customer that arrives and finds any type server available will be served by type 1 servers. Otherwise, if all type 1 servers are busy and there are any available type 2 servers, he will be served by a type 2 server.

Shumsky [56] gives an approximation scheme for the steady state distribution of the N Model under the same policy and with $\mu_2 = \mu_3$. Bell and Williams [9] analyzed the N Model in the context of conventional heavy traffic and established the asymptotic optimality of a certain threshold policy.

Yahalom and Mandelbaum [72] found that for certain settings of the N model, for which $\mu_2 = \mu_3$, the optimal policies are threshold policies combining thresholds on the number in queue and number of busy servers.

As mentioned above, [2] analyzes the $\bigwedge$ model (as described in Figure 5). The model, which constitutes a single customers class and multiple server types, is, in a sense, symmetric to the previously discussed V model.Armony and Mandelbaum [2] solve concurrently the asymptotically optimal staffing and routing problem in the context of Halfin-Whitt asymptotics as we did in the first part of this paper for the V Model. Our aim in this section is to combine both results to obtain the asymptotically optimal staffing, scheduling and routing for the N-Model. The work done in [2] turns out to be extremely useful for our setting since most of the methods employed there can be almost automatically adopted to our setting here.

As in [72], and for tractability, we limit our attention to settings of the N Model in which $\mu_2 = \mu_3$. The different settings we wish to consider vary in the value of the two customers classes (class 1 might be VIP or the opposite), in the cost of the different server types and in the relation between μ_1 and μ_2.

To illustrate this ongoing research we state here only results for a single specific setting. We limit ourselves to presentation of the asymptotically optimal policy and staffing, while omitting the propositions and their proofs. For the optimal policy we also have asymptotic analysis of performance measures (diffusion and steady state). These are also omitted.

The setting we present is the following: We consider the model in Figure 4 with $\mu_1 >= \mu_2 = \mu_3$, and where class 2 customers are the VIP customers. A motivation for this setting could be as follows. Class 2 are VIP customers requiring a certain type of service which can be given only by specialists of server type 2. Class 1 customers are regular customers who need a regular service which takes a shorter time. The experienced type 2 servers are capable of handling the regular service (but at a cost of slower service to class 1) while type 1 servers are not trained to handle the service required by the VIP customers. For this scenario the following policy was proved to be optimal in [72]:

- When a type 1 server completes a service, she will admit a class 1 customer to service if there are any waiting in queue 1.
- When a type 2 server completes a service, she will admit a class 2 customer to service, if any are waiting in queue 2. Otherwise, she will admit a class 1 customer if there are at least $q_1^*(I_2)$ waiting in queue 1, where the threshold $q_1^*(I_2)$ is a function of the system state.
- When a class 2 customer arrives to an empty queue 2, she will begin service with a type 2 server if any of them is idle.
- When a class 1 customer arrives to an empty queue 1, she will begin service with an idle type 1 server. If all type 1 servers are busy and there are at least I_2^* idle type 2 servers (or, equivalently, if $q_1^*(I_2) \leq 1$), she will begin service with one of them.

Since we wish to solve the control and staffing problems jointly, we consider the problem of minimizing staffing costs and waiting costs of the VIP customers. In addition, we impose constraints on the probability of delay for each class. Formally, we examine the following optimization problem:

$$\begin{aligned} &\text{minimize} \quad c_1 N_1 + c_2 N_2 + c_w \lambda_2 E[W_2 | W_2 > 0] \\ &\text{subject to} \\ &\qquad P_\pi(W_i > 0) \leq \alpha_i,\ 0 < \alpha_i < 1,\ i = 1, 2, \text{ for some } \pi \in \Pi. \end{aligned} \tag{168}$$

Here Π, as before, is the set of all non-preemptive non-anticipative scheduling policies, and c_w is the cost of a VIP customer waiting one unit of time. Also, we assume that $\alpha_2 \leq \alpha_1$.

Actually, we can limit ourselves to policies in which class 2 customers are served as soon as possible: When a type 2 server completes service he will choose to serve next a class 2 customer if there is one waiting in queue. To support this it can be shown that for any fixed $\bar{N} \stackrel{\triangle}{=} (N_1, N_2)$, $c_w E[W_2 | W_2 > 0]$ is minimized by such a policy.

The waiting cost can therefore be re-written as $c_w \frac{\lambda}{N_2 \mu_2 - \lambda_2}$ (since, given wait, the waiting of VIP customers has the same distribution as the waiting time given wait in a single class $M/M/1$ queue with service rate $N_2 \mu_2$ and arrival rate λ_2). We restate the optimization problem as follows:

$$\begin{aligned} &\text{minimize} \quad c_1 N_1 + c_2 N_2 + c_w \frac{\lambda_2}{N_2 \mu_2 - \lambda_2} \\ &\text{subject to} \\ &\qquad P_\pi(W_1 > 0) \leq \alpha_i,\ 0 < \alpha_i < 1, \text{ for some } \pi \in \Pi \end{aligned} \tag{169}$$

As done for the V Model, we consider a sequence of systems indexed by $r = 1, 2, ...$ such that $\lambda^r \to \infty$ and we assume non-negligibility of both classes, i.e $\frac{\lambda_i^r}{\lambda^r} \to a_i > 0,\ i = 1, 2$.

We consider a sequence of N models indexed by the superscript $r = 1, 2, ...$ For system r, we now present the asymptotically optimal policy, π^r. In contrast to the optimal policy of [72], the thresholds in this policy do not depend on the system state. Moreover, there are only thresholds on the number of idle servers and not on queue lengths. For fixed system size, π^r is as follows:

- When a type 1 server completes a service, she will admit a class 1 customer to service if there are any waiting in queue 1.
- When a type 2 server completes a service, she will admit a class 2 customer to service, if any are waiting in queue 2; otherwise she will admit a class 1 customer.
- When a class 2 customer arrives to an empty queue 2, she will begin service with a type 2 server if any of them is idle.
- When a class 1 customer arrives to an empty queue 1, she will begin service with an idle type 1 server. If all type 1 servers are busy, she will begin service with an idle type 2

server only if there are more than K^r type 2 servers idle.

Let $\bar{N}^r = (N_1^r, N_2^r)$ be the optimal solution to the following convex program:

$$\begin{array}{ll} \text{minimize} & c_1 N_1^r + c_2 N_2^r + c_w^r \frac{\lambda_2}{N_2\mu_2 - \lambda_2^r} \\ \text{subject to} & \mu_1 N_1^r + \mu_2 N_2^r \geq \lambda^r + \delta\sqrt{\lambda^r} \\ & \mu_2 N_2^r \geq \lambda_2^r \end{array} \tag{170}$$

where $\delta > 0$ is chosen so that:

$$\left[1 + \frac{(\delta/\sqrt{\mu_2})\Phi(\delta/\sqrt{\mu_2})}{\phi(\delta/\sqrt{\mu_2})}\right]^{-1} = \alpha_1. \tag{171}$$

Denote by $\underline{C}^r$ the solution to the above optimization problem when $\delta = 0$. Then $\underline{C}^r$ is the minimum cost when we only demand stability and it is clearly a lower bound for the cost of the original problem.

Let $K^r = \lceil \frac{\ln \alpha_2 - \ln \alpha_1}{\ln \rho_1^r} \rceil$

in which $\rho_1^r = \frac{\lambda_1^r}{N_2^r \mu_2}$.

Let $C^r(\bar{N}^r, \pi^r) = c_1 N_1 + c_2 N_2 + c_w^r \frac{\lambda_2}{N_2\mu_2 - \lambda_2^r}$.

Then, we can prove the following:

Assume that $c_w^r = \Theta(\lambda^r)$. Then $(\bar{N}^r, \pi^r)$ with K^r given above, are asymptotically optimal in the following sense:

- *Feasibility:* $\limsup_{r\to\infty} P\{W_i^r(\infty) > 0\} \leq \alpha_i,\ i = 1, 2.$
- *Optimality:* If we take any other feasible sequence $(\bar{\hat{N}}^r,$ $hat\pi^r)$, we have that

$$\frac{C^r(\bar{\hat{N}}_2^r, \hat{\pi}^r) - \underline{C}^r}{C^r(\bar{N}^r, \pi^r) - \underline{C}^r} \to \gamma \geq 1$$

The interesting fact is that, similarly to the V Model, the complicated state dependent optimal policy becomes asymptotically simple.

The above is just an illustration of what can be done in the context of the N Model. Our aim is to complete the picture by finding optimal staffing and controls for additional settings of the N Model, hopefully all of them.

8 Future Research

In this work we a have limited our attention mainly to the V-Model of Skills-Based Routing. We have tried to make the problem as realistic as possible. However, some more work is required to understand the effects of limited number of lines, retrials and feedback as well as the issue of different service requirements for different classes.

The V-Model is a simple example of Skills-Based Routing that isolates the scheduling problem. A natural continuation, as illustrated in Section 7.2, would be to combine the results of this work with other related works (such as [2]) to obtain results for networks of more general topology such as the N Model.

Also, the work done in this thesis lays the ground for the solution of several design problems. For example, the results of this thesis can be used to consider the question of pooling several V models (or I models) when the pooling incurs a cost as a result of the required cross-training of the CSRs.

We have dealt only briefly with abandonments. In this context, there is still a need for more general results to the two optimization problems that where presented in Section 7.1.

9 Appendix - Efficiency Driven $M/M/N$

In Section 5, we introduced the diffusion limit for the Efficiency Driven $M/M/\{K_i\}$ 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, i.e. we fix $\delta,\ 1/2 < \delta \leq 1$, and let λ^N grow in the following manner:

$$N^{\delta}(1-\rho^N) \to \beta,\ 0 < \beta < \infty. \tag{172}$$

Our aim is to prove convergence of the process $Q^N(t)$ (standing for the total number of customer in system N at time t) to a Reflected Brownian Motion. This result was proved in [70] 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 increased to one.

Essentially, in order to obtain convergence, all that we have to do is prove that the time that the process X spends below zero 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 just an adaptation of the proof used in Garnett's M.Sc. Thesis [26] for the proof of part 3 of Theorem 6.2 there (A brief version of Garnett's proof can be found in Garnett et al. [27], where most of the details are omitted).

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

$$X^N(t) = \frac{Q^N((N^{2\delta-1}t) - N}{N^{\delta}} \tag{173}$$

First we quote again proposition 5.1.1.

Proposition 5.1.1. *Consider a sequence of $M/M/N$ system indexed by $N = 1, 2, ...$, such that*

$$N^{\delta}(1-\rho^N) \to 0 < \beta < \infty, \tag{174}$$

Assume $\frac{Q^N(0)}{N^{\delta}} \Rightarrow X(0)$, where $X(0) \geq 0,$ a.s. Then,

$$X^N(t) \Rightarrow RBM(-\beta\mu, 2\mu) \tag{175}$$

Remark: 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 [26] and the references therein for more details on this kind of limits.

Proof:

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 [43]) 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.

$$\tau_+^N(t) = \int_0^t 1_{\{X^N(s) \geq 0\}} ds, \tag{176}$$

$$\tau_-^N(t) = \int_0^t 1_{\{X^N(s) < 0\}} ds, \tag{177}$$

Then,

$$X^N \circ \tau_+^N \Rightarrow RBM(-\beta\mu, 2\mu). \tag{178}$$

Where $f \circ g$ is the composition map (i.e. $f \circ g(t) = f(g(t))$). By the random time change theorem all that is left to prove is that

$$\tau_-^N(t) \Rightarrow 0. \tag{179}$$

Let us look at 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

$$\sigma^N(T) = \min\{n : \sum_{i=1}^{n} C_i^N > T\} \tag{180}$$

Then, $\sigma^N(T)$ is a stopping time with respect to the sequence $\{C_i^N\}$.

What we are seeking to prove is that

$$\lim_{N\to\infty} P\{\sum_{i=1}^{\sigma^N(T)} B_i^N > \epsilon\} = 0. \tag{181}$$

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,

$$E[B_i^N] = \frac{1}{N^{2\delta-1}(N\mu-\lambda)} \sim \frac{1}{\beta N^\delta} \tag{182}$$

$N^\delta(1-\rho^N) \to \beta$ also implies that $\sqrt{N}(1-\rho^N) \to 0$ and hence, following [26] and due to the time acceleration, we will have also 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 [26], 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 $X(0) = 0$

$$\lim_{N\to\infty} E[A_1^N] = 0, \tag{183}$$

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

$$V_N^{N-k} \leq E[A_i^N] \frac{1 - \left(\frac{\lambda^N}{\lambda^N + (N-k+1)\mu}\right)^k}{1 - \left(\frac{\lambda^N}{\lambda^N + (N-k+1)\mu}\right)} \tag{184}$$

The above is obtained by a simple adaptation of pages (67-68) in [26]. Now, $E[A_i^N] = o(\frac{1}{N^\delta})$ and the result follows. ■

References

[1] Aksin, O.Z., Karaesmen A.F., "Designing Flexibility: Characterizing the Value of Cross-Training Practices", submitted, 2002.

[2] Armony M., Mandelbaum A., "Routing, Staffing and Design of Large Service Systems: The Case Of a Single Customer Class and Heterogeneous servers", Draft, March 2004.

[3] Armony M., Bambos N., "Queueing Dynamics and Maximal Throughput Scheduling in Switched Processing Systems", *Queueing Systems*, July 2003, vol. 44(3), pp. 209-252.

[4] Armony M., Maglaras C., "On customer contact centers with a call-back option: customer decisions, routing rules and system design", 2003, *Oper. Res.*, to appear.

[5] Armony M., Maglaras C., "Contact centers with a call-back option and real-time delay information", 2003, *Operetaions Research*, to appear.

[6] Atar R., Mandelbaum A., Reiman M., "A Brownian control problem for a simple queueing system in the Halfin-Whitt regime" , Technical report, May 2002.

[7] Atar R., Mandelbaum A., Reiman M., "Scheduling a Multi-Class Queue with Many Exponential Servers: Asymptotic Optimality in Heavy Traffic", Submitted to *Annals of Applied Probability*, June 2002

[8] Bambos N., Walrand J., "Scheduling and stability aspects of a general class of parallel processing systems", *Advances in Applied Probability*, **25**, pp. 176–202, 1993.

[9] Bell S.L., Williams R.J., "Dynamic Scheduling of a System with Two Parallel Servers in Heavy Traffic with Complete Resource Pooling: Asymptotic Optimality of a Continuous Review Threshold Policy", *Annals of Applied Probability*, Vol. 11, 2001.

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

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

[12] Borst S.C., Seri P., "Robust algorithms for sharing agents with multiple skills", Working Paper, 2000.

[13] Brandt A., Brandt M., Spahl G. and Weber D., "Modelling and Optimization of Call Distribution Systems", Elsevier Science B.V., 1997.

[14] Brandt A. and Brandt M., "On a two-queue priority system with impatience and its application to a call center", *Methodology and Computing in Applied Probablity*, **1**, pp. 191-210, 1999.

[15] CC (2000). North american call center summit (NACCS), Call Center Statistics, Technical report, http://wwwcallcenternews.com/resourses/statistics.shtml.

[16] Chen H., Yao D., *Fundamentals of queueing networks : performance, asymptotics, and optimization*, Springer, New-York, 2001.

[17] De Véricourt F., Zhou, Y.-P., "Managing response time and service quality in a call allocation problem", submitted to *Operaions Research*, 2003.

[18] Erlang A.K. ,"On the rational determination of the number of circuits". In "The life and works of A.K. Erlang." E. Brockmeyer, H.L. Halstrom and A. Jensen, eds. Copenhagen: The Copenhagen Telephone Company, 4.1.1, 4.2.1, 1948.

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

[20] 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*, 34:1121-1138, September 1988.

[21] Fleming P., Stolyar A., Simon B., "Heavy traffic limit for a mobile phone system loss model", *Proceedings of 2nd Int'l Conf. on Telecomm. Syst. Mod. and Analysis*, Nashville, TN, 1994.

[22] Gans N., Koole G. and Mandelbaum A., "Telephone Call Centers: Tutorial, Review, and Research Prospects", Invited review paper by *Manufacturing and Service Operations Management (MSOM)*, 5(2), pp. 79-141, 2003.

[23] Gans N. and Van Ryzin G., "Optimal control of a multiclass, flexible queue system", *Operations Research*, **45**, pp. 677–693, 1997.

[24] Gans N.,Zhou Y.-P., "Managing learning and turnover in employee staffing", *Operations Research* 50(6), pp. 991 - 1006, 2002.

[25] Gans N. ,Zhou Y.-P., "A call-routing problem with service-level constraints", *Operations Research*, 51(2), pp. 255-271, 2003.

[26] Garnett, O. "Designing a telephone call center with impatient customers". M.Sc. thesis, Technion - Israel Institute of Technology, 1998.

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

[28] Glazebrook, K., Niño-Mora, J., "Parallel scheduling of multiclass M/M/m queues: approximate and heavy-traffic optimization of achievable performance", *Operations Research*, **49**:4, pp. 609-623, 2001.

[29] Glynn, P.W., "Upper bounds on Poisson tail probabilities", *Operations Research Letters*, 6(1), pp. 9-14, 1987.

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

[31] Harrison, J.M., "Heavy traffic analysis of a system with parallel servers: asymptotic analysis of discrete-review policies", *Annals of applied probability*, 8, pp. 822-848, 1998.

[32] Harrison J.M., Zeevi A., "Dynamic scheduling of a multiclass queue in the Halfin and Whitt heavy traffic regime", to appear in *Operations Research*, 2003.

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

[34] Hopp W.J., Van Oyen M.P., "Agile Workforce Evaluation: A Framework for Cross-training and Coordination". To appear *IIE Transactions*, 36, 2004.

[35] Jagerman D.L., "Some properties of the Erlang loss function", *Bell Systems Technical Journal*, **53**:3, pp. 525–551, 1974.

[36] Jelenkovic P., Mandelbaum A., and Momcilovic P., "The GI/D/N queue in the QED regime", Accepted to *Queueing Systems*, June 2003, available at `http://iew3.technion.ac.il/serveng/References/references.html`,

[37] Jennings O.B., Mandelbaum A., Massey W.A., and Whitt W., "Server staffing to meet time-varying demand", *Management Science*, **42**, pp. 1383–1394, 1996.

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

[39] 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.

[40] Luh, H.P., Viniotis, I., "Threshold Control Policies for Heterogeneous Server Systems", *Math Meth Oper Res*, **55**, pp 121-142, 2002.

[41] Maglaras C., Zeevi A., "Pricing and capacity sizing for systems with shared resources: Scaling relations and approximate solutions", *Management Science*, 49(8), pp. 1018-1038, 2003.

[42] Mandelbaum A., Sakov A. and Zeltyn S., “Empirircal Analysis of a Call Center”, Technical Report, 2000.

[43] Mandelbaum A., Pats G., “State-Dependent Queues: Approximations and Applications”.

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

[45] Mandelbaum A., “Call Centers: Research Bibliography with Abstracts”, Version 3, May 27, 2002.

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

[47] Meyn S.P. and Tweedie R.L., “Markov Chains and Stochastic Stabiliy”, Springer, 1993.

[48] Perry M., Nilsson A., “Performance modeling of automatic call distributors: assignable grade of service staffing”, In *XIV International Switching Symposium*, pp. 294–298, 1992.

[49] Puhalskii A. “On the Invariance Principle For the First Passage Time”, *Mathematics of Operations Research*, Vol. 19, Nov. 1994.

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

[51] Rege K.M., Sengupta B., “A Priority Based Admission Scheme for a Multiclass Queueing System”, AT & T Tech. J. 64, pp. 1731-1753, 1985.

[52] Resnick S. “Adventures in Stochastic Process”, Birkhäuser, 1992.

[53] Rykov, V.V., “Monotone Control of Queueing Systems with Heterogeneous Servers”, *Queueing Systems*, 37, pp. 391-403, 2001.

[54] Stanford D.A., Grassmann W.K., “Bilingual server call centres”, In *Analysis of Communication Networks: Call Centres, Traffic and Performance*, D.R. McDonald and S.R.E. Turner (eds.), Fields Institute Communications **28**, pp. 31–48, 1992.

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

[56] Shumsky R.A., “Approximation and Analysis of a Queueing System with Flexible and Specialized Servers”, *OR Spectrum*, 26(3), pp. 307-330, 2004.

[57] Simon B., "Priority Queues with Feedback", *Journal of the Association for Computing Machinery*, 31(1), pp. 134-149, 1984.

[58] Stolletz R., "Performance Analysis and Optimization of Inbound Call Centers", Springer, 2003.

[59] Sze, D.Y., "A queueing model for telephone operator staffing", *Operations Research*, 32, pp. 229–249, 1984.

[60] 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

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

[62] Wallace R.B., Whitt W., "Resource Pooling and Staffing in Call Centers with Skill-Based Routing". Submitted to *Operations Research*, 2004.

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

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

[65] Whitt W., "Understanding the efficiency of multi-server service systems", *Management Science*, **38**, pp. 708–723, 1992

[66] Whitt W., "Heavy-Traffic Approximation for Service Systems With Blocking", AT&T Bell Laboratories Technical Jornal. Volume 63, Number 5, pp. 689-708, May-June, 1984.

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

[68] Whitt W., "A diffusion approximation for the G/GI/n/m queue", 2002. *Operations Research*, to appear.

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

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

[71] Y.-Ch. Teh and A.R. Ward , "Critical thresholds for dynamic routing in queueing networks", *Queueing Systems*, 42, pp. 297–316, 2002

[72] Yahalom T., Mandelbaum A., “Optimal Scheduling of a Multi-Server Multi-Class Non-Preemptive Queueing System”, Preprint.

[73] 4 Call Centers Software. Downloadable from ie.technion.ac.il/serveng

תכן ובקרה של התור M/M/N עם לקוחות מסוגים שונים ומספר גדול של שרתים

איתי גורביץ'

תכן ובקרה של התור M/M/N עם לקוחות מסוגים שונים ומספר גדול של שרתים

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת תואר מגיסטר למדעים

בחקר ביצועים וניתוח מערכות

איתי גורביץ'

הוגש לסנט הטכניון – מכון טכנולוגי לישראל

תמוז תשס"ד – חיפה – יולי 2004

המחקר נערך בהנחייתו של פרופסור אבישי מנדלבאום מהפקולטה להנדסת תעשייה וניהול בטכניון. ברצוני להודות לו מקרב לב על מסירותו הראויה להערכה; על שיחות רבות וממושכות שבמהלכן סייע לי להתמודד עם בעיות שעלו במהלך המחקר, תוך מציאת האיזון הנכון בין מלאכת ההכוונה וההדרכה לבין הקידום של התפתחותי האישית והמקצועית; ועל התרומה הניכרת שתרם לי מידיעותיו ומרעיונותיו. עבודתנו המשותפת הייתה משמעותית ביותר בעבורי, ועל כך נתונה לו תודתי העמוקה.

כמו כן ברצוני להודות לפרופסור מור ערמוני מבית הספר למינהל עסקים ע"ש סטרן באוניברסיטת ניו-יורק, אשר במהלך שנת שהותה בטכניון הייתה לי בבחינת מנחה נוספת. לעזרתה הנדיבה תפקיד מרכזי בהצלחתו של מחקר זה.

תודתי נתונה גם לפרופסור חיה כספי מהפקולטה להנדסת תעשייה וניהול בטכניון, שעל אף עיסוקיה הרבים מצאה זמן לייעץ לי ולהשיב לשאלותיי באריכות ובסבלנות רבה.

לבסוף, אני מודה לבית הספר ללימודי מוסמכים בטכניון על התמיכה הכספית הנדיבה במהלך השתלמותי.

תוכן עניינים

רשימת טבלאות

רשימת הציורים

תקציר מורחב

מערכות שירות מודרניות מורכבות לעתים קרובות ממספר סוגי לקוחות ומספר סוגי שרתים. סוגי הלקוחות נבדלים אלה מאלה בדרישות השירות שלהם. סוגי השרתים מאופיינים על ידי סוגי הלקוחות שהם יכולים לשרת ועל ידי רמת השירות שהם יכולים להעניק להם. דוגמא חשובה למערכות שירות כאלה היא מוקדים טלפוניים. מוקדים רבים מטפלים במספר סוגי שיחות (הנבדלות בסוג השירות הנדרש, שפת המתקשר וכו') ומעסיקים מוקדנים מסוגים שונים. בתפעול של מוקדנים טלפוניים ישנן שלוש סוגיות מרכזיות. בהינתן תחזית של מופע הלקוחות למערכת וזמני השירות שלהם, שלוש הסוגיות הן:

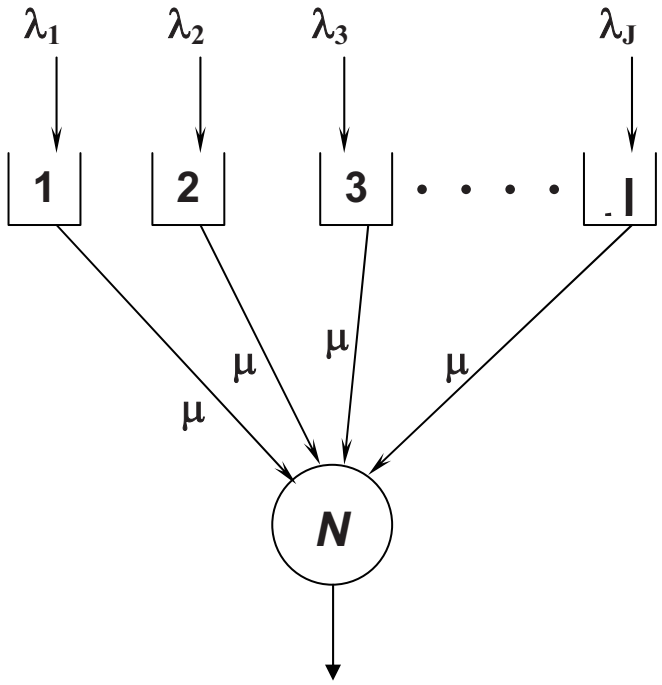
- **עיצוב**: זוהי החלטה בטווח הארוך שנועדה לתת מענה לשאלה: כיצד לפלח את סוגי הלקוחות, את סוגי השרתים וכיצד לקבוע אלו סוגי לקוחות יכול כל שרת לשרת.
- **איוש**: כמה שרתים נחוצים מכל סוג על מנת להתמודד עם הביקושים.
- **בקרה**: כיצד לנתב את הלקוחות במערכת בזמן אמת, דהיינו איך לצוות לקוחות לשרתים כאשר מתפנה שרת או כאשר מופיע לקוח למערכת.

שלוש הסוגיות האלה קשורות קשר הדוק אחת לשנייה ולכן יש לתת להן מענה במקביל. עם זאת, בגלל מורכבות הטיפול המשולב בכל הבעיות בו זמנית, הן מטופלות לרוב באופן היררכי ובלתי תלוי. גם אם מתעלמים מהתלות בין שלוש הסוגיות הללו הרי שפתרון סגור למקרה הכללי אינו בנמצא. לחלופין, אנו מנסים לפתור את הבעיה על ידי ניתוח של מודלים פשוטים יחסית במטרה לקבל תובנות לגבי המודל הכללי יותר. בעבודה זו אנו מתמקדים במודל ה-V (ראו תמונה 1). זהו מודל שבו מספר סוגי לקוחות משורתים על ידי קבוצה הומוגנית של שרתים. לכל סוגי הלקוחות אותן דרישות שירות באופן בלתי תלוי בסוגם.

ביחס למודל זה אנו שואלים את שתי השאלות הבאות:

- בהנתן מספר השרתים, כיצד לצוות שרתים לסוגי הלקוחות השונים על מנת להשיג ביצועים מיטביים של המערכת.
- כמה מוקדנים צריך על מנת להביא למינימום את עלות האיוש ואת זמני ההמתנה תוך עמידה באילוצים מסוימים על מדדי הביצוע של המערכת.

אנו עונים על שתי שאלות אלו על ידי אפיון של מדיניות ניתוב ואיוש שהם אופטימליים אסימפטוטית. המדיניות שאנו מציעים היא אופטימלית אסימפטוטית במובן שהיא מביאה למינימום את עלויות האיוש וההמתנה במצב יציב, תוך עמידה באילוץ על ההסתברות להמתין - ואת זאת היא משיגה אסימפטוטית ביחס לקצב המופע, דהיינו כאשר קצב המופע הולך וגדל.



תמונה 1 : מודל ה-V

המסגרת העיקרית שלנו לניתוח אסימפטוטי היא משטר העומס הגבוה שהוצג לראשונה על ידי Erlang בראשית המאה הקודמת, אבל נוסח מתמטית לראשונה על ידי Halffin and Whitt ב-1981. אנו מכנים תחום זו בשם QED – Quality and Efficiency Driven. מערכות שפועלות במשטר זה מאופיינות בשילוב נדיר של נצילות גבוהה ואיכות שירות גבוהה.

פורמלית, נתבונן בסדרה של מערכות בעלות מבנה קבוע וקצב מופע גדל λ. נניח כי קיבולת השירות של כל מערכת כזו עולה על λ במרווח ביטחון שהוא בסדרי גודל של $\sqrt{\lambda}$. אזי, בפרט, העומס על המערכת מתכנס ל-1 כאשר $\lambda \to \infty$ (כלומר המערכת פועלת בעומס גבוה). היבט איכותי של משטר זה ניתן על ידי האפיון הבא: כאשר

$\lambda \to \infty$, ההסתברות להמתין מתכנסת למספר לא טריוויאלי (דהיינו לערך באינטרוול הפתוח $(0,1)$). איכות שירות שכזו אינה אופיינית למערכות בעומס גבוה ומתאפשרת עקב יתרון לגודל, הבא לידי ביטוי במערכות עם מספר גדול של שרתים.

שני המאפניינים הללו של משטר ה- QED (העומס הגבוה מחד גיסא ואיכות השירות הגבוהה מאידך גיסא) הוכחו כשקולים במספר מודלים ואנו מוכיחים קיומה של שקילות זו גם במודל ה-V.

<u>סיכום התוצאות</u>: אנו מתבוננים במודל V שבו לקוחות מסוג i $(j=1,...,J)$ מופיעים למערכת לפי תהליך פואסון עם קצב λ_j . זמן השירות הוא אקספוננציאלי עם ממוצע $1/\mu$ באופן בלתי תלוי בסוג הלקוח.

אנו מעוניינים לפתור שתי בעיות אופטימיזציה. הראשונה היא בעיית מינימום עלות איוש תוך עמידה באילוצים על ההסתברויות להמתין. פורמלית, הבעיה מנוסחת כדלהלן:

$$\begin{aligned} &\min && \boldsymbol{N} \\ &\boldsymbol{s.t.} && \boldsymbol{P^{\pi}}\{\boldsymbol{W}_j > 0\} \le \boldsymbol{\alpha}_j, \quad 0 < \boldsymbol{\alpha}_j < 1, \quad \boldsymbol{j} = 1,...,\boldsymbol{J}; \text{ for some } \boldsymbol{\pi} \in \Pi \end{aligned}$$

כאן אנו משתמשים בסימון $\boldsymbol{P^{\pi}}$ עבור הסתברות המחושבת תחת מדיניות π. נניח לצורך הפשטות כי α_j מסודרים בסדר עולה (בפרט סוג 1 הם הלקוחות החשובים ביותר).

הבעיה השנייה היא בעיית מינימום עלות איוש והמתנה. פורמלית, הבעיה מנוסחת כדלקמן:

$$\min \quad \boldsymbol{N} + \sum_{\boldsymbol{j}=1}^{\boldsymbol{J}} \boldsymbol{c}_j \boldsymbol{\lambda}_j \boldsymbol{E}[\boldsymbol{W}_j]$$

כאשר $\boldsymbol{E^{\pi}}$ תוחלת המחושבת תחת מדיניות π ו- c_j עלות ההמתנה ליחידת זמן עבור לקוחות מסוג j. נניח לשם הפשטות כי c_j מסודרים בסדר יורד.

עבור שתי הבעיות אנו מראים כי מדיניות אסימפטוטית אופטימלית היא המדיניות הבאה:

צוות לקוח מסוג j לשירות רק אם יש יותר מ-K_j שרתים פנויים ואם התורים של כל סוגי הלקוחות מ-1 עד j-1 ריקים. הספים K_j מקיימים $0 = \boldsymbol{K}_1 \le \boldsymbol{K}_2 \le ... \le \boldsymbol{K}_J$.

לקוחות שלא יכולים להתקבל לשירות מייד עם הגעתם למערכת ממתינים בתור המתאים לסוג הלקוחות שלהם. כל תור כזה מטופל לפי FIFO.

המדיניות המוצעת מאפשרת לנו ליצור הבחנה עדינה בין ההסתברויות להמתין של סוגי הלקוחות השונים על ידי בחירה נכונה של הספים. יתרה מזאת, מסתבר שהספים האופטימליים הם פונקציה פשוטה של הפרמטרים של הבעייה.

יחד עם מדיניות זו אנו מציעים רמת איוש שגם היא אסימפטוטית אופטימלית ואשר לרוב מכוונת את המערכת במשטר ה QED. אנו מראים שתחת אילוצים סבירים על ההסתברויות להמתין רמת האיוש האופטימלית נקבעת בקלות כפונקציה של סך העומס על המערכת והפרמטרים של העדיפות הנמוכה ביותר. אנו מסיקים גם כי תחת עלויות המתנה c_j שאינן משתנות עם גודל המערכת המדיניות האסימפטוטית אופטימלית עבור בעיית העלויות היא מדיניות של עדיפויות סטטיות.

בנוסף אנו מציגים הרחבה של מודל ה-V למקרה עם נטישות. דהיינו, נניח כי בנוסף לנתונים הקודמים, ללקוחות מסוג j יש סבלנות אקספוננציאלית עם פרמטר θ_j. אנו מראים כי גם במקרה זה, תחת מבני עלויות מסוימים, המדיניות שהוצעה לעיל (מדיניות הספים) היא אסימפטוטית אופטימלית ומחשבים עבורה מדדי ביצוע אסימפטוטיים. בפרט, אם עלות הנטישה של סוג לקוחות גבוהה ככל שהסבלנות קצרה יותר, אזי מדיניות הסף שהוצגה לעיל היא אופטימלית במשטר ה- QED.

לבסוף, אנו מתארים בקצרה התקדמות שנעשתה במודל N, המשלב את בעיות ניתוב הלקוחות וציוות השרתים.