12-7-2015 12:00 AM

Recent Advances in Accumulating Priority Queues

Na Li, *The University of Western Ontario*

Supervisor: Dr. David A. Stanford, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Statistics and Actuarial Sciences

RECENT ADVANCES IN ACCUMULATING PRIORITY QUEUES

(Thesis format: Integrated Article)

by

Na Li

Graduate Program in Statistics and Actuarial Sciences

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

Abstract

This thesis extends the theory underlying the Accumulating Priority Queue (APQ) in three directions. In the first, we present a multi-class multi-server accumulating priority queue with Poisson arrivals and heterogeneous services. The waiting time distributions for different classes have been derived. A conservation law for systems with heterogeneous servers has been studied. We also investigate an optimization problem to find the optimal level of heterogeneity in the multi-server system. Numerical investigations through simulation are carried out to validate the model.

We next focus on a queuing system with Poisson arrivals, generally distributed service times and nonlinear priority accumulation functions. We start with an extension of the power-law APQ in Kleinrock and Finkelstein (1967), and use a general argument to show that there is a linear system of the form discussed in Stanford, Taylor and Ziedins (2014) which has the same priority ordering of all customers present at any given instant in time, for any sample path. Beyond the power-law case, we subsequently characterize the class of nonlinear accumulating priority queues for which an equivalent linear APQ can be found, in the sense that the waiting time distributions for each of the classes are identical in both the linear and nonlinear systems.

Many operational queuing systems must adhere to waiting time targets known as Key Performance Indicators (KPIs), particularly in health care applications. In the last aspect, we address an optimization problem to minimize the weighted average of the expected excess waiting time (WAE), so as to achieve the optimal performance of a system operating under KPIs. We then find that the Accumulating Priority queuing discipline is well suited to systems with KPIs, in that each class of customers progresses fairly towards timely access by its own waiting time limit. Due to the difficulties in minimizing the WAE, we introduce a surrogate objective function, the integrated weighted average excess (IWAE), which provides a useful proxy for WAE. Finally, we propose a rule of thumb in which patients in the various classes accumulate priority credit at a rate that is inversely proportional to their time limits.

Keywords: Accumulating priority queue, Heterogeneous servers, Waiting time distributions, Nonlinear priority function, Optimization.

Co-Authorship Statement

Paper title: Multi-server Accumulating Priority Queues with Heterogeneous Servers

Publication: Submitted to European Journal of Operational Research

List of authors: Na Li, David A. Stanford

I defined the heterogeneous multi-server accumulating priority queuing model, derived the waiting time distributions for different classes, and studied the optimization problem on the level of heterogeneity in the system. Dr. Stanford proposed a conservation law for $M/M_i/c$ systems. I performed the calculations, derivations and numerical investigations in this paper. Both authors contributed with corrections and recommendations to the contents and relevance.

Paper title: On Waiting Times for Nonlinear Accumulating Priority Queues

Publication: Submitted to *O*perations Research

List of authors: Na Li, David A. Stanford, Peter Taylor, Ilze Ziedins

I initiated an extension of Kleinrock and Finkelstein's model, and derived the waiting time distributions for various classes in the power-law APQs. Dr. Stanford introduced the idea of pairing the sample path of the power-law APQ with that of a corresponding linear APQ (named "linear proxy") to show the equivalence. Then, Dr. Taylor raised the question whether there exist other nonlinear APQs with linear proxies than the power-law APQ. Dr. Ziedins first stated the criterion (equation (4.15)) for the existence of a linear proxy. I proceeded the derivations and calculations for most of the proofs in the paper, expect for the necessary and sufficient proof for the criterion, which was derived by Dr. Taylor, whereas I proved its sufficient condition through constructing the maximum priority process for nonlinear APQs. I performed all the examples and numerical investigations. All authors contributed with corrections and recommendations to the contents and relevance.

Paper title: Optimization of Queues Operating under Waiting Time Limits

Publication: Submitted to *O*perations Research

List of authors: Na Li, David A. Stanford, Azaz B. Sharif, Richard J. Caron

The optimization problem to minimize the weighted average of the total expected excess waiting time (WAE) was formulated by Dr. Stanford, together with Azaz B. Sharif and Dr. Caron in 2014. A rule of thumb for the optimality of TEE (short for "total expected excess", which is the equal-weighted case) was introduced based on a variety of numerical calculations done by the other PhD candidate, Azaz B. Sharif. I joined this work in April 2015, and performed extensive numerical investigations combined with theoretical derivations to validate the rule of thumb. Then, I proposed a surrogate objective function, the integrated weighted average excess (IWAE), as a proxy for WAE, and obtained an explicit expression for the optimal solution of IWAE. Overall, I contributed to the paper with regards to the formulation of IWAE, as well as the derivations, calculations, and numerical investigations for both the WAE and IWAE optimization problems. A discussion of the convexity in the two-class APQ optimization problem was presented by Dr. Stanford. Dr. Caron provided the guidance with his expertise on optimization. All authors contributed with corrections and recommendations to the contents and relevance.

Acknowledgments

This thesis couldn't have been completed without the assistance of many individuals and I would like to express my deepest appreciation to them.

I would like to express my gratitude to my supervisor, Dr. Stanford, whose guidance, understanding, and caring, added considerably to my graduate experience. I appreciate his vast knowledge on queuing theory, and his assistance in formal English writing. His advice on both research and my career path have been valuable.

I would like to thank Dr. Taylor and Dr. Ziedins for their valuable suggestions and contributions to our paper, together with Dr. Stanford who inspired me to make further developments regarding more general problems on the nonlinear accumulating priority queues. I would also like to thank Dr. Caron for his guidance on optimization techniques and advice on writing.

A very special acknowledgment goes to Dr. John Braun, without whose motivation and encouragement I would not have considered to pursue a PhD.

I would also like to thank my husband and my son for keeping me company, along with endless touching and joyful moments.

Last but foremost, I would like to thank my parents for their perpetual and unconditional love and support through my life.

Contents

List of Figures

List of Tables

List of Appendices

List of Abbreviations

APQ	Accumulating priority queue
c.d.f.	Cumulative distribution function
CTAS	Canadian Triage and Acuity Scale
EE	Expected excess waiting time per unit of time
FCFS	First come first serve
FSF	Fast server first dispatch policy
GS	Gaver-Stehfest numerical inversion algorithm
i.i.d.	Independent and identically distributed
IWAE	Integrated delay of the weighted expected excess waiting time
KPI	Key Performance Indicators
LST	Laplace-Stieltjes Transform
p.d.f.	Probability density function
RBS	Rate balancing selection dispatch policy
RCS	Randomly chosen server dispatch policy
r.v.	Random variable
SDP	Service dispatch policies
SSF	Slowest server first dispatch policy
TEE	Total expected excess waiting time (equal-weighted)
WAE	Weighted average of the total expected excess waiting time

List of Symbols and Definitions

Queuing models and their descriptions

$M/G/1$ A single-server queue with Poisson arrivals and general service duration distributions.

$G/G/1$ A single-server queue with general inter-arrival distributions and general service duration distributions.

$GI/G/1$ A single-server queue with independent general inter-arrival distributions and general service duration distributions.

$M/M/c$ A multi-server queue with Poisson arrivals and a common exponential service distribution.

$M/M_i/c$ A multi-server queue with Poisson arrivals and heterogeneous exponential service distributions.

$GI/M_i/c$ A multi-server queue with independent general inter-arrival distributions and heterogeneous exponential service distributions.

Linear APQ An accumulating priority queue with linear priority accumulation functions.

Power-law APQ of order r An accumulating priority queue with the power-law priority accumulation functions with a parameter r.

Nonlinear APQ An accumulating priority queue with general nonlinear priority accumulation functions.

Commonly used symbols in the entire thesis

$\mathbb{N}$ Positive integers: $1, 2, \ldots$.

$\mathbb{R}$ Real numbers.

K K classes of customers.

c c servers.

b_k The linear accumulation rate of class-k customers; $k = 1, 2, \ldots, K$. Particularly, in the two-class case, $b = b_2/b_1$ where b_1 is usually set to be one. (Denoted by b_k^L in Chapter 4.)

$b_k^{(r)}$ The accumulation rate for class-k customers; $k = 1, 2, \ldots, K$ in the power-law APQ of order r.

λ_k The arrival rate of class-k customers; $k = 1, 2, \ldots, K$.

λ Total arrival rate of K classes, i.e., $\lambda = \sum_{k=1}^{K} \lambda_k$.

μ_i The service rate of i^{th} server; $i = 1, 2, \ldots, c$.

μ Total service rate of c servers, i.e., $\mu = \sum_{i=1}^{c} \mu_i$. (Denoted by μ_a in Chapter 3.)

ρ Server utilization of the system, i.e., $\rho = \lambda/\mu$.

π_n The stationary probability of n^{th} state; $n = 0, 1, 2, \ldots$.

m_k The average waiting time for class-k customers; $k = 1, 2, \ldots, K$.

$U(t)$ The unfinished workload process at time t, i.e., the total amount of work present in the system at time t still to be completed by the server.

$\bar{U}$ The limiting average of the unfinished work, i.e., $\bar{U} = \lim_{t\to\infty} E\{U(t)\}$.

$\overline{x_k^2}$ The second moment of service time for a customer from class k; $k = 1, 2, \ldots, K$.

W_0 The residual life of the customer found in service upon an arrival's entry.

$\boldsymbol{M}$ The maximum priority process at time t, i.e., $\boldsymbol{M} = \{M_i(t);\ t \geq 0,\ i = 1, \ldots, K\}$.

f_g The approximating function under Gaver-Stehfest method for numerical Laplace transform inversion for a given real-valued function $f(t); t \geq 0$.

$\tilde{f}$ The Laplace-Stieltjes transform of a distribution function.

Frequently used symbols in Chapter 3

μ_a Total service rate of c servers, i.e., $\mu_a = \sum_{i=1}^{c} \mu_i$.

$\boldsymbol{\mu}$ Service rate vector, i.e., $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_c)$.

C The set of idle servers for the current state.

r A given r-dispatch policy.

$p_i(C; r)$ The probability that server $i \in C$ is chosen to serve the newly-arriving customer under the r-dispatch policy.

$V(t)$ The release time at time t, i.e., the amount of time until a server can become idle, given the work present at time t.

$\bar{V}$ The limiting expected value of $V(t)$ over time, i.e., $\bar{V} = \lim_{t \to \infty} E\{V(t)\}$.

V_0 The expected residual cohort inter-completion time.

$\boldsymbol{\mathsf{G}}$ A vector of Gini-like coefficients to measure the level of heterogeneity in multi-server queues, i.e., $\boldsymbol{\mathsf{G}} = \{\mathsf{G}_i, i = 1, \ldots, c.\}$.

G The largest level of heterogeneity in the system, i.e., $\mathsf{G} = \mathsf{G}_c$.

$\pi(\boldsymbol{\mu}, c; r)$ The probability that all c servers are simultaneously busy in a stationary $M/M_i/c$ queue.

Frequently used symbols in Chapter 4

$\mathcal{F}$ A set of priority accumulation functions $\{f_k(.), k = 1, \ldots, K\}$ in nonlinear APQs.

r A given power-law APQ of order r.

b_k^L The linear accumulation rate of class-k customers; $k = 1, 2, \ldots, K$.

Frequently used symbols in Chapter 5

l_k The delay limit for class-k customers; $k = 1, 2, \ldots, K$.

p_k The compliance probability for class-k customers; $k = 1, 2, \ldots, K$.

$f_{1,k}$ The ratio of the delay limits between class-1 and class-k; $k = 1, 2, \ldots, K$. Particularly, in the two-class case, $f = f_{1,2}$.

α_k The weights for the expected excess waiting for class-k customers; $k = 1, 2, \ldots, K$.

$m_k^{(2)}$ The second moment of the waiting time distribution for class-k customers; $k = 1, 2, \ldots, K$.

G_k The expected excess waiting for class-k; $k = 1, 2, \ldots, K$ customers integrated over the whole range of the class-1 delay limit.

π_{busy} The probability that all servers are simultaneously busy in a stationary $M/M_i/c$ queue.

Chapter 1

Introduction

1.1 Background and motivation of this thesis

Long waiting time has been a serious problem for a very long time in many public systems, especially in health care systems. The issues caused by long waiting time are getting more and more attention from researchers. In most cases, a health care (or other) system inclines to separate patients (customers) into different priority groups according to their urgencies for commencement of service. In order to reduce waiting time, many different techniques have been studied by researchers, where classical priority queuing models appear to be the most popular approach in such a system. However, a classical priority queue, which only selects a customer of a given class when no customers of higher priority classes are waiting, can cause the customers from lower classes to experience extremely long waiting times which may result in serious outcomes. For example, the classical priority queuing discipline is commonly used by medical physicians and decision makers in the health care systems where the Canadian Triage and Acuity Scale (CTAS) [1] is applied. The CTAS classifies patients into five distinct priority classes according to the Key Performance Indicators (KPIs), which specify a time limit and a corresponding compliance probability for each class of patients. An extended wait, caused by the classical priority discipline, for the patients from lower priority classes may lead to critical consequences, for instance, a deterioration of a patient's condition or even death.

In 1964, Kleinrock [4] proposed a time-dependent priority queuing model, which was motivated by a problem in the context of computer processor design. He recommended that customers may accumulate priority as a linear function of their waiting time in the queue, at a rate that reflects their urgency or classification. Consequently, in such a queue, the customers from a lower-priority class will eventually earn enough priority to enter service, in all likelihood, at an earlier point in time than in a classical priority queue when their waiting times were ignored. He derived a set of expressions for the expected waiting times for different classes, under the assumptions of Poisson arrivals and a single server working at an exponential service rate.

Motivated by the applications of KPIs, Stanford, Taylor and Ziedins [11] reconsidered Kleinrock's model, renamed as the "accumulating priority queue" (APQ). They derived the waiting time distributions for different priority classes in a single server system with general service times. This discipline provides a more balanced approach to select customers for service, and consequently better regulates wait times for various classes of customers, by allowing customers to accumulate priority credit while they wait. As an extension of Stanford *et al.* [11], Sharif *et al.* [10] considered a multi-class multi-server APQ with Poisson arrivals and a common exponential service distribution, where waiting time distribution for each class was derived. Both Stanford *et al.* [11] and its successor Sharif *et al.* [10] showed that by varying the rates of priority accumulation for different classes, the time limits and corresponding compliance probabilities stated in KPIs can be met in an accumulating priority queue, which might not possibly to be achieved in a classical priority system.

This thesis extends the research on the accumulating priority queues to three directions: APQs with heterogeneous servers, APQs with nonlinear accumulation functions, and optimization problems for APQs. All three aspects are motivated by the issues appearing in the real systems.

The first direction in which this thesis extends current theory pertains to the case where heterogeneous servers are present. In many situations, most service sections have multiple servers working simultaneously. A common assumption in such multi-server queuing models, which is frequently violated in reality, is that all servers are working identically at the same speed. It is more reasonable to take the heterogeneity among different servers into account when constructing a queuing model to describe real systems, e.g. a health care system.

By taking account of the heterogeneity among servers in the analysis of APQs, a multi-class multi-server queue with Poisson arrivals and heterogeneous exponentially distributed service times under the APQ and related queuing disciplines is investigated in the first part of this thesis. Different service dispatch policies among the servers are considered in order to provide the decision makers more flexibility to manage the system. The waiting time distributions for different classes are derived in such a model. We also formulate an optimization problem to find the optimal level of heterogeneity, and discuss how the APQ approach would affect the optimal solution in terms of minimizing a well-defined cost function.

The second direction in which we build upon existing theory pertains to how priority is accumulated over time. We extend the accumulating priority queues of Stanford *et al.* [11] to allow a customer's priority to accumulate as a nonlinear function of its waiting time, which is motivated by Kleinrock and Finkelstein [5]. In 1967, Kleinrock and Finkelstein [5] proposed a queuing model where customers accumulate priority as a power r of their waiting time. We refer to such queues as the power-law APQs. They stated that an rth order system is equivalent to a linearly-increasing priority system, in the sense that the expected waiting times for customers of all classes are the same. Invoking the results of Kleinrock [4], they obtained a set of expressions for the expected waiting times for different classes in a given rth order system.

We then focus on the study of a multi-class APQ with Poisson arrivals, general service times and a class of nonlinear priority accumulation functions. The work is initiated by the analysis of the power-law APQ where the waiting time distributions of all classes of customers are obtained in an APQ setting with the power-law priority accumulation functions in Kleinrock and Finkelstein [5]. Subsequently, we show that when certain conditions are met, we can create an equivalent linear APQ with the same set of waiting time distributions of different classes. Such an extension provides the policy makers a much wider range of priority accumulation functions, where the patients' waiting times as well as their urgencies of treatments can be taken into consideration.

Table 1.1: CTAS Key Performance Indicators (KPIs)

Level	Level of acuity	Response time	Sample diagnosis	Targets
1	Resuscitation	Immediate	Cardiac arrest	98%
2	Emergent	< 15 mins	Chest pain	95%
3	Urgent	< 30 mins	Moderate asthma	90%
4	Less urgent	< 60 mins	Minor trauma	85%
5	Non urgent	< 120 mins	Common cold	80%

The last direction in which the thesis extends upon the existing work pertains to the matter of optimality of APQ systems. Optimization problems are formulated for the queues under waiting time limits, such as the KPIs in Table 1.1. As stated earlier, KPIs are widely used to regulate the health care systems in Canada, not only for "visible" queues but also for wait lists. The time limits and the corresponding compliance probabilities in KPIs were determined by medical professionals according to the clinical need of different patient classes from the history, prior to any consideration of the traffic characteristics of the patient classes (i.e., frequency of demand, treatment time distributions). A simple patten can be observed from Table 1.1, such that for the patients from Level 2 and onwards, the time limit for the next level doubles the one above and the compliance probability is simply reduced by 5%. The systems which meet the KPIs are considered to be in compliance. Even in such systems, the small percentage of patients who miss their time limits tend to be ignored, with no consequence specified for them. Certainly, it is irrational and unacceptable to neglect these patients; on the contrary, they should become a greater concern to the system managers.

Thus, proper optimization functions to minimize the delays beyond the time limits for various classes of patients are established to achieve an optimal performance of a KPI system. The optimization problems we define address this fundamental oversight with objective functions that seek to quantify how much excess is occurring under a given queuing discipline. We mainly consider three queuing disciplines: the first-come, first-served (FCFS) discipline, the classical priority discipline and the accumulating priority discipline. We do so initially by seeking to minimize the total expected amount of waiting time excess; in the sequel, we minimize a weighted sum of the expected amount of excess waiting for each class. We will be particularly interested in the performance of a "Rule of Thumb" that we propose, which ensures that all patients have the same total amount of accumulated priority credits when they reach their respective time limits, and are thereby considered to be equally urgent, as desired in the systems operating under KPIs.

1.2 Outline of this thesis

As discussed in the last section, this thesis consists of three major topics on the accumulating priority queues. The reminder of this thesis is arranged as follows.

A detailed review of related literature is presented in Chapter 2, including the work on accumulating priority queues, the researches on heterogeneous FCFS queues, the conservation laws in the single-server systems and a discussion of the Gaver-Stehfest numerical inversion algorithm.

In Chapter 3, the multi-class accumulating priority queuing model with heterogeneous servers is specified. A conservation law for the multi-class heterogeneous-server systems is presented, and following by a discussion of the calculations for the stationary probabilities in such a system. An optimization problem for the optimal level of heterogeneity is established for different service dispatch policies. The waiting time distribution for each priority class in the multi-class heterogeneous system under APQ and related queuing disciplines is derived. Finally, various numerical investigations are explored to address the impact of the level of heterogeneity in multi-server systems, as well as the advantage of the APQ approach.

In Chapter 4, the accumulating priority queue with nonlinear accumulation functions is described. With an initial discussion on the waiting times for the power-law APQ, the linkage between the APQs with a more general class of nonlinear accumulation functions and the linear APQ is discovered and proved. A set of recursive formulas for the waiting time distributions for the nonlinear APQ with a linear proxy are derived.

Chapter 5 focused on optimization problems of queues operating under waiting time limits. An optimization problem to minimize the weighted average of total expected excess is formulated, followed by a related optimization problem to resolve the difficulties of finding the optimality of the original objective function. Extensive numerical calculations have been conducted to investigate the behavior of the objective functions. Finally, a “rule of thumb” is introduced, which can be easily applied to the real systems by medical professionals.

The main contributions are summarized in Chapter 6, as well as some future research directions.

References

[1] Canadian Association of Emergency Physicians. The Canadian Triage and Acuity Scale (CTAS). From the website. *http ://www.calgaryhealthregion.ca/policy/docs/*1451*/Admission_over_capacity_AppendixA.eps*.

[2] Grassmann, W., & Zhao, Y. Q. (1997). Heterogeneous multiserver queues with a general input. INFOR. 35, 208–224.

[3] Gumbel, H. (1960). Waiting lines with heterogeneous servers. *O*perations Research. 8(4), 504–511.

[4] Kleinrock, L. (1964). A delay dependent queue discipline. Naval Research Logistics Quarterly. 11, 329–341.

[5] Kleinrock, L., & Finkelstein, R. (1967). Time dependent priority queues. *O*perations Research. 15, 104–116.

[6] Krishnamoorthi, B. (1963). On Poisson queue with two heterogeneous servers. *O*perations Research. 11, 321–330.

[7] Mokaddis, G. S., & Matta, C. H. (1998). On Poisson queue with three heterogeneous servers. Information and Management Sciences. 9, 53–60.

[8] Saaty, T. L. (1960). Time dependent solution of the many-server Poisson queue. *O*perations Research. 8, 768–771.

[9] Shalit, H. (1985). Calculating the Gini index of inequality for individual data. Oxford Bulletin of Economics and Statistics. 47, 185–189.

[10] Sharif, A. B., Stanford, D. A., Taylor, P. & Ziedins, I. (2014). A multi-class multi-server accumulating priority queue with application to health care. *O*perations Research for Health Care 3(2), 73–79.

[11] Stanford, D. A., Taylor, P. & Ziedins, I. (2014). Waiting time distributions in the accumulating priority queue. Queueing Systems 77(3), 297–330.

Chapter 2

Preliminaries

In this chapter, we start with a brief introduction of the Laplace-Stieltjes transforms, then following with a thorough review of the literature on the accumulating priority queues, including the time-dependent priority queue in Kleinrock [17], the *r*th order system in Kleinrock and Finkelstein [19] which we refer to as the power-law APQ, the single-server APQ in Stanford *et al.* [29], the homogeneous multi-server APQ in Sharif *et al.* [26] and the work on preemptive APQs in Fajardo's PhD thesis [7]. Previous researches on the FCFS queues with heterogeneous servers are reviewed in this chapter, as well as the conservation laws in single-server systems. Followed by a review of the Poisson mapping theorem in Kingman [15], the Gaver-Stehfest numerical inversion algorithm (Gaver [9], Stehfest [30]) and a modified version proposed in Abate and Whitt [1] are discussed in the last section.

The models we consider in this thesis operate under fairly standard assumptions in the queuing literature, which we now enumerate. First and foremost of these is that all of our models constitute non-preemptive and work-conserving queues. A work-conserving discipline is one in which the work requirements of customers remain unchanged by the passage of time, customers neither balk nor renege, and servers are never idle if there is anyone waiting. The non-preemptive discipline is such that a customer entering service remains in service without interruption until completion. The second most important is that all the models are operating in a stable regime; that is, the long-run service capacity exceeds the long-run demand. Other standard assumptions include an infinite customer population, an unbounded waiting area, and constant arrival and service rates over time. Moreover, we presume the queues have operated sufficiently long to have reached stationarity.

These assumptions need to be kept in mind when the models are applied in the health care setting. In particular, it is frequently the case that healthcare queues are operating close to 100% utilization, and may not operating in a stable regime, and furthermore that balking, reneging, and reordering of priorities frequently occur due to changes in patient health status. The decision maker needs to take account of such realities when making inferences from the numerical results these models provide.

2.1 Laplace-Stieltjes transforms

Laplace-Stieltjes transforms (LSTs) are widely used in probability, particularly the moment generating function. The LST is well suited when dealing with the distribution function of a nonnegative random variable (r.v.), which we will use extensively in the following chapters in this thesis.

The Laplace-Stieltjes transform of a function F defined in Feller [8] is given by

$$\tilde{f}(s) = \int_0^\infty e^{-sx} dF(x) \tag{2.1}$$

for all s for which this integral converges. It can be found immediately from the definition of the LST that $\tilde{f}(0) = 1$ for any distribution function F.

From equation (2.1), the LST of the distribution function F of a nonnegative r.v. X can be written as $\tilde{f}(s) = E(e^{-sX})$. Moreover, if $\tilde{f}(s)$ is n times differentiable at the origin, then $E(X^n) = (-1)^n \tilde{f}^{(n)}(s)$.

If F has a density function f, then equation (2.1) reduces to the Laplace transform

$$\tilde{f}(s) = \int_0^\infty e^{-sx} f(x) dx. \tag{2.2}$$

LSTs are very useful in queuing theory (e.g., the analysis for $M/G/1$ queues in Conway *et al.* [3]). One reason is that the waiting times in a queuing system generally have a point mass at zero, representing the probability of the system being idle, while the remaining mass is distributed across $(0, \infty]$. Furthermore, the original function F can be numerically computed to arbitrary precision from a given LST $\tilde{f}(s)$ using numerical inversion algorithms, which we would introduce in the last section of this chapter.

2.2 Accumulating priority queues

This queuing discipline was first proposed by Kleinrock [17] as "time-dependent priority queue", and recently developed by Stanford *et al.* [29] and Sharif *et al.* [26]. In all three papers, they assume:

There are $K \in \mathbb{N}$ classes of customers. Customers of class k arrive independently to the queue according to a Poisson process with rate $\lambda_k, k = 1, 2, \ldots, K$. From the moment of their arrival, a customer of class k accumulates priority at rate b_k where $b_1 \geq b_2 \geq \cdots \geq b_K \geq 0$. Thus a customer of class k arriving to the queue at time t will have accumulated priority as a linear function $b_k(t' - t)$ by time t'. At a service completion instant, the next waiting customer to be served will be selected according to the greatest accumulated priority in the queue at that instant.

Different assumptions of the service durations are made among the three models above. Kleinrock [17] studied a single-server queuing model with exponentially-distributed service times. Stanford *et al.* [29] extended Kleinrock's model to a single server working under general service distributions, while Sharif *et al.* [26] considered a multi-server model with homogeneous exponential service times.

On the other hand, Kleinrock and Finkelstein [19] extended the model in Kleinrock [17] to another direction, in which a customer's priority increases in proportion to some arbitrary nonnegative power r of its waiting time, and named it an "rth order delay dependent priority discipline". They, too, studied the single-server case with exponential service times.

In this thesis, we apply Kendall's notation, in the form of $a/b/c$, to classify the different queuing models we consider. The letter a represents the inter-arrival time distribution, b does the same for the service duration distribution, and c denotes the number of servers. Furthermore, we use "linear APQ" to indicate the system with linear accumulation functions, whereas "nonlinear APQ" for the system with general nonlinear accumulation functions. Thus, the models in Kleinroch [17], Stanford *et al.* [29] and Sharif *et al.* [26] can be represented as the $M/M/1$ linear APQ, $M/G/1$ linear APQ and $M/M/c$ linear APQ respectively, whereas we refer to the model in Kleinrock and Finkelstein [19] as the "$M/M/1$ power-law APQ of order r". The main results of these developments on accumulating priority queues are described below.

2.2.1 Average waiting times in APQs

The $M/M/1$ linear APQ

Kleinrock [17] derived a set of recursive formulas for the average waiting times of various customer classes in an $M/M/1$ linear APQ, where the customers from class-k has a required service time selected from an exponential distribution with mean $1/\mu_k$.

For $k = 1, 2, \ldots, K$, they obtained the average waiting time for class-k, denoted by m_k, recursively for $k = K, K-1, \ldots, 1$ as

$$m_K = \frac{W_0/(1-\rho)}{1-\sum_{j=1}^{K-1}\rho_j(1-b_K/b_j)}, \quad \text{and} \tag{2.3}$$

$$m_k = \frac{W_0/(1-\rho)-\sum_{j=k+1}^{K}\rho_j m_j(1-b_j/b_k)}{1-\sum_{j=1}^{k-1}\rho_j(1-b_k/b_j)}, \tag{2.4}$$

where $W_0 = \sum_{k=1}^{K}\rho_k/\mu_k$, $\rho_k = \lambda_k/\mu_k$ and $\rho = \sum_{k=1}^{K}\rho_k < 1$. Although Kleinrock [17] constructed this analysis for the $M/M/1$ APQ, it is equally applicable to the $M/G/1$ case where the form of W_0 is given by equation (2.14). Lastly, he performed a set of numerical calculations for the average waiting times in the accumulating priority and FSFC disciplines.

The $M/M/1$ power-law APQ of order r

Kleinrock and Finkelstein [19] studied the average waiting time for each class of customers in a single-server APQ with Poisson arrivals, exponential service times, and a set of power-law accumulation functions. The priority accumulation functions for the power-law APQ of order r is defined in terms of a sequence$\{b_k^{(r)}\}, k = 1, \ldots, K$ of positive constants such that $b_1^{(r)} \geq b_2^{(r)} \geq \cdots \geq b_K^{(r)} \geq 0$, and with a function form of $b_k^{(r)} t^r$ for all k.

They established in Kleinrock and Finkelstein [19, Theorem 1] that if one were to select the constants so that

$$\left(b_{k+1}^{(r)} / b_k^{(r)}\right)^{1/r} = \left(b_{k+1}^{(r')} / b_k^{(r')}\right)^{1/r'} \quad \text{for } k = 1, 2, \ldots, K, \tag{2.5}$$

then the expected waiting times of all customer classes in the corresponding power-law APQs of orders r and r' would be identical.

From this, using the results in Kleinrock [17] for the first order systems, they obtained the expected waiting times for different classes of customers in the power-law APQ of order r.

2.2.2 Waiting time distributions in APQs

The $M/G/1$ linear APQ

Stanford *et al.* [29] determined the waiting time distributions for each class of customers in a single-server linear APQ with Poisson-arrivals and general service distributions.

A key element in their derivation was the stochastic process named the *maximum priority process*, $\boldsymbol{M} = \{M_i(t);\ t \geq 0,\ i = 1, \ldots, K\}$, which gave the least upper bound of the accumulated priority $M_i(t)$ for each priority class at given instant in time.

They began with the accumulating priority queue in the two-class case with $\boldsymbol{M} = \{M_1(t), M_2(t)), t \geq 0\}$, where $M_i(t); i = 1, 2$ is an upper bound on the possible value of the maximal accumulated priority for a class-i customer by time t; moreover, $M_1(t) \geq M_2(t)$ for all t. A class-1 customer in the queue with accumulated priority at time t that lies in the interval $(M_2(t), M_1(t)]$ is referred to as "accredited relative to class-2", or simply, "accredited", since it is certain that there are no class-2 customers in the system with as much priority. Those customers with priority in the interval $[0, M_2(t))$ are referred to as "non-accredited". An accreditation interval is defined as consisting of the service time of a non-accredited customer followed by a sequence of service times of class-1 customers who have become accredited during the interval, until there are no accredited customers left. During an accreditation interval, the instants that customers become accredited constitute a Poisson process with rate $\lambda_1(1 - b_2/b_1)$. By analogy to similar constructs in the classical $M/G/1$ queue, they obtained the expression of the Laplace-Stieltjes transform (LST) of the duration of an accreditation interval and its mean duration.

Customers who are selected for service during an accreditation interval gain additional credit, up to the point in time when they enter service. Consider a random variable $\hat{V}$ to be the accumulated priority of a customer at the point that it enters service during an accreditation interval. Suppose the accreditation interval commences at time t_0. The random variable $\hat{V}$ can be written as $\hat{V} = V_{init} + V$ where V is the additional priority that the customer accumulates during the accreditation interval, after having accumulated priority V_{init}. Let $B^{(i)}$ denote the service time r.v. for the customers from class i, $i = 1, 2$. When the service times are the same for the two classes, set $B^{(1)} = B^{(2)} = B$. By modifying the delay cycle approach of Conway *et al.* [3, page 151], they obtained the LST of the distribution of V, given parameters b_1, b_2, λ_1 and B,

$$\tilde{V}^*(s; b_1, b_2, \lambda_1, B) = \frac{(\mu - \lambda_1(1 - \frac{b_2}{b_1}))(\tilde{\Gamma}(b_2 s) - \tilde{B}(b_1 s))}{(1 - \frac{b_2}{b_1})(b_1 s - \lambda_1(1 - \tilde{B}(b_1 s)))},$$

where $\tilde{\Gamma}(s)$ is the LST of the duration of an accreditation interval, which is the solution of $\tilde{\Gamma}(s) = \tilde{\Gamma}(s; b_1, b_2, \lambda_1, B) = \tilde{B}(s + \lambda_1(1 - b_2/b_1)(1 - \tilde{\Gamma}(s)))$. (They also discussed the case when $B^{(1)} \neq B^{(2)}$ which we choose not to review here.)

The non-accredited customers can be equivalently viewed as all arriving to the system in an aggregate Poisson process with rate $\lambda_2 + \lambda_1 b_2/b_1$ and all accumulating priority at rate b_2. Then, the LST of the stationary accumulated priority of the non-accredited customers at the time that they enter service, conditional on it being positive, is given by the accumulated priority distribution with parameters b_2, 0, $\lambda_2 + \lambda_1 b_2/b_1$, Γ as

$$\tilde{V}^{(2)}(s) = \tilde{V}(s; b_2, 0, \lambda_2 + \lambda_1 b_2/b_1, \Gamma).$$

Thus the LST of the stationary waiting time for class-2 customers is given by,

$$\tilde{W}^{(2)}(s) = (1 - \rho) + \rho \tilde{V}^{(2)}(s/b_2). \tag{2.6}$$

The probability that a class-1 customer, arriving during a busy period, becomes accredited is $(b_1 - b_2)/b_1$, while the probability that it enters service while unaccredited is b_2/b_1. Thus, the LST of the distribution of the priority of a class-1 customer when it enters service, conditional on this being positive, is

$$\tilde{V}^{(1)}(s) = \frac{b_2}{b_1}\tilde{V}^{(2)}(s) + \left(\frac{(1-\rho)(b_1-b_2)}{b_1(1-\sigma_1)} + \frac{(\rho-\sigma_1)(b_1-b_2)}{b_1(1-\sigma_1)}\tilde{V}^{(2)}(s)\right)\tilde{V}^{(1,0)}(s),$$

where $\tilde{V}^{(1,0)}$ is the LST of the stationary accumulated priority of a class-1 customer, and

$$\tilde{V}^{(1,0)}(s) = \tilde{V}(s; b_1, b_2, \lambda_1, B).$$

Finally, based upon the same logic as was used for class-2, the LST of the waiting time for class-1 customers is

$$\tilde{W}^{(1)}(s) = (1-\rho) + \rho\tilde{V}^{(1)}(s/b_1). \tag{2.7}$$

The authors also obtained a recursive equation to find the LST of the waiting time distribution for class-k, conditional upon it being positive, in a multi-class single server system,

$$\tilde{W}_+^{(k)}(s) = \left(\frac{b_{k+1}}{b_k}\right)\tilde{W}_+^{(k+1)}\left(\frac{b_{k+1}}{b_k}s\right) + \left(1 - \frac{b_{k+1}}{b_k}\right)\tilde{W}_{acc}^{(k)}(s), \tag{2.8}$$

to which we refer with greater detail later when we present our work to a multi-class APQ with heterogeneous servers.

The $M/M/c$ linear APQ

Sharif *et al.* [26] considered a multi-class multi-server linear APQ with Poisson arrivals and a common exponential service distribution for all classes of customers with equal service rates, i.e., $\mu_1 = \mu_2 = \cdots = \mu_K = \mu$. They also commented on how to choose feasible accumulation rates to satisfy KPI targets in CTAS [2].

They presented the LST of the waiting time distribution function, $\tilde{W}^{(k)}(s)$, for customers of class k; $k = 1, \ldots, K$,

$$\tilde{W}^{(k)}(s) = (1 - C(A,c)) + C(A,c)\tilde{W}_+^{(k)}(s; \mu, c), \tag{2.9}$$

where $C(A, c)$ denotes the Erlang-C delay probability; that is, the probability that all servers are simultaneously busy in a stationary $M/M/c$ queue with $A = \lambda/\mu$ and $\lambda = \sum_{k=1}^{K} \lambda_k$, and where $\tilde{W}_+^{(k)}(s)$ is the LST of the class-k waiting distribution, conditional on it being positive.

Sharif *et al.* [26] established that the LST of the waiting time distribution, conditional upon it being positive, in a homogeneous multi-server queue with common mean $1/\mu$ is the same as that in a single server queue with a exponential service rate of $c\mu$:

$$\tilde{W}_+^{(k)}(s; \mu, c) = \tilde{W}_+^{(k)}(s; c\mu, 1); \qquad k = 1, \ldots, K. \tag{2.10}$$

The results were evaluated through several numerical investigations, and an algorithm was presented for finding the maximum ρ and optimal b, which are the maximum amount of the utilization level and the optimal value of the accumulation rate for which a given set of KPI targets can be met.

2.2.3 Other work on APQs

In Fajardo's PhD thesis [7], they investigated two types of preemptive linear APQs. The first one they considered is the fully preemptive variant of the linear APQ in Stanford *et al.* [29], where a recursive procedure for obtaining the waiting time distributions was developed. They studied the waiting time distributions for this preemptive model under each of the three traditional preemption disciplines (i.e., resume, repeat-different and repeat-identical), as well as for a new hybrid-based preemption discipline which they called the Bernoulli-based decision for resumption of service discipline.

The second linear APQ they investigated is similar to the model in Stanford *et al.* [29], but incorporates the notion of urgent-type customers whose arrivals may preempt lower priority customers and whose priority is assigned in the classical (or static) sense. Furthermore, this model generalizes several previously-analyzed static priority models including the classical preemptive / non-preemptive models, the static priority model under a preemption distance rule, and the priority model under threshold-based discretion rules considered by Drekic and Stanford [5]. Finally, they established that the models can be treated as the $M/G/1$ model under a customer blocking policy, namely the q-policy. They found that the level-crossing methodology may provide a nice interpretation for each components in the LST of the additional priority accumulated by an accredited customer, by which they presented an alternate proof to the main theorem in Stanford *et al.* [29].

Haviv and Ravner [13] studied the strategic purchasing of priorities in an $M/G/1$ linear APQ. They formulated a non-cooperative game in which customers purchase priority coefficients (i.e., the slope of the linear accumulation function) with the goal of reducing waiting costs in exchange, where the unique pure Nash equilibrium was obtained explicitly for the case with homogeneous customers, and a general characterisation of the pure Nash equilibrium was provided for the heterogeneous-customer case.

2.3 FCFS queues with heterogeneous servers

Many papers have studied the queuing systems with heterogeneous servers. The pioneering work of Morse [23] in 1958 was the first to consider the analysis of multi-server systems (although his main focus was on homogeneous case). Saaty [24] found the LST of the transient probabilities of multi-server FCFS queues with Poisson arrivals and exponential service times at a common rate, where the explicit expressions for the stationary probabilities were derived for the two-server case. Subsequently, he discussed the LST of the transient probabilities in the two-server system with two different exponential service distributions operating under a specific service dispatch policy (i.e., when both servers are idle, the server is chosen by an arriving customer according to the proportion of its service rate to the total service rate).

In 1960, Gumbel [12] discussed the multi-server heterogeneous FCFS queue under the randomly chosen server (RCS) dispatch policy. He derived the expressions in closed form of the steady state probabilities and the expected queue length assuming the steady-state condition. He also pointed out there is no “equivalent” system with homogeneous servers present and analyzed the error incurred from the assumption that each server is working at an equal rate which is the average of all the service rates. If the heterogeneous exponential service rates for various servers are replaced by the average service rate, the probability of the system being busy (or, delayed probability) is inconsistent with the one in the original heterogeneous system; whereas if the delayed probabilities are the same in both systems, the common service rate in the respective homogeneous queue is not equal to the mean of the service rates in the heterogeneous system.

Krishnamoorthi [21] studied a Poisson queue with two heterogeneous servers with modified allocation disciplines. Singh [28] analyzed a Poisson queue with three heterogeneous servers, where an optimal combination of the service rates to minimize the performance measures of the system was presented. Sharma and Dass [27] considered the $M/M/2/N$ queue with heterogeneous servers to derive the probability density function of the busy period. Mokaddis and Matta [22] studied a Poisson-arrival queue with three heterogeneous servers under various allocation policies. Grassmann and Zhao [11] studied a queuing system with multiple heterogeneous servers and general inputs, where the steady state probabilities were calculated using different rules to allocate the arriving jobs.

In a queuing system with heterogeneous servers, the service dispatch policy, also known as the allocation policy, is defined as a method that determines which idle server is assigned to the next arriving job. While we consider a fairly general rule in this regard, we are particularly interested in four specific dispatch policies: 1) "randomly chosen server" (RCS), which assigns the next job to any of the idle servers with equal likelihood; 2) "fastest server first" (FSF); 3) "slowest server first" (SSF); and 4) "rate balancing selection" (RBS), where an idle server will be chosen according to its proportion of the total service rate among all idle servers. Steady state probabilities have been discussed in the past researches of the FCFS systems with heterogeneous servers under different dispatch policies, which are presented in what follows.

2.3.1 The classical heterogeneous queue under RCS

Gumbel [12] derived the stationary state probabilities under the assumptions of Poisson arrivals and exponentially distributed service times, with a different service rate for each server under RCS.

Let π_n be the steady state probability of n customers ($n \geq 0$) in the system. He derived that

$$\pi_n = \begin{cases} (c-n)! C^c_{c-n} \pi_c, & (0 \leq n < c) \\ \pi_c / (C^c_1)^{n-c}, & (n \geq c) \end{cases}$$

where

$$C^c_n = \sum_{a_1=1}^{c-n+1} \sum_{a_2=a_1+1}^{c-n+2} \cdots \sum_{a_{n-1}=a_{n-2}+1}^{c-n+n-1} \sum_{a_n=a_{n-1}+1}^{c} \frac{\mu_{a_1}\mu_{a_2}\cdots\mu_{a_n}}{\lambda^n};$$

where a_i are members of sets of k indices out of c indices. Note that $C^c_c = \Pi^c_{i=1}\frac{\mu_i}{\lambda}$, $C^c_1 = \sum^c_{i=1}\frac{\mu_i}{\lambda} = 1/\rho$ and $C^c_0 = 1$.

He also showed that an "equivalent" system with homogeneous servers, which has the same steady state probabilities as the heterogeneous case, does not exist. Thus, the device of replacing the unequal servers by an equal number whose mean service rate is the arithmetic mean of the individual service rates leads to computational errors. The error incurred in assigning each server the arithmetic mean of the service rates is analysed and illustrated using expected number in the system as the criterion of comparison.

2.3.2 Heterogeneous queues under general dispatch polices

In 1963, Krishnamoorthi [21] studied the Poisson queue with two heterogeneous servers, where the steady state probabilities were given. Subsequently, in 1998, Mokaddis and Matta [22] obtained the steady state probabilities of a Poisson queue with three heterogeneous servers. They both solved the probabilities via differential-difference equations in terms of state probabilities. Here, we present the work using global balance equations.

Figure 2.1 illustrates the state transition diagram with two heterogeneous servers, where customers arrive to servers according to Poisson distribution with rate λ and the two servers' service times follow exponential distribution with rates μ_1 and μ_2 respectively where $\mu_1 \geq \mu_2$. The states $(1,0)$ and $(0,1)$ in the figure means that one server is busy in the system, with the position of the 1 indicating which server is busy. When a customer arrives to an idle system, it can be served by server 1 with probability p_1 or by server 2 with probability p_2. By the

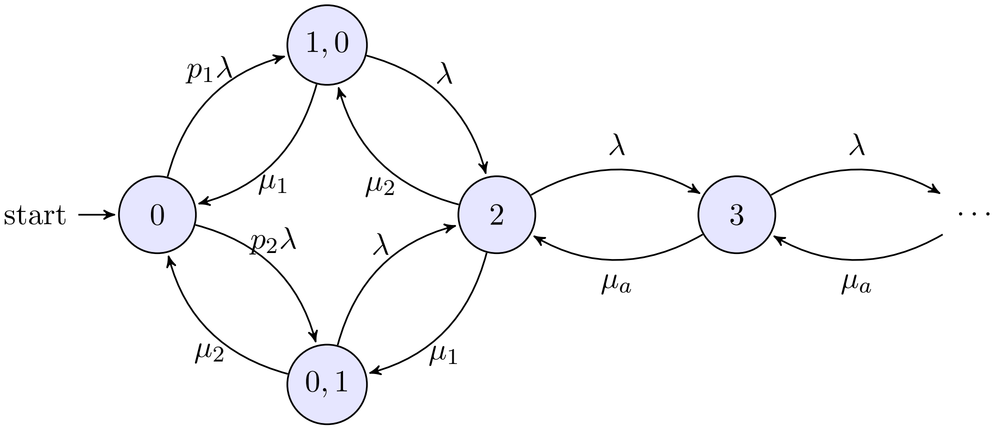


Figure 2.1: State-transition-rate diagram with two heterogeneous servers.

properties of Poisson distribution, the transition rate from state 0 to state $(1, 0)$ is $p_1\lambda$ and the transition rate from state 0 to state $(1, 0)$ is $p_2\lambda$. By global balance, we have the following equations:

$$\begin{aligned}
\lambda\pi_0 &= \mu_1\pi_{1,0} + \mu_2\pi_{0,1},\\
(\lambda + \mu_1)\pi_{1,0} &= p_1\lambda\pi_0 + \mu_2\pi_2,\\
(\lambda + \mu_2)\pi_{0,1} &= p_2\lambda\pi_0 + \mu_1\pi_2,\\
\pi_{n+1} &= \rho\pi_n, \qquad n \geq 2.
\end{aligned}$$

Then, the steady state probability π_n can be solved,

$$\begin{aligned}
\pi_0 &= \frac{(2\rho + 1)(1 - \rho)}{(\mu_1/\mu_2 + \mu_2/\mu_1)\rho^2 + (2 + p_1\mu_2/\mu_1 + p_2\mu_1/\mu_2)\rho + 1},\\
\pi_{1,0} &= \frac{\rho(1 + \mu_2/\mu_1)(\rho + p_1)(1 - \rho)}{(\mu_1/\mu_2 + \mu_2/\mu_1)\rho^2 + (2 + p_1\mu_2/\mu_1 + p_2\mu_1/\mu_2)\rho + 1},\\
\pi_{0,1} &= \frac{\rho(1 + \mu_1/\mu_2)(\rho + p_2)(1 - \rho)}{(\mu_1/\mu_2 + \mu_2/\mu_1)\rho^2 + (2 + p_1\mu_2/\mu_1 + p_2\mu_1/\mu_2)\rho + 1},\\
\pi_2 &= \frac{\lambda^2(\lambda + \mu_1 p_2 + \mu_2 p_1)}{\mu_1\mu_2(2\lambda + \mu_a)}\pi_0,\\
\pi_n &= \rho^{n-2}\pi_2, \qquad n \geq 2.
\end{aligned} \tag{2.11}$$

Let us consider the three heterogeneous servers case with Poisson arrivals and exponential service times. When a customer arrives to a totally idle system which is state 0 in Figure 2.2,

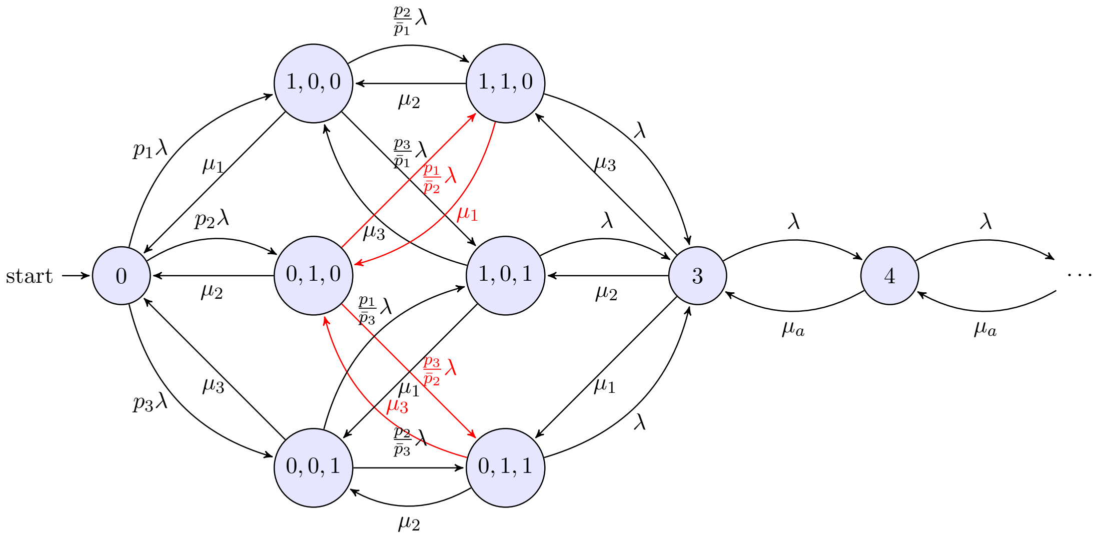


Figure 2.2: State-transition-rate diagram with three heterogeneous servers.

it can be assigned to server i with probability p_i for $i = 1, 2, 3$. A partially idle system in this case involves the six states from state $(1, 0, 0)$ to state $(0, 1, 1)$ in the figure. Assuming the busy server as a customer arrives is server 1, then the system is currently in state $(1, 0, 0)$. At the instant of a customer arrives to state $(1, 0, 0)$, it can be assigned to server 2 with probability $p_2/\bar{p}_1$ and to server 3 with $p_2/\bar{p}_1$ where $\bar{p}_1 = p_2 + p_3$. All the state transition rates are shown in Figure 2.2. Then, by global balance, we have,

$$\begin{aligned}
\lambda\pi_0 &= \mu_1\pi_{1,0,0} + \mu_2\pi_{0,1,0} + \mu_3\pi_{0,0,1}, \\
(\lambda + \mu_1)\pi_{1,0,0} &= p_1\lambda\pi_0 + \mu_2\pi_{1,1,0} + \mu_3\pi_{1,0,1}, \\
(\lambda + \mu_2)\pi_{0,1,0} &= p_2\lambda\pi_0 + \mu_1\pi_{1,1,0} + \mu_3\pi_{0,1,1}, \\
(\lambda + \mu_3)\pi_{0,0,1} &= p_3\lambda\pi_0 + \mu_1\pi_{1,0,1} + \mu_2\pi_{0,1,1}, \\
(\lambda + \mu_1 + \mu_2)\pi_{1,1,0} &= \frac{p_2}{\bar{p}_1}\lambda\pi_{1,0,0} + \frac{p_1}{\bar{p}_2}\lambda\pi_{0,1,0} + \mu_3\pi_3, \\
(\lambda + \mu_1 + \mu_3)\pi_{1,0,1} &= \frac{p_1}{\bar{p}_3}\lambda\pi_{0,0,1} + \frac{p_3}{\bar{p}_1}\lambda\pi_{1,0,0} + \mu_2\pi_3, \\
(\lambda + \mu_2 + \mu_3)\pi_{0,1,1} &= \frac{p_2}{\bar{p}_3}\lambda\pi_{0,0,1} + \frac{p_3}{\bar{p}_2}\lambda\pi_{0,1,0} + \mu_1\pi_3, \\
\pi_{n+1} &= \rho\pi_n, \qquad n \geq 3.
\end{aligned} \tag{2.12}$$

Mokaddis and Matta used Gauss-Jordan elimination method with 52 defined variables to solve the probabilities. The interested readers are directed to Mokaddis and Matta [22] for details. (However, we notice some errors may exist in their derivation.) The explicit expression for the stationary probability for each state has been solved with at most three heterogeneous servers in the queues under different dispatch policies, except for the RCS case considered above, in which an explicit solution is available for any number of servers.

In 2004, Grassmann and Zhao [11] studied a $GI/M_i/c$ queue with heterogeneous servers and general input, where equilibrium equations for the steady state probabilities were constructed, both at random-times and at the times preceding an arrival. They first determined the probabilities for all states in which there is no queue, and then calculated the probabilities in which customers are waiting. Three types of computational issues were addressed in their paper, namely the algorithm used to implement their formulas, the performance of the algorithm, and the accuracy of the results. Various algorithms were compared, however, they found that the correspondence between the results calculated by different methods was surprisingly high, normally, because of rounding errors. They stated that their algorithm run into performance problems as the number of servers increases. More specifically, the computational effort would increase by a factor of 8 if the number of servers increases by 1, which implies that even a large increase in computer speed would have only a marginal impact on the maximum size of the problem that can realistically be solved (e.g. less than 20 servers). In their paper, they performed the calculations for the systems with less than 10 servers. Some further discussion on the computational issues in Grassmann and Zhao [11] has been presented in Chapter 3.

2.4 Conservation laws

Conservation laws for the single-server queuing systems have been studied in the past literatures. The $M/G/1$ conservation law was first stated and proved in Kleinrock [16, 18] in 1964. Then in 1974, the conservation law was extended to the $G/G/1$ case in Schrage [25]. In 1980, Gelenbe and Mitrani [10] collected the works on the conservations laws and made a further development by introducing the “virtual load” concept.

In summary, the conservation laws state that in single-server work-conserving queues, the queuing discipline can only change the order of customers’s services, so that the total amount of work in the system is unaffected. Thus, the sum of mean waiting times for different classes of customers weighted by their respective occupancies is always a constant. Any attempt to improve the service for some class must inevitably come at the expense of some other class. In this section, we will review some important conservation laws.

2.4.1 The $M/G/1$ conservation law

Kleinrock [20, page 114] proved that for any $M/G/1$ system and any non-preemptive work-conserving queuing discipline it must be that

$$\sum_{k=1}^{K} \rho_k m_k = \begin{cases} W_0/(1-\rho) - W_0 = \rho W_0/(1-\rho), & \rho < 1; \\ \infty, & \rho \geq 1; \end{cases} \tag{2.13}$$

where $\rho_k = \lambda_k/\mu_k$ and $\rho = \sum_{k=1}^{K} \rho_k$. W_0, which represents the residual life of the customer found in service upon an arrival’s entry, is given by

$$W_0 = \sum_{k=1}^{K} \frac{\lambda_k \overline{x_k^2}}{2} \tag{2.14}$$

where $\overline{x_k^2}$ is the second moment of service time for a customer from class k. In other words, the conservation law states that the weighted sum of the average waiting times m_k for $k = 1, 2, \ldots, K$ can never change no matter how sophisticated or elaborate the queuing discipline may be.

2.4.2 The $G/G/1$ conservation law

The conservation law was extended (Schrage [25]) to the $G/G/1$ queues, where both the Poisson arrival assumption and the independence assumption were dropped. Kleinrock [20, page 117] presented the generalized version of the conservation, namely the $G/G/1$ conservation law: Given a specific work-conserving $G/G/1$ queuing system with a non-preemptive priority queuing discipline, then the linear equality constraint

$$\sum_{k=1}^{K} \rho_k m_k = \bar{U} - W_0 \tag{2.15}$$

must be satisfied regardless of the queuing discipline, where $\bar{U}$ is the limiting average of the unfinished work [20, page 114]. In the $M/G/1$ case, $\bar{U}$ equals $W_0/(1-\rho)$ as shown in equation (2.13).

2.4.3 A general conservation law

Gelenbe and Mitrani [10, page 174] introduced a stochastic process, named “virtual load”. For a particular queuing discipline S, the virtual load at time t, $V_S(t)$, was defined as the total amount of work in the system at time t, i.e., the sum of the remaining required service times for all jobs, including both the unfinished work and the residual service time, that are in the system at time t. They assumed the existence of the equilibrium distribution for $V_S(t)$ and denoted its steady-state average by V_S:

$$V_S = \lim_{t\to\infty} E[V_S(t)]. \tag{2.16}$$

They obtained the general conservation law in Theorem 6.1 [10, page 174]:

For any single-server queuing system in equilibrium there exists a constant V, determined only by the parameters of the arrival and required service times processes, such that

$$V_S = V \tag{2.17}$$

for all work-conserving queuing disciplines S.

By Theorem 6.1, with the condition that “only information about the current state and the past of the queuing process is used in making scheduling decisions”, they restated the $M/G/1$ conservation law in Kleinrock [20].

Theorem 6.2 [10, page 176] stated that when the required service times are distributed exponentially, there exists a constant V determined only by the inter-arrival time distributions and by the parameters μ, such that

$$\sum_{k=1}^{K} \rho_k m_k = V, \tag{2.18}$$

for all work-conserving queuing disciplines. Thus, by equation (2.13), the constant $V = \rho W_0/(1-\rho)$ for $\rho < 1$.

2.4.4 Modified $GI/G/1$ conservation law

With the references of Kleinrock [20] and Schrage [25], Theorem 6.3 in Gelenbe and Mitrani [10, page 177] presented the $GI/G/1$ conservation law:

For any multi-class $GI/G/1$ queuing system in the steady-state, there exists a constant V determined only by the inter-arrival and service time distributions, such that

$$\sum_{k=1}^{K} \rho_k m_k = V + \sum_{k=1}^{K} \rho_k \Big(\frac{1}{\mu_k} - \gamma_k\Big) \tag{2.19}$$

for all non-preemptive work-conserving queuing disciplines, where $\gamma_k = \mu_k \overline{x_k^2}/2$, for $k = 1, 2, \ldots, K$, is the average residual life of the class-k service time.

2.5 The Poisson mapping theorem

Kingman [15, page 17] introduced a great property of Poisson processes: in summary, if the state space is mapped into another space, the transformed random points again from a Poisson process.

He defined Π to be a Poisson process on the state space S, having mean measure μ, and f as a function from S into another (or the same) space T. They assumed that T, like S, is a measure space satisfying three conditions: 1) the empty set is measurable, 2) the complement of a measurable set is measurable, 3) the union of countably many measurable sets is measurable. They also assumed that f is measurable in the sense that

$$f^{-1}(B) = \{x \in S\,; f(x) \in B\} \tag{2.20}$$

is a measurable subset of S for every measurable $B \subseteq T$.

The points $f(x)$ for $x \in \Pi$ form a random countable set $f(\Pi) \subseteq T$. The number

$$N^*(B) = \#\{f(\Pi) \cap B\} \tag{2.21}$$

of points of $f(\Pi)$ in B, so long as the points $f(x)$ $(x \in \Pi)$ are distinct, and

$$N^*(B) = \#\{x \in \Pi; f(x) \in B\} = N(f^{-1}(B)) \tag{2.22}$$

which has distribution $\mathcal{P}(\mu^*)$, where

$$\mu^* = \mu^*(B) = \mu(f^{-1}(B)). \tag{2.23}$$

Moreover, if $B_1, B_2, \ldots, B_k$ are disjoint, so are their inverse images, so that the $N^*(B_j)$ are independent.

Finally, he proved the Poisson mapping theorem in Kingman [15, Mapping Theorem], which was stated as "Let Π be a Poisson process with σ-finite mean measure μ on the state space S, and let $f : S \to T$ be a measurable function such that the induced measure has no atoms. Then $f(\Pi)$ is a Poisson process on T having the induced measure μ^* as its mean measure." This property has profound implications, which has been applied in various journals (e.g. del Barrio *et al.* [4], Eliazar *et al.* [6], Holroyd *et al.* [14]). We will refer to this theorem in Chapter 4.

2.6 Gaver-Stehfest numerical inversion of LST

The Gaver-Stehfest (GS) technique for numerical inversion of LST was developed by Gaver [9] in 1966 in the probabilistic context of order statistics, and modified by Stehfest [30] through accelerating the convergence to obtain a better performance in 1970. It is becoming more refined and increasingly more acceptable in different areas because of its simplicity and good performance.

Given a real-valued function $f(t); t \geq 0$ whose LST is $\tilde{f}(s)$, then the GS method for numerical Laplace transform inversion at the point t is given by the following:

$$f_g(t) = \frac{ln2}{t} \sum_{j=1}^{N} V_j \tilde{f}\left(\frac{ln2}{t} \times j\right) \tag{2.24}$$

where the values V_j are the GS coefficients of order N (always even), half of which are positive and half negative numbers. These coefficients, as derived by Gaver, are combinatorial terms arising in order statistics, with the useful by-product that they always sum to 0. Typically $N = 8$ points provide two significant digits of accuracy, which is quite adequate for assessing waiting times. The table that provides the coefficients for $N = 2; 4; 6; 8$ is provided in Table 2.1.

Table 2.1: Coefficients for the Gaver-Stehfest Algorithm

V_2	V_4	V_6	V_8
2	-2	1	-1/3
-2	26	-49	145/3
	-48	366	-906
	24	-858	16394/3
		810	-43130/3
		-270	18730
			-35840/3
			8960/3

In 2006, Abate and Whitt [1] improved the original GS algorithm in equation (2.24) through introducing an inversion formula with $2M$ as the number of transform evaluations. For any $t > 0$ and positive integer M, the GS algorithm by encompassing the linear Salzer acceleration technique in Abate and Whitt [1] is given by

$$f_g(t, M) = \frac{ln(2)}{t} \sum_{j=1}^{2M} \zeta_j \tilde{f}\left(\frac{ln(2)}{t} \times j\right), \quad M \geq 1,\ t > 0, \tag{2.25}$$

where the coefficients ζ_i are

$$\zeta_j = \frac{(-1)^{M+j}}{M!} \sum_{i=\lfloor \frac{j+1}{2} \rfloor}^{\min(j,M)} i^{M+1} \begin{pmatrix} M \\ i \end{pmatrix} \begin{pmatrix} 2i \\ i \end{pmatrix} \begin{pmatrix} i \\ j-i \end{pmatrix}, \quad 1 \leq j \leq 2M. \tag{2.26}$$

Valko and Abate [31] concluded that the required system precision is about $2.2M$ when the parameter is M. From extensive experimentation, Valko and Abate [31] stated that about $0.90M$ significant digits were produced for $f(t)$ with good transforms. By $0.90M$ significant digits, they meant that

$$\text{relative error} = \left|\frac{f(t) - f_g(t, M)}{f(t)}\right| \approx 10^{-0.90M}. \tag{2.27}$$

Moreover, transforms are said to be "good" if the transforms have all their singularities on the negative real axis and the functions f are infinitely differentiable for all $t > 0$. If the transforms are not good, then the number of significant digits may not be so great and may not be proportional to M. Thus, the efficiency of the GS algorithm in equation (2.25), measured by the ratio of the significant digits produced to the precision required, was given by $0.9M/2.2M = 0.4$.

In summary, the significant digits produced by the formula of the parameter M is about $0.90M$, and the required system precision is about $2.2M$. For instance, if the system precision is 15 (8-byte floating point numbers), then the parameter M is 6 ($= \lfloor 15/2.2 \rfloor$), and the significant digits is about 5 ($= \lfloor 0.90 \times 6 \rfloor$).

References

[1] Abate J., & Whitt W. (2006). A unified framework for numerically inverting Laplace transforms. INFORMS Journal on Computing. 18, 408–421.

[2] Canadian Association of Emergency Physicians. The Canadian Triage and Acuity Scale (CTAS). From the website. *http ://www.calgaryhealthregion.ca/policy/docs/* 1451*/Admission_over_capacity_AppendixA.eps.*

[3] Conway, R. W., Maxwell, W. L., & Miller, L. W. (1967). Theory of scheduling. Addison-Wesley.

[4] del Barrio, E., Cuesta-Albertos, J. A., & Matrán, C. (2002). Asymptotic stability of the bootstrap sample mean. Stochastic Processes and their Applications. 97(2), 289–306.

[5] Drekic, S., & Stanford, D. A. (2000). Threshold-based interventions to optimize performance in preemptive priority queues. Queueing Systems. 35(1), 289–315.

[6] Eliazar, I., Klafter, J., & Cohen, M. H. (2009). A unified and universal explanation for Lévy laws and $1/f$ noises. Proceedings of the National Academy of Sciences of the United States of America. 106(30), 12251–12254.

[7] Fajardo, V. A. (2015). A generalization of $M/G/1$ priority models via accumulating priority (doctoral dissertation). Department of Statistics and Actuarial Science. University of Waterloo.

[8] Feller, W. (1957). An introduction to probability theory and its applications (2d ed.). New York: Wiley.

[9] Gaver, D. (1966) Observing stochastic process, and approcimate transform inversion. *O*perations Research. 14(3), 444–459.

[10] Gelenbe, E., & Mitrani, I. (1980). Analysis and synthesis of computer systems. Academic Press.

[11] Grassmann, W., & Zhao, Y. Q. (1997). Heterogeneous multiserver queues with a general input. INFOR. 35, 208-224.

[12] Gumbel, H. (1960). Waiting lines with heterogeneous servers. *O*perations Research. 8(4), 504–511.

[13] Haviv, M., & Ravner, L. (submitted in October 2015). Strategic bidding in an accumulating priority queue: equilibrium analysis. Computer Science and Game Theory.

[14] Holroyd, A. E., Lyons, R., & Soo, T. (2011). Poisson splitting by factors. The Annals of Probability. 39(5), 1938–1982.

[15] Kingman, J. F. C. (1993). Poisson Processes. Oxford University Press, Oxford.

[16] Kleinrock, L. (1964). Communication nets: stochastic message flow and delay. McGraw-Hill. New York.

[17] Kleinrock, L. (1964). A delay dependent queue discipline. Naval Research Logistics. Quarterly 11, 329–341.

[18] Kleinrock, L. (1965). A conservation law for a wide class of queueing disciplines. Naval Research Logistics. Quarterly 12, 181–192.

[19] Kleinrock, L., & Finkelstein, R. (1967). Time dependent priority queues. *O*perations Research. 15, 104–116.

[20] Kleinrock, L. (1976). Queueing systems Vol II: computer applications. Wiley. New York.

[21] Krishnamoorthi, B. (1963). On Poisson queue with two heterogeneous servers. *O*perations Research. 11, 321–330.

[22] Mokaddis, G. S., & Matta, C. H. (1998). On Poisson queue with three heterogeneous servers. Information and Management Sciences. 9, 53–60.

[23] Morse, P. M. (1958). Queues, inventories and maintenance. Wiley. New York, 82–84.

[24] Saaty, T. L. (1960). Time dependent solution of the many-server Poisson queue. *O*perations Research. 8, 768–771.

[25] Schrage, L. (1974). Optimal scheduling disciplines for a single machine under various degrees of information. Working Paper. Graduate School of Business, University of Chicago.

[26] Sharif, A. B., Stanford, D. A., Taylor, P. & Ziedins, I. (2014). A multi-class multi-server accumulating priority queue with application to health care. *O*perations Research for Health Care 3(2), 73–79.

[27] Sharma, O. P., & Dass, J. (1989). Initial busy period analysis for a multichannel Markovian queue. Optimization. 20, 317–323.

[28] Singh, V. P. (1971). Markovian queues with three heterogeneous servers. AIIE Transactions. 3(1), 45–48.

[29] Stanford, D. A., Taylor, P. & Ziedins, I. (2014). Waiting time distributions in the accumulating priority queue. Queueing Systems 77(3), 297–330.

[30] Stehfest, H. (1970). Numerical inversion of Laplace transforms. Communications of the ACM. 13(1), 47–49 (Algorithm 368 with correction (October 1970), 13, No. 10).

[31] Valko, P. P., & Abate, J. (2004). Comparison of sequence accelerators for the Gaver method of numerical Laplace transform inversion. Computers and Math. with Applics. 48, 629–636.

Chapter 3

Multi-server accumulating priority queues with heterogeneous servers

3.1 Introduction

Multi-server, multi-class queues (or service systems) have been used to model waiting times in areas as diverse as call centres, emergency departments, grocery stores, and other situations arising in daily life where randomly-arriving customers compete for limited resources. In a grocery store, a long waiting time is merely a matter of inconvenience, but in an emergency department, it might precipitate a change in a patient's acuity, or even be life-threatening. For these reasons of continuing importance, such queues remain an area of ongoing research interest and importance.

Most queuing models of multi-server systems make the assumption that all servers are equally capable, so that it does not matter which server is selected when several of them are idle. In reality, however, often it is not reasonable to assume every server can render service at the same speed. This reality is equally prevalent in health care systems as it is elsewhere, except the consequences can be more severe, in that an unanticipated long wait due to inaccurate modelling might impact a patient's health status markedly. One example would be that different doctors may treat patients at different speeds in an endoscopy suite, as they are distinct human servers. Another example would be two imaging machines in a hospital working at significantly different speeds, simply because one is based upon older technology than the other. Appropriate queuing models which explicitly model the heterogeneity among service speeds are particularly called for in such multi-server settings, so as to better reflect the reality of many health care (and other) systems.

Efforts to operate a heterogeneous multi-server system more efficiently fall into one of two broad strategies. The first strategy deals with the choice of server whenever there are multiple servers available to serve a newly-arriving customer. In these situations, various policies may be enforced so as to dictate the choice of server, among all of the idle servers, in a fashion that optimizes system performance. Such policies are known as "service dispatch" policies (e.g. the Randomly Chosen Server service dispatch policy). The second strategy addresses the choice of customer whenever there are multiple customers waiting to be served by the next available server. Various queuing disciplines can be applied in these situations. In particular, for systems where certain types of customers require faster access to the servers, priority queuing disciplines are appropriate. The goal of this paper is to make novel contributions in both types of strategies. With respect to the former, we develop a conservation law for the average waiting time, which depends upon the service dispatch policy. With respect to the latter, we determine the waiting time distributions for each class of customers, under a unifying model called the "accumulating priority queue" (APQ) which we define below. In particular, we can recover the waiting time distributions both for the first-come first-served (FCFS) queue and for traditional priority queues from this model.

There have been numerous advances in both areas in the literature. Efforts to determine who is to be selected for service among distinct classes of waiting customers originally resorted to what we call the "classical" priority queuing discipline (Kesten and Runnenberg [9]), in which a customer belonging to a given priority class is selected for service only when there are no waiting customers from higher priority classes. Customers from low priority classes in such a situation can be repeatedly overtaken by customers from higher priority classes whenever any are present in the queue. (By "overtaken", we are referring to those customers from higher-priority classes who arrive later than, yet enter service prior to, the specified customer from the low-priority class.) Thus, it can appear as if the customers from the lower-priority class are not making any progress towards entering service.

A further development that addressed this phenomenon occurred in 1964, when Kleinrock [10] first proposed a priority queuing discipline which he named the "time-dependent priority queue". In this discipline, waiting customers earn priority credits while they wait, at a rate that depends upon their priority class. Whenever a server becomes available, it selects the waiting customer with the greatest accumulated priority to that instant. In such a situation, a customer from a low priority class progressively earns more and more priority credits, thereby making it harder and harder to be overtaken as time progresses by customers from higher priority classes. For such a model, Kleinrock [10] obtained a recursive set of formulas to obtain the average waiting time for each class of customers, in the single-server case.

Recently, Stanford *et al.* [24] reconsidered the time-dependent priority queuing model, which they renamed the "Accumulating Priority Queue" (APQ). They derived the waiting time distributions for different priority classes in a single-server system. Stanford *et al.* [24] were motivated by applications in health care systems. As a specific example, they considered the Canadian Triage and Acuity Scale (CTAS) [3], which classifies patients according to the Key Performance Indicators (KPIs) described in Table 3.1. Despite being a clinical standard, as the acuity of the patient classes diminishes, so too does the clinical need.

It is clearly appropriate to assign absolute priority for severely ill patients (for example, KPI-1 & KPI-2). However, for the other classes where the potential for severe health degradation is not evident, it is more appropriate to reflect the incurred waiting time of a patient when assigning its current priority. (Furthermore, these respective groups of severely ill and less urgent patients are often treated in different areas of a hospital, which justifies modelling each at its own queue.) The APQ discipline provides a more balanced approach to select customers for service, and consequently better regulates wait times for various classes of customers. Under the APQ, a patient from a lower priority class who goes through an extraordinarily long waiting time will eventually accumulate enough priority to enter service before recently-arriving patients from some higher priority class.

Table 3.1: CTAS Key Performance Indicators (KPIs)

Level	Level of acuity	Response time	Sample diagnosis	Targets
1	Resuscitation	Immediate	Cardiac arrest	98%
2	Emergent	< 15 mins	Chest pain	95%
3	Urgent	< 30 mins	Moderate asthma	90%
4	Less urgent	< 60 mins	Minor trauma	85%
5	Non urgent	< 120 mins	Common cold	80%

Sharif *et al.* [21] extended the work of Stanford *et al.* [24] to the multi-class multi-server APQ with Poisson arrivals and a common exponential service time. They established the waiting time distributions for various classes of customers in such a homogeneous multi-server system. They also commented on how to select the accumulation rates to meet a specific goal for each priority class.

Returning to the issue of service dispatch policies, work in this area was initiated by Morse [18], who in 1958 was the first to consider the analysis of multi-server systems (although his main focus was on homogeneous cases). Saaty [19] found the Laplace-Stieltjes transform (LST) of the transient probabilities of multi-server FCFS queues with Poisson arrivals and exponential service times at a common rate, where the explicit expressions for the stationary probabilities were derived for the two-server case. Subsequently, he discussed the LST of the transient probabilities in the two-server system with two different exponential service distributions operating under a specific service dispatch policy (i.e., when both servers are idle, the server is chosen by an arriving customer according to the proportion of its service rate to the total service rate).

Gumbel [8] discussed an $M/M_i/c$ queue with c heterogeneous servers under the Randomly Chosen Server (RCS) dispatch policy. He derived expressions in closed form for the stationary probabilities and the expected queue length. He also pointed out that there is no equivalent system with homogeneous servers present. (By "no equivalent system", we mean there exists no homogeneous system which simultaneously has the same probability of all servers being busy, and the same aggregate service rate.) He then analyzed the degree of error incurred from the assumption that each server is working at an equal rate which is the average of all the service rates. Krishnamoorthi [15] studied an $M/M_i/2$ queue with two heterogeneous servers under a general service dispatch policy, where a newly-arriving customer chooses server 1 with probability p_1 and server 2 with probability $1 - p_1$ when both servers are idle. The LST of the transient state probability distribution of the queue length was derived.

Singh [23] considered an $M/M_i/2/N/(\beta)$ queue with two heterogeneous servers operating under the following two rules: (i) when both servers are idle, customers always choose the fast server and (ii) when both servers are busy, a customer joins the system with probability β and starts to wait. A cost model was discussed in his paper, and various tables and graphs representing the average characteristics of both the homogeneous and heterogeneous systems were given. Sharma and Dass [22] considered an $M/M_i/2/N$ queue with two heterogeneous servers under the general service dispatch policy considered in Krishnamoorthi [15] to derive the probability density function of the busy period. More recently, Mokaddis and Matta [17] studied an $M/M_i/3$ queue with three heterogeneous servers under the general service dispatch policy in Krishnamoorthi [15], where they used the Gauss-Jordan elimination method to solve the stationary probabilities. Grassmann and Zhao [7] studied a queuing system with multiple heterogeneous servers and general inputs, where the stationary probabilities were calculated using different rules to allocate the arriving jobs and computation issues were discussed.

All of the cited papers focus on the queue length distributions. However, it is possible to obtain the FCFS waiting time distribution from the queue length distribution. A newly-arriving customer's waiting time comprises one exponentially-distributed interval between successive service completions for every waiting customer found upon arrival. By conditioning upon the number of such waiting customers, and the fact that the queue length is geometric, it is possible to show that the waiting time distribution has an exponential tail.

In the present work, we investigate a multi-class multi-server queue with Poisson arrivals and heterogeneous exponentially-distributed service times under the APQ and related disciplines, where different service dispatch policies are considered. We extend the APQ approach to our heterogeneous multi-server, multi-class queue, obtaining the waiting time distribution for each class of customers. By taking various limits of the ratios of the accumulation rates involved, we are able to recover the waiting time distributions for each class under both the FCFS and classical priority queuing disciplines as well.

The rest of this paper is organized as follows. The model is introduced in Section 3.2, along with our notational conventions and relevant definitions. A conservation law for $M/M_i/c$ systems is presented in Section 3.3. The probability of all servers being busy under different dispatch policies is discussed in Section 3.4. An optimization problem of finding the optimal level of heterogeneity in an $M/M_i/2$ system to minimize a particular cost function is investigated in Section 3.5. The waiting time distributions under the APQ and related disciplines are derived in Section 3.6. Numerical investigations are carried out in Section 3.7.

3.2 Model description

We consider a multi-class multi-server queue with Poisson arrivals and heterogeneous servers, each with its own exponentially distributed service times. There are $K \in \mathbb{N}$ classes of customers. Customers of class k, $k = 1, 2, \ldots, K$ arrive independently to the queue according to a Poisson process with rate λ_k. A class-k customer accumulates priority while waiting at rate b_k, where $b_1 \geq b_2 \geq \cdots \geq b_K \geq 0$. A class-$k$ arriving at time t will have accumulated priority $b_k(t' - t)$ by time t'. At a service completion instant, the next waiting customer to be served is the one with the greatest accumulated priority at that instant. We observe that by setting $b_1 = b_2 = \cdots = b_K$, the FCFS case is obtained. Similarly, by setting $b_K = 1$ and $b_k = b_{k+1} * M$ for $k = K-1, K-2, \ldots, 1$, the classical priority queue is obtained in the limit as $M \to \infty$. Thus, the APQ discipline itself provides a unifying framework for all three queuing disciplines.

Let $A^{(k)}$ denote the inter-arrival time random variable (r.v.) for class-k customers, $k = 1, 2, \ldots, K$. Hence, $A^{(k)}$ has an exponential distribution with rate λ_k, which we simply write as $A^{(k)} \sim \text{Exp}(\lambda_k)$. Furthermore, if we let A denote the aggregate inter-arrival time r.v., then by the aggregation property of Poisson processes (see, for instance, Conway *et al.* [4]), it follows that $A \sim \mathrm{Exp}(\lambda)$. Conversely, from another well-known property of independent Poisson processes, we understand the probability that a randomly chosen customer belongs to class k to be λ_k/λ.

There are $c \in \mathbb{N}$ servers in the system, whose service times are independent and exponentially distributed with heterogeneous service rates $\mu_i, i = 1, 2, \ldots, c$. We let $\boldsymbol{\mu}$ denote the service rate vector $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_c\}$ and assume, without loss of generality, that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_c \geq 0$. We define the *cohort inter-completion time* to be the time between service completions when the full cohort of c servers is working in parallel. Whenever all servers are busy, the cohort inter-completion times are exponentially distributed at the sum of the rates $\mu_a = \sum_{i=1}^{c} \mu_i$.

The utilization levels, overall and by class, are defined by $\rho = \lambda/\mu_a$ and $\rho_k = \lambda_k/\mu_a$, so that $\rho = \sum_{k=1}^{K} \rho_k$. To ensure the system is stable we assume that $\rho < 1$, thereby ensuring that the system becomes idle occasionally, with probability one.

When a customer arrives during an idle period, it commences service with one of the idle servers based on the employed service dispatch policy. The family of service dispatch policies which we consider in this paper are those covered by the so-called "r-dispatch policy", introduced by Doroudi *et al.* [5]. Specifically, if we let C denote the set of idle servers for the current state, then, under the r-dispatch policy, the probability that server $i \in C$ is chosen to serve the newly-arriving customer is given by

$$p_i(C; r) = \frac{(\mu_i)^r}{\sum_{j \in C} (\mu_j)^r}, \qquad r \in \mathbb{R}. \tag{3.1}$$

The principle advantage of this approach is that it enables us to recover four popular server selection strategies by an appropriate choice of r. Setting $r = 0$ leads to the Randomly Chosen Server (RCS) policy; as $r \to \infty$, it approaches the Fastest Server First (FSF) policy; as $r \to -\infty$, it approaches the Slowest Server First (SSF) policy; and setting $r = 1$ results in the Rate Balancing Selection (RBS) policy.

Consider the case when all servers are idle. For simplicity, we write for $i = 1, 2, \ldots, c$, $p_i = p_i(\{1, 2, \ldots, c\}; r)$. The RCS implies that $p_1 = p_2 = \cdots = p_c = 1/c$, whereas the FSF implies that $p_1 = 1$, and $p_i = 0$ for $i = 2, 3, \ldots, c$, the SSF implies that $p_c = 1$, and $p_i = 0$ for $i = 1, 2, \ldots, c-1$, and RBS implies that $p_i = \mu_i/\mu_a$, for $i = 1, 2, \ldots, c$.

With multiple heterogeneous servers, we require a measure of the level of heterogeneity when comparing otherwise similar systems. The Gini index arose in the field of econometrics, as a measure of disparity in income distributions. Alves *et al.* [2] suggested the application of the Gini index in the multi-server queuing system to measure the degree of disparity (i.e., level of heterogeneity) in server speeds. We define a vector of Gini-like coefficients denoted by $\mathsf{\mathbf{G}}$ to evaluate the level of heterogeneity in the multi-server system, so that, $\mathsf{\mathbf{G}} = \{\mathsf{G}_i, i = 1, \ldots, c.\}$. The ith such index G_i is defined by

$$\mathsf{G}_i = \frac{\mu_1 - \mu_i}{\mu_1 + \mu_i} \quad i = 1, \ldots, c. \tag{3.2}$$

As the level of heterogeneity increases, it is reflected in an increased G_i: with $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_c$, it is readily seen that $0 = \mathsf{G}_1 \leq \mathsf{G}_2 \leq \cdots \leq \mathsf{G}_c \leq 1$. (When no subscript appears for G, it is to be assumed that $\mathsf{G} = \mathsf{G}_c$, the largest level of heterogeneity in the system.)

Whenever all the values in $\mathsf{\mathbf{G}}$ are zero, we are dealing with a homogeneous system of identical server speeds. If any value $\mathsf{G}_j \in \mathsf{\mathbf{G}}$; $j = 2, \ldots, c$ approaches one, it means that the fastest server is orders of magnitude faster than the jth server, which implies that for the slower servers $l > j$, G_l also approaches one.

3.3 A conservation law for $M/M_i/c$ systems

In this section, we derive a conservation law for the mean waiting times associated with any non-preemptive and work-conserving $M/M_i/c$ system such as the one considered in this paper. Note that our conservation law builds upon the well-known conservation law pertaining to the $M/G/1$ case which was first established by Kleinrock [11, 14]. To obtain this conservation law, one deals with the unfinished workload process $U(t)$, defined to be the total amount of work present in the system at time t still to be completed by the server. Specifically, if we let $\bar{U} = \lim_{t\to\infty} E\{U(t)\}$, then Kleinrock's conservation law can be expressed as

$$\sum_{k=1}^{K} \rho_k m_k = \bar{U} - W_0 \tag{3.3}$$

where W_0 denotes the average residual service time, ρ_k represents the long-run fraction of time that the server is occupied with work from class k, and m_k denotes the class-k mean waiting time, for $k = 1, 2, \ldots, K$. In the $M/G/1$ case, Kleinrock [11, 14] is able to establish that $\bar{U} = W_0/(1 - \rho)$ and $W_0 = (\sum_{k=1}^{K} \lambda_k E\{S_k^2\})/2$ where $E\{S_k^2\}$ denotes the second moment of the service time distribution for class-k customers. All work-conserving queuing disciplines satisfy these results.

Turning to the multi-server heterogeneous case, our extension of the unfinished workload is the *release time* at time t, denoted by $V(t)$, which we define as the amount of time until a server can become idle, given the work present at time t. It is readily seen that in a single-server system, the release time and the unfinished workload are the same. A typical sample path of the release time is shown in Figure 3.1.

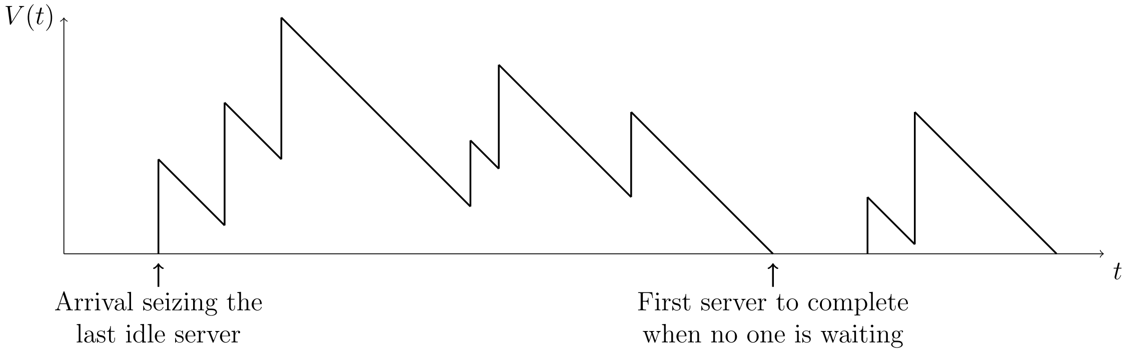


Figure 3.1: A sample path of the release time.

The proof which follows requires the system to be work-conserving, so that Figure 3.1 applies. This means that servers never become idle when there is work to be done, customers do not renege, and the release time process only increases at arrival instants by the amount of work the arriving customer brings into the system. For clarity of exposition, we assume that the system is non-preemptive, but this can in fact be relaxed to allow for preemptive resume systems. The resulting theorem can be stated as follows:

Theorem 3.3.1. *For any stable $M/M_i/c$ system operating under any non-preemptive work-conserving queuing discipline, we have*

$$\sum_{k=1}^{K} \rho_k m_k = \bar{V} - V_0 \tag{3.4}$$

where $\bar{V} = \lim_{t\to\infty} E\{V(t)\}$ represents the limiting expected value of $V(t)$ over time, and V_0 denotes the expected residual cohort inter-completion time.

Proof. Paralleling Kleinrock's derivation in Kleinrock [11, 14], we can express $V(t)$ as

$$V(t) = x_0 + \sum_{k=1}^{K} \sum_{i=1}^{N_{q_k}(t)} x_{ik} \tag{3.5}$$

where x_0 denotes the residual cohort inter-completion time if all servers are busy at time t, $N_{q_k}(t)$ represents the number of waiting customers present from class k at time t, and x_{ik} represents the cohort inter-completion time from when the ith waiting customer from class k commences service until the subsequent service completion instant by any server. The x_{ik}'s are non-overlapping and independent of each other, with each being exponentially distributed at rate $\mu_a = \sum_{l=1}^{c} \mu_l$. Therefore, letting $N_{q_k} = \lim_{t\to\infty} N_{q_k}(t)$ denote the limiting number of class-k customers in the queue, by taking expectations on both sides of equation (3.5) we obtain

$$\begin{aligned}
\lim_{t\to\infty} E\{V(t)\} &= V_0 + \lim_{t\to\infty} \sum_{k=1}^{K} E\{N_{q_k}(t)\}/\mu_a \\
&= V_0 + \sum_{k=1}^{K} E(N_{q_k})/\mu_a \\
&= V_0 + \sum_{k=1}^{K} \lambda_k m_k/\mu_a, \qquad (3.6)
\end{aligned}$$

where the latter equation arises as a result of Little's Law [16]. Therefore

$$\bar{V} = V_0 + \sum_{k=1}^{K} \rho_k m_k \tag{3.7}$$

and by rearrangement, equation (3.4) is obtained. □

In order to proceed further, we need to determine the probability that all servers are busy, which in turn depends upon the service dispatch policy. We therefore define $\pi(\boldsymbol{\mu}, c; r)$ as the probability that all c heterogeneous servers are busy under the particular r-dispatch policy in effect. Two facts facilitate the following analysis. In a multi-server work-conserving system,

1) The dispatch policy only affects the state transitions when at least two servers are idle, which in turn affects the global probability that all servers are busy. It has no effect once all servers are busy.

2) The APQ mechanism affects who is selected for service when a choice needs to be made. However, it does not affect the total number of customers present, as the service distributions depend upon the specific servers, but not upon any customer characteristic.

Once all servers are busy, aggregate customer arrivals to the system occur according to a Poisson process at rate λ, while the cohort inter-selection times are exponentially distributed at rate μ_a. Thus, when the system is busy, transitions between adjacent such states follow the classical birth-and-death behaviour (as seen from the rightmost part of Figure 3.2) in Section 3.4. For $n = 0, 1, 2 \ldots$, let π_n be the stationary probability of n customers in the system. The portion of the state transition diagram pertaining to busy states, when solved according to the birth-and-death equations, gives rise to a geometric tail relative to the state π_c when all servers are busy but no one is waiting, yielding $\pi_n = \pi_c \rho^{n-c}; n \geq c$, so that

$$\pi(\boldsymbol{\mu}, c; r) = \sum_{n=c}^{\infty} \pi_n = \frac{\pi_c}{1-\rho}. \tag{3.8}$$

The probability of c customers present in the system, π_c, depends on the dispatch policy in the system. Thus, the probability of all servers being busy is affected by the dispatch policy, but not by the APQ discipline.

Corollary 3.3.2. *In any $M/M_i/c$ system and any non-preemptive work-conserving queuing discipline, for a given r-dispatch policy, we have*

$$\sum_{k=1}^{K} \rho_k m_k = \frac{\pi(\boldsymbol{\mu}, c; r)}{\mu_a} \cdot \frac{\rho}{1-\rho}, \qquad \rho < 1. \tag{3.9}$$

Proof. Since the cohort inter-selection time is only positive once all servers are busy, it follows that

$$V_0 = \pi(\boldsymbol{\mu}, c; r)\Big(\frac{1}{\mu_a}\Big). \tag{3.10}$$

As Poisson arrivals see time averages (see Wolff [25]), the average release time will be the same as the mean waiting time for the aggregated class of all customers, which can be computed using Little's Law in Little [16] as

$$\bar{V} = \frac{1}{\lambda}\sum_{i=1}^{\infty} i\pi_{c+i} = \frac{1}{\lambda}\sum_{i=1}^{\infty} i\rho^i\pi_c = \frac{\pi_c}{\mu_a(1-\rho)^2} = \frac{\pi(\boldsymbol{\mu}, c; r)}{\mu_a(1-\rho)} = \frac{V_0}{1-\rho}. \tag{3.11}$$

Substitution for $\bar{V}$ and V_0 in equation (3.4) using equations (3.11) and (3.10) leads to equation (3.9). □

Remark In an $M/M_i/c$ queue, as $\mathsf{G}_i \to 1$, $\mu_1 \to \mu_a$ and $\mu_i \to 0$ for $i = 2, 3, \ldots, c$, then

$$\pi(\boldsymbol{\mu}, c; r) \to \rho, \qquad \Rightarrow \qquad \frac{\pi(\boldsymbol{\mu}, c; r)}{\mu_a} \cdot \frac{\rho}{1-\rho} \to \frac{\rho^2}{\mu_a(1-\rho)},$$

which means that the conservation law for $M/M_i/c$ queues approaches the conservation law for $M/M/1$ queues as the level of heterogeneity increases to one in the heterogeneous system.

Remark Kleinrock [12] derived the expressions for the average waiting times for different classes in a single-server APQ. By direct parallel in the multi-server case, the average waiting time for class-k in a multi-class APQ with heterogeneous servers, for $k = 1, 2, \ldots, K$, is readily found to be given by

$$m_k = \frac{\pi(\boldsymbol{\mu}, c; r)/(\mu_a - \lambda) - \sum_{j=k+1}^{K} \rho_j(1 - b_j/b_k)m_j}{1 - \sum_{j=1}^{k-1} \rho_j(1 - b_k/b_j)}. \tag{3.12}$$

Having determined the conservation law in terms of $\pi(\boldsymbol{\mu}, c; r)$, we turn in the next section to how to determine this probability in terms of the specified dispatch policy.

3.4 The stationary probability of the system being busy for $M/M_i/c$ queues

An arriving customer in our model may find that:

- At least one server is idle, in which case it will be assigned to a server according to the r-dispatch policy in effect (whenever two or more servers are idle).

- All servers are engaged, in which case it starts to wait and accumulate priority.

In what follows, we focus on the calculations of the stationary probabilities, $\pi_n; n = 0, 1, \dots, c$, when no one is waiting, where the dispatch policy can matter. The remaining state probabilities $\pi_n; n = c+1, c+2, \dots$ are related to π_c via $\pi_n = \pi_c \rho^{n-c}$, regardless of the service dispatch policy.

In 1960, Gumbel [8] derived the stationary probabilities under the assumptions of Poisson arrivals and exponentially distributed service times, with a different service rate for each server under RCS. The RCS policy satisfies the local-balance equations, and as such, it was possible for Gumbel to come up with an explicit product form solution for the stationary distribution, for all c. The stationary probability $\pi(\boldsymbol{\mu}, c; 0)$ can be directly calculated from his work (Gumbel [8]):

$$\begin{aligned} \pi(\boldsymbol{\mu}, c; 0) &= \frac{\pi_c}{1-\rho} = \frac{\left(\frac{1}{1-\rho} + \sum_{j=1}^{c} j! C_j^c\right)^{-1}}{1-\rho} \\ &= \frac{1}{1 + (1-\rho)\sum_{j=1}^{c} j! C_j^c}, \end{aligned} \tag{3.13}$$

where

$$C_j^c = \frac{1}{\lambda^j} \sum_{a_1=1}^{c-j+1} \sum_{a_2=a_1+1}^{c-j+2} \cdots \sum_{a_{j-1}=a_{j-2}+1}^{c-j+j-1} \sum_{a_j=a_{j-1}+1}^{c} \mu_{a_1}\mu_{a_2}\cdots\mu_{a_j}.$$

For example, $C_2^3 = \frac{1}{\lambda^2} \sum_{a_1=1}^{2} \sum_{a_2=a_1+1}^{3} \mu_{a_1}\mu_{a_2} = \left(\mu_1(\mu_2 + \mu_3) + \mu_2\mu_3\right)/\lambda^2$. Note that $C_c^c = \Pi_{i=1}^{c} \frac{\mu_i}{\lambda}$, $C_1^c = \sum_{i=1}^{c} \frac{\mu_i}{\lambda} = 1/\rho$ and $C_0^c = 1$.

As for non-RCS dispatch policies, in 1963, Krishnamoorthi [15] studied the Poisson queue with two heterogeneous servers, where the stationary probabilities were solved via differential-difference equations. The stationary probabilities can also be obtained from the global balance equations. Let states $(1, 0)$ and $(0, 1)$ denote the states where one server is busy in the system, with the position of the 1 indicating which server is busy, as well as $\pi_{1,0}$ and $\pi_{0,1}$ be their stationary probabilities respectively, where $\pi_{1,0} + \pi_{0,1} = \pi_1$. When a customer arrives to an idle system, it can be served by server 1 with probability p_1 or by server 2 with probability p_2. By the properties of the Poisson distribution, the transition rate from state 0 to state $(1, 0)$ is $p_1\lambda$ and the transition rate from state 0 to state $(1, 0)$ is $p_2\lambda$. Then, the stationary probability π_n

can be solved from the following global balance equations (i.e., probability flow into the state equals the flow out of the state):

$$\begin{aligned} \lambda\pi_0 &= \mu_1\pi_{1,0} + \mu_2\pi_{0,1}, \\ (\lambda + \mu_1)\pi_{1,0} &= p_1\lambda\pi_0 + \mu_2\pi_2, \\ (\lambda + \mu_2)\pi_{0,1} &= p_2\lambda\pi_0 + \mu_1\pi_2, \\ \pi_{n+1} &= \rho\pi_n, \qquad n \geq 2. \end{aligned} \tag{3.14}$$

The results of $\pi_n; n \geq 0$ are consistent with the ones in Krishnamoorthi [15]. By equation (3.2), $\mathsf{G} = (\mu_1 - \mu_2)/\mu_a$ in the two-server case, which implies $\mu_1 = \mu_a(1 + \mathsf{G})/2$ and $\mu_2 = \mu_a(1 - \mathsf{G})/2$. Thus, from equation (3.8), we are able to express the stationary probability $\pi(\boldsymbol{\mu}, 2; r)$ in terms of the particular r-dispatch policy as

$$\pi(\boldsymbol{\mu}, 2; r) = \frac{2\lambda^2\Big(1 + 2\rho - \mathsf{G}(p_1 - p_2)\Big)}{\mu_a^2(1 - \mathsf{G}^2) - \lambda\mu_a\Big(\mathsf{G}^2 + 2\mathsf{G}(p_1 - p_2) - 3\Big) + 2\lambda^2(1 + \mathsf{G}^2)}. \tag{3.15}$$

Subsequently, Mokaddis and Matta [17] derived the stationary probabilities of a Poisson queue with three heterogeneous servers. They used the Gauss-Jordan elimination method to solve for the probabilities. As the results involved 52 variables, the expressions are already marginally tractable in terms of being able to interpret the impact of various parameters on the performance measures of interest, so we do not replicate their results here. Interested readers are directed to Mokaddis and Matta [17] for the details.

While it would be of interest to extend this to a system with an arbitrary number $c > 3$ of heterogeneous servers, one can anticipate that the expression for $\pi(\boldsymbol{\mu}, c; r)$ would be even more complicated than the three server case. Thus, we need to find a suitable numerical method to compute the probability $\pi(\boldsymbol{\mu}, c; r)$ for an arbitrary number of servers under non-RCS policies. In Grassmann [6], several numerical methods are introduced, such as the state reduction method and other block-elimination methods. Grassmann and Zhao [7] also discussed several numerical methods for the solution of the heterogeneous systems and concluded that the state probabilities can be computed efficiently for up to 10 servers; however, all methods become problematic for a system with more than 20 servers. Interested readers are directed to

Grassmann and Zhao [7] for the details.

When extended to a Poisson queue with c heterogeneous servers assuming exponential service rates $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_c$, there are 2^c boundary states of interest. The structure of the state transition diagram is shown in Figure 3.2. In order to understand the relationships among the 2^c states, we chose three neighboring states (denoted by the three rectangular-shaped states in Figure 3.2) to describe the transition process between adjacent levels.

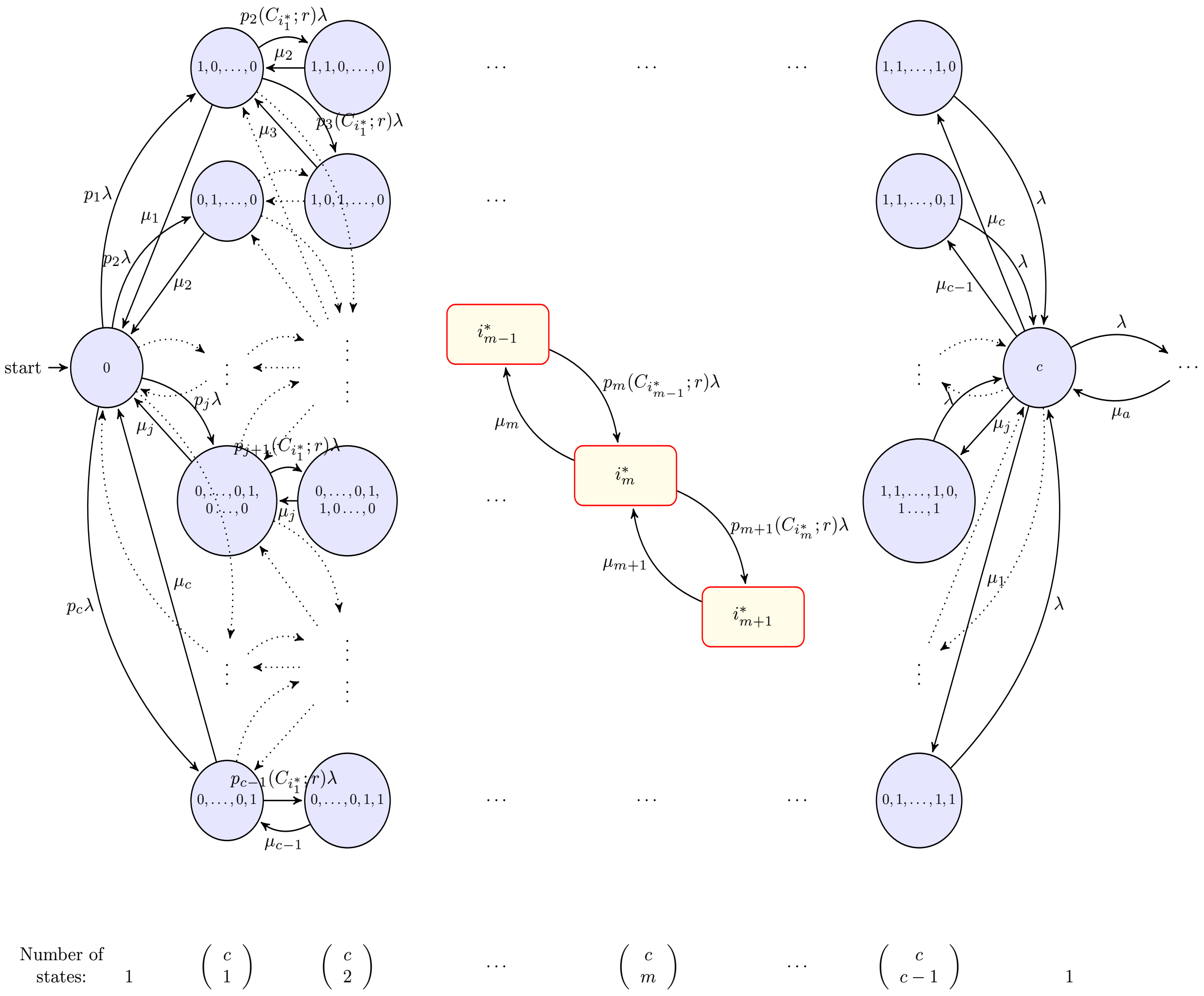


Figure 3.2: State-transition-rate diagram with arbitrary number of servers.

We define i_j to be an indicator function, such that

$$i_j = \begin{cases} 1 : & \text{server } j \text{ is busy.} \\ 0 : & \text{server } j \text{ is idle.} \end{cases}$$

Furthermore, let $\vec{i}_m$ represent the collection of state vectors $(i_1, \ldots, i_c)$ such that $\sum_{j=1}^{c} i_j = m$. An arrival finding $m-1$ busy servers where $m < c$, causes the system to transit to a new system state, in which m servers are busy. In other words, the system changes from being in state $i \in \vec{i}_{m-1}$ to a state $j \in \vec{i}_m$ such that $j - i$ is a vector whose entries are all zero except for the one corresponding to the server that attends to the newly-arriving customer. Similarly, if there is a service completion in a state with $(m + 1)$ busy servers, then the system changes from a state $i \in \vec{i}_{m+1}$ to a state $j \in \vec{i}_m$, such that $j - i$ is a vector of zeros except for the entry corresponding to the server that had just become idle. Moreover, the state vector $j \in \vec{i}_m$ indicates the set of idle servers at the current state, i.e., $C_j = \{s : i_s = 0, i_s \in j\}$. We choose to write the probability that idle server $s \in C_j$ is chosen under the r-dispatch policy as $p_s(C_j; r)$.

To illustrate, in Figure 3.2, we consider the cases for which the fastest servers are the ones that are busy during the states $\vec{i}_{m-1}$, $\vec{i}_m$ and $\vec{i}_{m+1}$. In particular, if we let $i^*_n, n < c$, denote the vector $i \in \vec{i}_n$ whose first n entries are all equal to one, then the transition rate from state i^*_{m-1} to state i^*_m is $\lambda p_m(C_{i^*_{m-1}}; r)$. Similarly, the transition rate from state i^*_m to state i^*_{m+1} is $\lambda p_{m+1}(C_{i^*_m}; r)$. The transition rates from state i^*_{m+1} to state i^*_m and from state i^*_m to state i^*_{m-1} are μ_{m+1} and μ_m respectively.

Then, we can write the global balance equation of state i^*_m as follows: for $i_s \in i^*_m$,

$$\left(\lambda + \sum_{s:i_s=1} \mu_s\right)\pi_{i^*_m} = \sum_{s:i_s=1} \lambda p_s(C_{i^*_{m-1}}; r)\pi_{i^*_{m-1}} + \sum_{s:i_s=0} \mu_s \pi_{i^*_{m+1}}. \tag{3.16}$$

The left hand side of the equation represents the rate of flow from state i^*_m, where $\lambda\pi_{i^*_m}$ is the rate of transition to state i^*_{m+1} due to an arrival, and $\sum_{s:i_s=1} \mu_s \pi_{i^*_m}$ is the rate of transition to state i^*_{m-1} due to a service completion. The right hand side of this equation presents the rate of flow into state i^*_m, where $\sum_{s:i_s=1} \lambda p_s(C_{i^*_{m-1}}; r)\pi_{i^*_{m-1}}$ is the sum of arrival rates causing a transition to state i^*_m, and $\sum_{s:i_s=0} \mu_s \pi_{i^*_{m+1}}$ is the rate of transition to state i^*_m due to a service completion.

The global balance equations for all the 2^c boundary states can be written out in the form of equation (3.16). To solve for $\pi(\boldsymbol{\mu}, c; r)$ requires one to solve 2^c such equations in as many variables. Thus, an increase of one server doubles the number of boundary states to be solved. Since solving n equations in n variables requires $O(n^3)$ operations, this implies that the computational effort increases by a factor of 8 as a result of increasing c by 1 (see Grassmann and Zhao [7]). This underscores even further our intent to resort to a numerical method when dealing with cases of $c > 3$. However, this is not the focus of this paper, so that in what follows, we consider only cases where $c \leq 3$.

3.5 Optimizing the level of heterogeneity in $M/M_i/2$ systems

In this section, we focus on finding the optimal level of heterogeneity G to minimize an appropriately defined cost, and the optimal range of G in which the r-dispatch policies have lower cost than the homogeneous system. We consider the following cost function:

$$F(c, r, \mathsf{G}) = \pi(\boldsymbol{\mu}, c; r) \cdot \frac{\rho}{1-\rho}, \tag{3.17}$$

where G is the decision variable. In light of equation (3.9), we can interpret $F(c, r, \mathsf{G})$ as the conservation law's right-hand side constant (scaled in units of $1/\mu_a$), which is the average number of waiting customers. Since μ_a is fixed, as a result of the specified service rates, any overall system improvement gained in terms of a lower $F(c, r, \mathsf{G})$ will be as a result of its particular r-dispatch policy.

We have previously observed when discussing Mokaddis and Matta [17] that the analytic expressions for the case of three or more servers are too complicated to ascertain the influence of specific parameters. However, in the two-server case, we are able to establish some analytical results about optimal levels of heterogeneity under a given r-dispatch policy.

Intuitively, one might anticipate that greater heterogeneity leads to greater congestion. This is certainly true under the SSF and RCS dispatch policies, but we shall see that, in fact, there exists a range of G values in the two-server case for which our cost function in fact decreases under the FSF dispatch policy. We will postpone discussion of the RBS case to the end of the section.

Lemma 3.5.1. *In a two-server queue under any work-conserving queuing discipline, for $r \leq 0$ (i.e., $p_1 \leq p_2$), the minimum value of function $F(2, r, \mathsf{G})$ occurs in the homogeneous case. For $r > 0$ (i.e., $p_1 > p_2$), the optimal value of G that minimizes the function $F(2, r, \mathsf{G})$, denoted by G^*, is given by*

$$\mathsf{G}^* = \frac{2\rho + 1 - \sqrt{(2\rho + 1)^2 - (p_1 - p_2)^2}}{p_1 - p_2}. \tag{3.18}$$

When r is sufficiently large, an optimal range of G can be found for which the function $F(2, r, \mathsf{G})$ is an improvement upon the homogeneous case, and the optimal range is given by

$$\mathsf{G} \in \Big(0, \frac{p_1 - p_2}{1 + 2\rho}\Big). \tag{3.19}$$

Proof. Consider the probability $\pi(\boldsymbol{\mu}, 2; r)$ as given by equation (3.15), where we have arbitrarily set $\mu_a = 2$. We choose to make explicit the dependence of the cost function upon G; this yields

$$F(2, r, \mathsf{G}) = \frac{2\rho^3\,(2\rho + 1 - \mathsf{G}\,(p_1 - p_2))}{(1 - \rho)\Big(2\,(\mathsf{G}^2 + 1)\,\rho^2 + (3 - \mathsf{G}^2 + 2\mathsf{G}\,(p_2 - p_1))\,\rho + 1 - \mathsf{G}^2\Big)}. \tag{3.20}$$

Taking the first partial derivative of $F(2, r, \mathsf{G})$ with respect to G, we get

$$\frac{\partial}{\partial \mathsf{G}} F(2, r, \mathsf{G}) = \frac{2(2\rho + 1)\rho^3\Big(4\rho\mathsf{G} + (\mathsf{G}^2 + 1)\,(p_2 - p_1) + 2\mathsf{G}\Big)}{\Big((2\,(\mathsf{G}^2 + 1)\,\rho^2 + (3 - \mathsf{G}^2 + 2\mathsf{G}\,(p_2 - p_1))\,\rho + 1 - \mathsf{G}^2)\Big)^2}. \tag{3.21}$$

It is readily seen that when $r \leq 0$, equation (3.21) is nonnegative for all $\rho \in (0, 1)$ and $\mathsf{G} \in [0, 1]$. Thus, no improvement can be gained over a homogeneous system when $r \leq 0$.

Turing to the case where $r > 0$, we compute the second partial derivative

$$\begin{aligned}
\frac{\partial^2}{\partial \mathsf{G}^2} F(2, r, \mathsf{G}) \quad = \quad & \frac{8\rho^3(2\rho + 1)}{\Big(2\,(\mathsf{G}^2 + 1)\,\rho^2 + (3 - \mathsf{G}^2 + 2\mathsf{G}\,(p_2 - p_1))\,\rho + 1 - \mathsf{G}^2\Big)^3} \cdot \\
& \Big(2\Big(1 - 3\mathsf{G}^2\Big)\rho^3 + \Big(4 + (p_1 - p_2)\,\mathsf{G}(\mathsf{G}^2 + 3)\Big)\rho^2 - \frac{1}{2}\Big(\,(p_1 - p_2)\,\mathsf{G}(\mathsf{G}^2 + 3) \\
& \quad -9\mathsf{G}^2 + 2(p_1 - p_2)^2 - 5\Big)\rho - \frac{1}{2}\,(p_1 - p_2)\,\mathsf{G}(\mathsf{G}^2 + 3) + \frac{1}{2}(1 + 3\mathsf{G}^2)\Big)
\end{aligned}$$

$$\begin{aligned} &\geq \frac{4\rho^3\left(4\rho^2(\rho+2) - \rho(2(p_1-p_2)^2-5)+1\right)}{(2\rho+1)^2(\rho+1)^3} \\ &> 0. \end{aligned} \tag{3.22}$$

Thus, the second partial derivative of $F(2, r, \mathsf{G})$ is positive for all $\rho \in (0, 1)$ and $\mathsf{G} \in [0, 1]$. (We note that the second last line above corresponds to the curvature in the homogeneous case.) To find the optimal value of G in this case, we set equation (3.21) equal to zero, and solve for G^*, obtaining equation (3.18). For example, the optimal level of heterogeneity that minimizes the function $F(2, \infty, \mathsf{G})$, the FSF policy, occurs at $\mathsf{G}^* = 1 + 2\rho - 2\sqrt{\rho(1+\rho)}$.

When r is sufficiently large (discussed in Corollary 3.5.2), it is possible to find a lower cost function $F(2, r, \mathsf{G})$ than the one in the homogeneous case which is $2\rho^3/(1-\rho^2)$. An optimal range of G obtained for which the function $F(2, r, \mathsf{G})$ is an improvement upon the homogeneous case is given by

$$F(2, r, \mathsf{G}) < \frac{2\rho^3}{1-\rho^2}, \quad \Rightarrow \quad 0 < \mathsf{G} < \frac{p_1 - p_2}{1+2\rho}. \tag{3.23}$$

□

Remark Equation (3.19) shows that the optimal range of the level of heterogeneity decreases as ρ increases.

Remark Lemma 3.5.1 shows that under the RCS and SSF cases where $p_1 \leq p_2$, the SSF and RCS dispatch rules have higher cost than the homogeneous case. The second half of the lemma establishes FSF has a range of values of the level of heterogeneity in the system which can result in a lower cost then the homogeneous case. While this would seem at first glance to be true as well for RBS since $p_1 > p_2$, in fact under RBS, it is seen that $\mathsf{G} = p_1 - p_2$ which exceeds the upper limit. A modification to RBS which attenuates sufficiently the probability of selecting the faster server could lead to improvement over the homogeneous case.

Corollary 3.5.2. *Consider a two-server queue operating under a work-conserving queuing discipline, and define*

$$r^* = \frac{\ln(\mu_1 + \rho(\mu_1 - \mu_2)) - \ln(\mu_2 - \rho(\mu_1 - \mu_2))}{\ln(\mu_1) - \ln(\mu_2)}. \tag{3.24}$$

Then an improvement can be gained over the homogeneous system with the same aggregate rate if and only if $r > r^*$.

Proof. In order to find the range of r which can lead to a lower cost than the homogeneous case, we take the difference, for simplicity denoted as D_{G}, between the Gini-like coefficient G from the definition in equation (3.2) and the upper bound of G in equation (3.19).

$$D_{\mathsf{G}} = \mathsf{G} - \frac{p_1 - p_2}{1 + 2\rho} = \frac{\mu_1 - \mu_2}{\mu_a} - \frac{\mu_1^r - \mu_2^r}{(\mu_1^r + \mu_2^r)(1 + 2\rho)}. \tag{3.25}$$

When $D_{\mathsf{G}} \geq 0$, the corresponding r-dispatch policy may have a higher cost than the homogeneous case; while, when $D_{\mathsf{G}} < 0$, a lower cost can be found. As $\mathsf{G} \in [0, 1]$ and $p_1 - p_2 \in [-1, 1]$, $-1/(1 + 2\rho) \leq D_{\mathsf{G}} \leq 2(1 + \rho)/(1 + 2\rho)$. Moreover, D_{G} is decreasing with r. Thus, there is a single point of r to separate the case of $D_{\mathsf{G}} \geq 0$ and $D_{\mathsf{G}} < 0$, and we denote this point by r^*. Clearly, r^* is the solution when $D_{\mathsf{G}} = 0$, as shown in equation (3.24). So, when $r > r^*$, $D_{\mathsf{G}} < 0$, which means that an improvement can be gained over the homogeneous case. □

Having obtained $\pi(\boldsymbol{\mu}, c; r)$ and the mean waiting times in terms of the dispatch policy, in the next section we turn to the determination of waiting time distributions for the APQ discipline in the multi-server heterogenous setting.

3.6 Waiting time distributions for $M/M_i/c$ queues under APQ and related disciplines

We seek the waiting time distribution before service commences under the APQ discipline in terms of its LST for each customer class. From these results, we will obtain the corresponding distributions under the FCFS and classical priority disciplines by appropriate choices of the accumulation rates. The numerical examples to follow in the next section will be obtained via numerical inversion of the corresponding LSTs.

For $k = 1, 2, \ldots, K$, let $\tilde{W}^{(k)}(s; \boldsymbol{\mu}, c, r)$ denote the LST of the stationary class-k waiting time distribution in an $M/M_i/c$ APQ under the r-dispatch policy and with service rate vector $\boldsymbol{\mu}$. Let

$\tilde{W}_+^{(k)}(s; \boldsymbol{\mu}, c)$ be the LST of the corresponding conditional waiting time distribution of a delayed class-k customer. Then for $k = 1, 2, \ldots, K$,

$$\tilde{W}^{(k)}(s; \boldsymbol{\mu}, c, r) = (1 - \pi(\boldsymbol{\mu}, c; r)) + \pi(\boldsymbol{\mu}, c; r)\tilde{W}_+^{(k)}(s; \boldsymbol{\mu}, c). \tag{3.26}$$

Lemma 3.6.1. *The LST of the conditional waiting time distribution for class-k operating under the APQ discipline in the $M/M_i/c$ queue can be related to the comparable measure in an $M/M/1$ APQ as follows:*

$$\tilde{W}_+^{(k)}(s; \boldsymbol{\mu}, c) = \tilde{W}_+^{(k)}(s; \mu_a, 1); \quad k = 1, 2, \ldots, K. \tag{3.27}$$

Proof. When the cohort is fully busy, the cohort inter-completion times are exponentially distributed with parameter $\mu_a = \sum_{j=1}^{c} \mu_j$. The same is true of a busy single exponential server working at rate μ_a. Thus, given the same arrival classes and distributions operating under the APQ discipline, the conditional waiting time distributions must be the same. □

Remark The same idea applied here was also used for the multi-server APQ with a common service rate in Sharif *et al.* [21].

The foregoing lemma enables us to invoke the conditional waiting time results of Stanford *et al.* [24] for the single server case with service rate μ_a, which allows for class-dependent service time distributions. The various conditional waiting time distributions therein were obtained in a recursive fashion, starting with the lowest priority class. Likewise, we start from that perspective in the heterogeneous c-server case.

One further concept is needed in order to state the relevant result. It is possible for customers of a given class (say, class k), or higher, to eventually accumulate enough priority to exceed the maximum possible credit of a class-$(k + 1)$ customer by that given point in time. Such customers are said to have gained "accreditation relative to class $(k + 1)$" (see Stanford *et al.* [24]). Let $\tilde{\Gamma}_k(s), k = 1, 2, \ldots, K$ be the LST of the duration of a busy period during which customers gain accreditation relative to class $(k + 1)$. In other words, it represents the LST of the time interval required to clear the system of those customers who have gained accreditation relative to class $(k+1)$. Its duration is the same as the duration of the busy period of an $M/M/1$ queue with arrivals rates $\Lambda_k = \sum_{i=1}^{k} \lambda_i(1 - b_{k+1}/b_i)$ and service rate μ_a. It is well-known (Conway *et al.* [4]) that the solution to the implicit equation for the busy period LST in this case is given by

$$\tilde{\Gamma}_k(s) = \frac{(\mu_a + \Lambda_k + s) - \sqrt{(\mu_a + \Lambda_k + s)^2 - 4\Lambda_k\mu_a}}{2\Lambda_k}. \tag{3.28}$$

Theorem 3.6.2. *The waiting time distribution for the lowest priority class, conditional on it being positive, has LST*

$$\tilde{W}_+^{(K)}(s; \boldsymbol{\mu}, c) = \frac{\mu_a(1-\rho)}{\mu_a(1-\rho) + g(s)} \tag{3.29}$$

where $g(s) = s + \Lambda_{K-1}(1 - \tilde{\Gamma}_{K-1}(s))$.

Proof. The waiting time for a delayed class-K customer can be viewed as consisting of two components. The first component is the virtual workload present upon their arrival, whose distribution is the same as the stationary waiting time of delayed customers in the equivalent $M/M/1$ queue with service rate μ_a. The second is the additional delay introduced by the customers of higher classes who overtake the marked class-K customer. This portion of the delay represents a "delay busy period" in the sense of Conway *et al.* ([4], page 151). When appropriate substitutions are made, equation (3.29) is obtained. The interested reader is directed to Stanford *et al.* [24] for further details on this approach. □

Corollary 3.6.3. *Under the FCFS discipline, the waiting time distribution for all classes, conditional on it being positive, is exponentially distributed with parameter* $\mu_a(1-\rho)$.

Proof. The result follows directly from Theorem 3.6.2. First of all, note that by setting $b_1 = b_2 = \cdots = b_K$, the APQ discipline becomes exactly the FCFS discipline. Furthermore, under this setting, it is readily observed that $\Lambda_{K-1} = 0$. Therefore, it follows from equation (3.29) that

$$\tilde{W}_+^{(K)}(s; \boldsymbol{\mu}, c) = \frac{\mu_a(1-\rho)}{\mu_a(1-\rho) + s}, \tag{3.30}$$

which is the LST of an exponential distribution with parameter $\mu_a(1-\rho)$. The extension of the above result to the remaining classes is due to the fact that $\tilde{W}_+^{(1)}(s; \boldsymbol{\mu}, c) = \tilde{W}_+^{(2)}(s; \boldsymbol{\mu}, c) = \cdots = \tilde{W}_+^{(K)}(s; \boldsymbol{\mu}, c)$ under the FCFS discipline. □

Returning to the general case, the recursion we employ to obtain the waiting time distributions for the higher priority classes proceeds in sequence for $k = K-1, K-2, \ldots, 1$ on account of equation (3.27) and the corresponding result in Stanford *et al.* [24] as follows:

$$\tilde{W}_{+}^{(k)}(s;\boldsymbol{\mu},c) \;=\; \left(\frac{b_{k+1}}{b_k}\right)\tilde{W}_{+}^{(k+1)}\left(\frac{b_{k+1}}{b_k}s;\boldsymbol{\mu},c\right)+\left(1-\frac{b_{k+1}}{b_k}\right)\tilde{W}_{acc}^{(k)}(s) \tag{3.31}$$

where $\tilde{W}_{acc}^{(k)}(s)$ is defined as the LST of the waiting time of a class-k customer who is served "at accreditation level k". A discussion of accreditation and the related formulas needed to determine the waiting time distributions can be found in Appendix A.1, as well as the waiting time expressions for the classical priority queue.

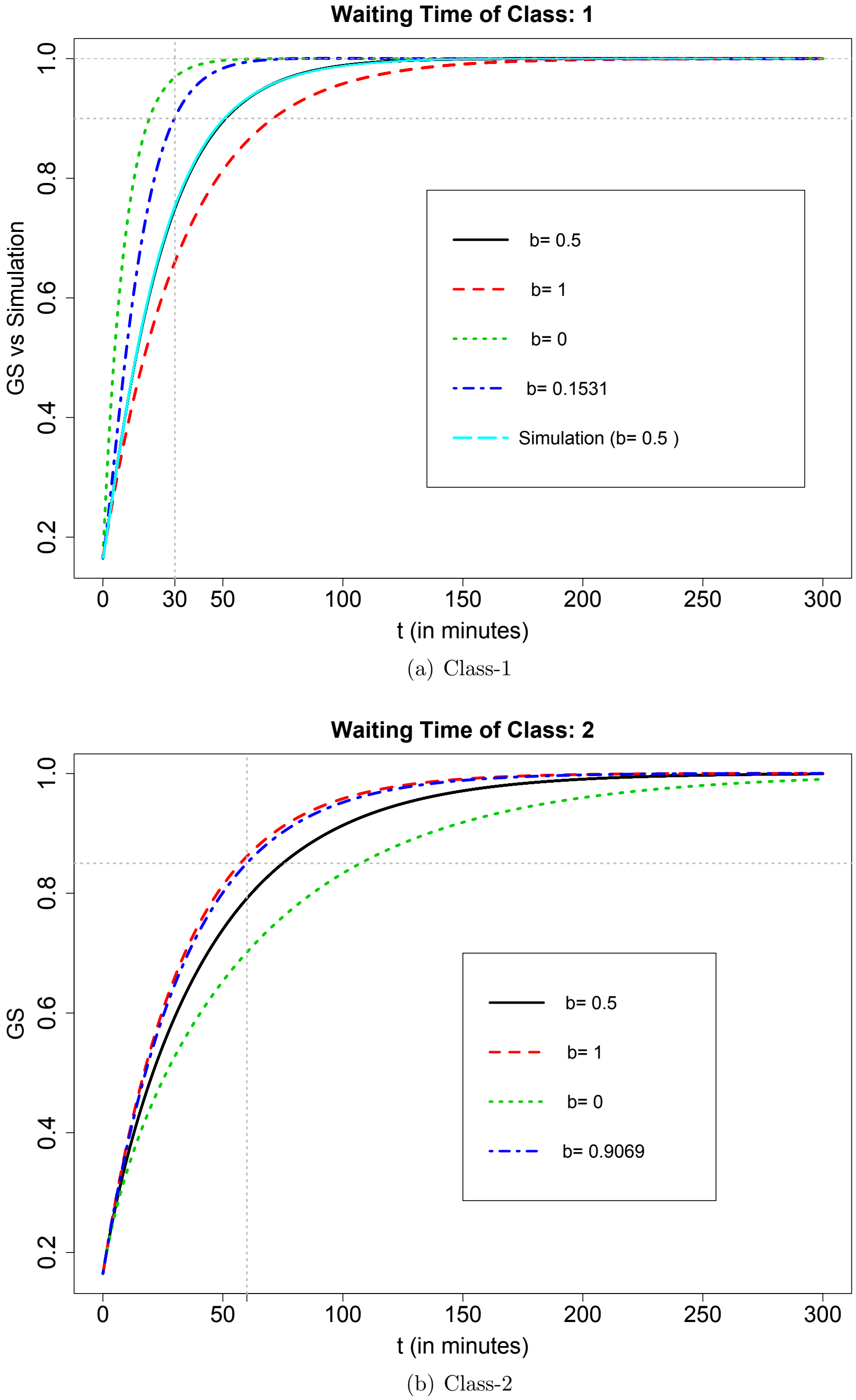
3.7 Numerical investigations

Our first numerical example is loosely drawn from the CTAS model, assuming servers working at heterogeneous rates. We mainly focus on CTAS-3 (urgent) and CTAS-4 (less urgent) patients with the APQ model. From Table 3.1, the KPI-3 suggests that CTAS-3 patients should be seen within 30 minutes at least 90% of the time; and the KPI-4 suggests that CTAS-4 patients be seen within 60 minutes at least 85% of the time.

We use the Gaver-Stehfest (GS) algorithm (Abate and Whitt [1]) with $M = 5$ to invert the LSTs of the waiting time distributions, which can provide four significant digits of accuracy. Simulation experiments were carried out to verify the results from GS. Each simulation run consisted of one million customers.

3.7.1 Gaver-Stehfest evaluation and simulation

We start with the results for a two-class, two-heterogeneous-server APQ model with 4 types of dispatch policies (RCS, FSF, SSF and RBS) comparing the class-1 (CTAS-3) with class-2 (CTAS-4) patients, based on the following parameters: the arrival rates are $\lambda_1 = 0.9$ and $\lambda_2 = 0.8$ for class-1 and class-2 patients, respectively, while the service times are exponentially distributed with $\mu_1 = 1.9$ and $\mu_2 = 0.1$ (resulting in $\mathsf{G} = 0.9$) for server-1 and server-2 respec-



(a) Class-1

(b) Class-2

Figure 3.3: The GS evaluation of the waiting time distribution function for class-1 and class-2 customers under RCS (G = 0.9).

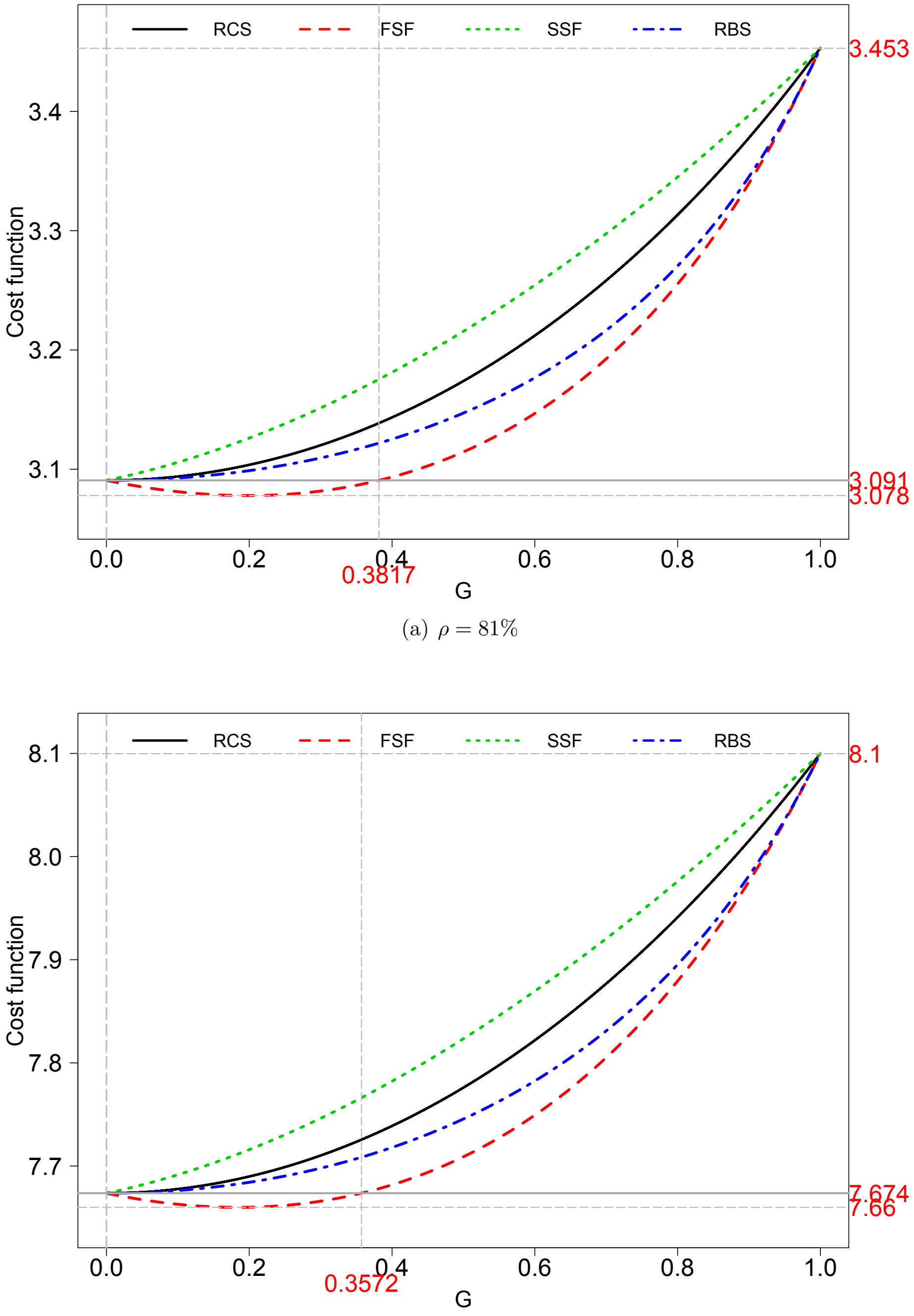
tively. Thus, the resulting occupancy level ρ is 85% with the assumed parameters. In all of our examples, we set $b_1 = 1$ and let $b_2 \equiv b$ which assumes one of the three values in $\{0, 0.5, 1\}$.

Figure 3.3 shows the waiting time distributions of class-1 and class-2 under RCS. We note that the GS evaluation and the simulation result for $b = 0.5$, which was run as a check, overlap each other, thereby confirming the accuracy of the results. In Figure 3.3 (a), we notice the cumulative distribution function (c.d.f.) of the waiting time for class-1 is stochastically smallest when $b = 0$ (i.e., classical priority) and stochastically largest when $b = 1$ (i.e., FCFS), as expected. To ensure class-1 patients meet their CTAS requirement, the maximum value of b is 0.1531. Similarly, in Figure 3.3 (b), the c.d.f. of the waiting time for class-2 is stochastically smallest when $b = 1$ and stochastically largest when $b = 0$. A minimum value of $b \geq 0.9069$ is required in order for class-2 patients to meet their target. Thus, there is no common value of b so that both classes can simultaneously meet their KPIs for the given arrival and service rates.

We also have tested our model in the homogeneous case with a common service rate $\mu_1 = \mu_2 = 1$ ($\mathsf{G} = 0$), where the parameters used are the same as Sharif *et al.* [21]. We remark that our results are in agreement with theirs. In the homogeneous case, the maximum value of b to allow class-1 patients to meet their requirement is 0.1647, and the minimum value of b for class-2 to meet their target is 0.825, so we can see that heterogeneity makes it harder to comply with the KPIs.

3.7.2 The impact of the level of heterogeneity

Next we address the impact of the level of heterogeneity on the chosen cost function $F(2, r, \mathsf{G})$ for the two-server case, as a measure to compare the system under different dispatch policies. We present a series of graphs with G to represent different levels of heterogeneity among the servers on the x-axis and the values of the cost function on the y-axis, under the assumption of $\lambda_1 = \lambda_2$ and $\mu_a = 2$. For the first scenario, we set $\rho = 81\%$, which is the maximum ρ to comply with the KPIs in the two-server case as shown in Table 3.2. The second scenario is for $\rho = 90\%$, which can be considered as an example of the heavily loaded case. In Figure 3.4, the dark grey horizontal line indicates the cost function of the homogeneous system. Several facts can be observed based on Figure 3.4:



(a) $\rho = 81\%$

(b) $\rho = 90\%$

Figure 3.4: Range of the level of heterogeneity that minimizes the cost function.

- As the level of heterogeneity increases, the value of the cost function increases and the system performance decreases.

- The performance of the heterogeneous system can be better than the homogeneous system at some levels of heterogeneity, under FSF. The optimal areas are bounded by the dark grey horizontal line and the red curve on Figure 3.4.

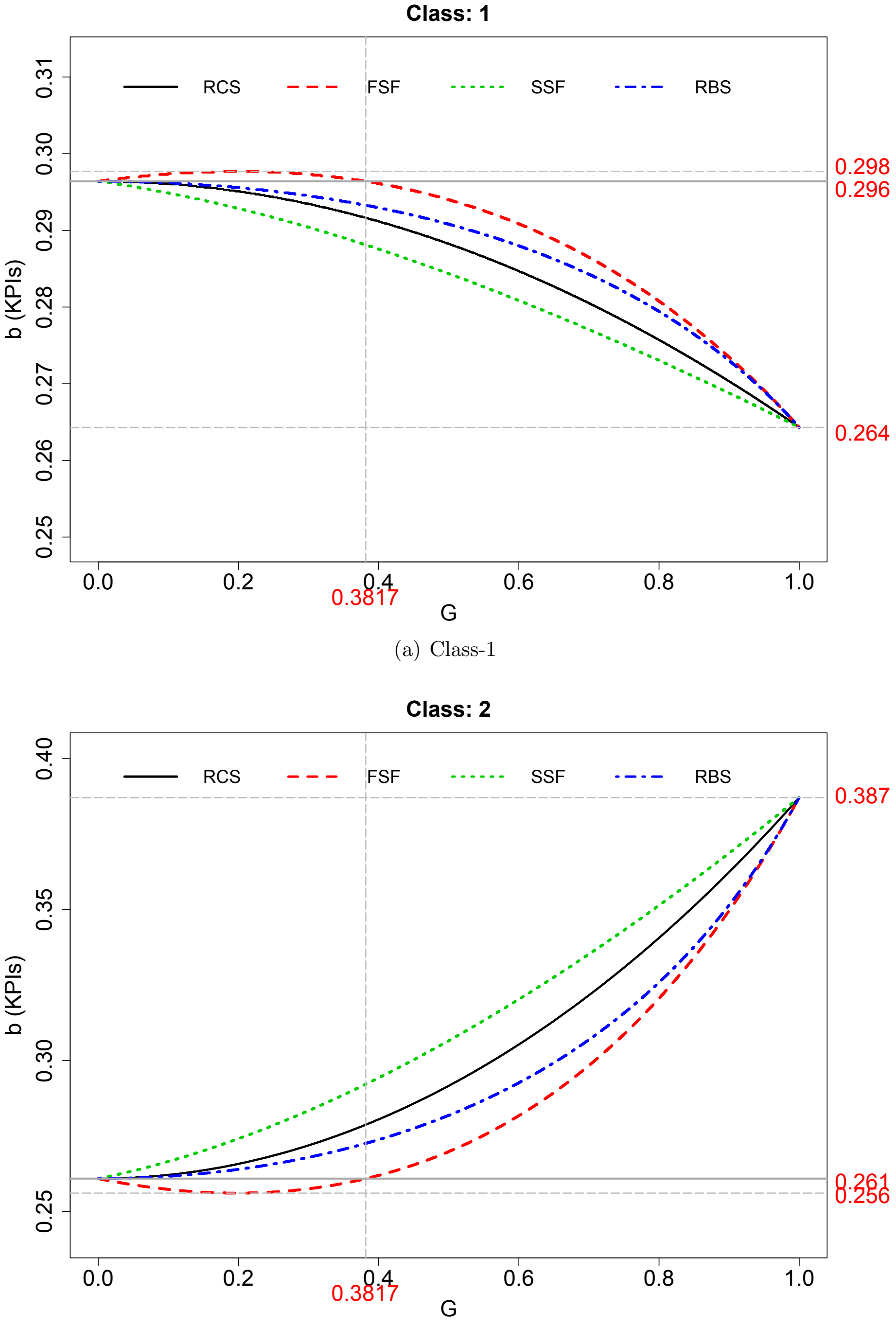
- The system performance can be ranked, with FSF being best, followed by RBS, and then RCS, and lastly SSF, as anticipated. RBS is more likely to select a faster server than RCS.

- When the utilization level increases, the optimal area for which FSF outperforms the homogeneous system is diminished in size.

We next explore the effect of the level of heterogeneity on the value of b for which both KPI targets can be met. To answer this question, we have plotted the extremal values of b against G for each class in the two-server case with $\lambda_1 = \lambda_2$, $\mu_a = 2$ and $\rho = \rho_{max} = 81\%$. In the case of Figure 3.5 (a) and (b), these correspond to the maximal values of b for which the class-1 and class-2 KPI targets are met, respectively. From Figure 3.5 we observe that

- The maximal values of b for the class-1 KPI target is decreasing as the level of heterogeneity increases, and the range of maximal values is bounded by FSF and SSF, while the minimal values of b for the class-2 KPI target is increasing with the level of heterogeneity.

- The FSF dispatch policy gives the greatest range of b values for which both classes meet their targets. In Figure 3.5, b needs to be less than 0.298 for class-1 patients, while b needs to be greater than 0.256 for class-2 patients. For both classes to meet their KPI targets when $\mathsf{G} = 0.19$, b must satisfy $0.256 \leq b \leq 0.298$, which is the greatest range of b under the utilization level $\rho = 81\%$. When G approaches one, the values of b for class-1 KPI is less than 0.264, and for class-2 KPI is greater than 0.387, which means



(a) Class-1

(b) Class-2

Figure 3.5: Relationship between the value of $b_2 = b$ for which KPI targets can be met and level of heterogeneity (G) with equal arrival rates and two servers ($\rho = 81\%$).

that there is no region of b for which both classes meet their targets. This is consistent with the results from Table 3.2.

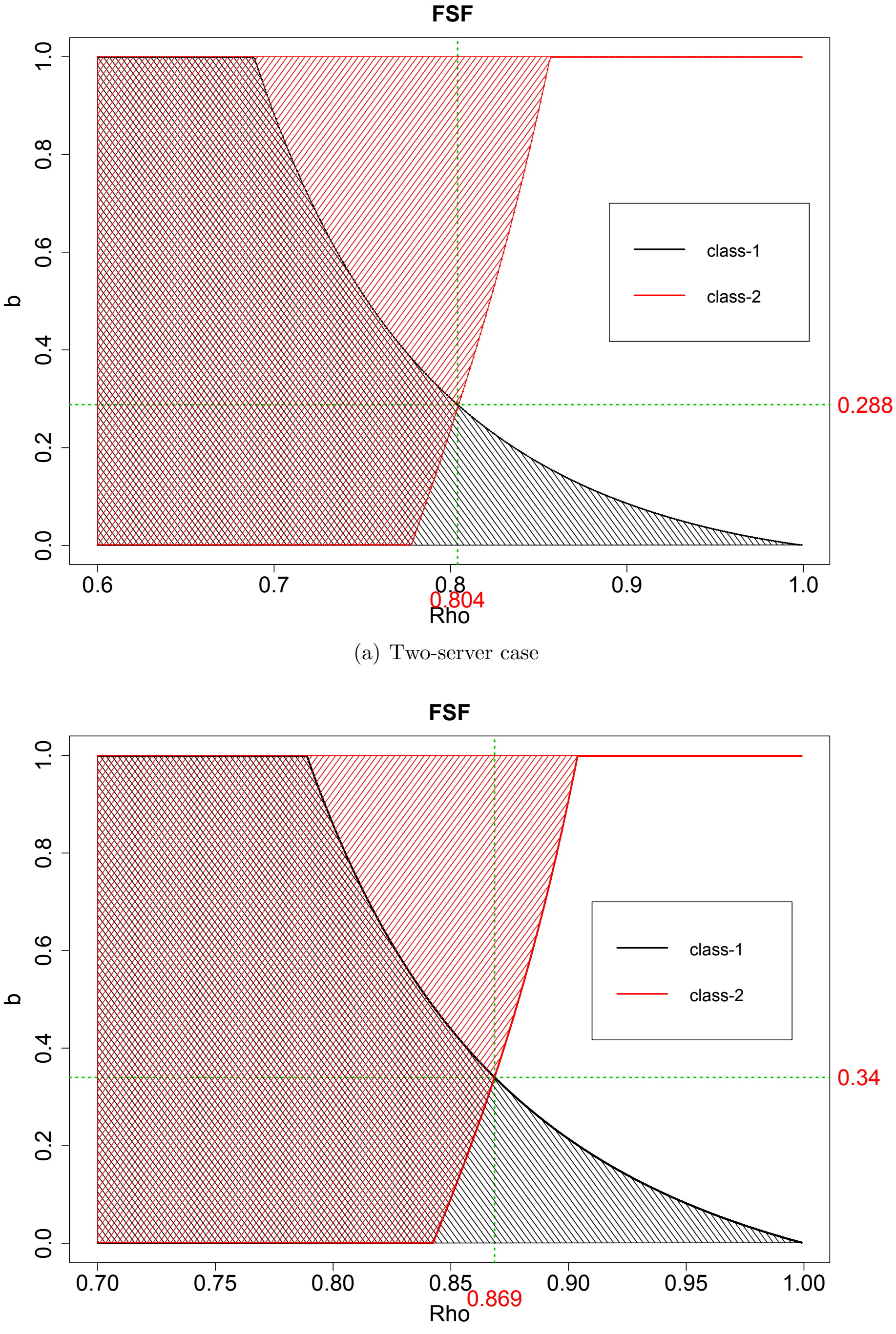
- The optimal range of the level of heterogeneity that promises the maximum range of b for the KPI targets is the same for both classes of patients, and the same as the one for the cost function in Figure 3.4 (a).

3.7.3 What are the optimal b and maximum ρ?

As we noted above with $\rho = 85\%$, we cannot find a common b so that both classes of customers can meet their KPI targets. It is therefore a matter of interest to wonder what values of b and utilizations ρ enable both classes to meet their KPIs. In order to answer this question, we illustrate a "feasible region" for such a goal by plotting the total utilization ρ on the x-axis and the values of b along the y-axis. The feasible region which contains the permissible combinations of b and ρ is the doubly-shaded area bounded by the maximal rate b for which the class-1 KPI can be met (black curve) and the minimum rate of b for which the class-2 KPI can still be met (red curve).

We compare the two-class two-server case and three-server case with equal arrival rates $\lambda_1 = \lambda_2$ and a fixed sum of service rates $\mu_a = 2$ with $\mathsf{G} = 0.9$ under FSF, that is, $\mu_1 = 1.9$ and $\mu_2 = 0.1$. Since $\rho = \lambda/\mu_a$ where $\mu_a = 2$ and $\lambda_1 = \lambda_2 = \lambda/2$, then $\lambda_1 = \lambda_2 = \rho$. To be noted, the stationary probability $\pi(\boldsymbol{\mu}, 3; r)$, for $r = \{-\infty, 0, 1, \infty\}$, are calculated by solving the global balance equations. The expressions are listed in Appendix A.3.

Figure 3.6 (a) presents the two-server case. We found that when $\rho > 80.4\%$, no value of b can allow both classes of customers to meet their KPI targets simultaneously; when $\rho \leq 80.4\%$, there is at least one b value for which both classes simultaneously meet their KPIs. Thus, we call the specific utilization level that distinguishes the single permissable value of b for which both classes meet their KPI targets as the *maximum* ρ, and the corresponding b as the *optimal* b. In this case, the *optimal* b is 0.288. Figure 3.6 (b) presents the three-server case with equal arrival rates, $\mu_a = 3$ and $\mathbf{G} = \{0, 0.31, 0.9\}$, that is, $\mu_1 = 1.9$, $\mu_2 = 1$ and $\mu_3 = 0.1$. The two



(a) Two-server case

(b) Three-server case

Figure 3.6: Permissible range of values of $0 < b_2 = b < 1$ to meet the class-1 and class-2 KPIs in the two-server and three-server case under FSF (G = 0.9).

Table 3.2: Maximum ρ and optimal b with different levels of heterogeneity

	$c = 2$			$c = 3$		
Heterogeneity	Policies	ρ_{max} (%)	Optimal b	Policies	ρ_{max} (%)	Optimal b
G = 0	All	81.19	0.2860	All	87.16	0.3370
	SSF	80.93	0.2862	SSF	87.09	0.3390
G = 0.4	RCS	81.05	0.2859	RCS	87.10	0.3387
	RBS	81.10	0.2860	RBS	87.12	0.3387
	FSF	81.20	0.2858	FSF	87.15	0.3389
	SSF	80.25	0.2858	SSF	86.86	0.3400
G = 0.9	RCS	80.31	0.2876	RCS	86.89	0.3410
	RBS	80.41	0.2876	RBS	86.91	0.3408
	FSF	80.42	0.2881	FSF	86.94	0.3413
G = 1	All	80.08	0.2868	All	80.08	0.2868

graphs appear similar; however, the values of the maximum ρ increased to 86.9% and optimal b increased to 0.34 as the number of servers increased.

Table 3.2 presents the maximum ρ and optimal b for the two-class two-server case and three-server case with equal arrival rates and a fixed sum of service rates at different levels of heterogeneity under four dispatch policies: SSF, RCS, RBS, FSF. We observe that the value of ρ_{max} decreases as the level of heterogeneity increases. Furthermore, the differences in the optimal b values for the various dispatch rules under a given level of heterogeneity occur at the third decimal place.

References

[1] Abate, J., & Whitt, W. (2006). A unified framework for numerically inverting Laplace transforms. INFORMS Journal on Computing. 18, 408–421.

[2] Alves, F. S. Q., Yehia, H. C., Pedrosa, L. A. C., Cruz, F. R. B., & Kerbache, L. (2011). Upper bounds on performance measures of heterogeneous $M/M/c$ queues. Mathematical Problems in Engineering. Vol. 2011.

[3] Canadian Association of Emergency Physicians. The Canadian Triage and Acuity Scale (CTAS). From the website. $http\ ://www.calgaryhealthregion.ca/policy/docs/1451/Admission\_over\_capacity\_AppendixA.eps$. 105, Table 2.1.

[4] Conway, R. W., Maxwell, W. L., & Miller, L. W. (2003). Theory of scheduling. Dover Publications. Reprint edition. Massachusetts.

[5] Doroudi, S., Gopalakrishnan, R., & Wierman, A. (2011). Dispatching to incentivize fast service in multi-server queues. ACM SIGMETRICS Perform. Eval. Rev. 39(3), 43–45.

[6] Grassmann, W. (2000). Computational probability. Kluwer Academic Publishers. Boston.

[7] Grassmann, W., & Zhao, Y. Q. (1997). Heterogeneous multiserver queues with a general input. INFOR. 35, 208–224.

[8] Gumbel, H. (1960). Waiting lines with heterogeneous servers. *O*perations Research. 8(4), 504–511.

[9] Kesten, H., & Runnenberg, J. Th. (1957). Priority in waiting line problems. Nederlandse Akademie van Wetenschappen. Proceedings. Series A. Indagationes Mathematicae.

[10] Kleinrock, L. (1964). A delay dependent queue discipline. Naval Research Logistics. Quarterly 11, 329–341.

[11] Kleinrock, L. (1965). A conservation law for a wide class of queueing disciplines. Naval Research Logistics. Quarterly 12, 181–192.

[12] Kleinrock, L., & Finkelstein, R. (1967). Time dependent priority queues. *O*perations Research. 15, 104–116.

[13] Kleinrock, L. (1975). Queueing systems Vol I: theory. Wiley. New York.

[14] Kleinrock, L. (1976). Queueing systems Vol II: computer applications. Wiley. New York.

[15] Krishnamoorthi, B. (1963). On Poisson queue with two heterogeneous servers. *O*perations Research. 11, 321–330.

[16] Little, J. D. C. (1961). A proof for the queuing formula: $L = \lambda W$. *O*perations Research. 9(3), 383–387.

[17] Mokaddis, G. S., & Matta, C. H. (1998). On Poisson queue with three heterogeneous servers. Information and Management Sciences. 9, 53–60.

[18] Morse, P. M. (1958). Queues, inventories and maintenance. Wiley. New York, 82–84.

[19] Saaty, T. L. (1960). Time dependent solution of the many-server Poisson queue. *O*perations Research. 8, 768–771.

[20] Shalit, H. (1985). Calculating the Gini index of inequality for individual data. Oxford Bulletin of Economics and Statistics. 47, 185–189.

[21] Sharif, A. B., Stanford, D. A., Taylor, P., & Ziedins, I. (2014). A multi-class multi-server accumulating priority queue with application to health care. Operations Research for Health Care. 3, 73–79.

[22] Sharma, O. P., & Dass, J. (1989). Initial busy period analysis for a multichannel Markovian queue. Optimization. 20, 317–323.

[23] Singh, V. P. (1970). Two-server Markovian queues with balking: heterogeneous vs. homogeneous servers. *O*perations Research. 18(1), 145–159.

[24] Stanford, D. A., Taylor, P., & Ziedins, I. (2014). Waiting time distributions in the accumulating priority queue, Queueing Systems. 77, 297–330.

[25] Wolff, R. W. (1982). Poisson arrivals see time averages. *O*perations Research. 30(2), 223-231.

Chapter 4

On waiting times for nonlinear accumulating priority queues

4.1 Introduction

Long waiting times for various groups of customers have been a recurring problem in many public systems, particularly in healthcare systems. Among the different techniques used to manage such queues are various sorts of priority queuing discipline, which are appropriate when diverse requirements of different classes of customer need to be met. However, in a classical priority queue in which customers of a given priority are selected for service only when there are no waiting higher priority customers, it is possible that customers from lower priority classes may be overtaken repeatedly, and have to wait for an extremely long time. Especially in a healthcare setting, where an extended wait might result in deterioration of a patient's condition, this is undesirable, and may well be unsustainable.

In order to deal with a similar problem in the context of computer processor design, a modification to the classical priority discipline was proposed in 1964 by Kleinrock [4]. He suggested that customers be allowed to accumulate priority as a linear function of their time in the queue, at a rate that depends upon their urgency or classification. Thus, a low-priority customer in a stable queue will eventually accrue enough priority to enter service no later, and typically at an earlier point in time, than if their waiting time were ignored. Kleinrock [4] termed such a system a *time dependent priority queue*, and he derived a set of recursive formulas for the expected waiting times for the different classes when arrivals are assumed to be Poisson.

Recently, Stanford *et al.* [7] reconsidered Kleinrock's model, renamed as the "accumulating priority queue" (APQ), obtaining the waiting time distribution for each class of customers in terms of its Laplace-Stieltjes transform (LST). Their work was motivated by many healthcare applications whose performance is governed by Key Performance Indicators (KPIs), see for instance CTAS [2], where the performance is measured in terms of the proportion of patients in a given acuity class who commence treatment before a specified time threshold. By manipulating the rates of priority accumulation in the various classes, Stanford *et al.* [7] and its successor Sharif *et al.* [6] showed that one can fine-tune an accumulating priority system to meet stated KPIs that might not be met by a classical priority system.

A natural extension to the models of Kleinrock [4] and Stanford *et al.* [7], is to allow a customer's priority to accumulate as a nonlinear function of its waiting time. Such an extension might seem to give decision makers more flexibility. For example, in a healthcare setting, it might be a good idea to allow priority to increase at an increasing rate as a patient waits, to reflect the fact that treatment could become even more urgent as the waiting time increases.

Figure 4.1 illustrates a set of priority accumulation functions that medical professionals might possibly consider. In the figure, the accumulated priority of the kth class of patients is in the form of $f_k(t) = 1/\left(1 + e^{-(c_k t-10)}\right) - 1/(1 + e^{10})$, $k = 1, 2, 3$. Class-1 patients become much more urgent around time $t = 10$, whereas class-2 patients do so around time $t = 20$. Class-3 patients become urgent more slowly over the time interval 20 – 50 mins.

Kleinrock and Finkelstein [5] proposed a class of APQs, which we shall call *power-law APQs*, in which priority accrues as a power r of the incurred waiting time, and called these *rth order* systems. Theorem 1 of Kleinrock and Finkelstein [5] stated that, for two systems of orders r_1 and r_2 respectively, parameters can be found that yield the same expected waiting times for customers from different classes, whichever system is used. The proof of their theorem

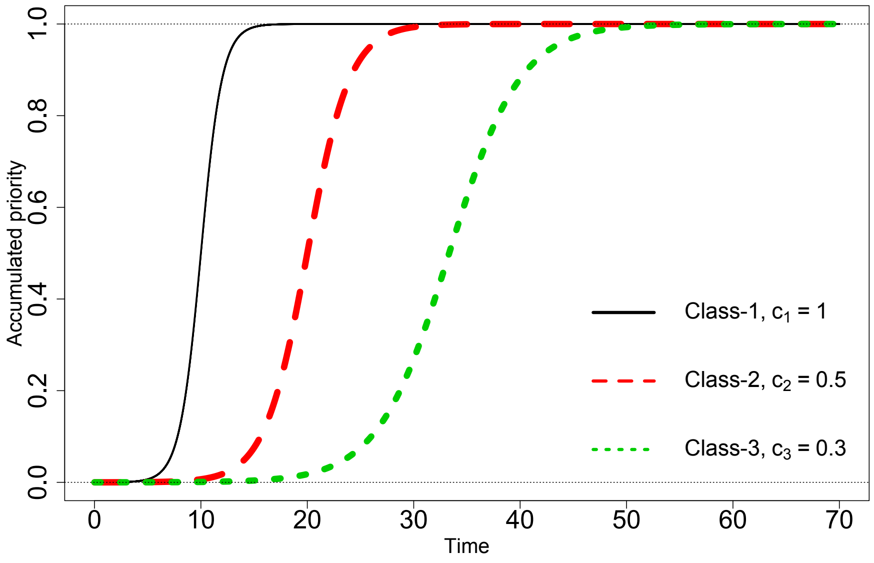


Figure 4.1: An example of priority accumulation function.

used sample path arguments for the three-customer case, and an extension of the reasoning to any number of customers. Taking $r_1 = 1$, their theorem established that an rth order system is equivalent to a linearly-increasing priority system in the sense that the expected waiting times for customers of each class in both systems are the same. Invoking the results of Kleinrock [4], Kleinrock and Finkelstein [5] used this observation to obtain a set of equations for the expected waiting times for all classes of customers in a given rth order system.

In this paper, we study a queuing system with Poisson arrivals, generally distributed service times and nonlinear priority accumulation functions. We start with the power-law APQ of Kleinrock and Finkelstein [5], and use a general argument to show that there is a linear system of the form discussed in Stanford *et al.* [7] which has the same priority ordering of all customers present at any given instant in time, for any sample path. (We shall refer to the linear system as the linear *proxy* of the nonlinear one.) Beyond the power-law case, we subsequently show that, if appropriate conditions are met for more general nonlinear APQs, we can construct an equivalent linear APQ in the sense that the waiting time distributions of all classes are the same in both systems. It will turn out that the priority functions illustrated in Figure 4.1 satisfy these conditions. Our method of proof employs only the priority accumulation functions, and as such it is equally applicable to multi-server queues such as those studied by Sharif *et al.* [6].

The remainder of this paper proceeds as follows. A description of our model and some preliminary definitions are given in Section 4.2. A discussion of the waiting time distributions for the power-law APQ appears in Section 4.3. The main theorems for nonlinear APQs and their proofs appear in Section 4.4, and a set of recursive expressions for the LST of the waiting time distributions for the nonlinear APQ with a linear proxy are shown in Section 4.5. Some examples of the nonlinear priority accumulation functions with linear proxies are illustrated in Section 4.7. The results of some numerical experiments are presented in Section 4.8.

4.2 Model description

We consider a multi-class single-server APQ with Poisson arrivals and general service time distributions and a nonlinear priority accumulation discipline. There are $K \in \mathbb{N}$ classes of customers, with lower-indexed classes requiring more favourable treatment in terms of waiting times than higher-indexed classes. (In a health care setting, this means that a lower index reflects a patient class of higher acuity.) For $k = 1, 2, \ldots, K$, customers of class k arrive according to a Poisson process with rate λ_k, independently of the arrival processes of all the other classes, and request a service time with distribution G_k and mean $1/\mu_k$.

Define

$$\begin{aligned} \lambda &= \sum_{k=1}^{K} \lambda_k, && (4.1)\\ 1/\mu &= \sum_{k=1}^{K} \lambda_k/(\lambda\mu_k), && (4.2)\\ \rho_k &= \lambda_k/\mu_k, && (4.3)\\ \rho &= \lambda/\mu = \sum_{k=1}^{K} \rho_k. && (4.4) \end{aligned}$$

The accumulating priority queuing discipline is defined in terms of a set $\mathcal{F}$ of *priority accumulation functions* $\{f_k(.), k = 1, \ldots, K\}$ that govern how waiting customers are selected when a service is completed. Specifically, a customer from class k who arrived at time t_0 has priority

$$q_k(t) = f_k(t - t_0) \tag{4.5}$$

at time $t > t_0$. At a service completion instant, the server chooses to serve that customer among those still present in the queue that has the greatest accumulated priority, provided that the queue is non-empty. A customer who arrives to an empty queue moves straight into service.

The functions in $\mathcal{F}$ have the following properties:

1. for $k = 1, \ldots, K$, f_k is a strictly increasing, differentiable function that maps $\mathbb{R}_+$ to $\Xi \subseteq \mathbb{R}_+$,

2. for $k = 1, \ldots, K - 1$, $f_k(0) = f_{k+1}(0)$,

3. for $j = 2, \ldots, K$ and $k < j$, $f_k(t) > f_j(t)$ for all $t > 0$.

4. for $k = 2, \ldots, K$, $j < k$ and $\delta > 0$, if there is a time $t^* > \delta$ such that $f_j(t^* - \delta) = f_k(t^*)$, then $f_j'(t^* - \delta) > f_k'(t^*)$.

Intuitively, we think of the priority of a customer from a lower-indexed class as increasing 'faster' than a customer from a higher-indexed class. This results in preferential treatment for the customer in the lower-indexed class. From a technical point of view, this concept is captured in Conditions 3 and 4.

A consequence of Condition 1 is that, for all k, the inverse function f_k^{-1} that maps Ξ to $\mathbb{R}_+$ exists. For some priority value $v \in \Xi$, $f_k^{-1}(v)$ is the time in the queue at which a class-k customer attains a priority v.

Condition 2 ensures that customers from all classes start with the same priority. Without loss of generality, we can take $f_k(0) = 0$ for $k = 1, \ldots, K$. This condition can be relaxed if we wish to model the *affine case* where customers in different classes start with different amounts of initial priority when they arrive. However, we choose not to do so in this work.

Condition 3 ensures that a customer with a lower priority index can never be caught up by a customer with a higher priority index that arrived later, while Condition 4 requires that the priority of the lower-indexed customer is increasing at a faster rate than that of the higher-indexed customer at any point where their priority curves meet. A consequence is that, if the priority of a class-k customer is caught up by that of a class $j < k$ customer who arrived later, then the class-j customer can never be re-overtaken by the class-k customer. We call the time t^* in 4 the *overtake time* of a class-k customer by a class $j < k$ customer that arrived δ time units later.

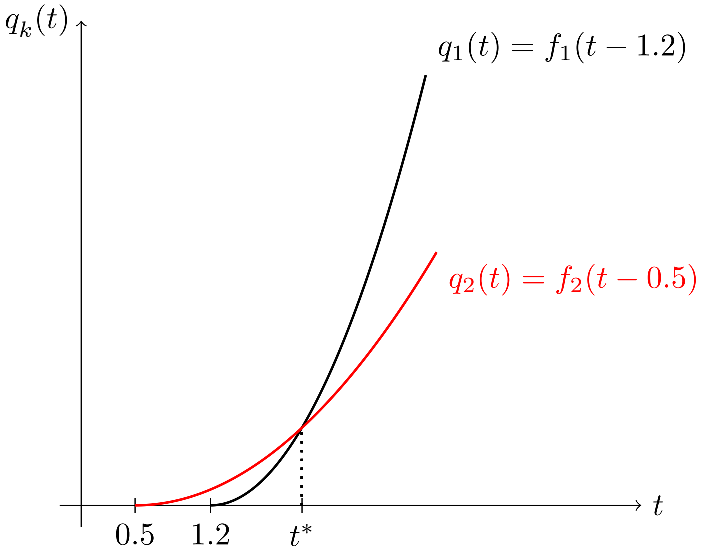


Figure 4.2: Accumulating priority functions for different priority classes.

First, we consider the case of Figure 4.2, which illustrates nonlinear accumulation functions for two customers from different priority classes. Here

$$\begin{aligned} f_1(t) &= t^2, \\ f_2(t) &= 0.3t^2. \end{aligned} \tag{4.6}$$

A customer from class 1 arrives at time point 1.2 and a customer from class 2 arrives at time point 0.5. Both customers accumulate priority according to equation (4.5) based on their waiting time in the system. However, the customer from class 1 accumulates priority more quickly. If the server becomes free before $t^* = 2.05$, the customer from class 2 will be selected into service before the class 1 customer, whereas if the server becomes free after $t^* = 2.05$, the reverse is true.

Note that it does not follow from Conditions 1 to 3 that there is *always* an overtake time for a lower-indexed customer that arrives after a higher-indexed customer. Consider, for example, the case where there are two classes of customers with respective priority functions

$$\begin{aligned} f_1(t) &= t + 1 - e^{-t}, \\ f_2(t) &= t. \end{aligned} \tag{4.7}$$

We have

$$\begin{aligned} f_1'(t) &= 1 + e^{-t}, \\ f_2'(t) &= 1 < f_1'(t). \end{aligned} \tag{4.8}$$

A waiting class-1 customer would eventually overtake a waiting class-2 customer if the former arrives within one unit of time of the arrival time of the latter, but not otherwise. See Figure 4.3.

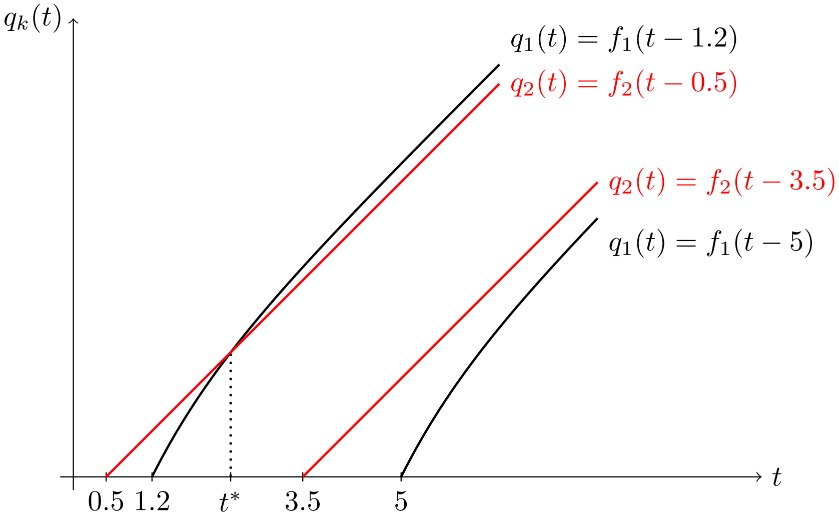


Figure 4.3: An example of two accumulating priority functions.

Of course, each customer that arrives initiates its own $q_k(t)$. Figure 4.4 illustrates the accumulated priorities of customers against time for a sample path of a process with two classes and the accumulation functions from equation (4.6). Customers of class 1 arrive at time points 1, 4 and 6.4, while customers of class 2 arrive at time points 2 and 5.2. The departure instants (4.4, 7, 8, 9.4, 11.6) are marked by vertical lines. At the first departure instant 4.4, the class-2 customer who arrived at time point 2 moves into service, since the class-1 customer who arrived at time point 4 has not accumulated enough priority to overtake it. The opposite case occurs when the third customer departs at time point 8. Here, the class-1 customer who arrived at time point 6.4 has overtaken the class-2 customer who arrived at time point 5.2. The overall sequence of services is: class-1, class-2, class-1, class-1, and class-2.

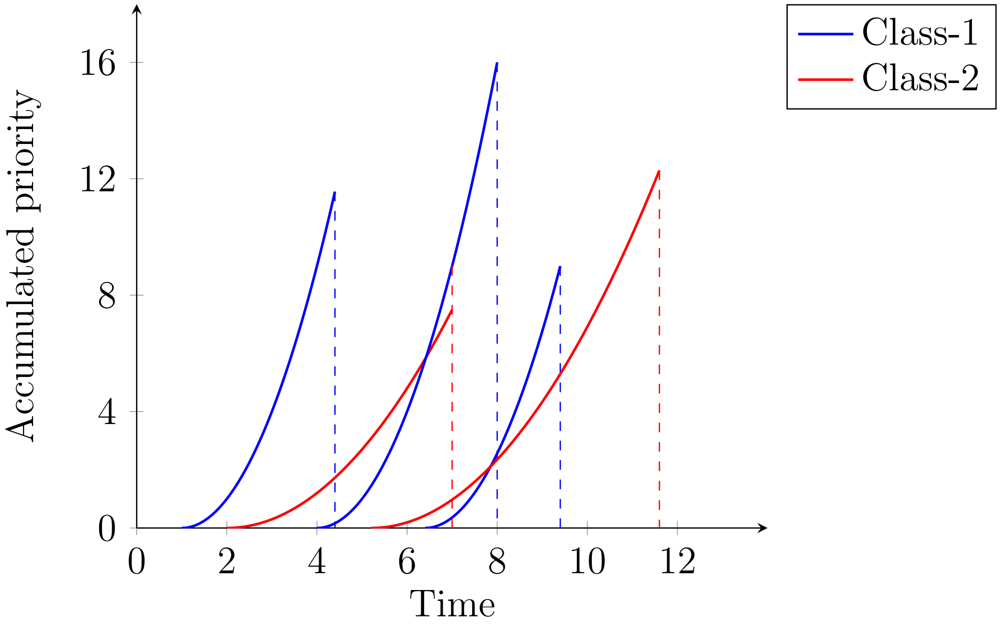


Figure 4.4: A sample path of an APQ with a number of different customers.

With these definitions, we are ready to consider the power-law APQ first considered by Kleinrock and Finkelstein [5].

4.3 Waiting times for power-law APQs

Kleinrock and Finkelstein [5] studied the expected waiting times in a multi-class single-server queue with exponential service times for different classes of customers, where priority accumulates as a power function of the incurred waiting time. They named it an "r-th order delay dependent priority discipline". We shall refer to a queue that employs this discipline as a *power-law accumulating priority queue (APQ)* of order r. In our model, customers from the different priority classes can have different generally-distributed service times.

The set $\mathcal{F}$ of priority accumulation functions for the power-law APQ of order r is defined in terms of a sequence$\{b_k^{(r)}\}, k = 1, \ldots, K$ of positive constants such that $b_1^{(r)} \geq b_2^{(r)} \geq \cdots \geq b_K^{(r)} \geq 0$. For $k = 1, 2, \ldots, K$, we take $f_k(t) = b_k^{(r)} t^r$. When $r = 1$, this reduces to the linear APQ, originally defined by Kleinrock [4] and discussed in Stanford *et al.* [7].

Kleinrock and Finkelstein [5, Theorem 1] established that if one were to select the constants so that

$$\left(b_{k+1}^{(r)}/b_k^{(r)}\right)^{1/r} = \left(b_{k+1}^{(r')}/b_k^{(r')}\right)^{1/r'} \quad \text{for } k = 1, 2, \ldots, K, \tag{4.9}$$

then the expected waiting times of all customer classes in the corresponding power-law APQs of orders r and r' would be identical.

We can invoke the results of Stanford *et al.* [7] by taking $r' = 1$ and writing $b_k^L = b_k^{(1)}$ for $k = 1, 2, \ldots, K$. Kleinrock and Finkelstein's equation (4.9) then states that if we put

$$b_k^L = (b_k^{(r)})^{1/r}, \tag{4.10}$$

the expected waiting times of all classes of customers in the rth order APQ are the same as those in the linear APQ.

At an arbitrary time instant $t \geq t_k$ in the power-law APQ with the $b_k^{(r)}$ given by equation (4.10), the priority of a customer from class k arriving at time t_k is $q_k^{(r)} = \left(b_k^L(t - t_k)\right)^r$ for $k = 1, 2, \ldots, K$, whereas in the corresponding linear APQ, the priority of the class-k customer is $q_k = b_k^L(t - t_k)$. We see that the priority accumulated under the linear APQ is at all times the rth root of the original power-law priority accumulation function.

We refer to this linear APQ as the linear "proxy" for the original power-law APQ and we show below that, at any instant in time, the accumulated priority ordering of all customers in the system will be the same under both disciplines.

For a given realisation of the arrival and service time process, let $\Gamma^{(r)}(t)$ and $\Gamma(t)$ denote the sets of customers in the various classes present in a power-law APQ of order r and its linear proxy at time t.

Lemma 4.3.1. *Consider a power-law APQ of order r and its linear proxy, both starting empty and driven by the same realisations of the arrival and service time processes. Then for any time $t \in \mathbb{R}$,*

$$\Gamma^{(r)}(t) = \Gamma(t), \tag{4.11}$$

and the ordering of the priorities of all customers is the same in both systems.

Proof. Since both queues are driven by the same realisation of the arrival process, the arrival times τ_n and the class $\chi(n)$ of the nth arriving customer are the same in both queues. Whether it is still in the queue or not at time t, the priority of the nth customer, with arrival time $\tau_n < t$, is $b^{(r)}_{\chi(n)}(t - \tau_n)^r$ in the power-law APQ and $b^L_{\chi(n)}(t - \tau_n)$ in its linear proxy. We observe that

- The ordering of these functions at time t is the same in both systems amongst all customers that have arrived by time t.

To complete the proof, we need to show that the same customers are selected for service in both systems. The service time of the customer initiating the first busy period will be the same in both systems and, when this customer completes service, the customer selected for the next service will be the same in both systems, by the above observation. The same set of customers is therefore present when this customer completes service, and so on. The conclusions of the lemma follow. □

The waiting time distributions for the linear APQ have been derived by Stanford *et al.* [7]. By applying their results via the linear proxy, we obtain the LST of the waiting time distribution for each class in the power-law APQ. The results are stated below.

Theorem 4.3.2. *Consider a single-server APQ which has a set $\mathcal{F}$ of priority accumulation functions given by $f_k(t) = b^{(r)}_k t^r$ for $k = 1, 2, \ldots, K$ and some positive parameter r. Conditional on class-k customer arriving to a non-empty queue, the LST $\tilde{W}^{(k)}_+(s; r)$ of its waiting time is* *given by*

$$\tilde{W}_{+}^{(k)}(s;r) \;=\left(1-\left(\frac{b_{k+1}^{(r)}}{b_{k}^{(r)}}\right)^{\frac{1}{r}}\right)\tilde{W}_{acc}^{(k)}(s;r)+\left(\frac{b_{k+1}^{(r)}}{b_{k}^{(r)}}\right)^{\frac{1}{r}}\tilde{W}_{+}^{(k+1)}((b_{k+1}^{(r)}/b_{k}^{(r)})^{\frac{1}{r}}s;r) \tag{4.12}$$

where

$$\begin{aligned}\tilde{W}_{acc}^{(k)}(s;r) &= \tilde{W}_{+}^{(k+1)}((b_{k+1}^{(r)}/b_{k}^{(r)})^{\frac{1}{r}}s;r)\sum_{j=1}^{k}\frac{\rho_j(b_{k+1}^{(r)}/b_{j}^{(r)})^{\frac{1}{r}}}{1-\delta_k}\tilde{W}_{acc}^{(k,j)}(s;r)\\ &+\sum_{j=k+1}^{K}\frac{\rho_j}{1-\delta_k}\tilde{W}_{+}^{(j)}((b_{j}^{(r)}/b_{k}^{(r)})^{\frac{1}{r}}s;r)\tilde{W}_{acc}^{(k,j)}(s;r)+\frac{1-\rho}{1-\delta_k}\tilde{W}_{acc}^{(k,0)}(s;r).\end{aligned} \tag{4.13}$$

and $\delta_k = \sum_{j=1}^{k}\rho_j\left(1-(b_{k+1}^{(r)}/b_{j}^{(r)})^{\frac{1}{r}}\right)$.

Proof. For a given parameter r, the $\tilde{W}_{acc}^{(k,0)}(s;r)$ in equation (4.13) denotes the LST of the distribution of the waiting time incurred by a class-k customer who becomes accredited during the initial accreditation interval in a busy period; while the $\tilde{W}_{acc}^{(k,j)}(s;r)$ for $j>0$ represents the LST of the distribution of the waiting time incurred by a class-k customer who becomes accredited during a later accreditation interval initiated by a class-j service time in a busy period.

The result follows directly after substituting for the linear accumulation rates $b_k^L = (b_k^{(r)})^{1/r}$ in the expression of Corollary 8.4 of Stanford *et al.* [7]. □

Combining this result with the fact that the probability that a customer arrives to find an empty queue with probability $1-\rho$, we see that the LST of the waiting time for class-k customers in the power-law APQ has LST given by

$$\tilde{W}^{(k)}(s;r) = (1-\rho)+\rho\tilde{W}_{+}^{(k)}(s;r). \tag{4.14}$$

Remark Note that there is a typographical error in the expression given in Stanford *et al.* [7]: in the part of that expression corresponding to the first sum on the right hand side of equation (4.13), the expression should have b_{k+1}/b_j, not b_{k+1}/b_k, as appears there. This has been corrected in equation (4.13).

4.4 Linear proxies for general nonlinear APQs

Having established that there is a linear proxy for the power-law APQ, it is natural to ask whether linear proxies can be found for other nonlinear APQs. Furthermore, it is of interest to determine what requirements these particular nonlinear APQs would need to satisfy. In this section, we consider a general set $\mathcal{F}$ of nonlinear priority accumulation functions satisfying Conditions 1 to 4.

Theorem 4.4.1. *Consider a nonlinear APQ with accumulation functions $\{f_k\}$. Assume that there exists a linear APQ with associated rates c_{jk} such that, for all possible values of the input arrival times $\boldsymbol{\tau} = \{\tau_n; n = 1, 2, \ldots\}$ and customer types $\boldsymbol{\chi} = \{\chi(n); n = 1, 2, \ldots\}$, the overtake times of the nonlinear APQ and the linear APQ are identical. Then, for $k \in \{2, \ldots, K\}$ and $j < k$,*

$$\frac{f_k'(t)}{f_j'(f_j^{-1}(f_k(t)))} = c_{jk}. \tag{4.15}$$

In particular it is constant for all time $t \in \mathbb{R}_+$.

Proof. Let $k \in \{2, \ldots, K\}$ and $j < k$ and consider the set of sample paths in which a class-k customer arrives at time 0 and a class-j customer arrives at some time τ_2 later, which we will regard as variable. Then the time in the linear APQ at which the type j customer overtakes the type k customer is $u = \tau_2/(1 - c_{jk})$. This time must satisfy the *overtake time equation*

$$\tau_2 + f_j^{-1}(f_k(u)) = u \tag{4.16}$$

for the nonlinear APQ. So we know that, for all $\tau_2 \in \mathbb{R}_+$,

$$\tau_2 + f_j^{-1}\left(f_k\left(\frac{\tau_2}{1 - c_{jk}}\right)\right) = \frac{\tau_2}{1 - c_{jk}}. \tag{4.17}$$

Equation (4.17) is equivalent to

$$f_j^{-1}\left(f_k\left(\frac{\tau_2}{1 - c_{jk}}\right)\right) = \frac{c_{jk}\tau_2}{1 - c_{jk}} \tag{4.18}$$

which is, in turn, equivalent to

$$f_k\left(\frac{\tau_2}{1 - c_{jk}}\right) = f_j\left(\frac{c_{jk}\tau_2}{1 - c_{jk}}\right). \tag{4.19}$$

Differentiating equation (4.19) with respect to τ_2, we derive

$$\frac{f_k'\left(\frac{\tau_2}{1-c_{jk}}\right)}{1-c_{jk}} = \frac{c_{jk}f_j'\left(\frac{c_{jk}\tau_2}{1-c_{jk}}\right)}{1-c_{jk}}, \tag{4.20}$$

which implies that

$$\frac{f_k'\left(\frac{\tau_2}{1-c_{jk}}\right)}{f_j'\left(\frac{c_{jk}\tau_2}{1-c_{jk}}\right)} = c_{jk} \tag{4.21}$$

and, by equation (4.18),

$$\frac{f_k'\left(\frac{\tau_2}{1-c_{jk}}\right)}{f_j'\left(f_j^{-1}\left(f_k\left(\frac{\tau_2}{1-c_{jk}}\right)\right)\right)} = c_{jk}. \tag{4.22}$$

This is true for all $\tau_2 \in \mathbb{R}_+$, which establishes that equation (4.18) holds for all $t \in \mathbb{R}_+$. □

Corollary 4.4.2. *A nonlinear APQ has an equivalent linear APQ in the sense explained in Theorem 4.4.1, if and only if, for all $k = 2, \dots, K$ and $j < k$, there exist numbers $c_{jk} \in (0,1)$ such that*

$$f_k(t) = f_j(c_{jk}t), \quad t > 0. \tag{4.23}$$

Remark Any set of numbers c_{jk} satisfying equation (4.15) must be such that, for $j < k < \ell$,

$$c_{j\ell} = c_{jk}c_{k\ell}. \tag{4.24}$$

To see this, from equation (4.23), we have $f_\ell(t) = f_j(c_{j\ell}t)$ and $f_\ell(t) = f_k(c_{k\ell}t) = f_j(c_{jk}c_{k\ell}t)$, $\Rightarrow c_{j\ell} = c_{jk}c_{k\ell}$.

If we take $\{c_{1k}, k = 1, 2, \dots, K\}$ to be any set of numbers that are strictly decreasing in k, then by writing $c_{1k} = \Pi_{i=1}^{k-1} c_{i,i+1} = c_{1j}c_{jk}$ for $1 < j < k$, we can ensure that equation (4.24) is satisfied. Furthermore, we can take $c_{11} = 1$ without loss of generality. Now putting $c_k = c_{1k}$, Corollary 4.4.2 leads immediately to the following corollary.

Corollary 4.4.3. *A nonlinear APQ has an equivalent linear APQ in the sense explained in Theorem 4.4.1 if and only if there is a decreasing sequence of numbers where $c_1 = 1$, $c_k \in (0,1)$; $k = 2, \dots, K$, and a function g such that*

$$f_k(t) = g(c_k t). \tag{4.25}$$

A suitable linear proxy for such an APQ is obtained by setting $b_k^L = c_k$.

We have used the insight gained from the accreditation processes to arrive at Corollary 4.4.3. Now that we see the result, we can go further, in the spirit of Lemma 4.3.1, and show that, if the priority accumulation functions are of the form as equation (4.25), then the orderings of all customers in the general APQ and it linear proxy are identical when driven by the same realisations of the arrival and service time processes. We state this result formally below.

Theorem 4.4.4. *Let* $\Gamma^N(t)$ *and* $\Gamma(t)$ *denote the sets of customers in the various classes present at time* t *in a general nonlinear APQ with priority accumulation functions of the form as equation (4.25) and its linear proxy, both starting empty and driven by the same realisations of the arrival and service time processes. Then for any time* $t \in \mathbb{R}$,

$$\Gamma^N(t) = \Gamma(t), \tag{4.26}$$

and the ordering of the priorities of all customers is the same in both systems.

Proof. Since both queues are driven by the same realisation of the arrival process, the arrival times τ_n and the class $\chi(n)$ of the nth arriving customer are the same in both queues. Whether it is still in the queue or not at time t, the priority of the nth customer, with arrival time $\tau_n < t$, is $g(c_k(t - \tau_n))$ in the nonlinear APQ and $c_k(t - \tau_n)$ in its linear proxy. By Condition 1, of Section 4.2, g must be monotone increasing, and so

- The ordering of these functions at time t is the same in both systems amongst all customers that have arrived by time t.

As with Lemma 4.3.1, to complete the proof we just need to show that the same customers are selected for service in both systems. The argument for this is identical to the one we used there. The service time of the customer initiating the first busy period will be the same in both systems and, when this customer completes service, the customer selected for the next service will be the same in both systems, and so on. □

4.5 Waiting time distributions for the nonlinear APQ with a linear proxy

In Section 4.4, we established conditions under which a nonlinear APQ has a linear proxy. In this formulation, the constant c_{jk} in equation (4.15) plays the role of the ratio b_k^L/b_j^L of slopes of the accumulation functions in the linear proxy. This observation enables us to derive the LST of the waiting time distributions in the nonlinear queue by applying the results of Stanford *et al.* [7] with b_k/b_j replaced by c_{jk}. This leads us to the following result.

Theorem 4.5.1. *For a single-server nonlinear APQ which has a linear proxy, then the LST of the delayed waiting time distribution for a class-k customer, $\tilde{W}_+^{(k)}(s)$, is given by*

$$\tilde{W}_+^{(k)}(s) = (1 - c_{k,k+1})\tilde{W}_{acc}^{(k)}(s) + c_{k,k+1}\tilde{W}_+^{(k+1)}(c_{k,k+1}s), \tag{4.27}$$

where

$$\begin{aligned}\tilde{W}_{acc}^{(k)}(s) = \frac{1-\rho}{1-\delta_k}\tilde{W}_{acc}^{(k,0)}(s) + \tilde{W}_+^{(k+1)}(c_{k,k+1}s)\sum_{j=1}^{k}\frac{\rho_j c_{j,k+1}}{1-\delta_k}\tilde{W}_{acc}^{(k,j)}(s) \\ + \sum_{j=k+1}^{K}\frac{\rho_j}{1-\delta_k}\tilde{W}_+^{(j)}(c_{kj}s)\tilde{W}_{acc}^{(k,j)}(s),\end{aligned} \tag{4.28}$$

and $\delta_k = \sum_{j=1}^{k}\rho_j(1 - c_{j,k+1})$.

The LST of the waiting time for class-k in the nonlinear APQ is

$$\tilde{W}^{(k)}(s) = (1-\rho) + \rho\tilde{W}_+^{(k)}(s). \tag{4.29}$$

4.6 Discussion on the maximum priority process for a nonlinear APQ

We start our discussion by defining the *maximum priority process* for such queues. At each instant in time, this process provides an upper bound on the accumulated priority of customers from each class present in the system. Using the maximum priority process, we are able to study the evolution of the APQ on a "need to know" basis, not tracking the identities of all customers present in the queue and all of their accumulated priorities, but instead taking advantage of the fact that, conditional on the maximum priorities, the customers' actual priorities are distributed as a Poisson process.

First, let us review some notation and definitions from Stanford *et al.* [7]. The sequence of intervals $\boldsymbol{T} = \{T_n; n = 1, 2, \dots\}$ is the process of inter-arrival times, with T_1 being the time of the first arrival and $\tau_n = \sum_{k=1}^{n} T_k$ being the time of the n^{th} arrival. For each n, $\chi(n)$ is the class and X_n is the service time of the n^{th} customer, with $\boldsymbol{\chi} = \{\chi(n); n = 1, 2, \dots\}$ and $\boldsymbol{X} = \{X_n; n = 1, 2, \dots\}$. For a positive integer m, $n(m)$ is defined as the position in the arrival sequence of the mth customer to be served. C_n is the time at which service commences for the nth arrival, and $D_n = C_n + X_n$ is the departure time of the nth customer to arrive, with $\mathbf{C} = \{C_n; n = 1, 2, \dots\}$ and $\mathbf{D} = \{D_n, n = 1, 2, \dots\}$. Thus, the departure of the mth customer to be served occurs at time $D_{n(m)} = C_{n(m)} + X_{n(m)}$.

The maximum priority process for a multi-class nonlinear APQ is defined as follows.

1. For all times t when the queue is empty, $M_k(t) = 0$ for all $k = 1, 2, \dots, K$.

2. At the sequence of successive departure times $D_{n(m)}$,

$$M_1(D_{n(m)}) = \max_{n \notin \{n(i): 1 \le i \le m\}} q_{\chi(n)}(D_{n(m)}), \tag{4.30}$$

   and, for $1 < k \le K$,

$$M_k(D_{n(m)}) = \min\{M_1(D_{n(m)}), f_k(X_{n(m)} + f_k^{-1}(M_k(C_{n(m)})))\}. \tag{4.31}$$

3. For $t \in [C_{n(m)}, D_{n(m)})$ with $\max_{m: D_{n(m)} > t} q_{\chi(m)}(t) > 0$,

$$M_k(t) = f_k(t - C_{n(m)} + f_k^{-1}(M_i(C_{n(m)}))), \quad 1 \le k \le K. \tag{4.32}$$

To illustrate the concept of the maximum priority process, let us consider a two-class nonlinear APQ. Figure 4.5 plots $M_1(t)$ and $M_2(t)$ against time t for an example sample path. $M_1(t)$ bounds the accumulated priorities at time t of all customers present in the queue. $M_2(t)$ bounds

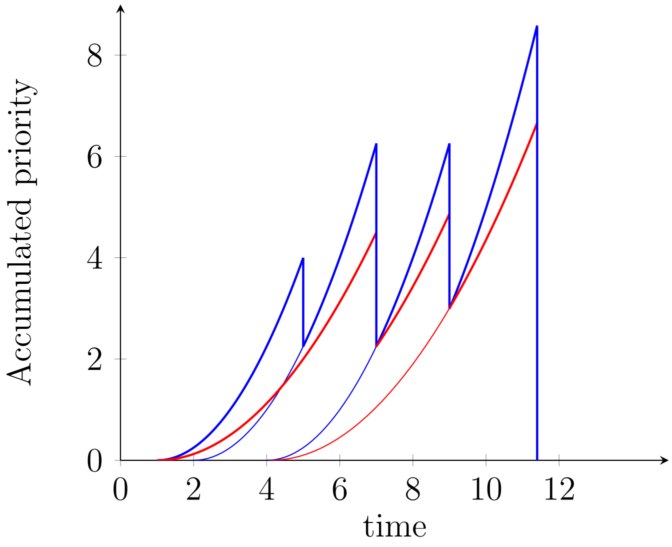


Figure 4.5: Maximum priorities.

the accumulated priorities of class-2 customers. Suppose that the queue starts empty and the first busy period begins at time $\tau_1 = C_{n(1)}$ with $M_1(\tau_1) = M_2(\tau_1) = 0$. At any time t during the first service time, any waiting class-2 customer (which must necessarily have arrived after τ_1) must have priority less than $f_2(t - \tau_1)$.

At the first departure instant, $D_{n(1)} = D_1 = \tau_1 + X_1$. Denoting the maximum priority of all the customers in the queue by V, one of the following three conditions must hold:

1. The queue is empty, and we set $M_1(D_{n(1)}) = M_2(D_{n(1)}) = 0$.

2. If $V \geq M_2(D_{n(1)}-)$, as in Figure 4.6, since $M_2(D_{n(1)}-)$ is an upper bound of the class-2 customers, the customer with priority V must be a class-1 customer. At any time t during the next service, the least upper bound for the class-1 customers is $f_1(t - D_{n(1)} + f_1^{-1}(V))$, reflecting the fact that the class-1 customer arrived at time $D_{n(1)} - f_1^{-1}(V)$, while the priority of the class-2 customers is still bounded by $f_2(t - \tau_1)$.

3. If $V < M_2(D_{n(1)}-)$, as in Figure 4.7, the customer with priority V can be of either class. At any time t during the next service, the least upper bound for class-1 customers is $f_1(t - D_{n(1)} + f_1^{-1}(V))$, and the priority of class-2 customers is bounded by $f_2(t - D_{n(1)} + f_2^{-1}(V))$.

If the customer is of class 1, it arrived at time $D_{n(1)} - f_1^{-1}(V)$, whereas if the customer is of class 2, it arrived at time $D_{n(1)} - f_2^{-1}(V)$.

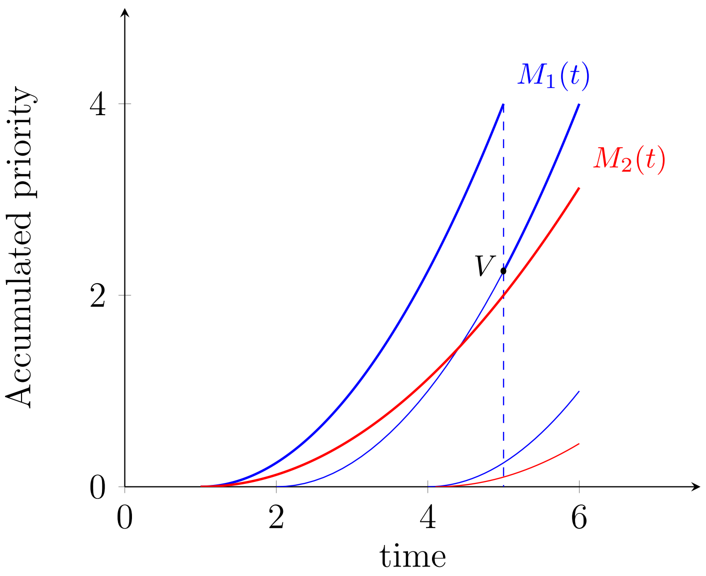


Figure 4.6: The priority process at a departure time: case 2.

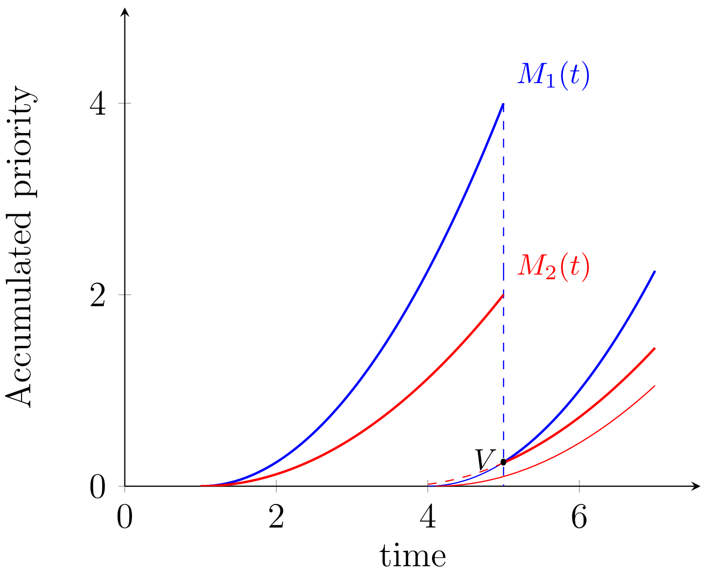


Figure 4.7: The priority process at a departure time: case 3.

Suppose a nonlinear APQ has a set $\mathcal{F}$ of priority accumulation functions $f_k(t)$ for $k = 1, 2, \ldots, K$. We say that a class $j < k$ customer *becomes accredited with respect to class k* when its priority $q_j(t)$ intersects the maximum priority function $M_k(t)$. So, customers from lower-indexed classes become accredited with respect to a higher-indexed class when their priority overtakes the maximum possible priority that a customer from the higher-indexed class can have. The following lemma describes the process via which customers become accredited.

Lemma 4.6.1. *For $k = 2, \ldots, K$, $j < k$, and $t \in [C_n, D_n)$, the process Ψ_{jk} of points at which customers of class j become accredited with respect to class k is a non-homogeneous Poisson process with intensity*

$$\lambda_{jk}(t) = \lambda_j(1 - c_{jk}(t)) \tag{4.33}$$

where

$$c_{jk}(t) = \frac{M_k'(t)}{f_j'(f_j^{-1}(M_k(t)))}. \tag{4.34}$$

Proof. By its definition in equation (4.32), the component $M_k(t)$ of the maximum priority process is given by

$$M_k(t) = f_k(t - C_n + f_k^{-1}(M_k(C_n))) \tag{4.35}$$

during the interval $[C_n, D_n)$. Observe that, conditional on $M_k(C_n)$, this process evolves deterministically.

It takes time $f_j^{-1}(v)$ for a customer of class j to attain a priority v. So customers of class j who become accredited with respect to class k during the interval $[C_n, t] \subseteq [C_n, D_n)$ must have arrived during the interval $[C_n - f_j^{-1}(M_k(C_n)), t - f_j^{-1}(M_k(t)))$. The arrival process of class-j customers on this latter interval is a homogeneous Poisson process with parameter λ_j and the process Ψ_{jk} is a transformation of this Poisson process that has no atoms. By the Poisson Mapping Theorem, see, for example, [3, page 17], Ψ_{jk} must also be a Poisson process, with mean measure

$$\lambda_j \left(t - f_j^{-1}(M_k(t)) - C_n - f_j^{-1}(M_k(C_n))\right) \tag{4.36}$$

on sets of the form $[C_n, t]$. Taking the derivative with respect to t, we see that Ψ_{jk} has intensity

$$\lambda_j \left(1 - \frac{d(f_j^{-1}(M_k(t)))}{dt}\right) = \lambda_j \left(1 - \frac{M_k'(t)}{f_j'(f_j^{-1}(M_k(t)))}\right) \tag{4.37}$$

at the point t. □

Note that equation 4.35 ensures that $M_k'(t)$ is the same as $f_k'(t-C_n+f_k^{-1}(M_k(C_n)))$. Condition 4 of our definition of the set $\mathcal{F}$ ensures that the ratio $c_{jk}(t)$ is less than one.

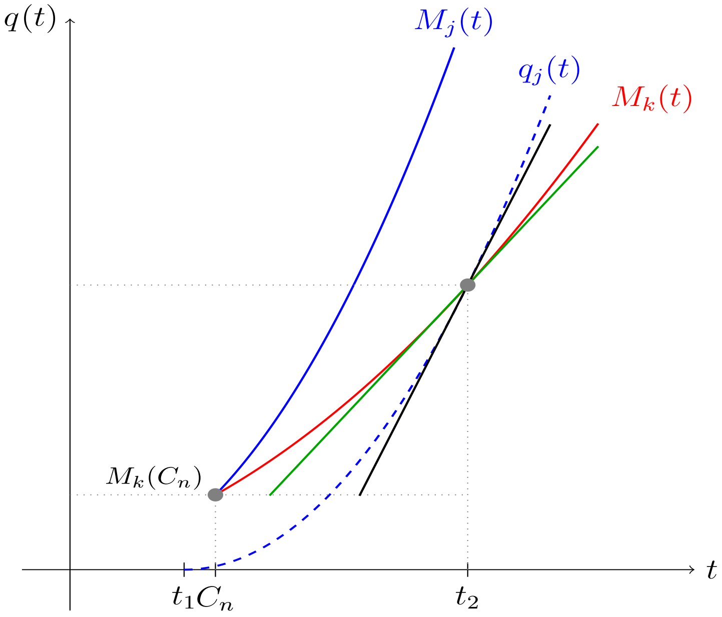


Figure 4.8: A customer of class j becoming accredited with respect to class k.

The process of a class-j customer becoming accredited with respect to class k is illustrated in Figure 4.8. The service of the current customer starts at the point C_n with (in this example) both $M_j(C_n)$ and $M_k(C_n)$ equal. The maximum priority functions $M_j(t)$ and $M_k(t)$ are shown in blue and red respectively. A class-j customer arrived at time t_1 becomes accredited with respect to class k at time t_2 when its priority function crosses the function $M_k(t)$. The tangents of the two functions are shown in black and green respectively.

Since any busy period can be decomposed into its constituent service times, Lemma 4.6.1 establishes that, during a busy period, the process of class-j customers becoming accredited with respect to class-$k; k > j$ is a non-homogeneous Poisson process. However, the intensity function (4.37) is randomly-varying because it depends on the random quantity $M_k(C_n)$ during each interval $[C_n, D_n)$. The key to our study of linear proxies for general nonlinear accumulating priority systems is to recognise that a linear proxy exists when the intensity function (4.37) is constant, independent of the realisation of the maximum priority process, which refers to Theorem 4.4.1.

4.7 Examples of nonlinear APQs

In this section, we list some examples of nonlinear APQs with linear proxies.

4.7.1 Power functions

In the framework of Section 4.4, the power-law APQ discussed in Section 4.3 can be modelled by taking $g(t) = t^r$.

By Corollary 4.4.3, for any decreasing sequence $c_k, k = 1, \dots, K$, the set $\mathcal{F}$ of functions $f_k(t) = g(c_k t) = (c_k t)^r = c_k^r t^r$ defines a nonlinear APQ with a linear proxy. The sequence of constants $\{b_k, k = 1, 2, \dots, K\}$ of Kleinrock and Finkelstein [5] is related to the sequence $\{c_k\}$ via the relation $b_k = c_k^r$.

4.7.2 Exponential functions

If we take $g(t) = \exp(t)$, then we see that, for any decreasing sequence $c_k, k = 1, \dots, K$, a nonlinear APQ with $f_k(t) = \exp(c_k t)$ for $k = 1, 2, \dots, K$ has a linear proxy. The waiting time distributions for this nonlinear APQ can be found using Theorem 4.5.1.

4.7.3 Logarithmic functions

A similar analysis can be performed by taking g to be a logarithmic function $g(t) = \log(t)$ and defining $f_k(t)$ via any decreasing sequence of numbers $\{c_k\}$ in $(0, 1)$ for $k = 1, \dots, K$.

4.8 Numerical examples

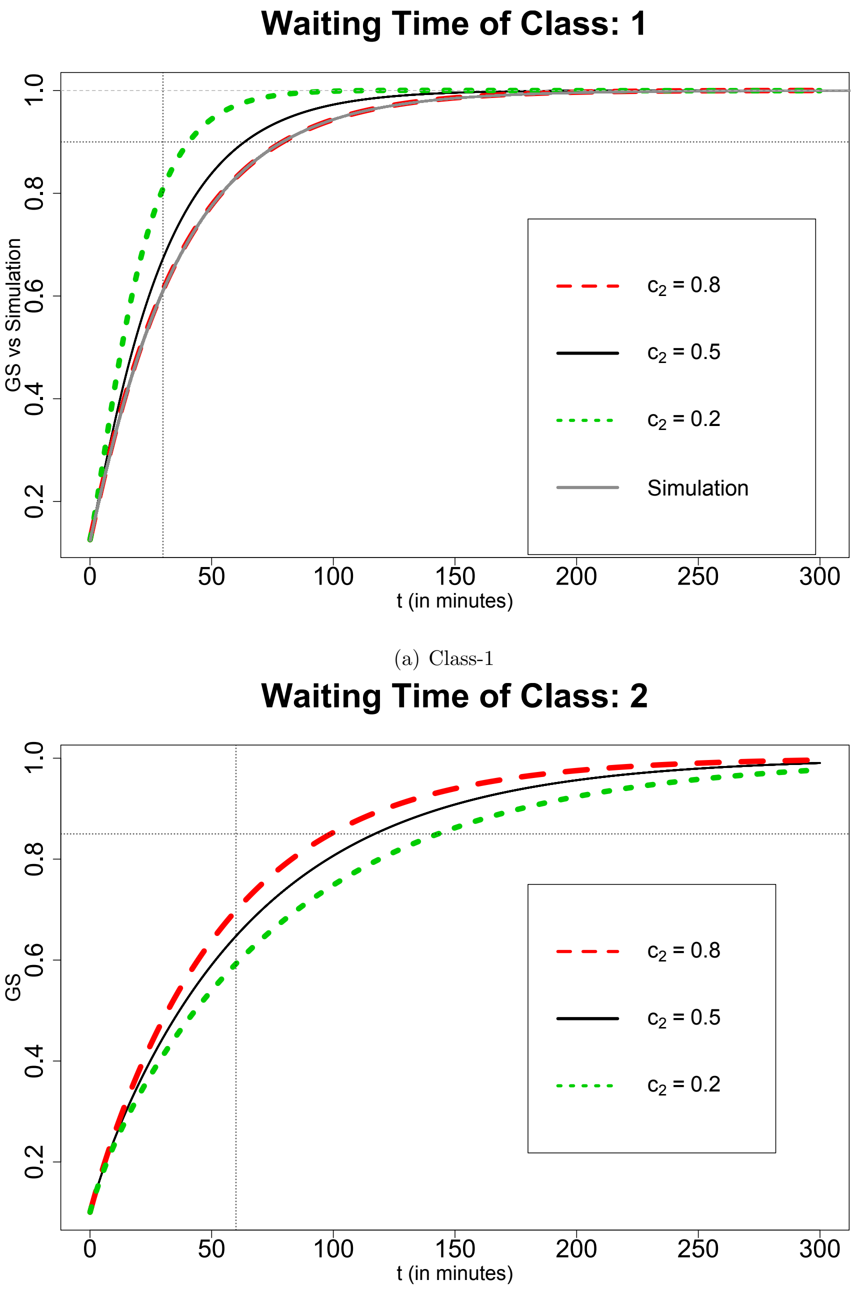
To verify the results of the LSTs of the waiting time distributions for a nonlinear APQ, we performed some numerical experiments for an APQ with the priority accumulation functions presented in the introduction and the power-law APQ. Our numerical examples are loosely drawn from the CTAS model [2], assuming that service times are exponential. We mainly focus on the urgent (CTAS-3) and less urgent (CTAS-4) patients. The key performance indicator (KPI) [2] suggests that CTAS-3 (urgent) patients should be seen within 30 mins at least 90% of the time; CTAS-4 (less urgent) patients should be seen within 60 mins at least 85% of the time.

The waiting time distributions were recovered from the LST expressions presented in Section 4.3 via numerical inversion using the Gaver-Stehfest (GS) algorithm in Abate and Whitt [1] with the parameter $M = 5$. Simulation experiments were carried out to verify the results from GS evaluation. Each simulation run consisted of two million customers.

Our first two numerical examples are for a two-class single-server APQ with arrival rates $\lambda_1 = 1$ and $\lambda_2 = 0.75$ for class-1 (CTAS-3) and class-2 (CTAS-4) respectively, while service times for both classes are exponentially distributed at a common rate $\mu_1 = \mu_2 = 2$.

In our first example of the nonlinear APQ, we take $g(t) = 1/\left(1 + e^{-(t-10)}\right) - 1/(1 + e^{10})$, and the accumulation functions $f_k(t) = g(c_k t)$ for $k = 1, 2$. The accumulation rate c_2 is evaluated at 0.2, 0.5 and 0.8, while c_1 is set to be one. The results of the waiting time distributions are shown in Figure 4.9. We note that the same curves would be obtained if we had used the linear functions $f_k(t) = c_k t$ for $k = 1, 2$ with c_k chosen as above, or for that matter, had we used any $f_k(t) = g(c_k t)$ for $k = 1, 2$ where $g(t)$ satisfies Corollary 4.4.3.

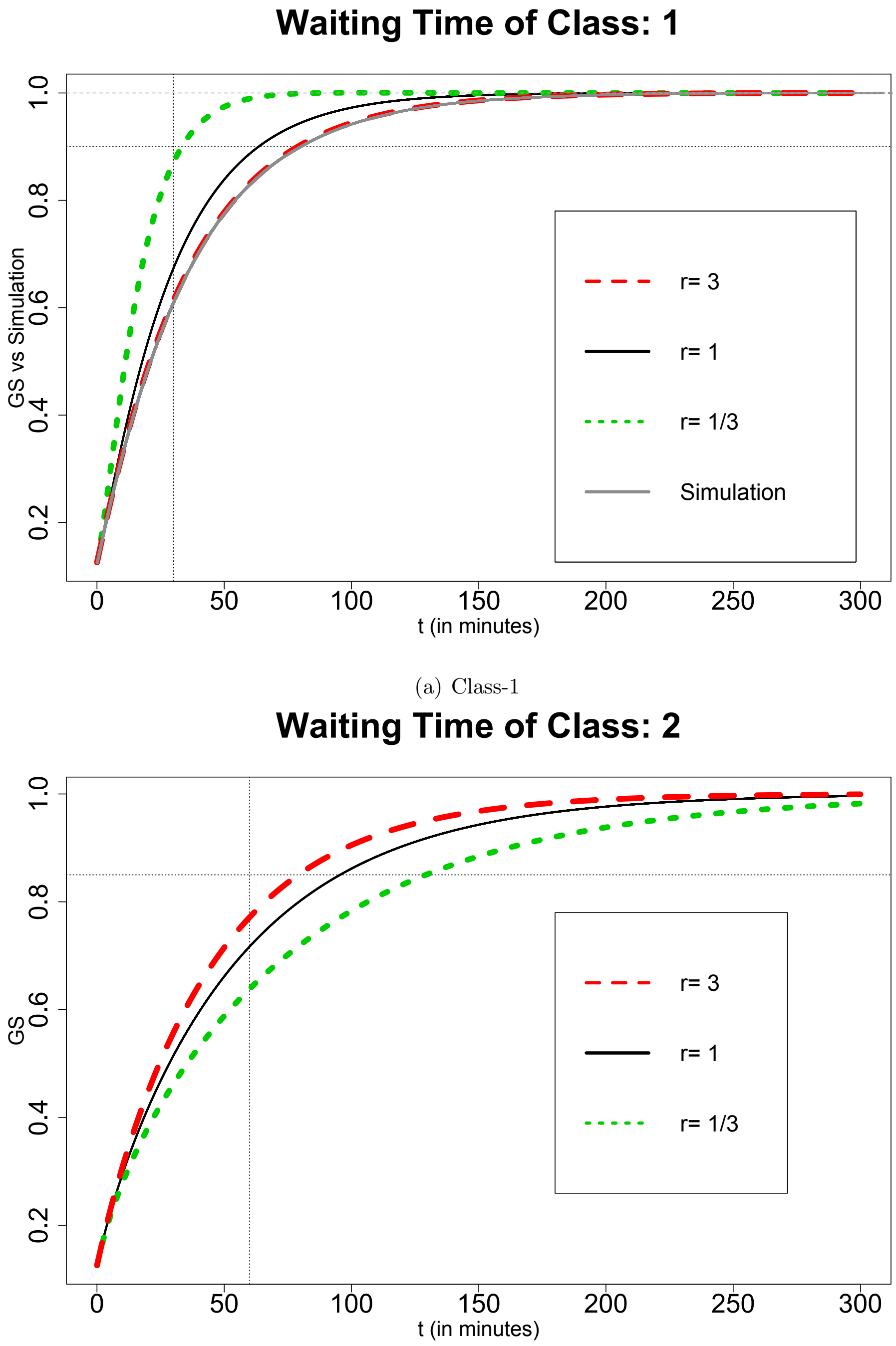
Figure 4.10 illustrates the waiting time distributions for class-1 and class-2 in the power-law APQ, where the proportionality constant b_1 for class-1 is set to unity, while $b_2 = 0.5$, and the power of the priority accumulation function, r, assumes one of three values: $r = 1/3, 1$ and 3. By comparing these two figures, we note that when r is smaller than one, the system favours class-1, whereas when r is greater than one, the reverse is true. If $r > 1$, then time is valued



(a) Class-1

(b) Class-2

Figure 4.9: Waiting time distributions for a two-class APQ with the accumulation functions proposed in Section 4.1.



(a) Class-1

(b) Class-2

Figure 4.10: Waiting time distributions for the power-law APQ.

supra-linearly, and one might anticipate this would lead to a disproportionate favouring of high acuity cases, where time is really critical. Instead, the opposite occurs: since $b_k^L = (b_k^{(r)})^{1/r}$, it favours the lower acuity classes. (This favouring is readily seen when $b_1^{(r)} = 1$ and the other $b_k^{(r)}$'s are all less than 1, but is equally true when this is not the case). As the value of r increases, the system favours class-2 more relative to a linear system with the same rate. The greater the value of r, the greater the value of the linear proxy's accumulation rate, which means the class-2 patients accumulate priority at a faster rate.

Observe that in both examples, we first note that the GS evaluation and simulation for class-1 with $c_2 = 0.8$ (Figure 4.9 (a) in red and gray respectively) and $r = 3$ (Figure 4.10 (a) in red and gray respectively) are virtually indistinguishable from the numerically-produced distributional curves. Second, at this utilization level of $\rho = 87.5\%$, it is not possible for both KPI targets to be met simultaneously.

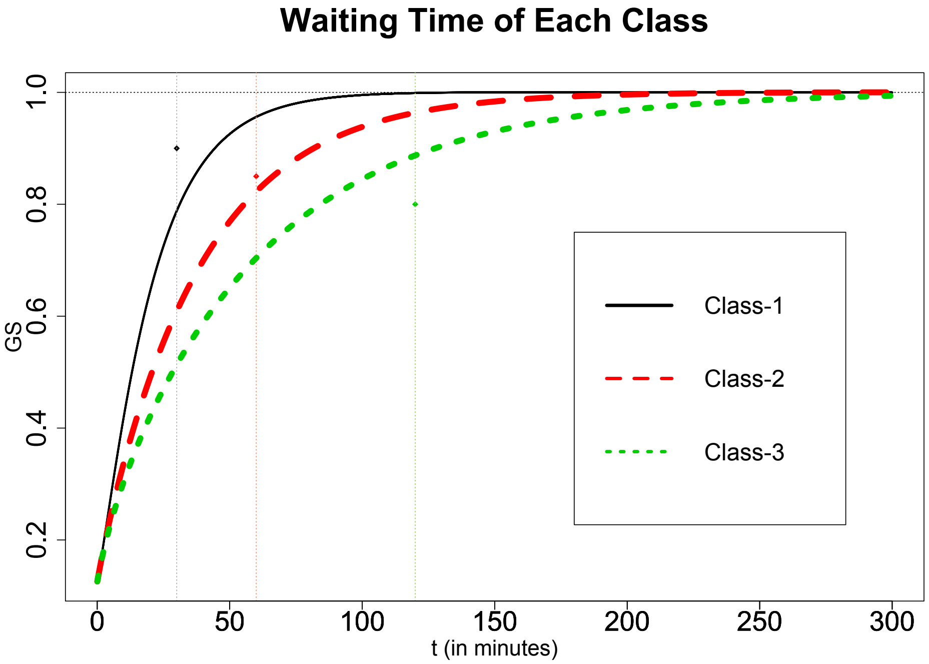


Figure 4.11: Waiting time distributions for a three-class APQ with the accumulation functions proposed in Section 4.1.

The last example we present is a three-class single-server APQ with arrival rates $\lambda_1 = 1$, $\lambda_2 = 0.7$ and $\lambda_3 = 0.4$ for class-1 (CTAS-3), class-2 (CTAS-4) and class-3 (CTAS-5) respectively, while service times for all classes are exponentially distributed at a common rate $\mu_1 = \mu_2 = \mu_3 = 2.4$. The set of accumulation functions are the one proposed in the introduction, and the accumulation rate are $c_1 = 1$, $c_2 = 0.5$ and $c_3 = 0.3$ as in Figure 4.1. It shows that under $\rho = 87.5\%$ only the patients from class-3 can meet their KPI target.

References

[1] Abate J., & Whitt W. (2006). A unified framework for numerically inverting Laplace transforms. INFORMS Journal on Computing. 18, 408-421.

[2] Canadian Association of Emergency Physicians. The Canadian Triage and Acuity S-cale (CTAS). From the website. $http\ ://www.calgaryhealthregion.ca/policy/docs/1451/Admission\_over\_capacity\_AppendixA.eps$.

[3] Kingman, J. F. C. (1993). Poisson Processes. Oxford University Press, Oxford.

[4] Kleinrock, L. (1964). A delay dependent queue discipline. Naval Research Logistics Quarterly. 11, 329–341.

[5] Kleinrock, L., & Finkelstein, R. (1967). Time dependent priority queues. Operations Research. 15, 104–116.

[6] Sharif, A. B., Stanford, D. A., Taylor, P. & Ziedins, I. (2014). A multi-class multi-server accumulating priority queue with application to health care. Operations Research for Health Care 3(2), 73–79.

[7] Stanford, D. A., Taylor, P. & Ziedins, I. (2014). Waiting time distributions in the accumulating priority queue. Queueing Systems 77(3), 297–330.

Chapter 5

Optimization of queues operating under waiting time limits

5.1 Introduction

In many situations, a service system attends to a number of distinct customer classes with differing urgencies for commencement of service. One ready example arises in the field of emergency medicine, where the customer priority classification is determined by the urgency for treatment; in medical literature, the term "patient acuity" is often used to describe such distinctions. The Canadian Triage and Acuity Scale (CTAS) [7] (see Table 5.1), as well as the Australasian Triage Scale (ATS) [3], identifies five distinct patient classes for patients in Emergency Departments, and sets a standard for commencement of treatment for each group. These standards specify a time limit and a corresponding compliance probability p for each class, such that the chance is at least p that a patient from the given class will have commenced service by the time limit (i.e., commence treatment).

CTAS is an example of a set of Key Performance Indicators (KPIs) comprising time limits l_k and corresponding compliance probabilities p_k; $k = 1, 2, \ldots, 5$. KPIs are widely used in health care, both for "visible" queues such as those in Emergency departments, as well as "invisible" queues or waiting lists, such as in Arnett *et al.* [2], which pertains to hip and knee

Table 5.1: CTAS Key Performance Indicators (KPI)

Category	Classification	Time Limit	Compliance Level
1	Resuscitation	Immediate	98%
2	Emergency	15 minute	95%
3	Urgent	30 minute	90%
4	Less urgent	60 minute	85%
5	Not urgent	120 minute	80%

replacement surgery. Britain's National Health Service (NHS) uses KPIs for a diverse range of health services, including, for example, mental health (Dodd [9]), Accident & Emergency department (NHS-Stockport [16]), Cancer time to treatment (NHS-Leeds [15]), and Diagnostic tests (NHS-Stockport [16]), to name a few. Similar trends can be found in most Organization for Economic Cooperation and Development countries.

At face value, there may appear to be a close link between systems operating to KPI time limits and service systems offering specified lead times for the delivery of a particular service (see, for instance, Keskinocak *et al.* [11], Çelik and Maglaras [5], Akan *et al.* [1]). However, lead time problems are typically characterized by a revenue stream, and are typically concerned with the "right" lead time to offer for a specified request as a function of the orders presently in the service system, in order to maximize profit or minimize a penalty function. In contrast to this, the time limits in KPI problems are fixed, and in the health care field where they predominate, they have usually been set by medical professionals in response to the perceived clinical need of the various patient classes. Furthermore, they are typically set prior to any consideration of the traffic characteristics of the patient classes (frequency of demand, treatment time distributions, etc.). It then falls to the health care professionals responsible for the operation of a particular facility to determine a patient selection rule (in queuing terms, a queuing discipline) so that the KPIs are met.

What KPI systems typically lack is an objective to be optimized. At face value, they comprise a set of delay constraints by which the particular queue or wait list needs to operate by.

Any system complying with them is considered equally good, and no consequence is given to the small fraction of patients who fail to gain access in time. This inability of KPIs to reflect the increased (rather than diminished) urgency of patients whose wait exceeds the specified limit was one of the points commented upon by Dr. Chris Baggoley of Australia's Expert Panel Review of Elective Surgery and Emergency Access Targets in a 2012 keynote address (Davies and Little [8]). As we shall establish, the minimization of the extent to which patients' wait times exceed their time limit is an appropriate goal to help ensure KPI compliance.

When waits beyond the time limits are considered to be equally bad for all patient classes, what is needed is a queuing discipline that considers patients from the various classes to be equally urgent as they approach their respective time limits. However, the best known queuing disciplines fail to do this: both the first-come, first-served (FCFS) discipline and what we refer to as "classical" priority disciplines operate irrespective of how long customers have waited. A queuing discipline that tracks patient waiting times is called for, which assigns priority credit as a function of how long patients have waited. Such "accumulating priority" systems, originally proposed by Kleinrock [12] in 1964, and recently extended in Stanford *et al.* [18] and Sharif *et al.* [17], are ideally suited to the task. We will be particularly interested in the performance of a "Rule of Thumb" that we propose, in which patients in the various classes accumulate priority credit at a rate that is inversely proportional to their time limits. In so doing, the Rule of Thumb ensures that all patients have the same total amount of accumulated priority credits when they reach their respective time limits, and are thereby considered to be equally urgent, as desired. In systems in which a weighted average of the expected excesses in each class is the appropriate criterion, an extension to our Rule of Thumb will be proposed.

In a nutshell, a queue operating under the APQ discipline treats customers whose various service requirements need to be delivered on different time scales. The advantage of an APQ approach for systems operating under KPIs is that it enables each class of patients (customers) to progress fairly towards timely access to a server by its own waiting time limit. Customers can still be overtaken by others of greater urgency or acuity, but they will not be overtaken indefinitely, due to the growing priority accumulated as each customer waits.

This paper presents an optimization model whose objective is to minimize the weighted average of total expected excess delay beyond the various limits for health systems operating under KPIs; we call this the "weighted average of the total expected excess" (WAE). When the weights are equal, the goal reduces to the minimization of the total expected excess waiting time (TEE) over all classes of customers. We seek solutions to the model using the Accumulating Priority Queue (APQ) discipline which specify the optimal priority accumulation rates. We note that the ability to specify these rates for each class provides an extra margin of flexibility over static queuing disciplines, so as to identify the best accumulation rates for the observed traffic patterns for the greatest chance that all classes of customers adhere simultaneously to the respective KPIs.

After some preliminary work which reveals the difficulty in minimizing the WAE, we present a related objective function which we define as the "integrated weighted expected excess" (IWAE). Focussing on the two-class case, we then establish that one can readily optimize the IWAE objective in terms of the optimal ratio of the two priority accumulation rates. Through further analysis and several numerical examples, we show that the optimal IWAE ratio of accumulation rates is near-optimal for many instances of the WAE problem, and that when desired, the optimal ratio can be used as a starting point in an iterative scheme for a WAE problem with particular time limits.

Of course, there is no obligation upon a health facility governed by a set of KPIs to implement the same strategy for all its patient classes. In Emergency departments, one would never make a Resuscitation or Emergency patient wait for someone of lower acuity. Nonetheless, this can still be achieved in an APQ setting merely by allowing for relatively huge accumulation rates for the truly urgent patient classes.

The rest of the paper is organized as follows. In section 5.2, we present the optimization problems for both the general case, and its restriction to the Accumulating Priority queuing discipline. We address the matter of convexity of the corresponding two-class APQ optimization problem in section 5.3. Section 5.4 introduces the IWAE objective function for which the convexity and the optimal solution can be established.

Section 5.5 relates the Laplace-Stieltjes transforms of the waiting time distributions and expected excesses. We show that when one resorts to a numerical inversion of the pertinent waiting time transform to compute the probabilities, the corresponding numerical inversion of the weighted expected excess waiting time is obtained with minimal additional effort, where the Gaver-Stehfest algorithm (Gaver [10], Stehfest [19]) is chosen to obtain the numerical results.

In section 5.6, we present a series of numerical examples which explore the optimal behaviour of these optimization problems. The nature and impact of our various results and insights are summarized in the Conclusions section.

5.2 Formulation of the optimization problem

Consider a queue featuring either a single server or many servers, which attend(s) to K independent classes of customers with distinct KPIs of the form discussed in the introduction. Arrivals for the kth class of customers are from a Poisson process at rate $\lambda_k; k = 1, 2, \ldots, K$. All service time distributions are known, and they may differ from class to class.

We presume that the queue is operating under a particular queuing discipline $\mathbf{a} \in \mathcal{A}$ (that is, a rule for selecting the next customer to enter service), where $\mathcal{A}$ denotes the set of permissible work-conserving disciplines. The set $\mathcal{A}$ includes among others first-come, first-served (FCFS), last-come, first-served (LCFS), Random Order of Service (ROS) and both Non-preemptive (N-P) and Preemptive Resume (PR) disciplines, as well as the Accumulating Priority Queue (APQ) which we will specify shortly. We presume that the queue has been operating sufficiently long to have reached stationarity.

Letting $\mathcal{W}_k$ denote the stationary class-k waiting time random variable, define the following for $k = 1, 2, \ldots, K$:

- $W_k(x) = P(\mathcal{W}_k \leq x)$ is the cumulative waiting time distribution function,
- $S_k(x) = P(\mathcal{W}_k > x) = 1 - W_k(x)$ is the survival function of the waiting time, and
- $w_k(x) = dW_k(x)/dx$ is the probability density function of the stationary class-k waiting

time distribution.

We will denote the respective Laplace-Stieltjes transforms (LSTs) of these quantities, and all others to be defined below, by a $\widetilde{}$ throughout the paper.

The expected amount of excess waiting time $H_k(l_k)$ for a typical class-k customer beyond a specified delay l_k under an employed queuing discipline in the set $\mathcal{A}$, henceforth abbreviated the "expected excess", can be determined from

$$H_k(l_k) = \int_{l_k}^{\infty} (x - l_k) w_k(x) dx = \int_{l_k}^{\infty} S_k(x) dx. \tag{5.1}$$

The quantity $\lambda_k H_k(l_k)$ can be interpreted as the expected amount of excess waiting for class-k customers per unit of time.

For such a queue as described above, the general form of the optimization problem can be stated as follows. We seek to minimize a weighted average of the total expected excess waiting per unit time (henceforth denoted WAE) over all permissible customer selection strategies,

$$\begin{aligned} \min_{\mathbf{a} \in \mathcal{A}} \quad & Z = \sum_{k=1}^{K} \alpha_k \lambda_k H_k(l_k) \qquad (5.2) \\ \text{subject to} \quad & W_k(l_k) \geq p_k; \quad k = 1, 2, \ldots, K, \end{aligned}$$

where $\alpha_k, k = 1, 2, \ldots, K$ denote the given respective weights for the expected excess waiting for class-k customers. A priori, we observe that the stated problem might be infeasible. If feasible, we observe that the optimal solution may not be unique.

5.2.1 Formulating the WAE optimization problem for APQ systems

In the introduction, we observed that the APQ discipline is better suited than other well-known disciplines such as FCFS and Classical Priority since customer selection depends upon priority credits earned while waiting. Stanford *et al.* [18] has determined the stationary waiting time distributions for each class of customer under the APQ discipline, so that one can determine the quantiles needed to assess compliance in a KPI system.

Waiting customers of the kth priority class accrue priority credit at a linear accumulation rate $b_k \geq 0$, where $b_k \geq b_j$ if class $k < j$, such that, the customers from a higher priority class earn priority faster than ones from a lower class. Setting $b_k = B \geq 0 \ \forall k$, a FCFS queue that aggregates all customer classes is obtained. Setting $b_K = 1$ while $b_k = b_{k+1} * M;\ k = 1, 2, \ldots, K-1$, a classical non-preemptive priority model is obtained in the limit as $M \rightarrow \infty$. Thus, not only is the APQ better suited than FCFS and Classical Priority to the task at hand, but these disciplines can each be recovered from the APQ discipline, so that it can be viewed as a unifying queuing discipline for all three.

Let $\mathcal{A}_{APQ}$ denote the set of APQ disciplines that satisfy the restrictions above. The resulting WAE optimization problem then reduces to the selection of the best accumulation rates $b_k;\ k = 1, 2, \ldots, K$ in order to minimize the weighted average excess:

$$\begin{aligned} \min_{\mathbf{a}\in\mathcal{A}_{APQ}:\ \{b_1,b_2,\ldots,b_K\}} \quad & Z = \sum_{k=1}^{K} \alpha_k \lambda_k H_k(l_k) && (5.3)\\ \text{subject to} \quad & W_k(l_k) \geq p_k;\ \ k = 1, 2, \ldots, K, \\ & b_k \geq b_{k+1}\ ;\ \ k = 1, 2, \ldots, K-1, \\ & b_K \geq 0, \end{aligned}$$

where $\alpha_k, k = 1, 2, \ldots, K$ again denote the respective weights for the expected excess waiting for customers from the various classes.

5.3 Convexity in the two-class APQ optimization problem

Clearly, the waiting time distributions in an APQ system depend upon the ratios of the accumulation rates; a system with K classes would have $K-1$ such ratios which can act as decision variables in order to craft waiting time distributions that meet specified waiting time goals. In particular, a two-class problem has a single decision variable, which we denote by $b = b_2/b_1$, and we shall henceforth make the dependence of all waiting-time-related quantities upon b explicit.

It is immediately a matter of interest as to whether the WAE objective function Z is convex in b in a two-class APQ system. If it could be established that each of the expected excess functions $H_1(b, l_1)$ and $H_2(b, l_2)$ is convex in b, it would follow that Z in equation (5.3) would be as well, since it is a linear combination of the two, with positive coefficients.

It is possible to establish that $H_2(b, l_2)$ is convex in b. This is done by first establishing that the integrand on the right-hand-side of

$$H_2(b, l_2) = \int_{x=l_2}^{\infty} S_2(b, x)dx \tag{5.4}$$

is convex:

Theorem 5.3.1. *Given the foregoing definitions, the function $S_2(b, x)$ is a monotonically decreasing convex function in b for $0 \leq b \leq 1, \forall x \geq 0$.*

Proof: See Appendix B.1.

Corollary 5.3.2. *The excess function $H_2(b, l_2)$ above is monotonically decreasing and strictly convex in b for $0 \leq b \leq 1$ for every fixed $l_2 \geq 0$.*

Proof. This follows immediately from Theorem 5.3.1 and equation (5.4). □

In fact, the foregoing corollary is equally true for the lowest-priority class in an K-class APQ system:

Corollary 5.3.3. *The excess function $H_K(b, l_K)$, where $b = b_K/b_{K-1}$, for the lowest priority class in a stable K-class APQ is strictly convex in b for $0 \leq b \leq 1$, for every fixed $l_K \geq 0$.*

Proof: See Appendix B.1.

However, as some of the numerical examples will reveal in the two-class case which we present later, the expected class-1 excess $H_1(b, l_1)$ can have a negative curvature over some or all of its range as a function of b. As a consequence we cannot infer convexity of the resulting weighted average even in this simplest two-class case, based upon properties of mixtures of convex functions.

In contrast to the difficulties we have encountered by dealing directly with the weighted expected excess objective WAE, there is a related objective function we can work with, for which both its convexity, and its optimal solution as a function of b can be readily established. This is explored in the next section.

5.4 Optimizing IWAE

In Section 5.3, we discussed the difficulties relating to the optimization problem of minimizing the WAE. In order to gain further insight into the expected excess wait and the heaviness of the tail of the waiting time distribution, we introduce the integrated weighted average excess waiting (IWAE), defined as

$$Z^* = \int_0^\infty Z dl_1 = \sum_{k=1}^{K} \alpha_k \lambda_k G_k \tag{5.5}$$

where G_k is the expected excess waiting for class-k integrated over the range of the class-1's delay limit, such that $G_k = \int_0^\infty H_k(l_k) dl_1$.

Define $f_{1,k} = l_1/l_k$ for $k = 1, 2, \ldots, K$ as the ratio of the delay limits between class-1 and class-k. Note that $f_{1,1} = 1$ and $0 < f_{1,k+1} \leq f_{1,k} \leq 1$ for $k = 1, 2, \ldots, K-1$. Let $m_k^{(2)}$ be the second moment of the waiting time distribution for class-k customers. Since $H_k(t) = \int_t^\infty S_k(x)dx$ (see Klugman, Panjer and Willmot [13, page 40]), it follows that

$$\int_0^\infty H_k(t)dt = \int_0^\infty x S_k(x)dx = \frac{m_k^{(2)}}{2}. \tag{5.6}$$

Then, G_k can be written as

$$G_k = \int_0^\infty H_k(l_k)dl_1 = f_{1,k} \int_0^\infty H_k(l_1)dl_1 = \frac{f_{1,k} m_k^{(2)}}{2}. \tag{5.7}$$

The resulting IWAE optimization problem can be stated as

$$\begin{aligned} \min_{\mathbf{a} \in \mathcal{A}_{APQ}:\ \{b_1, b_2, \ldots, b_K\}} \quad & Z^* = \sum_{k=1}^{K} \alpha_k \lambda_k G_k = \frac{1}{2} \sum_{k=1}^{K} \alpha_k \lambda_k f_{1,k} m_k^{(2)} \\ \text{subject to} \quad & W_k(l_k) \geq p_k; \;\; k = 1, 2, \ldots, K, \\ & b_k \geq b_{k+1}\;; \;\; k = 1, 2, \ldots, K-1, \\ & b_K \geq 0. \end{aligned} \tag{5.8}$$

At this point, we will focus on the properties and optimal solution of the unconstrained objective function, i.e., without reference to the constraints, both for the IWAE and the WAE. With respect to the constrained versions of the problems, we can observe in general that at sufficiently light load, none of the constraints will be binding, and at sufficiently high load, all of the constraints will be violated. In between, as the load increases, progressively more and more constraints will be violated. In terms of KPI systems, usually the tightest probability constraints are associated with the highest acuity patients, so one might anticipate it will often be the first constraint to become binding. We will comment later upon how one can proceed from the optimal unconstrained solution to identifying optimal solutions in the two-class case.

By considering the unconstrained IWAE objective function, we find that the difficulties encountered when optimizing the WAE can be resolved. Moreover, we obtain an explicit equation for the optimal accumulation rate that minimizes IWAE, which we propose as a proxy for the optimality of minimizing WAE. This is because the IWAE represents an aggregation of all the specific cases we can expect to encounter with the WAE.

For the remainder of this section, we restrict our attention to the optimization problem for the two-class $M/M_i/c$ APQ. Define π_{busy} to be the probability of the system being busy in an $M/M_i/c$ queue (Li and Stanford [14]), and $\rho = \lambda/\mu$ as the occupancy level, where $\lambda = \lambda_1 + \lambda_2$ and μ is the total service rate of c servers. For simplicity, we set $b_1 = 1$, $b_2 = b$, $f_{1,2} = f < 1$, $\hat{\lambda} = \lambda_1(1-b)$ and $\theta = \lambda_2/\lambda_1$. Furthermore, the second moment of the waiting time distribution for class-$k; k = 1, 2$ is a function of b in the two-class case, thus, here we denote it as $m_k^{(2)}(b)$.

We present the LST, $\widetilde{w}_k(s)$, of the waiting time distribution for class-k; $k = 1, 2$ customers under the accumulating priority discipline, in Appendix B.2 equations (B.4)–(B.8). After calculating $m_k^{(2)}(b) = d^2\widetilde{w}_k(s)/ds^2|_{s=0}; k = 1, 2$, we obtain the following:

Define $c_0 = \pi_{\text{busy}}/(1-\rho)^2 > 0$, which does not depend on b. The second moment of the waiting time for class-1 customers is given by

$$m_1^{(2)}(b) = 2c_0 \times m_1^{(2)*}(b) \tag{5.9}$$

where

$$m_1^{(2)*}(b) = \frac{\rho^2(1-b)(\mu - \lambda_1(1-b)^2) - \rho(1-b)((3b-2)\lambda_1 + (b+2)\mu) - \lambda_1(1-b)^2 + \mu}{(\mu - \hat{\lambda})^3}, \tag{5.10}$$

while the corresponding expression for class-2 is

$$m_2^{(2)}(b) = 2c_0 \times m_2^{(2)*}(b) \tag{5.11}$$

where

$$m_2^{(2)*}(b) = \frac{\mu - \rho\hat{\lambda}}{(\mu - \hat{\lambda})^3}. \tag{5.12}$$

Lemma 5.4.1. *In the two-class $M/M_i/c$ APQ, the IWAE is*

$$Z^*(b) = c_0 \cdot \left[\alpha_1\lambda_1 m_1^{(2)*}(b) + f\alpha_2\lambda_2 m_2^{(2)*}(b)\right]. \tag{5.13}$$

Proof. In the two-class $M/M_i/c$ APQ, by equation (5.8), the IWAE can be written as a function of b,

$$\begin{aligned} Z^*(b) &= \alpha_1\lambda_1 G_1 + \alpha_2\lambda_2 G_2 \\ &= \frac{1}{2}\left[\alpha_1\lambda_1 m_1^{(2)}(b) + f\alpha_2\lambda_2 m_2^{(2)}(b)\right] \\ &= c_0 \cdot \left[\alpha_1\lambda_1 m_1^{(2)*}(b) + f\alpha_2\lambda_2 m_2^{(2)*}(b)\right]. \end{aligned}$$

□

Theorem 5.4.2. *Denote b^* as the optimal b value that minimizes IWAE in equation (5.13), and let r_1 be one of roots of a quadratic equation arising from $dZ^*(b)/db = 0$, specified in equation (5.18) below. For $\alpha_1 \geq \alpha_2$, then $r_1 \leq f$, and*

$$b^* = \begin{cases} 0, & r_1 \leq 0; \\ r_1, & r_1 > 0. \end{cases} \tag{5.14}$$

Proof. To minimize IWAE in equation (5.13), we first take the first derivative of $Z^*(b)$ respect to b.

$$\frac{\mathrm{d}}{\mathrm{d}b} Z^*(b) = \frac{c_0\lambda_1\lambda_2}{\mu(\mu - \lambda_1(1-b))^4}\left[c_1 b^2 + c_2 b + c_3\right], \tag{5.15}$$

where

$$c_1 = \alpha_1\lambda_1(3\lambda - \mu),$$

$$
\begin{aligned}
c_2 &= 2\Big(\alpha_1\mu(\mu+\lambda_1) - \lambda_1\lambda(2\alpha_1+\alpha_2 f)\Big), \text{ and} \qquad (5.16)\\
c_3 &= \mu\Big((\mu-\lambda)(\alpha_1-\alpha_2 f) - (\lambda_1\alpha_1 + 2\mu\alpha_2 f)\Big) + \lambda_1\lambda(\alpha_1+2\alpha_2 f).
\end{aligned}
$$

Since $dZ^*(b)/db$ is a quadratic function of b, for $0 < \rho < 1, 0 < f < 1, \theta > 0$, and $\alpha_1 \geq \alpha_2 > \alpha_2 f$, it follows that

$$
\begin{aligned}
\triangle &\equiv c_2^2 - 4c_1c_3\\
&= \frac{4\mu^4}{(1+\theta)^4}\Big\{(\alpha_1-\alpha_2 f)^2\rho^4 + \alpha_1^2(\theta+1)^2 + 3\rho(\theta+1)\alpha_1(\alpha_1-\alpha_2 f)\Big[(\rho-\frac{4}{3})^2 - \frac{7}{9}\Big]\Big\}\\
&\geq \frac{4\mu^4}{(1+\theta)^4}(\alpha_1\theta+\alpha_2 f)^2\\
&> 0. \qquad (5.17)
\end{aligned}
$$

The roots of the quadratic equation $dZ^*(b)/db = 0$ are

$$
r_1 = \frac{-c_2 + \sqrt{\triangle}}{2c_1} \qquad \text{and} \qquad r_2 = \frac{-c_2 - \sqrt{\triangle}}{2c_1},
$$

where $r_2 \notin [0,1]$ (i.e., it is out of the suitable range at all times) is not the optimal solution for IWAE. After some rearrangement, r_1 can be expressed as

$$
r_1 = \frac{2\mu^2\Big[(2\alpha_1+\alpha_2 f)\rho^2 - \alpha_1\rho - (\theta+1)\alpha_1\Big] + (\theta+1)\sqrt{\triangle}}{2\alpha_1\rho\mu^2(3\rho-1)}. \qquad (5.18)
$$

We find that r_1 increases with ρ; $\rho \in (0,1)$, so that

$$
r_1 \leq \lim_{\rho\to 1} r_1 = \frac{f\alpha_2}{\alpha_1} \leq f, \quad \text{for } \alpha_2 \leq \alpha_1. \qquad (5.19)
$$

Thus, $b^* = r_1$ when $r_1 \in [0, f]$ and $b^* = 0$ when $r_1 < 0$. □

When $b^* > 0$, the system under the accumulating priority discipline outperforms that under the classical priority queuing discipline. We are interested in determining the specific occupancy where the classical priority gives away to the APQ as the optimal queuing discipline. We denote this occupancy by ρ_{sp} (i.e., *sp* for "switching point").

Corollary 5.4.3. *The switching point ρ_{sp} where the accumulating priority discipline starts to outperform the classical priority discipline is given by*

$$
\rho_{sp} = \frac{\alpha_1(\theta+2) - \alpha_2 f(\theta+1) - \sqrt{(\theta+25)(\theta+1)\alpha_2^2 f^2 - 2\theta(\theta+1)\alpha_1\alpha_2 f + \alpha_1^2\theta^2}}{2\alpha_1 + 4\alpha_2 f}. \qquad (5.20)
$$

Proof. The solution is by setting $r_1 = 0$ in equation (5.18) and solving for the corresponding value of ρ. □

5.5 Optimizing WAE

Having addressed the IWAE optimization problem, we return to the original WAE optimization problem which involves specific values of $l_k; k = 1, 2, \ldots, K$, since KPI systems have stipulated delay limits, ostensibly for clinical reasons. By optimizing the WAE, we are able to identify the optimal accumulation rate for each class, which minimizes the consequence of the customers missing their specific delay limit.

The expected excess wait function for each class, under a fixed delay limit, is unavailable in a tractable closed form. Similarly, even the waiting time distributions are only known in terms of their LSTs. Since a closed-form analytical expression for the waiting time distribution is unavailable, the compliance probabilities will have to be recovered from the LSTs via a numerical inversion technique.

In contrast to this, we are able to establish a useful relationship between the transform of the expected excess waiting time in terms of the waiting time LST, which is readily available. Furthermore, this transform relationship will be used to show that when we use our preferred numerical inversion method (Gaver-Stefest inversion) to obtain the compliance probabilities, the numerically inverted expected excesses are found via a trivial extension of that calculation.

By definition, the LSTs of $w_k(t)$, $W_k(t)$, $S_k(t)$, and $H_k(t)$ are given by

$$\begin{aligned}
\widetilde{w}_k(s) &= \int_{t=0}^{\infty} e^{-st} w_k(t)dt = \int_{t=0}^{\infty} e^{-st} dW_k(t), \\
\widetilde{W}_k(s) &= \int_{t=0}^{\infty} e^{-st} W_k(t)dt, \\
\widetilde{S}_k(s) &= \int_{t=0}^{\infty} e^{-st} S_k(t)dt, \\
\widetilde{H}_k(s) &= \int_{t=0}^{\infty} e^{-st} H_k(t)dt.
\end{aligned}$$

Lemma 5.5.1. *The transforms defined above for the waiting time distribution and the expected excess per customer are related by*

$$\widetilde{H}_k(s) = \frac{m_k}{s} - \frac{1}{s^2} + \frac{\widetilde{w}_k(s)}{s^2}. \tag{5.21}$$

Proof: Standard properties of LST imply that

$$\widetilde{W}_k(s) = \frac{\widetilde{w}_k(s)}{s}, \tag{5.22}$$

and

$$\widetilde{S}_k(s) = \frac{1}{s} - \widetilde{W}_k(s). \tag{5.23}$$

From equation (5.1), it follows that $H_k'(t) = -S_k(t)$, so that when integrating by parts,

$$\begin{aligned}\widetilde{H}_k(s) &= \int_{t=0}^{\infty} H_k(t)e^{-st}dt \\ &= \left[\frac{H_k(t)e^{-st}}{-s}\right]_0^{\infty} + \int_{t=0}^{\infty} \frac{e^{-st}}{s} H_k'(t)dt \\ &= \frac{H_k(0)}{s} - \frac{1}{s}\widetilde{S}_k(s).\end{aligned}$$

Letting m_k be the mean class-k waiting time, (5.21) is obtained, as $H_k(0) = \int_{x=0}^{\infty} S_k(x)dx = m_k$.
□

Equation (5.21) immediately implies that $\widetilde{H}_k(s)$ is readily obtained for any value of s, once the corresponding evaluation has been carried out for the $\widetilde{w}_k(s)$.

We are now ready to apply numerical inversion to evaluate the LST of the compliance probabilities and the expected excess functions. The numerical inversion of LSTs has been an alternative to analytical inversion since the Fast Fourier Transform (FFT) technique gained popularity (Brigham and Morrow [4]). Whereas the FFT can require hundreds of evaluations for a single time point of interest, there are alternatives that require only a handful of evaluations. The one we choose to employ is Gaver-Stehfest (GS) numerical inversion; it is so named because the pioneering probabilistic work of Gaver [10] was later refined algorithmically by Stehfest [19].

Lemma 5.5.2. *The GS numerical evaluation of the class-k expected excess waiting time per customer for $k = 1, 2, \dots, K$ is given by*

$$H_{g,k}(t) \;=\; \sum_{j=1}^{N} \frac{V_j}{j} \left[\frac{\widetilde{w}_k(\frac{ln2}{t} \times j)}{(\frac{ln2}{t} \times j)} - \frac{1}{(\frac{ln2}{t} \times j)} + m_k \right], \tag{5.24}$$

where the V_j, $j = 1, 2, \dots, N$ are combinatorial terms related to order statistics as derived by Gaver [10].

The proof of Lemma 5.5.2 is give in Appendix B.3.

Remark: The GS coefficients V_j have the useful properties that $\sum_{j=1}^{N} V_j = 0$ and $\sum_{j=1}^{N} V_j/j = 1$.

Theorem 5.5.3. *The WAE in a multi-class $M/M_i/c$ APQ as evaluated by GS approximation is given by*

$$Z_g = \sum_{k=1}^{K} \sum_{j=1}^{N} \frac{\alpha_k \pi_{\text{busy}} \lambda_k l_k V_j}{j^2 ln2} \Big(\widetilde{w}_k^{+}(\frac{ln2}{l_k} \times j) - 1 \Big) + \sum_{k=1}^{K} \alpha_k \lambda_k m_k. \tag{5.25}$$

When $\alpha_k = 1, k = 1, 2, \dots, K$, the corresponding total expected excess (TEE) is

$$Z_g = \sum_{k=1}^{K} \sum_{j=1}^{N} \frac{\pi_{\text{busy}} \lambda_k l_k V_j}{j^2 ln2} \Big(\widetilde{w}_k^{+}(\frac{ln2}{l_k} \times j) - 1 \Big) + M, \tag{5.26}$$

where M is the constant in the conservation law for an $M/M_i/c$ APQ (Li and Stanford [14]), such that $M = \sum_{k=1}^{K} \lambda_k m_k$. Moreover, M does not depend on the accumulation rates $b_k; k = 1, 2, \dots, K$.

The proof of Theorem 5.5.3 is derived in Appendix B.3.

Corollary 5.5.4. *In any multi-class APQ with c heterogeneous servers, each working at an exponential service rate μ_i for $i = 1, 2, \dots, c$, the optimality of the TEE in equation (5.26) is the same as the one in a single-server APQ with an exponentially-distributed service time at rate $\mu = \sum_{i=1}^{c} \mu_i$.*

The equivalence claimed by Corollary 5.5.4 is established in Appendix B.3.

According to Theorem 5.5.3, the (equal) weighted average of the total expected excess in the two-class multi-server APQ can be written out by equations (B.4) – (B.9) in Appendix B.2. In the next section, we turn to the numerical investigation of various quantities arising from the minimizing of the TEE, WAE, and IWAE in the two-class multi-server APQ.

5.6 Numerical investigations

A series of calculations were performed for the two-class single-server APQ with a total service rate $\mu = 12/hr = 0.2/min$, to determine the WAE and IWAE for various configurations of the parameters b, ρ, f, θ, and $l_k; k = 1, 2$. While we present results for a wide range of parameter values, we will be primarily interested in realistic parameter values, e.g., $1/6 < f < 1$ and $\rho > 60\%$ in KPI systems which we might encounter in reality. In all cases, the GS numerical inversion algorithm with 8 points is used, which provides two significant digits of accuracy.

Our focus is initially on the optimal solutions to the unconstrained problems; at the end we will comment on the corresponding constrained problems.

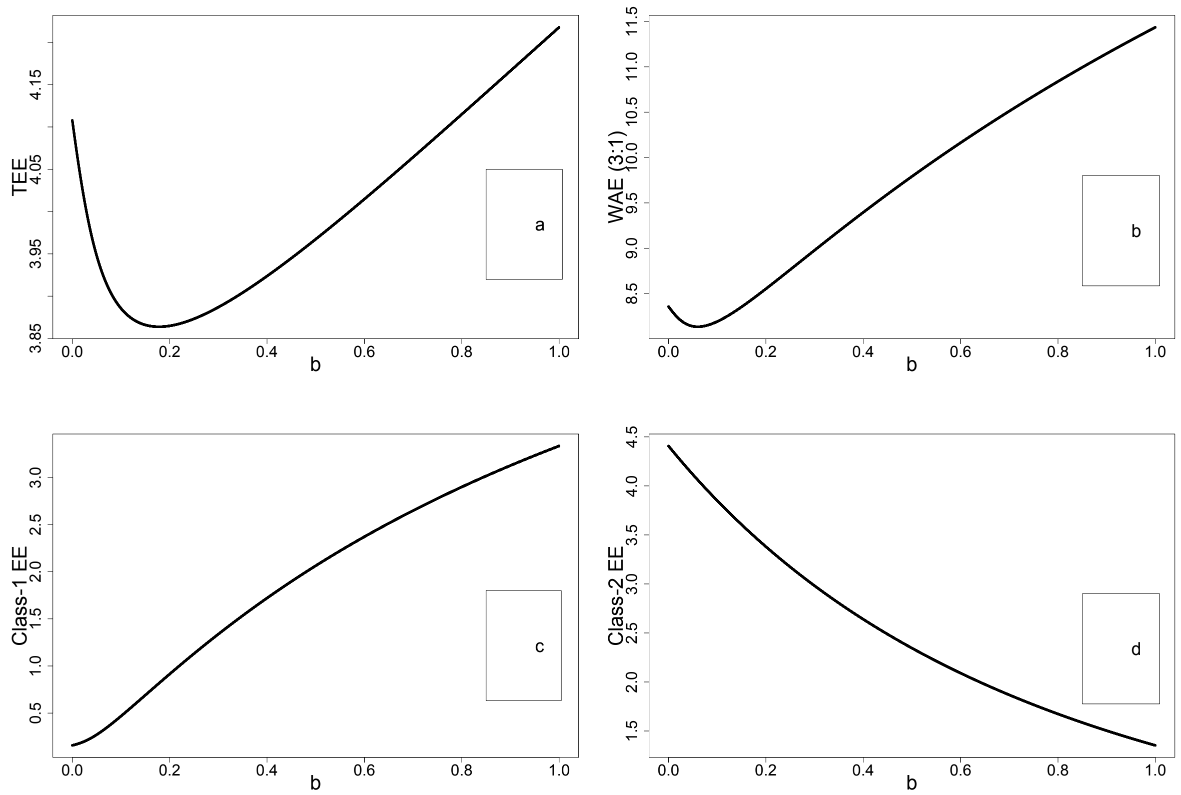


Figure 5.1: WAE ($\rho = 90\%$) with the parameters $\mu = 2$, $\lambda_1 = \lambda_2$, $l_1 = 15$ mins and $f = 1/4$.

In Figure 5.1 we present the total expected excess (TEE), the weighted average of the expected excess (WAE), and the expected excess waiting time per customer (EE) for class 1 and class 2, which were calculated as a function of the priority accumulation rate for class-2

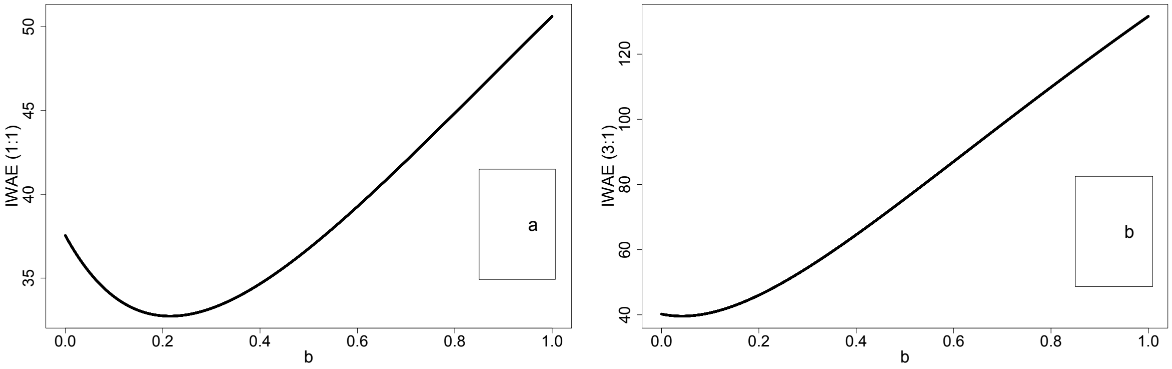


Figure 5.2: IWAE ($\rho = 90\%$) with the parameters $\mu = 2$, $\lambda_1 = \lambda_2$, and $f = 1/4$.

customers, $b_2 = b$ with fixed $b_1 = 1$. In Figure 5.2, we present the corresponding cases of the IWAE as Figure 5.1 images (a) and (b), to provide a contrast to the TEE and WAE. In both cases, $f = l_1/l_2 = 1/4$. These figures assume equal arrival rates and a 90% occupancy level. The specific delay limits for class-1 and class-2 are 15 mins and 60 mins respectively in Figure 5.1.

Figure 5.1 (a) and Figure 5.2 (a) compare the TEE and the IWAE when $\alpha_1 = \alpha_2$. We observe that both curves appear to be convex, and that the optimal b values for both curves are very close to $0.2 < 1/4$. The optimal b that minimizes TEE is slightly less than 0.2, whereas the optimal b that minimizes IWAE (1:1) is slightly greater than it. In the remainder of this section, we will use $(\alpha_1 : \alpha_2)$ to denote the proportion of the weights between class-1 and class-2, so that for instance, IWAE (1:1) means that the IWAE was computed with $\alpha_1 = \alpha_2 = 1$.

Figure 5.1 (b) and Figure 5.2 (b) compare the WAE (3:1) and the IWAE (3:1). Once again, only a tiny difference between the optimal b values can be observed. Moreover, both optimal b values are less than $f\alpha_2/\alpha_1 = 1/12$. The WAE curve in this unequal weights case is clearly not convex over the range $b \in [0, 1]$, whereas the IWAE is convex. To further investigate the non-convexity of WAE, we plot the expected excess waiting time per customer for each class, the results of which are presented in Figure 5.1 (c) and (d). We see that the expected excess curve for class-2 is convex, as we proved earlier, whereas the curve for class-1 is not convex for all b, which results in the non-convex behaviour of WAE (3:1) in Figure 5.1 (b).

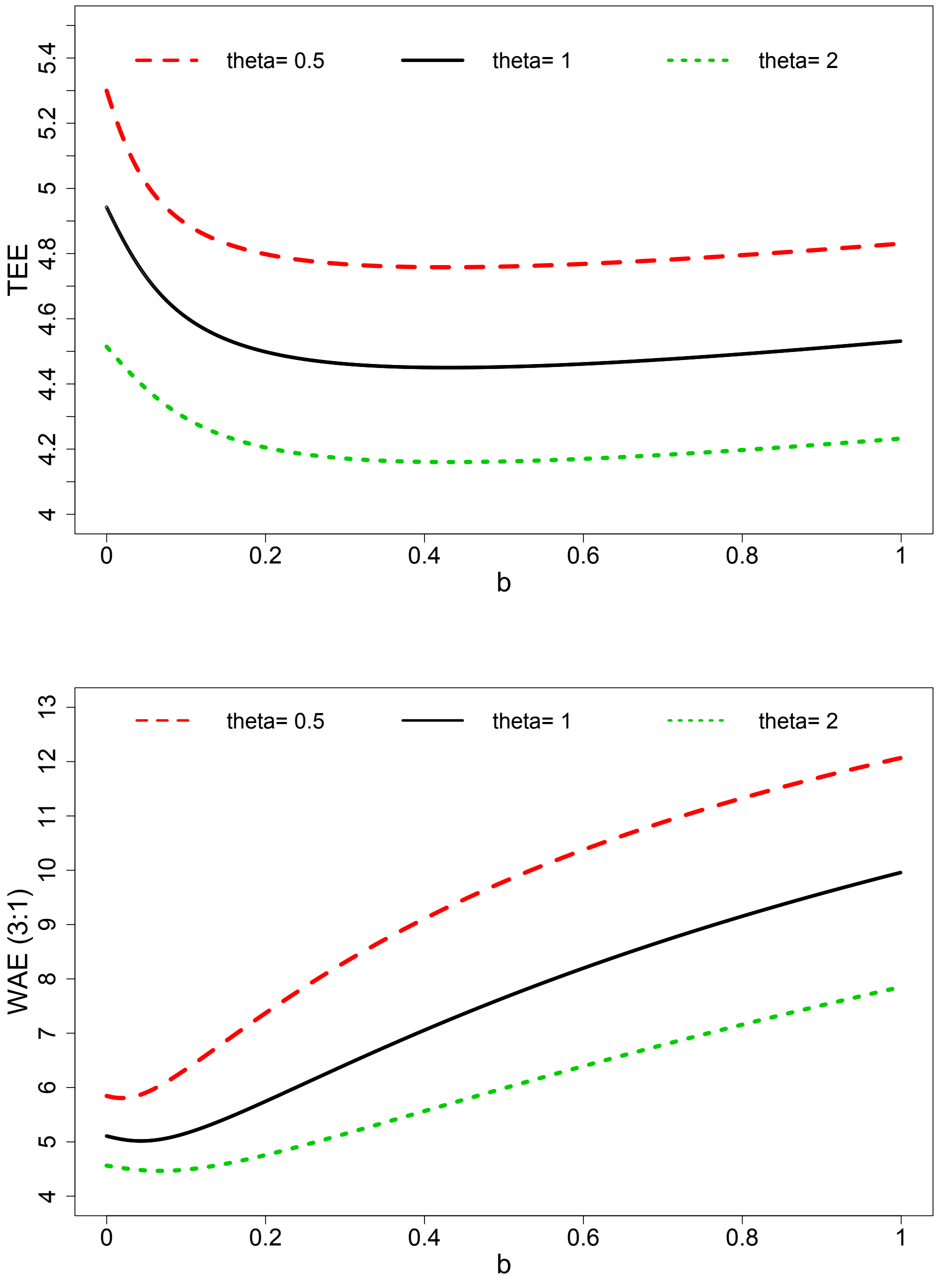


Figure 5.3: TEE and WAE (ρ = 90%) with parameters: $l_1 = 20$ mins, $f = 1/2$.

Figure 5.3 presents the values of TEE and WAE (3:1) respectively when the ratio of class-2 to class-1 arrivals $\theta = \lambda_2/\lambda_1$ assumes one of three values $\theta = 0.5$, 1 & 2. The occupancy is fixed at 90%, while $l_1 = 20$ mins and $l_2 = 40$ mins. For both the TEE and WAE in Figure 5.3, we notice that the curves for different values of θ display similar curvatures, with the optimal

b values for all three curves close for the WAE, and very close for the TEE. Furthermore, we find that the amounts of TEE and WAE are always greater for small θ than they are for large. This is because a small value of θ for a fixed ρ means there are more class-1 and fewer class-2 customers in the system, meaning that more patients are subject to the tighter time limit l_1. In addition, the APQ priority structure in such a scenario leads to longer delays for the lower-indexed class. The converse argument explains why the reverse is true for $\theta = 2$.

Figure 5.4 shows the optimal b values that minimize TEE against different occupancy levels with $l_1 = 20$ mins, $f = 1/2$ and $\theta = 0.1, 0.5, 1, 2, 10$. We observe that when $\rho > 60\%$, the differences among the optimal b values are small for all the ratios considered. Consequently, in a KPI system we can realistically expect to arise, the mix of traffic from the various classes has little impact upon the optimal value of b.

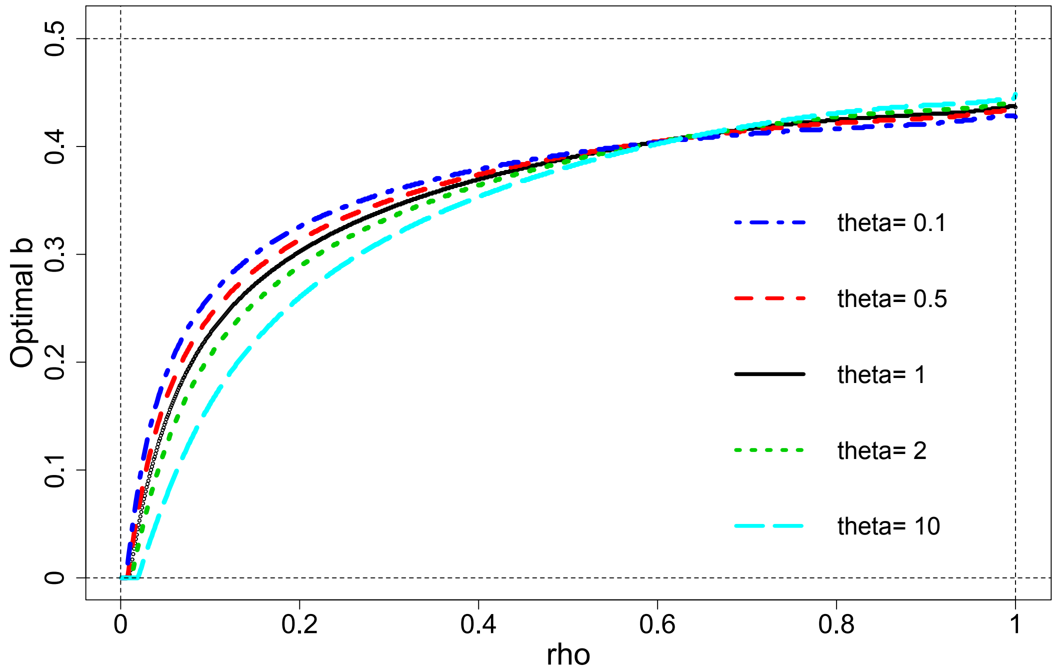


Figure 5.4: The optimal b to minimize TEE with parameters: $l_1 = 20$ mins and $f = 1/2$.

We performed a set of calculations to find the optimal b value to minimize TEE, denoted by $b^{\dagger}$, as a function of ρ. The results are presented in Figures 5.5, where we have considered

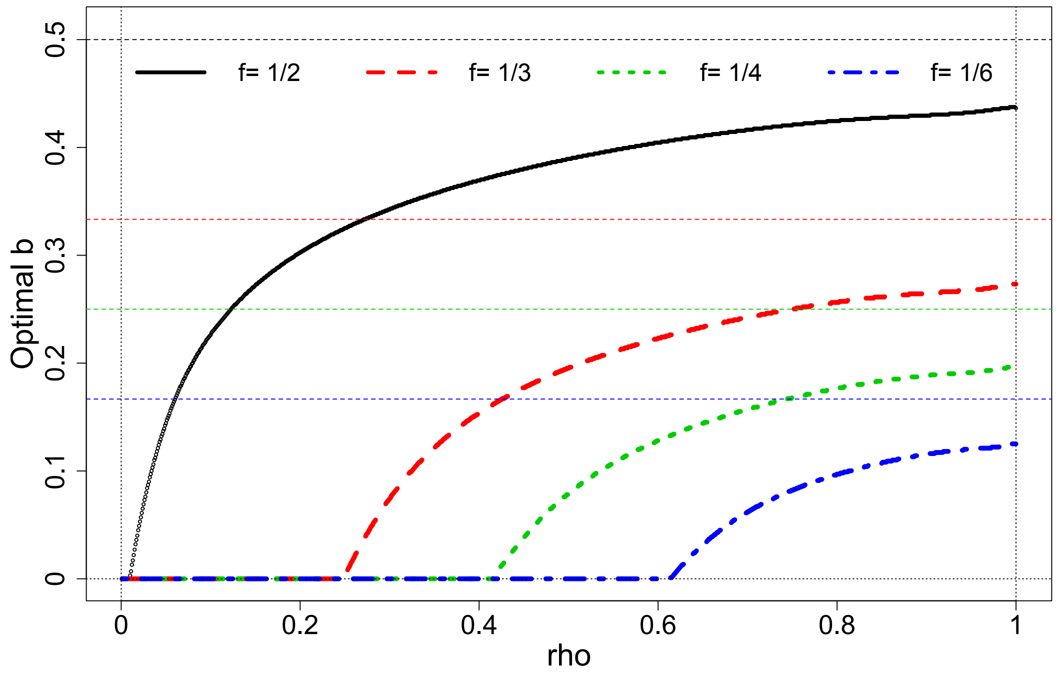


Figure 5.5: The optimal b to minimize TEE with parameters: $\lambda_1 = \lambda_2$ and $l_1 = 20$ mins.

equal customer arrival rates for both classes, the delay limit for class-1 $l_1 = 20$ mins, the values of $f = 1/2, 1/3, 1/4$ and $1/6$. We observe the following, based on Figure 5.5:

- At sufficiently light load ρ, $b^\dagger = 0$; that is, in light traffic, a classical priority queue is the optimal policy for minimizing total expected excess. This is because at light load, the chance (small though it is) of a high priority patient exceeding l_1 and incurring excess waiting time is much greater, relatively speaking, than the chance of a lower priority customer exceeding l_2. As f decreases, $l_2 = l_1/f$ increases, so that we observe a larger range of occupancies for which a classical priority situation is optimal. For any fixed f value, however, there is eventually a large enough ρ_{sp} where the optimal queuing discipline switches from classical priority to an APQ discipline with positive $b^\dagger$.

- For a fixed f value, as $\rho > \rho_{sp}$ increases, the rate of increase in $b^\dagger$ decreases, and the greatest optimal b value occurs when the the occupancy level approaches one; this value is always smaller than f. We denote the greatest optimal b value by $b^\dagger_{\max}$.

Table 5.2: The values of ρ_{sp} and $b^{\dagger}_{\max}$ for different values of f

	l_1 = 20 mins		l_1 = 30 mins	
f	ρ_{sp}	$b^{\dagger}_{\max}$	ρ_{sp}	$b^{\dagger}_{\max}$
1/2	0.010	0.437	0.099	0.452
1/3	0.249	0.273	0.317	0.287
1/4	0.416	0.197	0.472	0.209
1/6	0.615	0.125	0.656	0.135
1/12	0.824	0.059	0.843	0.065
1/100	0.983	0.007	0.984	0.007
1/600	0.998	0.001	0.998	0.001

Table 5.2 presents the numerical values of ρ_{sp} and $b^{\dagger}_{\max}$ for different values of f. The results in the table are consistent with the observations stated above. In addition to this, we observe that a relaxation in the value of l_1 leads to an increase in both ρ_{sp} and $b^{\dagger}_{\max}$. For instance, when $f = 1/3$, $\rho_{sp} = 0.249$ when $l_1 = 20$ mins, which is smaller than 0.317 when $l_1 = 30$ mins. Likewise, $b^{\dagger}_{\max} = 0.273$ in the former case is smaller than 0.287 in the latter; we note that both of them are smaller than 1/3. We also note that as f approaches zero, ρ_{sp} goes to one while $b^{\dagger}_{\max}$ goes to zero, since there is no excess class-2 waiting in the limit. The table indicates that in realistic situations where f is in the range $1/6 < f < 1$ and for occupancies $\rho > 60\%$ in a two-class system, an APQ with positive $0 < b < f$ will be preferable to both classical priority and FCFS queuing disciplines.

Similar results as what we have discussed above were observed in Figure 5.6 which considers the optimal b to minimize WAE (3:1) with the same set of parameters. The main difference is that for a fixed f value, as $\rho > \rho_{sp}$ increases, the optimal b increases from zero and reaches its maximum at a particular level of occupancy, then starts to decrease until it returns to zero when the occupancy level approaches one. The occupancy level corresponding to the greatest optimal b increases as f decreases. When ρ increases, there are more arrivals from both classes in the system. As we increase b, we incur greater delays for class-1, which may result in a big impact on WAE at high occupancy levels, especially since we have assigned more weight to class-1 whose delay limit is much tighter. Thus, the optimal b to minimize WAE decreases at high occupancy.

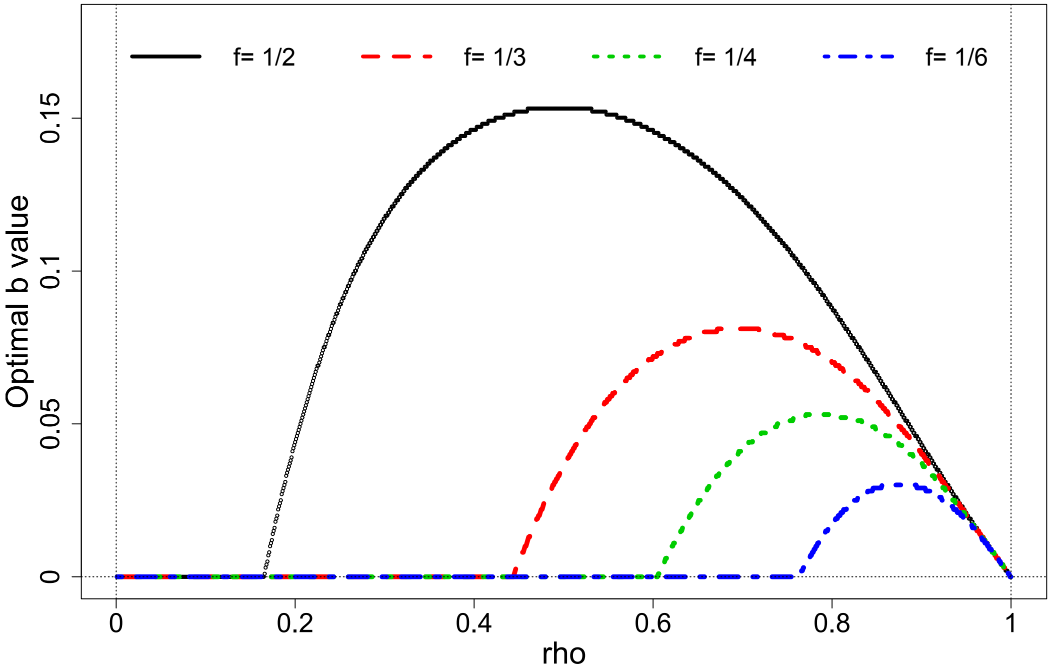


Figure 5.6: The optimal b to minimize WAE (3:1) with $\lambda_1 = \lambda_2$ and $l_1 = 20$ mins.

Figure 5.7 presents the optimal b values to minimize IWAE (1:1) against different levels of occupancy with equal arrivals from both classes. Compared to Figure 5.5, the optimal b for IWAE (1:1) also increases with ρ and reaches its maximum at f when ρ approaches one, whereas the optimal b for TEE is always smaller than f in Figure 5.5. For a fixed f, the ρ_{sp} for IWAE (1:1) is smaller than that for TEE, which again implies that the APQ structure is preferable to both the classical priority and FCFS disciplines.

Figure 5.8 is plotted to minimize IWAE (3:1) as a comparison both with Figures 5.6 (which has the same weights, but considers WAE) and 5.7 (which considers IWAE with equal weights). A similar trend as in Figure 5.7 is observed from Figure 5.8. Unlike Figure 5.6, the optimal b

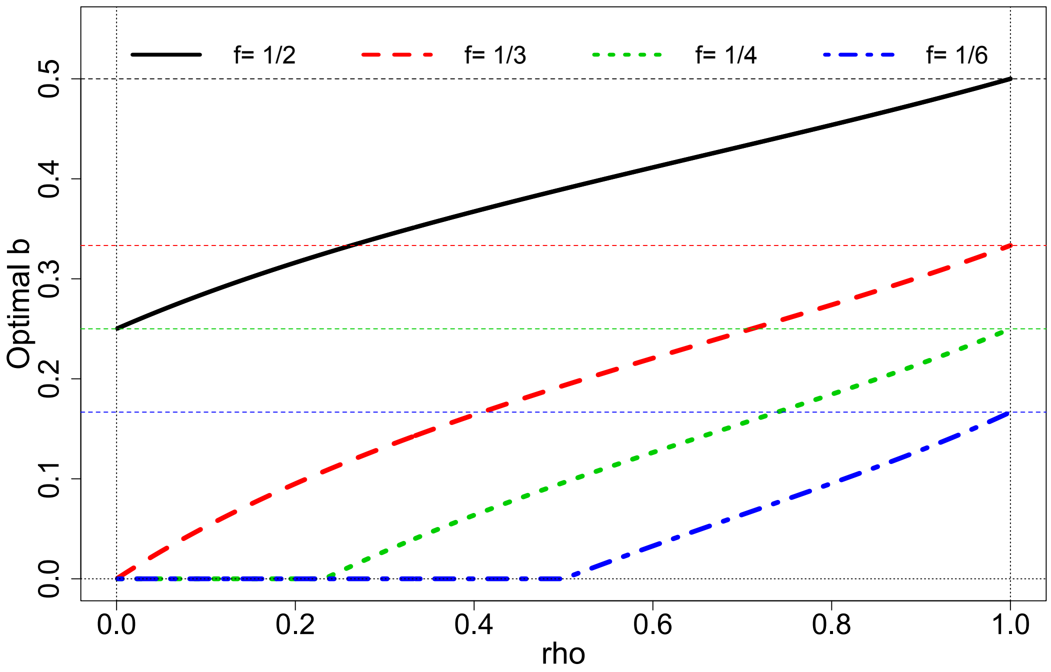


Figure 5.7: The optimal b to minimize IWAE (1:1) with equal arrivals.

for IWAE (3:1) keeps increasing with ρ, and the greatest value is $f\alpha_2/\alpha_1$ when ρ approaches one. Compared to Figure 5.6, for a fixed f, the ρ_{sp} for IWAE (3:1) is greater than that for WAE (3:1).

A comparison of Figures 5.5 and 5.7 reveals that the difference between the optimal b for TEE and that for IWAE (1:1) appears small for occupancy levels one is likely to encounter in practice. In order to investigate this more closely, we plot the difference between the optimal b for IWAE (1:1) and for TEE against different occupancy levels in Figure 5.9. Firstly, the maximum difference occurs at ρ_{sp} since the optimal b for TEE is zero when $\rho < \rho_{sp}$, whereas that for IWAE (1:1) is positive. Secondly, at high occupancy (i.e., $\rho > 80\%$), the absolute difference is less than 0.065 for all f considered in the figure. Therefore, it seems reasonable to use the optimal b for IWAE with equal weights to approximate the one for TEE where a specific delay limit is imposed on each class. In specific cases where a user is seeking greater precision in the optimal ratio, an iterative numerical method, such as Newton's method, with

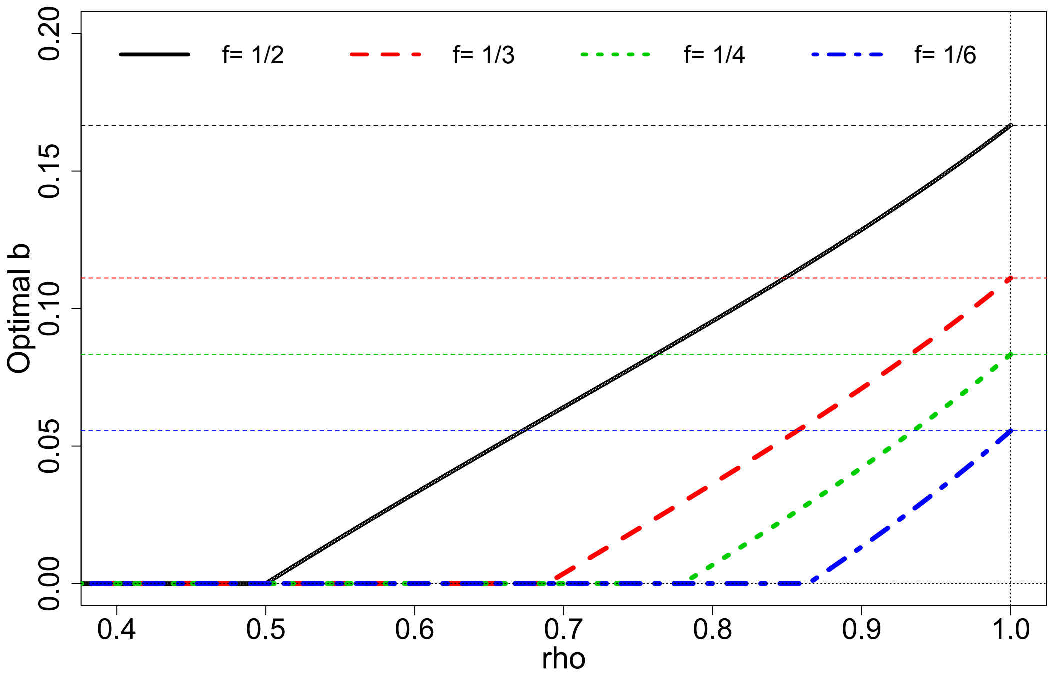


Figure 5.8: The optimal b to minimize IWAE (3:1) with equal arrivals.

the optimal IWAE value of b^* as a starting point can be employed.

The difference between the optimal b for IWAE (3:1) and for WAE (3:1; $l_1 = 20$ mins) against different occupancy levels is plotted in Figure 5.10. When unequal weights are considered, the difference between the optimal b values is relatively large. Since the difference increases with ρ, at high occupancy the optimal b for IWAE is not likely to be a good proxy of that for WAE when unequal weights are considered.

We conclude this section with some comments about the optimal solutions of the corresponding constrained two-class problems. We can readily dismiss the case where both KPIs are satisfied and the case where both are violated: in the former, we know the optimal solution, and in the latter, no adjustment to the accumulation rates will lead to compliance of both KPIs. Either the volume of traffic would have to be reduced, or service capacity increased, in order for compliance to occur.

This leaves the situation where one constraint is satisfied and one is violated. In such a

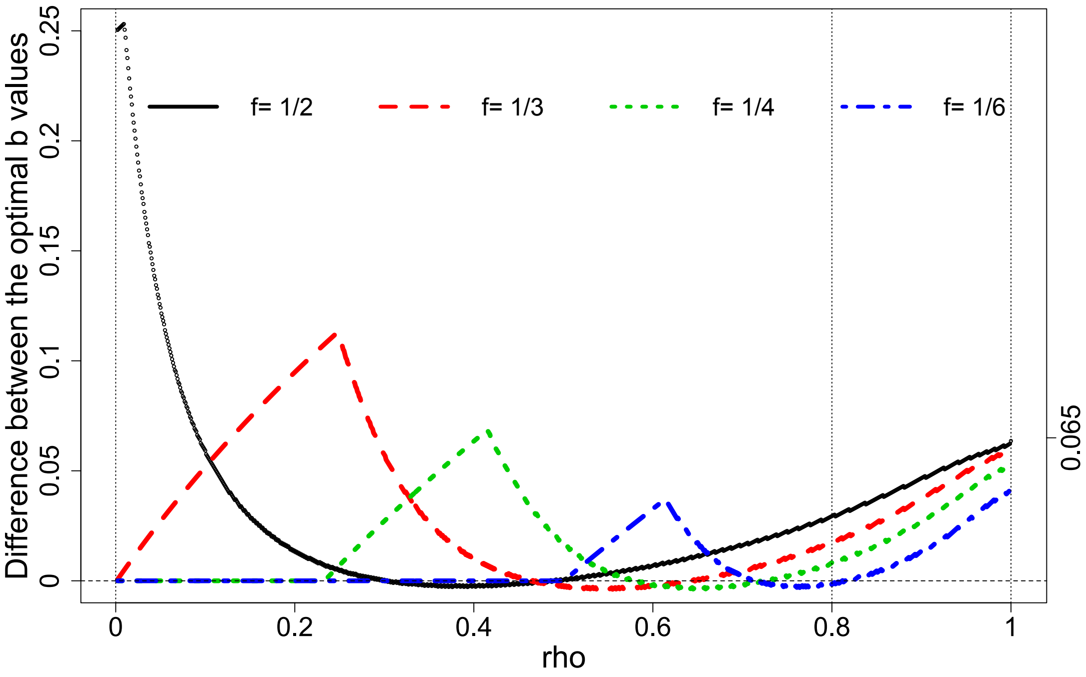


Figure 5.9: The difference between the optimal b to minimize IWAE (1:1) and the optimal b to minimize TEE ($l_1 = 20$ mins) with equal arrivals.

situation, an adjustment to the optimal class-2 priority accumulation rate b with b_1 fixed will improve the degree of compliance for one class, while degrading it for the other. In terms of the numerical examples we carried out, it was generally the case that the constraint for class-1 waiting times was the first to become binding as traffic increased. In such a situation, if the lower priority constraint is still satisfied, one can progressively reduce the level of b, which will lead to greater class-1 compliance and less class-2 compliance. If the class-1 constraint becomes satisfied before the class-2 constraint becomes violated, then the largest value of b for which this is achieved would be the optimal solution. Conversely, if it were to occur that the class-2 constraint was the first to become binding, the reverse procedure would be applied (increasing b in the hope that the class-2 constraints would become satisfied).

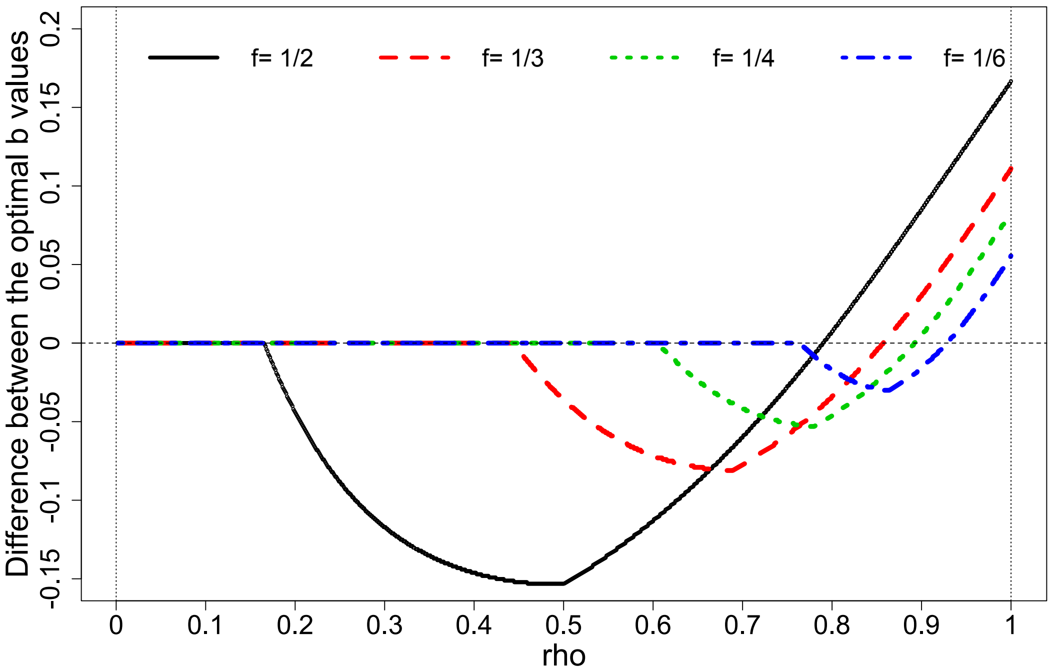


Figure 5.10: The difference between the optimal b to minimize IWAE (3:1) and the optimal b to minimize WAE (3:1; l_1 = 20 mins) with equal arrivals.

5.7 Discussion

In this paper, we have presented a general optimization problem for queuing systems operating under waiting time limits. Our optimization problem reflects the fact that customers who miss their waiting time limits are in fact of greater concern than those who satisfy them, whereas Key Performance Indicators (KPIs) on their own do not address this fact. Formally, our goal was to minimize a weighted average of the total expected excess waiting per unit time over all permissible customer selection strategies, subject to the constraints that the waiting time limits are met for each customer class.

We have also established that the individual terms in our objective function, which represent the expected excess waiting time per customer for each class, are readily obtained as a by-product of the computation of the probabilities of KPI compliance whenever the latter are being evaluated via numerical inversion of their LSTs.

While the aforementioned optimization problem can be applied to any work conserving queuing discipline, we have established that the Accumulating Priority Queue is well-suited to the goal of minimizing the excess waiting beyond the delay limits, since customers are selected for service based both on their priority class and the amount of time they have spent waiting for service. By providing the decision maker with the added flexibility of determining the best priority accumulation rates for each class, one can better achieve the desired balance between the amount of excess waiting that occurs in each class.

We have studied the convexity of the expected excess waiting time for each class in the two-class APQ optimization problem. We established that the expected excess waiting time for class-2 is convex in the accumulation rate of class-2, b; as an extension, we have proved that it is equally true for the lowest-priority class in the multi-class APQ. However, we have seen that the weighted expected excess objective WAE is not always convex in b, since a considerable number of numerical examples suggest that the expected excess waiting time for class-1 has a negative curvature over some range of b. Nonetheless, a related objective function is available, the integrated weight average of the total expected excess (IWAE), which we can work with easily. The convexity of the IWAE has been established, and the optimal solution for IWAE in terms of the optimal ratio of accumulation rates has been presented in this paper.

Extensive numerical experiments have been conducted to study the optimality behaviour of WAE in the two-class APQ based on the GS numerical inversion method. We first conclude that in high occupancy systems, the optimality behavior for TEE can be approximated by that for the integrated objective function IWAE with equal weights. The numerical examples for two classes seem to suggest a "rule of thumb" in which near-optimal performance was achieved by accumulation rates in inverse proportion to the delay limits, especially in heavy-loaded systems (as the occupancy level approaches one). In fact, all our examples (both those shown, and those not) are such that the "rule of thumb" provided an upper bound for the optimal ratio of the accumulation rates for TEE, and it is the exact limit as the occupancy approaches one for the optimal ratio for IWAE with equal weights in the two-class case.

We also find that in low-occupancy systems, the classical priority queuing discipline minimizes expected excess waiting time, whereas in more heavily-loaded systems, the accumulating priority queuing discipline is optimal. The numerical examples suggest that the cross-over occupancy point between these two regimes increases with the inverse ratio of the delay limits. An explicit solution for the switching point of the occupancy level, ρ_{sp}, has been found when considering the integrated objective IWAE. Finally, in the case where the APQ discipline is optimal, the optimal ratio b^* for IWAE (1:1) seems to provide a good approximation for the optimal ratio to minimize TEE, and the inverse ratio of the delay limits is typically a tight upper bound.

The "rule of thumb" ensures that all customers approaching their waiting time limits will have accumulated comparable amounts of priority, so that each customer class has roughly the same probability of exceeding their respective limits. Consequently, all acuity classes will observe comparable levels of compliance, suggesting that different compliance standards are are at best unnecessary and at worst inconsistent with a TEE goal. Based on our numerical investigations, we conjecture that the "rule of thumb" is also nearly optimal in the multi-class APQ system, however the complexity of the mathematical analysis in the multi-class case prevented us from obtaining similar optimality results. In the more general case where the weights placed on the excess waiting times are different, no rule of thumb has been found for WAE; however, the optimal solution has been derived for IWAE with unequal weights, which could be used as a rough starting point for WAE. To find successively better approximations for WAE, some numerical algorithms need to be considered, such as Newton's method.

We conclude this paper with some observations on the use of differing compliance probabilities for the various KPIs. In situations where the Rule of Thumb performs well, we have noted that the consequence would be that the various classes of customers would attain comparable amounts of credit as they approach their respective time limits, and thus we can anticipate similar waiting time compliance. As it is typically the case that the KPI for the top priority class has the most stringent compliance standards, any KPI with a lower standard becomes unnecessary. In other cases where the optimal accumulation rates for the various classes are notably lower than the Rule of Thumb would suggest, it may be the case that the KPI for a lower class would be the first to become binding. However, this rarely occurred in the examples we ran. In any case, it is important to remember that the level of compliance is clearly a secondary measure to be considered, relative to the desirable goal that all patients access treatment by their specified time limit.

References

[1] Akan, M., Ata, B. s. & Olsen, T. (2012). Congestion-based lead-time quotation for heterogenous customers with convex-concave delay costs: Optimality of a cost-balancing policy based on convex hull functions. *O*perations research. 60(6), 1505–1519.

[2] Arnett, G., Hadorn, D. & the Steering Committee of the Western Canada Waiting List Project. (2003). Developing priority criteria for hip and knee replacement surgery: Results from the Western Canada Waiting List Project. Canadian Journal of Surgery. 46(4), 290–296.

[3] ATS (2000). The Australasian Triage Scale. From the website. *http : //www.acem.org.au/media/policies_and_guidelines/P*06*_Aust_Triage_Scale_Nov_*2000*.eps*.

[4] Brigham, E. & Morrow, R. (1967). The fast fourier transform. Spectrum, IEEE. 4(12), 63–70.

[5] Çelik, S. & Maglaras, C. (2008). Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. Management Science. 54(6), 1132–1146.

[6] Conway, R. W., Maxwell, W. L., & Miller, L. W. (1967). Theory of scheduling. Addison-Wesley.

[7] Canadian Association of Emergency Physicians. The Canadian Triage and Acuity Scale (CTAS). From the website. *http ://www.calgaryhealthregion.ca/policy/docs/*1451*/Admission_over_capacity_AppendixA.eps*.

[8] Davies, J. & Little, L. (2012). The Taming of the Queue: Looking back, going forward: reframing timely access as part of health system transformation. From the website. *http : //www.cfhi − fcass.ca/sf − docs/default − source/taming − of − the − queue− english/TQ*2012 − *FinalReport − EN.pdf?sfvrsn* = 0.

[9] Dodd, G. (2011). Key performance indicator (KPI) report. South West London and St George's, Mental Health NHS Trust. Quarterly 1, 1–19.

[10] Gaver, D. (1966). Observing stochastic processes and approximate transform inversion. *O*perations Research. 14(3), 444–459.

[11] Keskinocak, P., Ravi, R. & Tayur, S. (2001). Scheduling and reliable lead-time quotation for orders with availability intervals and lead-time sensitive revenues. Management Science. 47(2), 264–279.

[12] Kleinrock, L. (1964). A delay dependent queue discipline. Naval Research Logistics. Quarterly 11, 329–341.

[13] Klugman, S. A., Panjer H. H. & Willmot G. E. (2008). Loss models from data to decisions. Third edition. Weiley.

[14] Li, N. & Stanford, D. A. (Submitted). Multi-server accumulating priority queues with heterogeneous servers. European Journal of Operational Research.

[15] NHS-Leeds (2013). NHS Leeds West Clinical Commissioning Group quality & performance report. From the website. *http : //www.leedswestccg.nhs.uk/downloads/about %*20*us/gb%*20*may%*202013/68%20 − %20*performance%*20*report%*20*leeds%*20*west %*20*april%*20*minus%*20*deep%*20*dive.pdf.*

[16] NHS-Stockport (2014). Accident & emergency clinical quality indicators. From the website. *https : //www.stockport.nhs.uk/webdocs/AE_Clinical_Quality_Indicators_*20140 3.*pdf.*

[17] Sharif, A. B., Stanford, D. A., Taylor, P. & Ziedins, I. (2014). A multi-class multi-server accumulating priority queue with application to health care. *O*perations Research for Health Care. 3(2), 73–79.

[18] Stanford, D. A., Taylor, P. & Ziedins, I. (2014). Waiting time distributions in the accumulating priority queue. Queueing Systems. 77(3), 297–330.

[19] Stehfest, H. (1970). Numerical inversion of Laplace transforms. Communications of the ACM. 13(1), 47–49 (Algorithm 368 with correction (October 1970), 13, No. 10).

[20] Wolff, R. W. (1982). Poisson arrivals see time averages. *O*perations Research. 30(2), 223–231.

Chapter 6

Conclusions and future work

6.1 Main contributions

1. The main contributions of the first paper "Multi-server Accumulating Priority Queues with Heterogeneous Servers" as follows:

   We have addressed two distinct aspects when studying a heterogeneous multi-server queue. In the first instance, we considered the queuing systems operating under the so-called "r-dispatch policy" to determine which of the available servers is to be used to server a newly arriving customer. This policy is quite flexible and is able to capture several traditional service dispatch policies including RCS, FSF, SSF and RBS. We then ponder the question of how to determine the optimal service dispatch policy, among the four popular service dispatch policies above, in order to improve the system performance relative to the corresponding homogeneous system. To answer this question, we first presented a conservation law for a heterogeneous multi-server system, under any non-preemptive work-conserving queuing discipline, with a given service dispatch policy. Then, we were able to find an optimal level of heterogeneity to minimize an appropriate cost which arises as a result of this conservation law.

   In the second instance, we addressed the question of who to select next for service when there are different classes of customers present. The accumulating priority queue is the modern name for the “delay dependent queue discipline” introduced by Kleinrock in 1964. This model allows customers to accumulate priority while they wait, at a rate that reflects their class’s relative importance. The customer with the greatest accumulated priority at service completion epochs is the one selected for service. The principle advantage of the accumulating priority queuing discipline is that it balances the advantages of the FCFS and the classical priority queuing disciplines — priority matters, but long waiting times do as well. Further, we can regain each of these limiting cases by appropriate choices of the accumulation rate parameters. We then derived the waiting time distributions for all customer classes under the accumulating priority queue, and obtained the waiting times for the two other cases as a by-product. We concluded by carrying out several numerical experiments for validating our model and investigating the impact of the level of heterogeneity under different service dispatch policies.

2. The main contributions of the second paper “On Waiting Times for Nonlinear Accumulating Priority Queues” as follows:

   In the first instance, motivated by the Kleinrock and Finkelstein approach, we managed to link any power-law APQ with an equivalent linear APQ such that the order of all customers present in both systems is the same at all points in time. This enables us to invoke the results of Stanford *et al.* to determine the waiting time distributions in all power-law APQs.

   In the second instance, we ponder the question as to how general a nonlinear priority accumulation system can be, yet still give rise to an equivalent linear system. We answered that question, and determined the necessary and sufficient condition which says, essentially, that so long as the priority accumulation functions are based upon scaled versions of a common, differentiable, monotonically increasing function, we can find a linear equivalent, and come up with the corresponding waiting time distributions. Beyond these requirements, any such function will do the job.

   From a practical viewpoint, this provides great breadth, in that a decision maker can select any priority accumulation shape they like, so long as it monotonically increasing and differentiable. From a theoretical viewpoint, it means that no further generality is thus obtained, and so the waiting time results in Stanford *et al.* can still be invoked. We illustrated the results through three numerical examples.

3. The main contributions of the third paper "Optimization of Queues Operating Under Waiting Time Limits" as follows:

   This paper pertains to the trend for health care systems to respond to so-called "Key Performance Indicators (KPIs)". The KPI approach specifies for each class of customers both a time target for customers to commence service, and a compliance probability indicating the proportion of customers that meet the target. The main problem with working solely on the basis KPIs is that no consequence is specified for customers who miss their target, when in fact a customer who misses their KPI target is of greater, not lesser importance.

   We addressed this fundamental oversight with an objective function that seeks to quantify how much excess is occurring under a given queuing discipline. We have done so initially by seeking to minimize the weighted average of the total expected excess waiting time (WAE) in KPI-based systems, which is an ideal setting for the Accumulating Priority queuing discipline, where each class of customers progresses fairly towards timely access by its own delay limit. By selecting the right priority accumulation rate for each class, one can arrange the system so that customers from different classes may accumulate the same amount of priority credits when they reach their respective time limits. Issues relating to convexity for WAE have been presented and discussed.

   Due to the difficulties of the WAE optimization problem, we introduced a surrogate objective function, the integrated weighted average excess (IWAE), which provides a useful proxy for WAE. We also proposed a rule of thumb in which near-optimal performance can be achieved by accumulation rates in the inverse proportion to the delay limits, especially in heavy loaded systems. Extensive numerical investigations were carried out for validating the rule of thumb under various conditions; we also compared numerically the differences in optimal performance under the WAE and IWAE, both in equally and unequally weighted environments. Furthermore, a comprehensive numerical study, combined with theoretical derivations, on the convexity of the WAE objective function was presented in Appendix B.4, which provides us a better understanding of its optimal behaviour.

6.2 Future work

1. Extension of the multi-class APQ with heterogeneous servers:

   We have presented the expressions in closed form of the waiting time distributions and the conservation law in a multi-class heterogeneous multi-server APQ. We anticipate two major directions for future developments:

   - Heterogeneous multi-server APQ optimization problems, such as, how to select the parameter set $\{b_k; k = 1, \ldots, K\}$ to minimize an appropriately defined cost or meet certain waiting time targets, as well as optimal control of heterogeneous APQs using different service dispatch policies.
   - Extension to the situation where the service time distributions are non-exponential. Such a problem is non-trivial, although some results might be possible in the case of the service time distributions built upon the exponential. Other types of service dispatch policies can be considered in such systems, for instance, the idle-time-based dispatch policies including Longest Idle Server First (LISF), and Shortest Idle Server First (SISF).

2. Accumulating priority queues with other types of priority accumulation functions:

   We have presented conditions for a general nonlinear APQ to have a linear proxy, in which customers of all classes have the same waiting time distribution as in the nonlinear APQ. There are two major sets of priority accumulation functions left to be explored:

- An APQ with affine priority accumulation functions, in which customers of different classes start with different amounts of initial priority.
- An APQ with more general nonlinear priority accumulation functions, for which a linear proxy does not exist.

3. Multi-class $M/M_i/c$ APQs with class-dependent service rates:

   We are currently studying a multi-class $M/M_i/c$ APQ where the service rate depends on the customer's class type, rather than the server. However, this problem is complicated by the possible combinations of customers' types in service at the beginning of an accreditation interval. Hopefully, we may solve the problem in the near future.

4. Extension of the optimization problems of queues operating under waiting time limits:

   We have formulated and studied the WAE and IWAE optimization problems for the systems operating under waiting time limits, where the investigations and derivations are primarily considered for the two-class multi-server APQs in this thesis. We have made a further exploration of the case with multiple classes of customers. Based on some fundamental derivations we have carried out, we foresee that explicit solutions would be obtained for the IWAE in the multi-class queues.

Appendix A

Additional materials in Chapter 3

A.1 Waiting times of accredited customers in an $M/M_i/c$ APQ

Stanford *et al.* [24] introduces the concept of the "the maximum priority process" $\{M_k(t); k = 1, 2, \ldots, K\}$. For $k = 1, 2, \ldots, K$, $M_k(t)$ provides an upper bound on the possible accumulated priority of class-k customers at time t. Within a single busy period, class-k (and higher) customers can eventually exceed the maximum priority of customers from classes with priorities lower than k. Such customers are said to be "served at accreditation level k" if their priority upon entry to service at time t lies in the interval $(M_{k+1}(t), M_k(t)]$. Moreover, the authors defined the accreditation interval at level k, which is a period of time that starts either at the beginning of a busy period or when a customer is served at some accreditation level l_1 for $l_1 > k$, and finishes either at the end of a busy period or when another customer is served at some accreditation level l_2 for $l_2 > k$. Whenever a customer is served at accreditation level k, accreditation intervals at all level $l < k$ commence.

Stanford *et al.* [24] showed that an accreditation interval at level k can be thought of as a delay cycle in the sense of Conway *et al.* [4] that starts with the service time of the initiating customer and continues as long as there are customers at accreditation level $l \leq k$. In our case, within the delay cycle, the service times are exponentially distributed with rate μ_a, and the instants at which customers of all classes $i; i \leq k$ become accredited at level k are distributed as a Poisson process with rate $\Lambda_k = \sum_{i=1}^k \lambda_i(b_i - b_{k+1})/b_i$. Therefore, the LST $\tilde{\Gamma}_k(s)$ of the duration of the accreditation interval at level k is obtained by solving the functional equation

$$\tilde{\Gamma}_k(s) = \tilde{B}(s + \Lambda_k(1 - \tilde{\Gamma}_k(s))) \tag{A.1}$$

(see Conway *et al.* [4, page 150, equation (7)]), where $\tilde{B}(s) = \mu_a/(\mu_a + s)$, which results in equation (3.28).

Let σ_k denote the stationary proportion of time that the server spends on customers served at all accreditation levels $l = 1, 2, \ldots, k$, so that $\sigma_k = \sum_{j=1}^k \rho_j(b_j - b_{k+1})/b_j$ (see Stanford *et al.* [24]). Then the LST of the waiting time of a class-k customer who is served at accreditation level k; $\tilde{W}_{acc}^{(k)}(s)$, is shown in Stanford *et al.* [24] to be given by

$$\begin{aligned}\tilde{W}_{acc}^{(k)}(s) = \Bigg[\frac{1-\rho}{1-\sigma_k} &+ \tilde{W}_+^{(k+1)}\left(\frac{b_{k+1}}{b_k}s; \boldsymbol{\mu}, c\right)\sum_{j=1}^{k}\frac{\rho_j(b_{k+1}/b_j)}{1-\sigma_k} \\ &+ \sum_{j=k+1}^{K}\frac{\rho_j}{(1-\sigma_k)}\tilde{W}_+^{(j)}\left(\frac{b_j}{b_k}s; \boldsymbol{\mu}, c\right)\Bigg]\tilde{W}_{acc}^{(k,0)}(s).\end{aligned} \tag{A.2}$$

The term $\tilde{W}_{acc}^{(k,0)}(s)$ in equation (A.2) is given by

$$\tilde{W}_{acc}^{(k,0)}(s) = \frac{[\mu_{\Gamma_{k-1}} - \sum_{i=1}^k \lambda_i(b_k - b_{k+1})/b_i][\tilde{\phi}_k(sb_{k+1}/b_k) - \tilde{\Gamma}_{k-1}(s)]}{(1 - b_{k+1}/b_k)[s - (\sum_{i=1}^k \lambda_i b_k/b_i)(1 - \tilde{\Gamma}_{k-1}(s))]}, \tag{A.3}$$

where $1/\mu_{\Gamma_k}$ is the mean of a random variable with LST $\tilde{\Gamma}_k(s)$; that is, $1/\mu_{\Gamma_k} = -\frac{d\tilde{\Gamma}_k(s)}{ds}|_{s=0}$. The LST, $\tilde{\phi}_k(s)$, is the solution to the functional equation $\tilde{\phi}_k(s) = \tilde{\Gamma}_{k-1}[s + (\sum_{i=1}^k \lambda_i(b_k - b_{k+1})/b_i)(1 - \tilde{\phi}_k(s))]$. The derivation of these expressions can be found in Stanford *et al.* [24].

Remark Assuming $b_{k+1}/b_k = 1/M$; $k = 1, 2, \ldots, K-1$. Under the classical priority queuing discipline, $\lim_{M\to\infty} b_{k+1}/b_k = 0$, then we have $\lim_{M\to\infty} \Lambda_k = \sum_{i=1}^k \lambda_i$, $\lim_{M\to\infty} \sigma_k = \sum_{j=1}^k \rho_j$, and $\lim_{M\to\infty} \mu_{\Gamma_{k-1}} = \mu_a - \Lambda_{k-1}$. By equations (3.31), (A.2) & (A.3), the LST of the delayed waiting time distribution for each class under the classical priority queue is

$$\begin{aligned}\lim_{M\to\infty} \tilde{W}_+^{(k)}(s; \boldsymbol{\mu}, c) &= \lim_{M\to\infty} \tilde{W}_{acc}^{(k)}(s) = \lim_{M\to\infty} \tilde{W}_{acc}^{(k,0)}(s) \\ &= \frac{(\mu_{\Gamma_{k-1}} - \lambda_k)\big(1 - \tilde{\Gamma}_{k-1}(s)\big)}{s - \lambda_k(1 - \tilde{\Gamma}_{k-1}(s))} = \frac{(\mu_a - \Lambda_k)\big(1 - \tilde{\Gamma}_{k-1}(s)\big)}{s - \lambda_k(1 - \tilde{\Gamma}_{k-1}(s))}.\end{aligned} \tag{A.4}$$

We observe that it can be shown readily that(A.4), after accounting for customers who do not wait, is consistent in the single server case with the unconditional waiting time LST found in Conway *et al.* [4] page 164, equation (29).

A.2 Additional materials for the average waiting times in a linear APQ

Kleinrock [12] derived the expressions for the average waiting times for different classes in a single-server APQ:

$$m_k = \frac{M_0 - \sum_{j=k+1}^{K} \rho_j (1 - b_j/b_k) m_j}{1 - \sum_{j=1}^{k-1} \rho_j (1 - b_k/b_j)}. \tag{A.5}$$

In the $M/G/1$ case, $M_0 = W_0/(1-\rho)$. However, it can be extended to the $M/M_i/c$ case in which $M_0 = \pi(\boldsymbol{\mu}, c; r)/(\mu_a - \lambda)$, as shown in equation (3.12). Furthermore, the following equations can be derived. For $k = 1, 2, \ldots, K$,

$$A_k \;=\; \frac{m_k}{M_0} = \frac{1 - \sum_{j=k+1}^{K} \rho_j (1 - b_j/b_k) A_j}{1 - \sum_{j=1}^{k-1} \rho_j (1 - b_k/b_j)}, \tag{A.6}$$

where $A_K = [1 - \sum_{j=1}^{K-1} \rho_j (1 - b_k/b_j)]^{-1}$ and $A_1 = 1 - \sum_{j=2}^{K} \rho_j (1 - b_j/b_1) A_j$. Immediately, we may have, for $k, j \in [1, 2, \ldots, K]$,

$$\frac{m_k}{m_j} = \frac{A_k}{A_j}. \tag{A.7}$$

A.3 The stationary probability $\pi(\boldsymbol{\mu}, 3; r)$ for RCS, FSF, SSF and RBS

For RCS, $\pi(\boldsymbol{\mu}, 3; 0)$ calculated from Gumbel [8] is given by

$$\pi(\boldsymbol{\mu}, 3; 0) = \frac{\lambda^3 \mu_a}{\left(\mu_1{}^2+\mu_2{}^2+\mu_3{}^2\right)\lambda^2+\left((2\mu_3+2\mu_2)\mu_1{}^2+\left(2\mu_2{}^2+2\mu_3{}^2\right)\mu_1+2\mu_2\mu_3(\mu_3+\mu_2)\right)\lambda+6\mu_1\mu_2\mu_3\mu_a}. \tag{A.8}$$

For FSF, $\pi(\boldsymbol{\mu}, 3; \infty)$ calculated by the global balance equations using Maple is given by

$$\pi(\boldsymbol{\mu}, 3; \infty) = (\lambda + \mu_2)\mu_a \lambda^3 \left(\mu_3{}^2 + (\mu_1 + 2\lambda)\mu_3 + \lambda^2\right)(\mu_2 + \mu_3 + \lambda)\Big[(\lambda + \mu_1)\left(\mu_2{}^2 + (\mu_1 + 2\lambda)\mu_2 + \lambda^2\right)\mu_3{}^5$$

$$\begin{aligned}
&+ 3\ (\lambda + \mu_1)\Big(\mu_2{}^2 + (\mu_1 + 2\,\lambda)\,\mu_2 + \lambda^2\Big)(\mu_1 + \mu_2 + \lambda)\,\mu_3{}^4 + \Big((3\,\mu_1 + 3\,\lambda)\,\mu_2{}^4 + \Big(13\,\lambda^2 + 23\,\lambda\,\mu_1 \\
&+ 10\,\mu_1{}^2\Big)\mu_2{}^3 + \Big(19\,\lambda^3 + 46\,\lambda^2\mu_1 + 37\,\lambda\,\mu_1{}^2 + 10\,\mu_1{}^3\Big)\mu_2{}^2 + \Big(12\,\lambda^4 + 32\,\lambda^3\mu_1 + 33\,\lambda^2\mu_1{}^2 \\
&+ 15\,\lambda\,\mu_1{}^3 + 3\,\mu_1{}^4\Big)\mu_2 + 3\,\mu_1{}^3\lambda^2 + 9\,\mu_1{}^2\lambda^3 + 8\,\lambda^4\mu_1 + 3\,\lambda^5\Big)\mu_3{}^3 + \Big((\lambda + \mu_1)\,\mu_2{}^5 + (6\mu_1 + 7\,\lambda)\,(\lambda + \mu_1)\,\mu_2{}^4 \\
&+ \Big(16\,\lambda^3 + 41\,\lambda^2\mu_1 + 35\,\lambda\,\mu_1{}^2 + 10\,\mu_1{}^3\Big)\mu_2{}^3 + \Big(14\,\lambda^4 + 47\,\lambda^3\mu_1 + 57\,\lambda^2\mu_1{}^2 + 29\,\lambda\,\mu_1{}^3 + 6\,\mu_1{}^4\Big)\mu_2{}^2 \\
&+ \Big(5\lambda^5 + 18\lambda^4\mu_1 + 31\mu_1{}^2\lambda^3 + 20\mu_1{}^3\lambda^2 + 6\lambda\mu_1{}^4 + \mu_1{}^5\Big)\mu_2 + \lambda^2\Big(\lambda^4 + 3\lambda^3\mu_1 + 7\lambda^2\mu_1{}^2 + 4\lambda\mu_1{}^3 + \mu_1{}^4\Big)\Big)\mu_3{}^2 \\
&+ \Big((\lambda + \mu_1)\,\mu_2{}^3 + \Big(\lambda^2 + 3\,\lambda\,\mu_1 + 2\,\mu_1{}^2\Big)\mu_2{}^2 + \mu_1{}^2\,(\mu_1 + 2\,\lambda)\,\mu_2 + \lambda^2\mu_1{}^2\Big)\Big((\lambda + \mu_1)\,\mu_2{}^2 + (\mu_1 + 2\,\lambda)^2\,\mu_2 \\
&+ 2\,\lambda^2\,(\mu_1 + 2\,\lambda)\Big)\mu_3 + \lambda^3\Big((\lambda + \mu_1)\,\mu_2{}^3 + \Big(\lambda^2 + 3\,\lambda\,\mu_1 + 2\,\mu_1{}^2\Big)\mu_2{}^2 + \mu_1{}^2\,(\mu_1 + 2\,\lambda)\,\mu_2 + \lambda^2\mu_1{}^2\Big)(\lambda + \mu_2)\Big]^{-1}.
\end{aligned} \tag{A.9}$$

For SSF, $\pi(\boldsymbol{\mu}, 3; -\infty)$ calculated by the global balance equations using Maple is given by

$$\begin{aligned}
\pi(\boldsymbol{\mu}, 3; -\infty) = {}& \mu_a\,(\lambda + \mu_1 + \mu_2)\,(\mu_2 + \lambda)\,\lambda^3\,\Big(\mu_1{}^2 + (\mu_3 + 2\,\lambda)\,\mu_1 + \lambda^2\Big)\Big[\Big(\mu_2{}^2 + (\mu_3 + 2\,\lambda)\,\mu_2 + \lambda^2\Big)(\mu_3 + \lambda)\,\mu_1{}^5 \\
&+ 3\ (\mu_2 + \mu_3 + \lambda)\Big(\mu_2{}^2 + (\mu_3 + 2\,\lambda)\,\mu_2 + \lambda^2\Big)(\mu_3 + \lambda)\,\mu_1{}^4 + \Big((3\,\mu_3 + 3\,\lambda)\,\mu_2{}^4 + \Big(13\,\lambda^2 + 23\,\lambda\,\mu_3 \\
&+ 10\,\mu_3{}^2\Big)\mu_2{}^3 + \Big(19\,\lambda^3 + 46\,\lambda^2\mu_3 + 37\,\lambda\,\mu_3{}^2 + 10\,\mu_3{}^3\Big)\mu_2{}^2 + \Big(12\,\lambda^4 + 32\,\lambda^3\mu_3 + 33\,\lambda^2\mu_3{}^2 + 15\,\lambda\,\mu_3{}^3 \\
&+ 3\,\mu_3{}^4\Big)\mu_2 + 3\,\lambda^2\mu_3{}^3 + 9\,\lambda^3\mu_3{}^2 + 8\,\lambda^4\mu_3 + 3\,\lambda^5\Big)\mu_1{}^3 + \Big((\mu_3 + \lambda)\,\mu_2{}^5 + (6\mu_3 + 7\lambda)\,(\mu_3 + \lambda)\,\mu_2{}^4 \\
&+ \Big(16\,\lambda^3 + 41\,\lambda^2\mu_3 + 35\,\lambda\,\mu_3{}^2 + 10\,\mu_3{}^3\Big)\mu_2{}^3 + \Big(14\,\lambda^4 + 47\,\lambda^3\mu_3 + 57\,\lambda^2\mu_3{}^2 + 29\,\lambda\,\mu_3{}^3 + 6\,\mu_3{}^4\Big)\mu_2{}^2 \\
&+ \Big(5\,\lambda^5 + 18\,\lambda^4\mu_3 + 31\,\lambda^3\mu_3{}^2 + 20\,\lambda^2\mu_3{}^3 + 6\,\lambda\,\mu_3{}^4 + \mu_3{}^5\Big)\mu_2 + \lambda^2\Big(\lambda^4 + 3\,\lambda^3\mu_3 + 7\,\lambda^2\mu_3{}^2 + 4\,\lambda\,\mu_3{}^3 \\
&+ \mu_3{}^4\Big)\Big)\mu_1{}^2 + \Big((\mu_3 + \lambda)\,\mu_2{}^3 + \Big(\lambda^2 + 3\,\lambda\,\mu_3 + 2\,\mu_3{}^2\Big)\mu_2{}^2 + \mu_3{}^2\,(\mu_3 + 2\,\lambda)\,\mu_2 + \lambda^2\mu_3{}^2\Big)\Big((\mu_3 + \lambda)\,\mu_2{}^2 \\
&+ (\mu_3 + 2\,\lambda)^2\,\mu_2 + 2\,\lambda^2\,(\mu_3 + 2\,\lambda)\Big)\mu_1 + \lambda^3\Big((\mu_3 + \lambda)\,\mu_2{}^3 + \Big(\lambda^2 + 3\,\lambda\,\mu_3 + 2\,\mu_3{}^2\Big)\mu_2{}^2 \\
&+ \mu_3{}^2\,(\mu_3 + 2\,\lambda)\,\mu_2 + \lambda^2\mu_3{}^2\Big)(\mu_2 + \lambda)\Big]^{-1}.
\end{aligned} \tag{A.10}$$

For RBS, $\pi(\boldsymbol{\mu}, 3; 1)$ calculated by the global balance equations using Maple is given by

$$\begin{aligned}
\pi(\boldsymbol{\mu}, 3; 1) = {}& \mu_a\lambda^3\Big[2\,\mu_2\mu_3\,(\mu_3 + \lambda + \mu_2)\,(2\,\mu_2 + 2\,\mu_3 + \lambda)\,\mu_1{}^4 + \Big(4\,\mu_2{}^4\mu_3 + \Big(4\,\lambda^2 + 22\,\mu_3\lambda + 16\,\mu_3{}^2\Big)\mu_2{}^3 \\
&+ \Big(4\,\lambda^3 + 27\,\mu_3\lambda^2 + 46\,\mu_3{}^2\lambda + 16\,\mu_3{}^3\Big)\mu_2{}^2 + (\mu_3 + \lambda)\Big(\lambda^3 + 9\,\mu_3\lambda^2 + 18\,\mu_3{}^2\lambda + 4\,\mu_3{}^3\Big)\mu_2 \\
&+ \mu_3\lambda^2\,(2\,\mu_3 + \lambda)^2\Big)\mu_1{}^3 + \Big(\Big(6\,\mu_3\lambda + 8\,\mu_3{}^2\Big)\mu_2{}^4 + \Big(4\,\lambda^3 + 27\,\mu_3\lambda^2 + 46\,\mu_3{}^2\lambda + 16\,\mu_3{}^3\Big)\mu_2{}^3 \\
&+ \Big(2\,\lambda^4 + 22\,\lambda^3\mu_3 + 66\,\lambda^2\mu_3{}^2 + 46\,\lambda\,\mu_3{}^3 + 8\,\mu_3{}^4\Big)\mu_2{}^2 + \Big(4\,\mu_3\lambda^4 + 22\,\mu_3{}^2\lambda^3 + 27\,\mu_3{}^3\lambda^2 + 6\,\mu_3{}^4\lambda\Big)\mu_2 \\
&+ 2\,\lambda^3\mu_3{}^2\,(2\,\mu_3 + \lambda)\Big)\mu_1{}^2 + ((2\,\mu_3 + \lambda)\,\mu_2 + \mu_3\lambda)\Big(2\,\mu_3\,(\mu_3 + \lambda)\,\mu_2{}^3 + (\mu_3 + \lambda)\Big(\lambda^2 + 7\,\mu_3\lambda + 2\,\mu_3{}^2\Big)\mu_2{}^2 \\
&+ \Big(3\,\lambda^3\mu_3 + 8\,\lambda^2\mu_3{}^2 + 2\,\lambda\,\mu_3{}^3\Big)\mu_2 + \mu_3{}^2\lambda^3\Big)\mu_1 + ((2\,\mu_3 + \lambda)\,\mu_2 + \mu_3\lambda)^2\,\mu_2\lambda^2\mu_3\Big] \times \Big[\mu_2\mu_3\,(\mu_2 + \mu_3)^2\,\mu_1{}^8 \\
&+ (5\,\mu_2 + 5\,\mu_3 + 4\lambda)\,\mu_2\,(\mu_2 + \mu_3)\,(\mu_3 + \lambda + \mu_2)\,\mu_3\mu_1{}^7 + \Big(10\,\mu_2{}^5\mu_3 + \Big(36\,\mu_3\lambda + 42\,\mu_3{}^2\Big)\mu_2{}^4
\end{aligned}$$

$$
\begin{aligned}
&+ (2\lambda^3 + 43\mu_3\lambda^2 + 112\mu_3^2\lambda + 64\mu_3^3)\mu_2^3 + (\lambda^4 + 24\lambda^3\mu_3 + 86\lambda^2\mu_3^2 + 112\lambda\mu_3^3 + 42\mu_3^4)\mu_2^2\\
&+ (4\mu_3\lambda^4 + 24\mu_3^2\lambda^3 + 43\mu_3^3\lambda^2 + 36\mu_3^4\lambda + 10\mu_3^5)\mu_2 + \lambda^3\mu_3^2(2\mu_3 + \lambda)\Big)\mu_1^6 + \Big(10\mu_2^6\mu_3\\
&+ (54\mu_3\lambda + 58\mu_3^2)\mu_2^5 + (6\lambda^3 + 109\mu_3\lambda^2 + 246\mu_3^2\lambda + 124\mu_3^3)\mu_2^4 + (10\lambda^4 + 105\lambda^3\mu_3 + 355\lambda^2\mu_3^2\\
&+ 384\lambda\mu_3^3 + 124\mu_3^4)\mu_2^3 + (5\lambda^5 + 49\mu_3\lambda^4 + 204\mu_3^2\lambda^3 + 355\mu_3^3\lambda^2 + 246\mu_3^4\lambda + 58\mu_3^5)\mu_2^2\\
&+ (\lambda^6 + 12\lambda^5\mu_3 + 49\lambda^4\mu_3^2 + 105\lambda^3\mu_3^3 + 109\lambda^2\mu_3^4 + 54\lambda\mu_3^5 + 10\mu_3^6)\mu_2 + \lambda^3\mu_3(\mu_3 + \lambda)(\lambda^2\\
&+ 4\mu_3\lambda + 6\mu_3^2)\Big)\mu_1^5 + \Big(5\mu_2^7\mu_3 + (36\mu_3\lambda + 42\mu_3^2)\mu_2^6 + (6\lambda^3 + 109\mu_3\lambda^2 + 246\mu_3^2\lambda + 124\mu_3^3)\mu_2^5\\
&+ (18\lambda^4 + 162\lambda^3\mu_3 + 546\lambda^2\mu_3^2 + 562\lambda\mu_3^3 + 174\mu_3^4)\mu_2^4 + (11\lambda^5 + 113\mu_3\lambda^4 + 514\mu_3^2\lambda^3 + 878\mu_3^3\lambda^2\\
&+ 562\mu_3^4\lambda + 124\mu_3^5)\mu_2^3 + (2\lambda^6 + 33\lambda^5\mu_3 + 198\lambda^4\mu_3^2 + 514\lambda^3\mu_3^3 + 546\lambda^2\mu_3^4 + 246\lambda\mu_3^5 + 42\mu_3^6)\mu_2^2\\
&+ (4\lambda^6\mu_3 + 33\lambda^5\mu_3^2 + 113\lambda^4\mu_3^3 + 162\lambda^3\mu_3^4 + 109\lambda^2\mu_3^5 + 36\lambda\mu_3^6 + 5\mu_3^7)\mu_2 + 6\mu_3^5\lambda^3 + 18\mu_3^4\lambda^4\\
&+ 11\mu_3^3\lambda^5 + 2\mu_3^2\lambda^6\Big)\mu_1^4 + \Big(\mu_2^8\mu_3 + (9\mu_3\lambda + 15\mu_3^2)\mu_2^7 + (2\lambda^3 + 43\mu_3\lambda^2 + 112\mu_3^2\lambda + 64\mu_3^3)\mu_2^6\\
&+ (10\lambda^4 + 105\lambda^3\mu_3 + 355\lambda^2\mu_3^2 + 384\lambda\mu_3^3 + 124\mu_3^4)\mu_2^5 + (11\lambda^5 + 113\mu_3\lambda^4 + 514\mu_3^2\lambda^3 + 878\mu_3^3\lambda^2\\
&+ 562\mu_3^4\lambda + 124\mu_3^5)\mu_2^4 + (2\lambda^6 + 42\lambda^5\mu_3 + 304\lambda^4\mu_3^2 + 828\lambda^3\mu_3^3 + 878\lambda^2\mu_3^4 + 384\lambda\mu_3^5 + 64\mu_3^6)\mu_2^3\\
&+ (5\lambda^6\mu_3 + 64\lambda^5\mu_3^2 + 304\lambda^4\mu_3^3 + 514\lambda^3\mu_3^4 + 355\lambda^2\mu_3^5 + 112\lambda\mu_3^6 + 15\mu_3^7)\mu_2^2 + (5\mu_3^2\lambda^6 + 42\mu_3^3\lambda^5\\
&+ 113\mu_3^4\lambda^4 + 105\mu_3^5\lambda^3 + 43\lambda^2\mu_3^6 + 9\lambda\mu_3^7 + \mu_3^8)\mu_2 + 2\lambda^6\mu_3^3 + 11\lambda^5\mu_3^4 + 10\lambda^4\mu_3^5 + 2\lambda^3\mu_3^6\Big)\mu_1^3\\
&+ \Big(2\mu_2^8\mu_3^2 + (4\mu_3\lambda^2 + 18\mu_3^2\lambda + 15\mu_3^3)\mu_2^7 + (\lambda^4 + 24\lambda^3\mu_3 + 86\lambda^2\mu_3^2 + 112\lambda\mu_3^3 + 42\mu_3^4)\mu_2^6\\
&+ (5\lambda^5 + 49\mu_3\lambda^4 + 204\mu_3^2\lambda^3 + 355\mu_3^3\lambda^2 + 246\mu_3^4\lambda + 58\mu_3^5)\mu_2^5 + (2\lambda^6 + 33\lambda^5\mu_3 + 198\lambda^4\mu_3^2\\
&+ 514\lambda^3\mu_3^3 + 546\lambda^2\mu_3^4 + 246\lambda\mu_3^5 + 42\mu_3^6)\mu_2^4 + (5\lambda^6\mu_3 + 64\lambda^5\mu_3^2 + 304\lambda^4\mu_3^3 + 514\lambda^3\mu_3^4 + 355\lambda^2\mu_3^5\\
&+ 112\lambda\mu_3^6 + 15\mu_3^7)\mu_2^3 + (6\mu_3^2\lambda^6 + 64\mu_3^3\lambda^5 + 198\mu_3^4\lambda^4 + 204\mu_3^5\lambda^3 + 86\lambda^2\mu_3^6 + 18\lambda\mu_3^7 + 2\mu_3^8)\mu_2^2\\
&+ (5\lambda^6\mu_3^3 + 33\lambda^5\mu_3^4 + 49\lambda^4\mu_3^5 + 24\lambda^3\mu_3^6 + 4\lambda^2\mu_3^7)\mu_2 + 2\lambda^6\mu_3^4 + 5\mu_3^5\lambda^5 + \lambda^4\mu_3^6\Big)\mu_1^2 + \Big(\mu_2^4\mu_3^2\\
&+ (3\mu_3\lambda^2 + 6\mu_3^2\lambda + 2\mu_3^3)\mu_2^3 + (\mu_3 + \lambda)(\lambda^3 + 8\mu_3\lambda^2 + 5\mu_3^2\lambda + \mu_3^3)\mu_2^2 + (3\mu_3\lambda^4 + 9\mu_3^2\lambda^3 + 3\mu_3^3\lambda^2)\mu_2\\
&+ \lambda^4\mu_3^2\Big)(\mu_2^3\mu_3 + (2\mu_3 + \lambda)(\mu_3 + \lambda)\mu_2^2 + (3\mu_3^2\lambda + \mu_3^3)\mu_2 + \lambda^2\mu_3^2)(\mu_2 + \mu_3)\mu_1 + ((2\mu_3 + \lambda)\mu_2 + \mu_3\lambda)\\
&(\mu_2^3\mu_3 + (2\mu_3 + \lambda)(\mu_3 + \lambda)\mu_2^2 + (3\mu_3^2\lambda + \mu_3^3)\mu_2 + \lambda^2\mu_3^2)\mu_2\lambda^3(\mu_2 + \mu_3)\mu_3\Big]^{-1}.
\end{aligned}
\tag{A.11}
$$

Table A.1: The stationary probability $\pi(\boldsymbol{\mu}, 3; r)$ when $\rho = 40\%$

$\{\mu_1, \mu_2, \mu_3\}$	RCS	FSF	SSF	RBS
$\{1, 1, 1\}$	0.14118	0.14118	0.14118	0.14118
$\{1.2, 1, 0.8\}$	0.14354	0.13266	0.15333	0.14201
$\{1.5, 1, 0.5\}$	0.15738	0.12758	0.17943	0.14723
$\{1.8, 1, 0.2\}$	0.19169	0.14215	0.21865	0.16497

Table A.2: The stationary probability $\pi(\boldsymbol{\mu}, 3; r)$ when $\rho = 75\%$

$\{\mu_1, \mu_2, \mu_3\}$	RCS	FSF	SSF	RBS
$\{1, 1, 1\}$	0.56776	0.56776	0.56776	0.56776
$\{1.2, 1, 0.8\}$	0.57074	0.56093	0.57930	0.56938
$\{1.5, 1, 0.5\}$	0.58696	0.56274	0.60404	0.57897
$\{1.8, 1, 0.2\}$	0.61965	0.59001	0.63692	0.60449

Table A.3: The stationary probability $\pi(\boldsymbol{\mu}, 3; r)$ when $\rho = 90\%$

$\{\mu_1, \mu_2, \mu_3\}$	RCS	FSF	SSF	RBS
$\{1, 1, 1\}$	0.81706	0.81706	0.81706	0.81706
$\{1.2, 1, 0.8\}$	0.81861	0.81412	0.82253	0.81798
$\{1.5, 1, 0.5\}$	0.82684	0.81617	0.83446	0.82332
$\{1.8, 1, 0.2\}$	0.84258	0.83048	0.84996	0.83646

Table A.4: The stationary probability $\pi(\boldsymbol{\mu}, 3; r)$ when $\rho = 98\%$

$\{\mu_1, \mu_2, \mu_3\}$	RCS	FSF	SSF	RBS
$\{1, 1, 1\}$	0.96245	0.96245	0.96245	0.96245
$\{1.2, 1, 0.8\}$	0.96280	0.96185	0.96362	0.96267
$\{1.5, 1, 0.5\}$	0.96462	0.96242	0.96621	0.96389
$\{1.8, 1, 0.2\}$	0.96803	0.96561	0.96954	0.96681

Appendix B

The proofs and additional materials in Chapter 5

B.1 The proofs in Section 5.3

Theorem 5.3.1. *Given the foregoing definitions, the function* $S_2(b, x)$ *is a monotonically decreasing convex function in b for* $0 \leq b \leq 1, \forall x \geq 0$.

The proof of Theorem 5.3.1 requires the following lemma:

Lemma B.1.1. *The stationary waiting time random variable* $\mathcal{W}_2$ *for class-2 customers in a stable two-class APQ can be expressed as the sum of two dependent random variables*

$$\mathcal{W}_2 = \mathcal{W} + Y \tag{B.1}$$

where $\mathcal{W}$ *denotes the stationary waiting time random variable in the* $M/G/1$ *FCFS comparator queue, and Y refers to a compound Poisson random variable*

$$Y = \eta_1 + \eta_2 + \ldots + \eta_N \tag{B.2}$$

where the random variable N denotes the number of class-1 customers to accredit during $\mathcal{W}$, *and* $\eta_i;\ i = 1, 2, \ldots$ *denote a sequence of i.i.d. busy period random variables for an* $M/G/1$

queue with arrivals at rate $\lambda_1(1-b)$ and whose service times are drawn from the class-1 service time distribution. (In (B.2)*, it is to be understood that if $N = 0$, then $Y = 0$.)*

Proof. We employ the same viewpoint to express the class-2 stationary waiting time $\mathcal{W}_2$ as was employed in Theorem 9.1 of Stanford *et al.* [18]. This is done by rearranging the service discipline in such a way that preserves the total waiting time experienced by a tagged arrival from that class.

Consider such a tagged class-2 customer. It will wait for all of the work present in the system upon its arrival, plus that which arrives later but will be served ahead of it. Since the tagged customer is from the lowest priority class, and is an arrival from a Poisson process, the distribution of the work it finds in system at such instants equals the distribution of the stationary unfinished workload, by the "Poisson-arrivals-see-time averages" (PASTA) property of the Poisson process (Wolff [20]). However, the stationary unfinished workload is invariant for all work-conserving service disciplines. As shown in Stanford *et al.* [18] Theorem 9.1, its distribution is the same as the stationary waiting time distribution for $\mathcal{W}$ in the $M/G/1$ FCFS comparator queue.

During this initial period $\mathcal{W}$, the process of later arrivals that will gain accreditation (and therefore be served ahead of the tagged class-2 customer) constitutes a Poisson process at rate $\lambda_1(1-b)$, as established in Stanford *et al.* [18] Lemma 4.2. While some of these later class-1 arrivals may enter service according to the APQ service discipline ahead of some of the class-2 customers already present at the arrival instant of the tagged customer, the actual order of service does not matter, so long as all such work is completed prior to the tagged customer's entry into service.

Thus we can write $\mathcal{W}_2 = \mathcal{W} + Y$, where Y represents the total of all of the service times for class-1 customers that gain accreditation relative to the tagged customer prior to its entry into service. We are able to characterize Y in terms of $\mathcal{W}$ as follows. We do so by resorting to the following discipline, which parallels the rearrangement of service times in the derivation by Conway *et al.* [6] of the implicit transform equation for the duration of a busy period in an $M/G/1$ queue.

Our rearrangement places all of the N class-1 customers who gain accreditation during $\mathcal{W}$ in a special queue. Upon completion of $\mathcal{W}$, the first (if any) of these N customers is selected for service. The server then selects newly-accredited class-1 customers in the main queue until none remain, at which point the next (if any) of the class-1 customers in the special queue is selected, and so on. We observe that, by construction, each such "sub-busy period" comprising the service of one customer from the special queue and the subsequent accredited arrivals to the main queue is identically distributed to the busy period η in an $M/G/1$ queue with arrivals at rate $\lambda_1(1-b)$ and whose service times are drawn from the class-1 service time distribution.

In this way, Y can be expressed according to equation (B.2), with N representing the number of Poisson events at rate $\lambda_1(1-b)$ to occur during $\mathcal{W}$, and the ith such accrediting customer, $i = 1, 2, \ldots, N$, adding an i.i.d. busy period duration η_i from such an $M/G/1$ queue. □

Proof of Theorem 5.3.1:

Proof. We proceed by calculating $(\partial/\partial b)S_2(b, x)$ from first principles:

$$(\partial/\partial b)[S_{q2}(b, x)] = \lim_{\triangle b \to 0} \left\{ \frac{S_2(b + \triangle b, x) - S_2(b, x)}{\triangle b} \right\}$$

Reconsider the revised service discipline used in Lemma B.1.1. From Figure B.1, which presumes that we have conditioned upon $\mathcal{W} = t$, closer inspection of the accreditation process reveals that the N Poisson events underpinning Y are those that occur at rate λ_1 during the first $(1-b)$ portion of $\mathcal{W}$. If the class-2 accreditation rate were to change to $b + \triangle b$, then only those events occurring during the first $(1 - b - \triangle b)$ portion of $\mathcal{W}$ would contribute to the compound Poisson process delaying our tagged customer.

Due to the independence of compound Poisson processes arising from Poisson events in non-overlapping intervals, we can write $Y = Y_1 + Y_2$ where Y_1 is the compound Poisson process corresponding to the N_1 accreditation events that occurred during the first $(1 - b - \triangle b)$ portion of $\mathcal{W}$, and Y_2 is the compound Poisson process corresponding to the N_2 accreditation events occurring during the subsequent $(\triangle b)$ portion of $\mathcal{W}$. Thus we obtain immediately

$$S_2(b, x) = P(\mathcal{W} + Y_1 + Y_2 > x); \quad S_2(b + \triangle b, x) = P(\mathcal{W} + Y_1 > x).$$

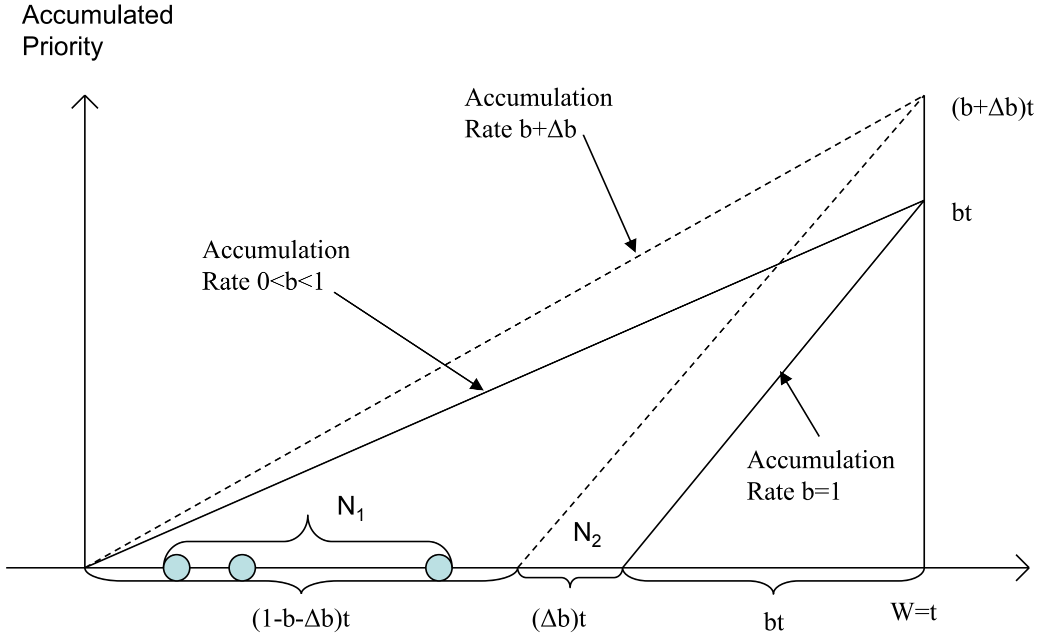


Figure B.1: Impact of change in accumulation rate.

Therefore

$$\begin{aligned} S_2(b + \triangle b, x) - S_2(b, x) &= -[P(\mathcal{W} + Y_1 + Y_2 > x) - P(\mathcal{W} + Y_1 > x)] \\ &= -[P(\mathcal{W} + Y_1 \leq x;\ \mathcal{W} + Y_1 + Y_2 > x)] \ < \ 0. \end{aligned}$$

Now for sufficiently small $\triangle b$, the probability of two or more events in a small period of duration $\triangle b\mathcal{W}$ is $o(\triangle b\mathcal{W})$. Similarly, the probability of one such event is $\lambda_1 \triangle b\mathcal{W} + o(\triangle b\mathcal{W})$. Substituting the terms as appropriate and removing the conditioning upon $\mathcal{W}$, one readily obtains

$$(\partial/\partial b)[S_2(b, x)] = -\lambda_1 E(\{\mathcal{W} \times I(\mathcal{W} + Y + \eta > x)\} < 0. \tag{B.3}$$

In equation (B.3), $I(A)$ denotes the indicator function for the event A, equal to 1 if the event occurs, and 0 otherwise.

Since the partial derivative is negative for all x, this establishes that $S_2(b, x)$ is a monotonically decreasing function of b. Furthermore, as b increases, Y decreases in distribution, since the rate $\lambda_1(1 - b)$ of the corresponding underlying Poisson decreases. Consequently, the probability associated with the indicator function decreases, so that the derivative is a strictly increasing function of b, and the convexity immediately follows. □

Corollary 5.3.3. The excess function $H_K(b, l_K)$, where $b = b_K/b_{K-1}$, for the lowest priority class in a stable K-class APQ is strictly convex in b for $0 \leq b \leq 1$, for every fixed $l_K \geq 0$.

Proof. The lowest priority class (class-K) observes the unfinished workload in the corresponding $M/G/1$ FCFS queue at its arrival instants. By Corollary 7.3 in Stanford *et al.* [18], the higher classes $1, 2, \ldots, K-1$ gain accreditation over class-K customers according to a Poisson process at rate $\sum_{i=1}^{K-1} \lambda_i(1 - b_K/b_i)$. The rest follows by direct analogy to the proof of the two-class case. □

B.2 Waiting times in the two-class multi-server APQ

The LST $\widetilde{w}_2(s)$ of the class-2 waiting time distribution can be written as

$$\widetilde{w}_2(s) = [1 - \pi_{\text{busy}}] + \pi_{\text{busy}}\widetilde{w}_2^+(s), \tag{B.4}$$

where $\widetilde{w}_2^+(s)$ is the LST of the class-2 waiting time distribution, conditional upon an arrival finding all servers busy, which is given by

$$\widetilde{w}_2^+(s) = \frac{2\mu(1-\rho)}{\mu + \hat{\lambda} + s - 2\lambda + \sqrt{(\mu + \hat{\lambda} + s)^2 - 4\mu\hat{\lambda}}}. \tag{B.5}$$

The mean waiting time of the class-2 customers can be derived from equations (B.4) & (B.5), which is

$$m_2(b) = \frac{\pi_{\text{busy}}}{(1-\rho)(\mu - \hat{\lambda})}. \tag{B.6}$$

The LST $\widetilde{w}_1(s; b)$ of the class-1 waiting time distribution can be written as

$$\widetilde{w}_1(s) = (1 - \pi_{\text{busy}}) + \pi_{\text{busy}}\widetilde{w}_1^+(s), \tag{B.7}$$

and the LST of the waiting time distribution for the class-1 customers conditional on it being positive, $\widetilde{w}_1^+(s)$, is

$$\widetilde{w}_1^+(s;b) = \frac{(\mu+s)\tilde{\Gamma}_1(bs;b)-\mu}{s(\mu+s-\lambda_1)}\left[(\mu-\lambda)+(\lambda-\hat{\lambda})\widetilde{w}_2^+(bs;b)\right]+b\widetilde{w}_2^+(bs;b), \tag{B.8}$$

where $\tilde{\Gamma}_1(s;b) = (\mu+\hat{\lambda}+s-\sqrt{(\mu+\hat{\lambda}+s)^2-4\mu\hat{\lambda}})/2\hat{\lambda}$.

The mean waiting time of the class-1 customers can be derived from equations (B.7) & (B.8), which is

$$m_1(b) = \frac{\pi_{\text{busy}}\left[1-(1-b)\rho\right]}{(1-\rho)(\mu-\hat{\lambda})}. \tag{B.9}$$

B.3 The proofs in Section 5.5

Lemma 5.5.2. The GS numerical evaluation of the class-k expected excess waiting time per customer for $k = 1, 2, \ldots, K$ is given by

$$H_{g,k}(t) = \sum_{j=1}^{N}\frac{V_j}{j}\left[\frac{\widetilde{w}_k(\frac{ln2}{t}\times j)}{(\frac{ln2}{t}\times j)} - \frac{1}{(\frac{ln2}{t}\times j)} + m_k\right]. \tag{5.24}$$

where the $V_j, j = 1, \ldots, N$ are combinatorial terms related to order statistics as derived by Gaver [10].

Proof. Given a real-valued function $f(t); t \geq 0$ whose LST is $\widetilde{f}(s)$, then the GS method for numerical Laplace transform inversion at the point t is given by the following:

$$f_g(t) = \frac{ln2}{t}\sum_{j=1}^{N} V_j\,\widetilde{f}\left(\frac{ln2}{t}\times j\right), \tag{B.10}$$

where the values V_j are the GS coefficients of order N (always even), half of which are positive and half negative numbers. These coefficients, as derived by Gaver, are combinatorial terms arising in order statistics, with the useful by-product that they always sum to zero. Typically $N = 8$ points provides two significant digits of accuracy, which is quite adequate for assessing waiting times. The table that provides the coefficients for $N = 2; 4; 6; 8$ is provided in Table 2.1.

In light of equations (5.22) and (B.10), the GS evaluation of the class-k waiting time distribution, $k = 1, 2, \dots, K$ is achieved via

$$W_{g,k}(t) = \frac{ln2}{t} \sum_{j=1}^{N} V_j \frac{\widetilde{w}_k\left(\frac{ln2}{t} \times j\right)}{\left(\frac{ln2}{t} \times j\right)} = \sum_{j=1}^{N} \frac{V_j}{j} \widetilde{w}_k\left(\frac{ln2}{t} \times j\right). \tag{B.11}$$

Meanwhile, the GS numerical evaluation of the class-k expected excess per customer function, $k = 1, 2, \dots, K$ is given by

$$\begin{aligned} H_{g,k}(t) &= \frac{ln2}{t} \sum_{j=1}^{N} V_j \widetilde{H}_k\left(\frac{ln2}{t} \times j\right) \\ &= \frac{ln2}{t} \sum_{j=1}^{N} V_j \left[\frac{m_k}{(\frac{ln2}{t} \times j)} - \frac{1}{(\frac{ln2}{t} \times j)^2} + \frac{\widetilde{w}_k(\frac{ln2}{t} \times j)}{(\frac{ln2}{t} \times j)^2}\right] \\ &= \sum_{j=1}^{N} \frac{V_j}{j} \left[m_k - \frac{1}{(\frac{ln2}{t} \times j)} + \frac{\widetilde{w}_k(\frac{ln2}{t} \times j)}{(\frac{ln2}{t} \times j)}\right], \end{aligned}$$

which is equation (5.24).

It is readily apparent from a comparison of (5.24) and (B.11) that minimal extra effort is involved in determining the expected excess per customer beyond the specified thresholds $H_k(l_k); k = 1, 2, \dots, K$ once the evaluations of the corresponding compliance probabilities $W_k(l_k)$ have been performed. □

Theorem 5.5.3. The WAE in a multi-class $M/M_i/c$ APQ as evaluated by GS approximation is given by

$$Z_g = \sum_{k=1}^{K} \sum_{j=1}^{N} \frac{\alpha_k \pi_{\text{busy}} \lambda_k l_k V_j}{j^2 ln2} \left(\widetilde{w}_k^{+}(\frac{ln2}{l_k} \times j) - 1\right) + \sum_{k=1}^{K} \alpha_k \lambda_k m_k. \tag{5.25}$$

When $\alpha_k = 1, k = 1, 2, \dots, K$, the corresponding total expected excess (TEE) is

$$Z_g = \sum_{k=1}^{K} \sum_{j=1}^{N} \frac{\pi_{\text{busy}} \lambda_k l_k V_j}{j^2 ln2} \left(\widetilde{w}_k^{+}(\frac{ln2}{l_k} \times j) - 1\right) + M, \tag{5.26}$$

where M is the constant in the conservation law for an $M/M_i/c$ APQ (Li and Stanford [14]), such that $M = \sum_{k=1}^{K} \lambda_k m_k$. Moreover, M does not depend on the accumulation rates $b_k; k = 1, 2, \dots, K$.

Proof. As the LST of the waiting time for class-k in the multi-class $M/M_i/c$ APQ is given by

$$\widetilde{w}_k(s) = (1 - \pi_{\text{busy}}) + \pi_{\text{busy}} \widetilde{w}_k^{+}(s). \tag{B.12}$$

From equations (5.3), (5.24) & (B.12), we have

$$
\begin{aligned}
Z_g &= \sum_{k=1}^{K} \alpha_k \lambda_k H_{g,k}(l_k) \\
&= \sum_{k=1}^{K} \alpha_k \lambda_k \sum_{j=1}^{N} \frac{V_j}{j} \left[m_k - \frac{1}{(\frac{ln2}{l_k} \times j)} + \frac{\widetilde{w}_k(\frac{ln2}{l_k} \times j)}{(\frac{ln2}{l_k} \times j)} \right] \\
&= \sum_{k=1}^{K} \alpha_k \lambda_k \sum_{j=1}^{N} \left[\frac{l_k V_j}{j^2 ln2} \Big(\widetilde{w}_k(\frac{ln2}{l_k} \times j) - 1 \Big) + \frac{V_j}{j} m_k \right] \\
&= \sum_{k=1}^{K} \alpha_k \lambda_k \sum_{j=1}^{N} \left[\frac{l_k V_j}{j^2 ln2} \Big(1 - \pi_{\text{busy}} + \pi_{\text{busy}} \widetilde{w}_k^{+}(\frac{ln2}{l_k} \times j) - 1 \Big) + \frac{V_j}{j} m_k \right] \\
&= \sum_{k=1}^{K} \alpha_k \lambda_k \sum_{j=1}^{N} \Big[\frac{\pi_{\text{busy}} l_k V_j}{j^2 ln2} \Big(\widetilde{w}_k^{+}(\frac{ln2}{l_k} \times j) - 1 \Big) + \frac{V_j}{j} m_k \Big] \\
&= \sum_{k=1}^{K} \sum_{j=1}^{N} \frac{\alpha_k \pi_{\text{busy}} \lambda_k l_k V_j}{j^2 ln2} \Big(\widetilde{w}_k^{+}(\frac{ln2}{l_k} \times j) - 1 \Big) + \sum_{k=1}^{K} \alpha_k \lambda_k m_k,
\end{aligned}
$$

which is equation (5.25). By setting $\alpha_k = 1 \; \forall \; k$ in equation (5.25), equation (5.26) follows immediately.

The constant M in equation (5.26) refers to the constant in the conservation law in Li and Stanford [14], where $M = \lambda \pi_{\text{busy}} / (\mu - \lambda)$ does not depend on the accumulation rates $b_k; k = 1, 2, \ldots, K$. □

Corollary 5.5.4. In any multi-class APQ with c heterogeneous servers, each working at an exponential service rate μ_i for $i = 1, 2, \ldots, c$, the optimality of the TEE in equation (5.26) is the same as the one in a single-server APQ with an exponentially-distributed service time at rate $\mu = \sum_{i=1}^{c} \mu_i$.

Proof. It is apparent that the optimality of the TEE in equation (5.26) is eventually driven by the function $\widetilde{w}_k^{+}$, which is the only part that depends on the accumulation rates. However, Lemma 6.1 in Li and Stanford [14] have proved that the LST, $\widetilde{w}_k^{+}(s)$, of the conditional waiting time distribution for class-k in the $M/M_i/c$ APQ can be related to that in an $M/M/1$, by setting the single service rate as the sum of the individual service rates in the $M/M_i/c$ system (this also follows from Sharif *et al.* [17]). Thus, it immediately follows that the optimal solution for the TEE in $M/M_i/c$ APQ is the same as the one in the $M/M/1$ APQ with $\mu = \sum_{i=1}^{c} \mu_i$. □

B.4 Further investigations on the TEE for a two-class APQ

Theorem B.4.1. *The TEE in a two-class $M/M_i/c$ APQ as evaluated by GS approximation is given by*

$$
\begin{aligned}
Z_g(b) &= \lambda_1 H_{g,1}(l_1) + \lambda_2 H_{g,2}(l_2) \\
&= d_1 \sum_{j=1}^{N} \frac{V_j}{j^2} J(d_0 j; b, f) + d_2 + M, \qquad \text{(B.13)}
\end{aligned}
$$

where $d_0 = ln2/l_1$, $d_1 = (\mu-\lambda)\pi_{\text{busy}}\lambda_1 l_1/ln2 > 0$ and $d_2 = -(\lambda_1 l_1+\lambda_2 l_2)\sum_{j=1}^{N} V_j/j^2$ are constants that do not depend on the parameter b. The function $J(d_0 j; b, f)$ is defined in equation (B.14).

Proof. From equations (5.26), (B.4)–(B.8), we have

$$
\begin{aligned}
Z_g(b) &= \frac{\pi_{\text{busy}}}{ln2} \sum_{j=1}^{N} \frac{V_j}{j^2}\Big[\lambda_1 l_1\Big(\widetilde{w}_1^{+}(d_0 j) - 1\Big) + \lambda_2 l_2\Big(\widetilde{w}_2^{+}(f d_0 j) - 1\Big)\Big] + M \\
&= \frac{(\mu-\lambda)\pi_{\text{busy}}\lambda_1 l_1}{ln2} \sum_{j=1}^{N} \frac{V_j}{j^2(\mu-\lambda)}\Big(\widetilde{w}_1^{+}(d_0 j) + \frac{\theta}{f}\widetilde{w}_2^{+}(f d_0 j)\Big) + d_2 + M \\
&= d_1 \sum_{j=1}^{N} \frac{V_j}{j^2} J(d_0 j; b, f) + d_2 + M,
\end{aligned}
$$

where

$$
\begin{aligned}
J(d_0 j; b, f) &= \frac{1}{\mu-\lambda}\Big(\widetilde{w}_1^{+}(d_0 j; b) + \frac{\theta}{f}\widetilde{w}_2^{+}(f d_0 j; b)\Big) \\
&= \frac{\hat{\lambda}(\mu+d_0 j)(\tilde{\Gamma}_1(bd_0 j;b))^2 - \Big((\mu+d_0 j)(\mu+bd_0 j)+\mu\hat{\lambda}\Big)\tilde{\Gamma}_1(bd_0 j;b)+bd_0 j(\lambda_1-d_0 j)+\mu^2}{d_0 j(\mu-\lambda_1+d_0 j)(\hat{\lambda}\tilde{\Gamma}_1(bd_0 j;b)+(\theta+b)\lambda_1-(\mu+bd_0 j))} \\
&\quad + \frac{\theta/f}{\mu-\lambda+d_0 jf+\hat{\lambda}(1-\tilde{\Gamma}_1(d_0 jf;b))}. \qquad \text{(B.14)}
\end{aligned}
$$

To minimize TEE, we may set the first derivative of the function $Z_g(b)$ with respect to b to be zero. From equation (B.13), we have

$$
\frac{\partial}{\partial b} Z_g(b) = d_1 \sum_{j=1}^{N} \frac{V_j}{j^2} \frac{\partial}{\partial b} J(d_0 j; b, f). \qquad \text{(B.15)}
$$

As the constant d_1 is positive, in order to find the optimal solution $b^{\dagger}$ for TEE, we can study the optimal behaviour of the function $J(d_0 j; b, f)$ for each j under the GS numerical inversion method. □

Lemma B.4.2. *For $j > 0$, the first derivative of the function $J(d_0 j; b, f)$ evaluated at $b = f$ is always greater than zero.*

Proof. With Theorem B.4.1, under the framework of the GS numerical inversion algorithm, $b^{\dagger}$ is determined by the function $J(d_0 j; b, f)$.

Let $J(d_0 j; b, f) = J_1(d_0 j; b) + J_2(d_0 j; b, f)$, where

$$J_1(d_0 j; b) = \frac{X(d_0 j; b) + d_0 jb + \mu - \lambda_1(1 + b)}{(\mu - \lambda_1 + d_0 j)\Big(d_0 jb - (1 + 2\theta + b)\lambda_1 + X(d_0 j; b) + \mu\Big)}, \tag{B.16}$$

$$J_2(d_0 j; b, f) = \frac{\theta/f}{\mu - \lambda + d_0 jf + \hat{\lambda}(1 - \tilde{\Gamma}_1(d_0 jf; b))}, \tag{B.17}$$

where $X(d_0 j; b) = \sqrt{(\mu + \hat{\lambda} + bd_0 j)^2 - 4\mu\hat{\lambda}}$. Take the first derivative of $J_1(d_0 j; b)$ with respect to b,

$$\frac{\partial}{\partial b} J_1(d_0 j; b) = \frac{2\lambda_1\theta\left((\lambda_1 - d_0 j)X(d_0 j; b) + (1 - b){\lambda_1}^2 - \Big((1 - 2b)d_0 j + \mu\Big)\lambda_1 + d_0 j(bd_0 j + \mu)\right)}{X(d_0 j; b)\Big(X(d_0 j; b) - (b + 2\theta + 1)\lambda_1 + bd_0 j + \mu\Big)^2 (\mu - \lambda_1 + d_0 j)}. \tag{B.18}$$

Take the first derivative of $J_2(d_0 j; b, f)$ with respect to b,

$$\frac{\partial}{\partial b} J_2(d_0 j; b, f) = \frac{2\theta\lambda_1\Big(Y(d_0 j; b, f) + \hat{\lambda} - \mu + fd_0 j\Big)}{fY(d_0 j; b, f)\Big(Y(d_0 j; b, f) + \mu - (1 + b + 2\theta)\lambda_1 + fd_0 j\Big)^2}, \tag{B.19}$$

where $Y(d_0 j; b, f) = \sqrt{(\mu + \hat{\lambda} + fd_0 j)^2 - 4\mu\hat{\lambda}}$. We then evaluate

$$\begin{aligned}
\frac{\partial}{\partial b} J(d_0 j; b, f)|_{b=f} &= \frac{\partial}{\partial b} J_1(d_0 j; b)|_{b=f} + \frac{\partial}{\partial b} J_2(d_0 j; b, f)|_{b=f} \\
&= \frac{2\theta\lambda_1}{fX(d_0 j; f)(\mu - \lambda_1 + d_0 j)\Big(X(d_0 j; f) + \mu - (1 + f + 2\theta)\lambda_1 + fd_0 j\Big)^2} \times \\
&\quad \Big[(\mu - \lambda_1 + d_0 j)(X(d_0 j; f) + \lambda_1 - \mu) - f^2(d_0 j - \lambda_1)^2 \\
&\quad + f\Big(2\lambda_1^2 + (X(d_0 j; f) - 3d_0 j - 2\mu)\lambda_1 + d_0 j(d_0 j - X(d_0 j; f))\Big)\Big] \\
&= A \times B,
\end{aligned} \tag{B.20}$$

where

$$A = \frac{2\theta\lambda_1}{fX(d_0 j; f)(\mu - \lambda_1 + d_0 j)\Big(X(d_0 j; f) + \mu - (1 + f + 2\theta)\lambda_1 + fd_0 j\Big)^2} > 0, \tag{B.21}$$

and

$$\begin{aligned} B &= -f^2(d_0 j - \lambda_1)^2 + f\Big(2\lambda_1^2 + (X(d_0 j; f) - 3d_0 j - 2\mu)\lambda_1 + d_0 j(d_0 j - X(d_0 j; f))\Big) \\ &\quad + (\mu - \lambda_1 + d_0 j)(X(d_0 j; f) + \lambda_1 - \mu) \\ &> (\mu - (1-f)\lambda_1)((1-f)\lambda_1 - \mu + (\mu - (1-f)\lambda_1)) \\ &= 0. \end{aligned} \tag{B.22}$$

Thus, $\partial J(d_0 j; b, f)/\partial b|_{b=f} = A \times B > 0$ for all $j > 0$ in the GS numerical inversion algorithm.

□

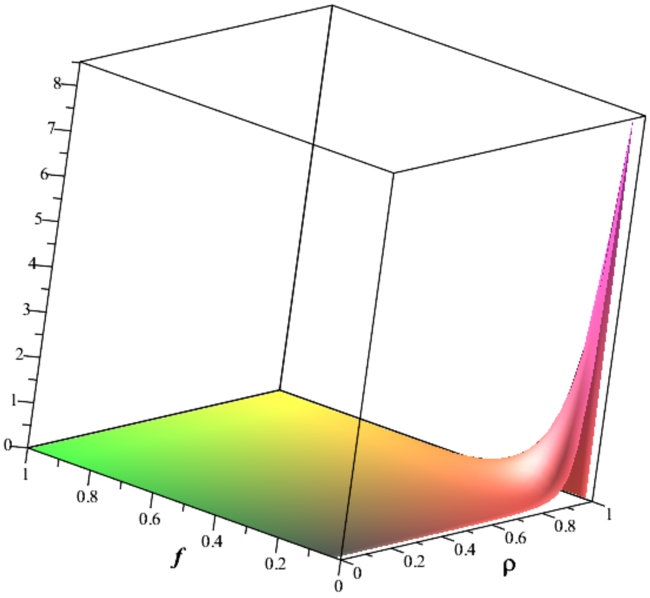


Figure B.2: Numerical values of $\partial Z_g(b)/\partial b|_{b=f}$ under the GS algorithm with 8 points.

Figure B.2 shows the numerical values of $\partial Z_g(b)/\partial b|_{b=f}$ using the GS algorithm with 8 points (see Table 2.1) when $\theta = 1$, $\mu = 2$ and $l_1 = 30$ mins. We observe that the values are all positive for $0 < \rho < 1$ and $0 < f < 1$.

To study the convexity of the function $J(d_0 j; b, f)$, our next step is to investigate the second derivative of the function $J(d_0 j; b, f)$ respective to b, which is given by

$$
\begin{aligned}
\frac{\partial^2}{\partial b^2} J(d_0 j; b, f) = {} & \frac{2\theta\lambda_1\Big((X(d_0 j; b) + \mu + bd_0 j)d_0 j + \big(\mu - X(d_0 j; b) + (1-2b)d_0 j\big)\lambda_1 - \lambda_1\hat{\lambda}\Big)}{X(d_0 j; b)^3(\mu + d_0 j - \lambda_1)\big(X(d_0 j; b) - (1+b+2\theta)\lambda_1 + bd_0 j + \mu\big)^3} \times \\
& \Big\{\Big[\mu^2 + 4\mu X(d_0 j; b) - X(d_0 j; b)^2 + 2d_0 j\big((\theta + 2 - 2b)X(d_0 j; b) - \mu(\theta + b)\big) \\
& - bd_0^2 j^2(3b + 2\theta)\Big]\lambda_1 + \Big[2(b - \theta - 2)X(d_0 j; b) - 2\mu(1+\theta) + bd_0 j(3b + 4\theta) \\
& - d_0 j(2\theta - 1) - \hat{\lambda}(1 + b + 2\theta)\Big]\lambda_1^2 + d_0 j\big(bd_0 j + X(d_0 j; b) + \mu\big)^2\Big\} \\
& + \frac{2\theta\lambda_1^2(Y(d_0 j; b, f) + d_0 jf - \mu + \hat{\lambda})}{fY(d_0 j; b, f)^3\big(Y(d_0 j; b, f) + \mu + d_0 jf - (1+b+2\theta)\lambda_1\big)^3} \times \\
& \Big\{(Y(d_0 j; b, f) + d_0 jf)^2 - 4\mu Y(d_0 j; b, f) - \mu^2 + 2\big((1+\theta)\mu \\
& - d_0 jf(\theta + b) + Y(d_0 j; b, f)(2 + \theta - b)\big)\lambda_1 - (1 + b + 2\theta)\hat{\lambda}\lambda_1\Big\}. \qquad \text{(B.23)}
\end{aligned}
$$

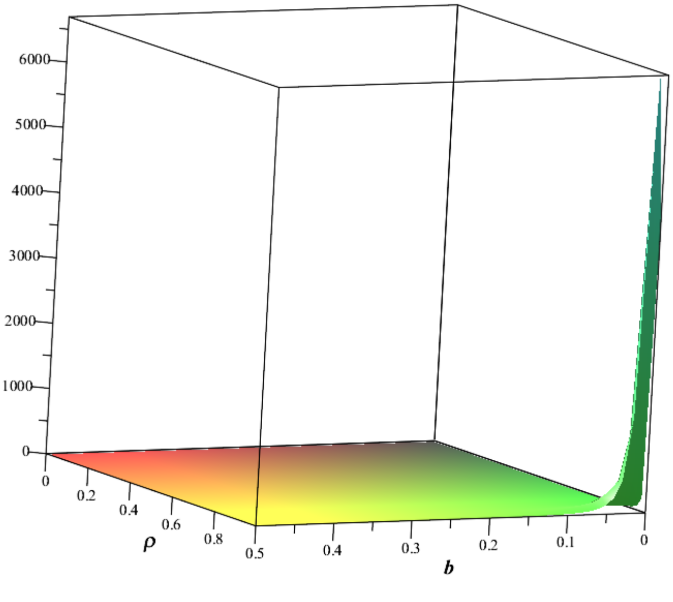


Figure B.3: Numerical values of $\partial^2 Z_g(b)/\partial b^2$ under the GS algorithm with 8 points.

Figure B.3 presents the numerical values of the function $\partial^2 Z_g(b)/\partial b^2$ with $\theta = 1$, $\mu = 2$, $l_1 = 30$ mins and $l_2 = 60$ mins. When $b \in [0, 1/2]$ and $\rho \in (0, 1)$, the values of the function is greater than zero. Moreover, we notice that when ρ approaches one and b goes to zero, the value approaches infinity. Similarly, we have examined the second derivative at various points of $f \in (0, 1)$. Based on extensive numerical investigations under the GS algorithm, we find that $\partial^2 Z_g(b)/\partial b^2$ is always positive for $b \in [0, f]$. Numerically speaking, the $b^\dagger$ that minimizes TEE is bounded by the ratio of the delay limits, f. Clearly, $b^\dagger$ should not be negative.

However, we have concluded from Figure 5.5 that for a fixed f, $b^\dagger$ is zero which means that the optimal strategy for TEE is the classical priority discipline when $0 < \rho < \rho_{sp}$, whereas, $b^\dagger$ is positive when $\rho_{sp} < \rho < 1$. Moreover, ρ_{sp} increases as f decreases. To study the behaviour of the point ρ_{sp} that distinguishes the optimal disciplines, let us take the first derivative of $J(d_0 j; b, f)$ with respect to b, then evaluate b at zero.

$$\frac{\partial}{\partial b} J_1(d_0 j; b)|_{b=0} = \frac{-\mu\lambda_1 d_0 j\theta}{(\mu - \lambda_1)(\mu + d_0 j - \lambda_1)(\mu - (\theta + 1)\lambda_1)^2} < 0, \tag{B.24}$$

$$\frac{\partial}{\partial b} J_2(d_0 j; b, f)|_{b=0} = \frac{2\theta\lambda_1\Big(Y(d_0 j; 0, f) + \lambda_1 + f d_0 j - \mu\Big)}{fY(d_0 j; 0, f)\Big(Y(d_0 j; 0, f) + \mu - (2\theta + 1)\lambda_1 + f d_0 j\Big)^2} > 0. \tag{B.25}$$

As $\partial J(d_0 j; b, f)/\partial b|_{b=0} = \partial J_1(d_0 j; b)/\partial b|_{b=0} + \partial J_2(d_0 j; b, f)/\partial b|_{b=0}$, the value of $\partial J(d_0 j; b, f)/\partial b|_{b=0}$ increases as f decreases (see equation (B.25)). When f is small enough, $\partial J(d_0 j; b, f)/\partial b|_{b=0}$ is possibly greater than zero. ρ_{sp} is the solution of $\partial Z_g(b)/\partial b|_{b=0} = 0$ when other parameters are fixed.

Thus, based on a thorough investigation, we are confident that the $b^\dagger$ for TEE in equation (B.13) is in the range of $[0, f)$. We have mentioned that in reality we mostly consider the systems with $f \in (1/6, 1)$ and $\rho \in (0.6, 1)$, and we have found in these situations the APQ discipline outperforms the classical priority discipline, which implies that $b^\dagger$ is greater than zero. What would be the lower bound of $b^\dagger > 0$ when we consider the situation that $f \in (1/6, 1)$ and $\rho \in (0.6, 1)$? To answer this question, we take the first derivative of $J(d_0 j; b, f)$ with respect to b, then evaluate b at $f/2$.

$$\frac{\partial}{\partial b} J(d_0 j; b, f)|_{b=f/2} = \frac{-4\theta\lambda_1}{X(d_0 j; f/2)(\mu - \lambda_1 + d_0 j)\Big(2X(d_0 j; f/2) + 2\mu + f(d_0 j - \lambda_1) - 2(2\theta + 1)\lambda_1\Big)^2} \times \\ \Big[(f - 2)\lambda_1^2 + 2\Big(\mu - X(d_0 j; f/2) + (1 - f)d_0 j\Big)\lambda_1 + d_0 j(2\mu + 2X(d_0 j; f/2) + d_0 j f)\Big]$$

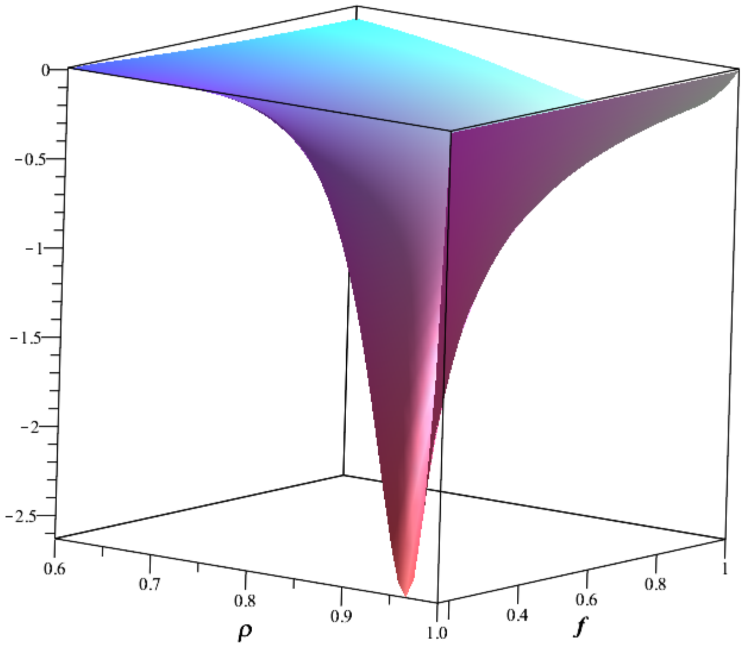


Figure B.4: Numerical values of $\partial Z_g(b)/\partial b|_{b=f/2}$ under the GS algorithm with 8 points.

$$+\frac{2\theta\lambda_1\Big(Y(d_0 j;f/2,f)+d_0 jf-\mu+(1-f/2)\lambda_1\Big)}{fY(d_0 j;f/2,f)\Big(Y(d_0 j;f/2,f)+\mu+f(d_0 j-\lambda_1/2)-(2\theta+1)\lambda_1\Big)}. \qquad (B.26)$$

Figure B.4 illustrates the numerical values of the function $\partial Z_g(b)/\partial b|_{b=f/2}$ using the GS algorithm with 8 points when $\theta = 1$, $\mu = 2$, and $l_1 = 30$ mins. When $\rho \in (0.6, 1)$ and $f \in (1/6, 1)$, the values of the function is always smaller than zero. Based on the numerical investigation under the GS algorithm, $\partial Z_g(b)/\partial b|_{b=f/2}$ is smaller than zero. Overall, for a two-class $M/M_i/c$ APQ with $\rho \in (0.6, 1)$ and $f \in (1/6, 1)$, $b^\dagger$ is bounded within the range of $(f/2, f)$.

Curriculum Vitae

Name:	Na Li
Post-Secondary Education and Degrees:	Beijing Institute of Graphic Communication Beijing, P.R. China 2005 - 2009 B.Eng.
	University of Western Ontario London, ON, Canada 2011 - 2015 Ph.D.
Honours and Awards:	Ontario Graduate Scholarship (OGS) 2014-2015
Related Work Experience:	Research Assistant & Teaching Assistant University of Western Ontario 2011 - 2015
	Statistical Consultant University of Western Ontario 2014 - 2015

Publications:

Multi-server Accumulating Priority Queues with Heterogeneous Servers (Submitted)

On Waiting Times for Nonlinear Accumulating Priority Queues (Submitted)

Optimization of Queues Operating Under Waiting Time Limits (Submitted)

Other researches:

Two-class Two-server Accumulating Priority Queue with Class-dependent Service Rates

Methods to Obtain Wait Time 1 Using Billing Data in Saskatchewan